Об изогональном сопряжении, точках Микеля, прямых Гаусса и др.

Н.Белухов, А.Заславский, П.Кожевников

1 Вводные задачи

1. **Изогональное сопряжение.** Дан треугольник ABC и точка P.

a)Докажите, что прямые, симметричные AP, BP, CP относительно биссектрис соответствующих углов, пересекаются в одной точке или параллельны. Полученная точка P' называется *изогонально сопряженной* P относительно треугольника ABC.

b)Докажите, что P' — бесконечно удаленная точка (т.е. три соответствующие прямые параллельны) тогда и только тогда, когда P лежит на описанной окружности треугольника ABC.

c)Найдите образ при изогональном сопряжении окружности, проходящей через две вершины треугольника.

d)[1] Докажите, что проекции P и P' на стороны ABC лежат на одной окружности. Как звучит это утверждение, если P' бесконечно удалена?

e)Если X, X' и Y, Y' — две пары изогонально сопряженных точек, то $XY \cap X'Y'$ и $XY' \cap X'Y$ изогонально сопряжены.

Дан четырехугольник $ABCD$ и точка P.

f)Докажите, что, если три из четырех прямых, симметричных AP, BP, CP, DP относительно биссектрис соответствующих углов, пересекаются в одной точке, то четвертая также проходит через эту точку.

g)Докажите, что точка, изогонально сопряженная P относительно четырехугольника, существует тогда и только тогда, когда проекции P на стороны лежат на одной окружности (причем на этой окружности лежат и проекции изогонально сопряженной точки).

Коникой, вписанной в многоугольник, называется коника, касающаяся всех прямых, содержащих стороны многоугольника.

h)Докажите, что фокусы вписанной в треугольник коники изогонально сопряжены.

i)Докажите, что фокус любой параболы, касающейся прямых AB, BC, CA, лежит на описанной окружности треугольника ABC.

2. **Точка Микеля.** Дан четырехугольник $ABCD$. Прямые AB и CD пересекаются в точке E, AD и BC — в точке F.

a)Докажите, что в обозначениях предыдущей задачи окружности, описанные около треугольников ABF, CDF, ADE и CDE, пересекаются в одной точке M (точке Микеля четверки прямых AB, BC, CD, DA).

b)Докажите, что M — центр поворотной гомотетии, переводящей отрезок BE в FD (или DE в FB, и т.д.)

c)Два таракана B и C ползут с постоянными скоростями по двум прямым, пересекающимся в точке A. Докажите, что окружности ABC проходят через фиксированную точку, а прямые BC касаются фиксированной параболы.

d)(IMO2005) Дан выпуклый четырехугольник $ABCD$, стороны BC и AD которого равны, но не параллельны. Пусть E и F – внутренние точки отрезков BC и AD соответственно такие, что $BE = DF$. Прямые AC и BD пересекаются в точке P, прямые BD и EF пересекаются в точке Q, прямые EF и AC пересекаются в точке R. Рассмотрим треугольники PQR, получаемые для всех таких точек E и F. Докажите, что окружности, описанные около всех этих треугольников, имеют общую точку, отличную от P.

e)Выясните связь точки Микеля со вписаннымиониками.

[1] Здесь и далее мелким шрифтом набраны задачи, которые не используются при получении основных результатов, приведенных в разделе 2.

f)Докажите, что проекции точки Микеля на стороны четырехугольника лежат на одной прямой, перпендикулярной прямой Гаусса. Как связана эта прямая со вписанной в четырехугольник параболой?

3. **Прямая Гаусса.** Дан четырехугольник $ABCD$. Прямые AB и CD пересекаются в точке E, AD и BC — в точке F.

a)Докажите, что середины отрезков AC, BD и EF лежат на одной прямой (прямой Гаусса четырехугольника $ABCD$, или четверки прямых AB, BC, CD, DA).

b)Докажите, что центры окружностей, проходящих через проекции пары изогональных относительно четырехугольника точек, лежат на прямой Гаусса четырехугольника.

c)Докажите, что точка Микеля четырехугольника изогонально сопряжена бесконечно удаленной точке его прямой Гаусса.

d)Докажите, что центры коник, вписанных в четырехугольник, лежат на его прямой Гаусса.

e)(Всероссийская олимпиада 2009) На сторонах AB и BC параллелограмма $ABCD$ выбраны точки A_1 и C_1 соответственно. Отрезки AC_1 и CA_1 пересекаются в точке P. Описанные окружности треугольников AA_1P и CC_1P вторично пересекаются в точке Q, лежащей внутри треугольника ACD. Докажите, что $\angle PDA = \angle QBA$.

2 Три Микеля для квартетов.

В задачах этого раздела рассматривается следующая конструкция и используются следующие обозначения. A, B, C, D — четыре точки общего положения. X — точка Микеля прямых AB, AC, BD, CD, Y — точка Микеля прямых AB, AD, BC, CD, Z — точка Микеля прямых BC, AC, BD, AD. $P_X = AD \cap BC$, $P_Y = AC \cap BD$, $P_Z = AB \cap CD$. K_X, L_X — середины BC и AD, K_Y, L_Y — середины AC и BD, K_Z, L_Z — середины AB и CD. Γ_X, Γ_Y, Γ_Z — прямые (Гаусса для соответствующих четверок прямых) K_XL_X, K_YL_Y, K_ZL_Z.

4. Докажите, что прямые AX, BY, CZ пересекаются в одной точке D' или параллельны. Аналогично определяются точки A', B', C' как центры перспективы треугольника XYZ с треугольниками DCB, CDA, BAD.

5. Докажите, что A', B', C', D' — точки, изогонально сопряженные A, B, C, D относительно треугольника XYZ.

6. Докажите, что X, Y, Z — точки Микеля для четверки точек A', B', C', D'.

7. Докажите, что прямые AA', BB', CC' и DD' параллельны.

8. Докажите, что прямые AD, $A'D'$ и YZ пересекаются в одной точке (и аналогичные пересечения).

9.

a)Докажите, что точки X, Z, P_Y, K_Y, L_Y лежат на одной окружности — ω_Y. Окружности ω_X, ω_Z определяются аналогично.

b)Докажите, что окружности ω_X, ω_Y, ω_Z проходят через одну точку T (или совпадают).

c)Докажите, что прямые XP_X, YP_Y, ZP_Z проходят через T.

3 Квартеты для трех Микелей.

Дан треугольник XYZ. Определим преобразование ψ_X, как композицию симметрии относительно биссектрисы угла X и инверсии с центром X и таким радиусом R, что $R^2 = XY \cdot XZ$. Аналогично определим ψ_Y, ψ_Z.

10. Докажите, что

a)$\psi_X(Y) = Z$, $\psi_X(Z) = Y$;

b)ψ_X^2 — тождественное преобразование;

c)Композиция ψ_X, ψ_Y и ψ_Z — тождественное преобразование.

Пусть D — произвольная точка, $A = \psi_X(D)$, $B = \psi_Y(D)$, $C = \psi_Z(D)$.

11. Докажите, что $\triangle XDZ \sim \triangle XYA$ и $\triangle XDY \sim \triangle XZA$.

12. Докажите, что каждое из преобразований ψ_X, ψ_Y, ψ_Z переводит набор из четырех точек A, B, C, D в себя.

Будем называть набор из четырех (не обязательно различных) точек A, B, C, D *квартетом*. Из последней задачи следует, что вся плоскость может быть разбита на квартеты.

13. Докажите, что четверка точек, изогонально сопряженных квартету — квартет.

14. Найдите квартеты, содержащие

a)центр I вписанной окружности треугольника X, Y, Z;

b)центр O его описанной окружности.

c)Найдите неподвижные точки преобразования ψ_Z и соответствующие квартеты.

15.

a)Докажите, что X — точка Микеля прямых AB, AC, BD, CD.

b)Докажите обратное утверждение: если X, Y, Z — точки Микеля, определяемые точками A, B, C, D, то A, B, C, D образуют квартет.

16. Докажите, что каждое из преобразований ψ_X, ψ_Y, ψ_Z коммутирует с изогональным сопряжением относительно треугольника XYZ.

17. Пусть точки A, B, C, D образуют квартет относительно треугольника XYZ, A', B', C', D' изогонально сопряжены им. Тогда существуют вписанные в треугольник коники с фокусами A и A', B и B', C и C', D и D'.

a)Докажите, что эти коники гомотетичны друг другу.

b)Докажите, что середины шести отрезков, соединяющих центры этих коник, лежат на гомотетичной им конике, описанной около треугольника XYZ.

18. Внутри треугольника ABC лежат две изогонально сопряженные точки M и N. Известно, что $AM \cdot AN \cdot BC = BM \cdot BN \cdot AC = CM \cdot CN \cdot AB = k$.

a)Докажите, что середина MN совпадает с центром тяжести треугольника.

b)Выразите k через стороны треугольника.

4 Дополнительные задачи.

19.
a)Пусть A, B, C, D — квартет, A', B', C', D' — изогонально сопряженный квартет; P_X — точка пересечения AD и BC, P_Y — AC и BD, P_Z — AB и CD. Точки Q_X, Q_Y, Q_Z определяются аналогично по точкам A', B', C', D'. Докажите, что прямые P_XQ_X, P_YQ_Y, P_ZQ_Z пересекаются в одной точке, лежащей на описанной окружности треугольника XYZ (из предыдущих обозначений).
b)В обозначениях предыдущего пункта докажите, что прямые P_XQ_Y, P_YQ_X и XY пересекаются в одной точке.
c)Обозначим точку, полученную в предыдущем пункте через Z'. Докажите, что ZZ' параллельна AA', BB', CC', DD'.
d)Пусть D_1, D_1' и D_2, D_2' — две пары изогонально сопряженных точек такие, что $D_1D_1' \parallel D_2D_2'$. Докажите, что прямые $A_1A_2, B_1B_2, C_1C_2, D_1D_2$ пересекаются в одной точке (A_1, B_1, C_1, D_1 и A_2, B_2, C_2, D_2 — квартеты).
20. Даны точки A, B, C, D. Известно, что треугольник XYZ перспективен каждому из треугольников ABC, BCD, CDA, DAB (именно в таком порядке вершин). Точки D', A', B', C' — соответствующие центры перспективы. Докажите, что прямые AA', BB', CC', DD' пересекаются в одной точке.

Об изогональном сопряжении, точках Микеля, прямых Гаусса и др. Решения

Н.Белухов, А.Заславский, П.Кожевников

1 Вводные задачи

1.
a)Непосредственно следует из теоремы Чевы в форме синусов.
b)**Указание.** Доказывается счетом углов.
c)Пусть P лежит на этой окружности. Тогда угол APB имеет фиксированную величину. Но так как сумма углов четырехугольника $CAPB$ равна 360°, то фиксирована и сумма углов CAP и CBP. А значит будет фиксирован угол $AP'B$, то есть P' лежит на фиксированной окружности, проходящей через точки A и B.
d)Обозначим через P_A, P_B, P_C точки, симметричные P относительно BC, CA, AB. Так как $P_AC = PC = P_BC$, серединный перпендикуляр к отрезку P_AP_B проходит через C и совпадает с биссектрисой угла P_ACP_B. Но легко видеть, что этой биссектрисой является луч CP'. Следовательно, P' — центр описанной окружности треугольника $P_AP_BP_C$. Применив гомотетию с центром P и коэффициентом $1/2$, получим, что середина отрезка PP' является центром окружности, проходящей через проекции P на стороны. Аналогично получаем, что эта же точка является центром окружности, проходящей через проекции P', а поскольку она равноудалена от проекций P и P' на любую прямую, обе окружности совпадают.
Если P' бесконечно удалена, получаем теорему Симсона: основания перпендикуляров, опущенных из точки на описанной окружности треугольника на стороны треугольника, лежат на одной прямой.
e)Обозначим точку пересечения XY' с YX' через P, а XY с $X'Y'$ через Q. Пусть $XA \cap YX' = X_A$, а $Y'A \cap YX' = Y_A$. Также обозначим через Q_A точку пересечения QA с YX'. Покажем, что прямые AQ и AP симметричны относительно биссектрисы угла A. Тогда, так как аналогичное утверждение верно и для остальных углов треугольника, точки P и Q будут изогонально сопряжены, что и требуется.
Рассмотрим двойное отношение точек (Y, Q_A, X_A, X'). Спроецируем его из точки Q на прямую AX'. Тогда Y перейдет в X', Q_A перейдет в A, X_A перейдет сама в себя, а X' перейдет в пересечение прямых $X'Y'$ и XA. Спроецируем получившееся двойное отношение из точки Y' на прямую YX'. Тогда A перейдет в Y_A, X перейдет в P, X_A опять останется на месте, а образ точки X' вернется в X'. То есть $(Y, Q_A, X_A, X') = (P, Y_A, X_A, X') = (Y_A, P, X', X_A)$. А значит при симметрии относительно биссектрисы угла A прямая AQ перейдет в прямую AP, ч.т.д.
Другое доказательство так же можно прочитать в статье А.Акопяна и А.Заславского "Разные взгляды на изогональное сопряжение" "Математическое просвещение" №11, 2007.
f)Обозначим эту точку пересечения P'. Пусть прямые AB и CD пересекаются в точке K. Тогда для одного из треугольников KBC и KDA точка P' будет точкой пересечения хотя бы двух прямых симметричных прямым, соединяющим P с вершинами треугольников. А следовательно она будет являться точкой, изогонально сопряженной точке P. А следовательно, она лежит на прямой, симметричной KP относительно биссектрисы угла K. Но тогда и для оставшегося треугольника она является изогонально сопряженной точкой точке P, а следовательно, лежит на всех четырех прямых, симметричных AP, BP, CP и DP относительно биссектрис соответствующих углов
g)Пусть проекции точки P лежат на одной окружности. Тогда, рассуждая, как в п.d),

получаем, что точка P', симметричная P относительно центра окружности, изогонально сопряжена P. Обратное утверждение доказывается аналогично.

h)См. указанную выше статью.

i)Из обратной теоремы Симсона получаем, что нам надо доказать, что основания перпендикуляров из фокуса параболы на стороны треугольника лежат на одной прямой. А это верно, поскольку, если симметрично отразить фокус параболы относительно любой касательной к этой параболе, то он попадет на директрису этой параболы.

2.

a)Пусть описанный окружности треугольников ABF и CDF вторично пересекаются в точке M. Тогда из вписанности получаем $\angle(AM,\, MD) = \angle(AM,\, MF) + \angle(MF,\, MD) = \angle(BA,\, BF) + \angle(CF,\, CD) = \angle(BA,\, CD) = \angle(AE,\, ED)$. То есть M так же лежит и на описанной окружности треугольника ADE. Аналогично для оставшегося треугольника.

b)Для этого нам достаточно доказать, что треугольники MBE и MFD подобны. Для этого покажем, что угол MBE равен углу MFD. Тогда углы MEB и MDF равны по аналогичным соображениям, а следовательно, треугольники подобны по двух углам. $\angle(EB,\, BM) = \angle(CE,\, CM) = \angle(CD,\, CM) = \angle(FD,\, FM)$, ч.т.д.

c)Пусть вектор скорости таракана B равен $\overrightarrow{b}$, а таракана C — $\overrightarrow{c}$. Тогда пусть $\overrightarrow{CC'} = \overrightarrow{c}$, а $\overrightarrow{BB'} = \overrightarrow{b}$. Тогда точка пересечения описанных окружностей треугольников ABC и $AB'C'$ (точка P) будет центр поворотной гомотетии, переводящий отрезок BB' в CC'. А следовательно и переводящий всю прямую AB в прямую AC с коэффициентом отношения скоростей тараканов. Следовательно, так как в какой-то момент эта гомотетия переводит одного таракана в другого, то она всегда будет делать это. А тогда для любого другого положения тараканов B_0, C_0 для четверки прямых B_0B, BC, CC_0, C_0B_0 точка P всегда будет являться центром такой поворотной гомотетии, а следовательно и точкой Микеля.
Заметив еще, что если отразить точку P симметрично относительно прямой BC то получится треугольник PBP', который так же при движении B будет оставаться все время подобным самому себе (так как треугольник PBC таковым остается). Следовательно всевозможные точки P' можно получить поворотной гомотетией из прямой AB с центром в точке P. То есть это тоже будет прямая. Обозначим ее через l. Тогда прямая CB все время будет касаться параболы с фокусом в точке P и директрисой l. Точкой касания будет являться точка пересечения прямой CB с перпендикуляром, восстановленным в точке P' к прямой l.

d)Пусть точка E ползет с постоянной скоростью от точки B к точке C, а точка F от D к A с такой же скоростью. Тогда условие сохраняется в любой момент времени. С другой стороны посмотрим на то как будут двигаться точки R и Q. Покажем, что точка Q двигается с постоянной скоростью, тогда и точка R по аналогичным причинам будет двигаться с постоянной скоростью, и задача сведется к предыдущей. Заметим, что углы EQB и FQD равны. Так же равны стороны EB и FD треугольников EBQ и FDQ. А следовательно, по теореме синусов равны и их описанные окружности. А следовательно, так как углы EBQ и FDQ фиксированы, то постоянно и отношение EQ к QF. А следовательно, точка Q так же двигается с постоянной скоростью, ч.т.д.

e)Точка Микеля — фокус вписанной в четырехугольник параболы.

f)Заметим, что если взять любые три из четырех проекций, то они лежат на одной прямой по теореме Симсона для соответственного треугольника. Следовательно, все они лежат на одной прямой. Несложно показать, что если сделать гомотетию с центром в точке, для которой применяется теорема Симсона, с коэффициентом 2, то образ прямой Симсона для этой точки пройдет через ортоцентр треугольника. Поэтому если сделать гомотетию с центром в точке Микеля и коэффициентом 2 то полученная прямая пройдет через ортоцентры

всех треугольников. А если взять три окружности с диаметрами на диагоналях (всех трех) четырехугольника, то все ортоцентры будут иметь одинаковые степени относительно этих трех окружностей. А следовательно, это их общая радикальная ось и эта прямая перпендикулярна линии центров, то есть прямой Гаусса. Для вписанной параболы ее фокус будет лежать на описанных окружностях всех треугольников. Следовательно, точка Микеля и будет фокусом этой параболы. А основания перпендикуляров будут лежать на прямой, являющейся касательной к этой параболе в ее вершине. Прямая, проходящая через ортоцентры будет директрисой этой параболы. А значит прямая Гаусса будет параллельна главной оси параболы.

3.

a)Обозначим рассматриваемые середины через M (середина AC), N (середина BD) и T (середина EF). Пусть середины треугольника ABF — точки F', A' и B'. Заметим, что M лежит на $F'B'$, N на $F'A'$, T на $A'B'$. Если сделать гомотетии с центрами в вершинах треугольника ABF и коэффициентом два получаем $\frac{\overrightarrow{F'M}}{\overrightarrow{MB'}} = \frac{\overrightarrow{BC}}{\overrightarrow{CF}}$, $\frac{\overrightarrow{B'T}}{\overrightarrow{TA'}} = \frac{\overrightarrow{AE}}{\overrightarrow{EC}}$, $\frac{\overrightarrow{A'N}}{\overrightarrow{NF'}} = \frac{\overrightarrow{FD}}{\overrightarrow{DA}}$. Перемножив эти три равенства получим справа по теореме Менелая -1, а следовательно, по этой же теореме примененной в обратную сторону получаем требуемое.

b)**Указание.** Прямая Гаусса четырехугольника $ABCD$ является геометрическим местом точек, для которых X $S_{XAB} + S_{XCD} = S_{XBC} + S_{XDA}$ (площади ориентированные).

c)Следует из задачи 2f.

d)Это переформулировка п.b)

e)Заметим, что Q — точка Микеля для четверки прямых AB, BC, CA_1, AC_1. Ввиду 3c достаточно понять, что DP параллельна прямой Гаусса этой четверки прямых. Но это верно, так как при гомотетии с центром B и коэффициентом $1/2$ DP переходит в прямую Гаусса.

2 Три Микеля для квартетов.

4. Из задач 13, 15 следует, что прямые AX, BY, CZ проходят через точку D', изогонально сопряженную D относительно XYZ.

5. Непосредственно следует из задачи 13.

6. Непосредственно следует из задач 13, 15.

7. По задаче 11 треугольники XDZ и XYA, $XD'Z$ и XYA' подобны. Следовательно, $XA : XD' = (XA : XZ)(XZ : XD') = (XY : XD)(XA' : XY) = XA' : XD$, что равносильно утверждению задачи.

8. Следует из предыдущей задачи и теоремы о трех центрах гомотетии, примененной к отрезкам AA', DD' и $B'B$. Действительно, Z является центром гомотетии, переводящей A в B', а A' — в B и т.д.
Другое решение можно сразу получить из утверждения задачи 1e).

9.
a)Точка X является центром поворотной гомотетии, переводящей C в D, а A — в B. Поскольку K_Y при этой гомотетии переходит в L_Y, угол K_YXL_Y равен углу между прямыми AC и BD. Следовательно, X лежит на окружности $P_YK_YL_Y$. Аналогично получаем, что Z тоже лежит на этой окружности.
b)Так как X лежит на окружности AP_YB, $\angle XP_YB = \angle XAB$. Аналогично, $\angle BP_YZ = \angle BCZ$. Из этих и четырех аналогичных равенств получаем, что $\angle XP_YZ + \angle ZP_XY + \angle YP_ZX = \pi$, откуда, очевидно, следует утверждение задачи.
c)Из решения задачи 15 видно, что $\psi_X(P_Y) = P_Z$ и т.п. Это означает, что точки P_X, P_Y, P_Z входят в один квартет, т.е. $\psi_X(P_X) = \psi_Y(P_Y) = \psi_Z(P_Z)$. Утверждение задачи означает, что эта точка изогонально сопряжена T или $\psi_Z(T)$ изогонально сопряжена P_Z. Заметим, что ψ_Z переводит проходящие через T окружности ZXP_Y и ZYP_X в прямые YP_X и XP_Y, так что $\psi_Z(T)$ — точка пересечения этих прямых. Из равенств $\angle P_XYX = \angle ZYP_Z$, $\angle P_YXY = \angle ZXP_Z$ получаем искомое утверждение.

3 Квартеты для трех Микелей.

10. a)-b) Непосредственно следует из определения.
c) Каждое из преобразований ψ_X, ψ_Y, ψ_Z является круговым (т.е. переводит любую окружность в окружность или прямую) и сохраняющим ориентацию. Значит, их композиция также обладает этими свойствами. Кроме того, из пп. a)-b) следует, что она оставляет неподвижными точки X, Y, Z. Но круговое преобразование, сохраняющее ориентацию, однозначно определяется образами трех точек. Заметим, что утверждение задачи верно независимо от порядка применения преобразований ψ_X, ψ_Y, ψ_Z. Следовательно, эти преобразования коммутируют друг с другом.
Второе решение. Положим $\psi_X(D) = A$, $\psi_Y(A) = C$. Достаточно доказать, что треугольники YDZ и CXZ совмещаются поворотной гомотетией.
Имеем (многократно пользуемся подобиями $YXD \sim AXZ$ и аналогичными): $\angle(YD, DZ) = \angle(YD, DX)+\angle(DX, DZ) = \angle(AZ, ZX)+\angle(YX, AY) = \angle(AZ, ZY)+\angle(ZY, ZX)+\angle(YX, AY) = \angle(XC, CY)+\angle(ZY, ZX)+\angle(CY, ZY) = \angle(CX, ZX)$; $\frac{YD}{CX} = \frac{YD}{AZ}\cdot\frac{AZ}{CX} = \frac{DX}{XZ}\cdot\frac{YA}{YX} = \frac{DX}{XZ}\cdot\frac{DZ}{DX} = \frac{DZ}{XZ}$, что и требовалось.
11. По определению ψ_X $\angle ZXD = \angle AXY$ и $XD \cdot XA = XY \cdot XZ$, откуда сразу следует первое подобие. Второе доказывается аналогично

12. Из задачи 10 следует, что, например, $\psi_X \circ \psi_Y = \psi_Z^{-1} = \psi_Z$. Следовательно, $\psi_Y(A) = \psi_Y(\psi_X(D)) = \psi_Z(D) = C$, т.е. ψ_Y меняет местами точки A и C, B и D. Аналогично получаем, что ψ_X меняет местами A и D, B и C, а ψ_Z — A и B, C и D.
13. Пусть D', A' — точки, изогонально сопряженные D, A. Тогда A' лежит на прямой XD, а D' — на прямой XA. Кроме того, $\angle XD'Z = \pi - \angle ZXD' - \angle D'ZX = \pi - \angle DXY - \angle YZD = \angle ZDX + \angle XYZ - \pi$. Но по задаче 11 $\angle ZDX = \angle AYX$, т.е. $\angle XD'Z = \angle XYA'$. Значит, треугольники $XD'Z$ и $XA'Y$ подобны и $A' = \psi_X(D')$.
14.
a)Центры вписанной и трех вневписанных окружностей треугольника XYZ.
b)Точка O и три точки, симметричные вершинам треугольника XYZ относительно противоположных сторон.
c)Из определения ψ_Z следует, что его неподвижные точки должны лежать на биссектрисе угла Z и окружности с центром Z и радиусом $\sqrt{ZX \cdot ZY}$. Таких точек две, обозначим их U, V. Из задачи 10 получаем, что $\psi_X(U) = \psi_Y(\psi_Z(U)) = \psi_Y(U) = \psi_Z(\psi_X(U))$, т.е. $\psi_X(U)$ — тоже неподвижная точка ψ_Z. Очевидно, $\psi_X(U) \neq U$, следовательно, $\psi_X(U) = \psi_Y(U) = V$, и искомый квартет (как для U, так и для V) состоит из дважды повторенных точек U, V. Более того, ясно, что точки, изогонально сопряженные U, V, также являются неподвижными точками ψ_Z. Значит, U и V изогонально сопряжены, а полученный квартет совпадает со своим сопряженным.
15.
a)Из определения ψ_X и результата задачи 12 следует, что X — центр поворотной гомотетии, переводящей A в B, а C в D. По задаче 2b) этот центр совпадает с точкой Микеля
b)Так как X — точка Микеля, то , например, треугольники XBD и XAC подобны. Пусть P_X, P_Y, P_Z — точки пересечения AD и BC, AC и BD, AB и CD. Тогда треугольники XBP_Y и XP_ZC подобны, следовательно, $XA \cdot XD = XB \cdot XC = XP_Y \cdot XP_Z = R_X^2$ и углы AXD, BXC, P_YXP_Z имеют общую биссектрису l. Композиция инверсии с центром X и радиусом R_X и симметрии относительно l переводит треугольники ADP_Y и BCP_Y соответственно в DAP_Z и CBP_Z. Следовательно, общая точка Z описанных окружностей двух первых треугольников переходит в Y, т.е. эта композиция совпадает с ψ_X.
16. Из задачи 13 следует, что композиция ψ_X и изогонального сопряжения, примененных в любом порядке, переводит точку D в A'.

On isogonal conjugacy, Miquel points, (Newton-)Gauss lines, etc.

N.Beluhov, A.Zaslavsky, P.Kozhevnikov

1 Introductory problems

1. **Isogonal conjugacy.** Given a triangle ABC and a point P.
a)Prove that lines symmetric to AP, BP, CP in the bisectors of corresponding angles are concurrent or parallel. The common point P' of these lines is called *isogonal conjugate* to P with respect to ABC.
b)Prove that P' is a point at infinity (i.e. three corresponding lines are parallel) iff P lies on the circumcircle of ABC.
c)Determine the image isogonal conjugacy of a circle passing through two of three points A, B, C.
d)[1] Prove that all projections of P and P' to the sidelines of ABC are concyclic. Reformulate the statement above for the case when P' is a point at infinity.
e)For two pairs X, X' and Y, Y' of isogonal conjugate points, prove that $XY \cap X'Y'$ and $XY' \cap X'Y$ are isogonal conjugates.

Given a quadrilateral $ABCD$ and a point P.
f)Suppose that three of four lines symmetric to AP, BP, CP, DP in the bisectors of corresponding angles are concurrent. Prove that all four lines are concurrent.
g)Prove that for a point P there exists an isogonal conjugate P' iff projections of P to the sidelines of $ABCD$ are concyclic (if P' exists, then all projections of P and P' to the sidelines of $ABCD$ are concyclic).

A conic is said to be inscribed to a polygon if it touches all the sidelines of this polygon.
h)Prove that foci of a conic inscribed to a triangle are isogonal conjugates.
i)Prove that focus of a parabola inscribed to a triangle lies on its circumcircle.

2. **Miquel point.** Given a quadrilateral $ABCD$. Let $E = AB \cap CD$, $F = AD \cap BC$.
a)Prove that (in notation of the previous problem) circumcircles of triangles ABF, CDF, ADE, CDE have a common point M (Miquel point for a quadruple of lines AB, BC, CD, DA).
b)Prove that M is a center of spiral similarity that takes segment BE to FD (or DE to FB, etc.)
c)Two bugs B and C move, each at a constant speed, along two lines intersecting at A. Prove that all the circles ABC have a common point, and * all the lines BC touch fixed parabola.
d)(IMO2005) Let $ABCD$ be a convex quadrilateral with sides BC and AD equal in length and not parallel. Let E and F be interior points of the sides BC and AD such that $BE = DF$. The lines AC and BD meet at P, the lines BD and EF meet at Q, the lines EF and AC meet at R. Consider all triangles PQR as E and F vary. Prove that the circumcircles of these triangles have a common point other than P.
e)Establish a connection between Miquel point and inscribed conics.
f)Prove that the projections of Miquel point to the sidelines of a quadrilateral lie on a line perpendicular to Gauss line. Establish a connection between this line and a parabola inscribed to the quadrilateral.

3. **Gauss line.** Given a quadrilateral $ABCD$. Let $E = AB \cap CD$, $F = AD \cap BC$.
a)Prove that the midpoints of the segments AC, BD, EF lie on a line (that is called Gauss line of $ABCD$, or Gauss line of quadruple of lines AB, BC, CD, DA).
b)Prove that the center of the circle passing through the projections of a pair of isogonal conjugates lies on Gauss line.
c)Prove that Miquel point is isogonal conjugate to the infinite point of Gauss line.
d)Prove that centers of conics inscribed to a quadrilateral lie on Gauss line.

[1] Here and further we footnotesize the statements that not used in the proofs of results from section 2.

e)(All-Russian Olympiad 2009) Let A_1 and C_1 be points on the sides AB and BC of parallelogram $ABCD$. Let $P = AC_1 \cap CA_1$. Circumcircles of triangles AA_1P and CC_1P meet for the second time at point Q lying inside triangle ACD. Prove that $\angle PDA = \angle QBA$.

2 Three Miquels for a Quartet.

In this section we use the following notation. Let A, B, C, D be four points such that no three of them are collinear. Let X be Miquel point for the quadruple of lines AB, AC, BD, CD, let Y be Miquel point for the quadruple of lines AB, AD, BC, CD, let Z be Miquel point for the quadruple of lines BC, AC, BD, AD. We set $P_X = AD \cap BC$, $P_Y = AC \cap BD$, $P_Z = AB \cap CD$. Let K_X and L_X be midpoints of the segments BC and AD respectively, similarly, let K_Y, L_Y be midpoints of AC, BD, let K_Z, L_Z be midpoints of AB, CD. Let $\Gamma_X = K_XL_X$, $\Gamma_Y = K_YL_Y$, $\Gamma_Z = K_ZL_Z$ be Gauss lines for the corresponding quadruples of lines.

4. Prove that AX, BY, CZ have a common point D', or parallel. Similarly define A', B', C'.

5. Prove that A', B', C', D' are isogonal conjugates to A, B, C, D with respect to triangle XYZ.

6. Prove that X, Y, Z are Miquel points for quadruples of lines joining A', B', C', D'.

7. Probe that lines AA', BB', CC', DD' are parallel.

8. Prove that AD, $A'D'$, YZ are concurrent (find other analogous intersections).

9.

a)Prove that points X, Z, P_Y, K_Y, K_Y lie on a certain circle ω_Y. Similarly define circles ω_X, ω_Z.

b)Prove that ω_X, ω_Y, ω_Z have a common point T.

c)Prove that XP_X, YP_Y, ZP_Z meet at T.

3 Quartets for three Miquels.

Let XYZ be a triangle. Define a transformation ψ_X as the symmetry in the bisector of angle X followed by the inversion with center X and radius $R = \sqrt{XY \cdot XZ}$. Similarly define transformations ψ_Y, ψ_Z.

10. Prove that

a)$\psi_X(Y) = Z$, $\psi_X(Z) = Y$;

b)ψ_X^2 is the identity transformation;

c)Product $\psi_Z\psi_Y\psi_X$ is the identity transformation.

Let D be an arbitrary point, let $A = \psi_X(D)$, $B = \psi_Y(D)$, $C = \psi_Z(D)$.

11. Prove that $\triangle XDZ \sim \triangle XYA$ and $\triangle XDY \sim \triangle XZA$.

12. Prove that each of the transformations ψ_X, ψ_Y, ψ_Z takes the 4-element set $\{A, B, C, D\}$ to itself. A 4-element set of points $\{A, B, C, D\}$ defined as above is said to be a *quartet*. From the previous problem it follows that all the plane except X, Y, Z is partitioned into quartets.

13. Prove that four isogonal conjugates to points of a quartet is a quartet.

14. Find all the quartets containing

a)the incenter I of triangle XYZ;

b)the circumcenter O of triangle XYZ.

c)Find the invariant points for ψ_Z, and corresponding quartets.

15.

a)Prove that X is Miquel point for the quadruple of lines AB, AC, BD, CD.

b)Formulate similar statements for Y, Z.

c)Prove the converse: if X, Y, Z are Miquel points defined by A, B, C, D, then A, B, C, D us a quartet (for X, Y, Z).

16. Prove that each of transformations ψ_X, ψ_Y, ψ_Z commutes with the isogonal conjugacy with respect to XYZ.

17. Suppose A, B, C, D be a quartet with respect to XYZ, let A', B', C', D' be isogonal conjugates to A, B, C, D respectively. Consider four conics having pairs of foci A and A', B and B', C and C', D and D'.

a)Prove that these conics are homothetic to each other.

b)Prove that midpoints of six segments joining centers of these conics lie on a certain conic that is homothetic to them and passing through X, Y, Z.

18. Let M, N be a pair of isogonal conjugates with respect to triangle ABC lying inside ABC. It appears that $AM \cdot AN \cdot BC = BM \cdot BN \cdot AC = CM \cdot CN \cdot AB = k$.

a)Prove that the midpoint of MN is the gravity center of A, B, C.

b)Find k in terms of side lengths of ABC.

4 Additional problems.

19.
a)Let A, B, C, D be a quartet, A', B', C', D' be conjugated quartet; let P_X be intersection point of AD and BC, P_Y — of AC and BD, P_Z — of AB and CD. Points Q_X, Q_Y, Q_Z are defined similarly by points A', B', C', D'. Prove that lines P_XQ_X, P_YQ_Y, P_ZQ_Z are concurent in the point, which lie on the circumcircle of triangle XYZ (notations as above).
b)In previous notations prove that lines P_XQ_Y, P_YQ_X and XY concur.
c)Let Z' be the point obtained in b). Prove that line ZZ' is parallel to AA', BB', CC', DD'.
d)Let D_1, D_1' and D_2, D_2' be two pairs of isogonally conjugated points such that $D_1D_1' \parallel D_2D_2'$. Prove that lines $A_1A_2, B_1B_2, C_1C_2, D_1D_2$ concur (A_1, B_1, C_1, D_1 and A_2, B_2, C_2, D_2 are quartets).
20. Given points A, B, C, D. It is known that triangle XYZ is perspective to each of triangles ABC, BCD, CDA, DAB (with indicated order of vertices). Points D', A', B', C' are respective centers of perspective. Prove that lines AA', BB', CC', DD' concur.

On isogonal conjugacy, Miquel points, (Newton-)Gauss lines, etc. Solutions.

N.Beluhov, A.Zaslavsky, P.Kozhevnikov

1 Introductory problems

1.

a)Follows from sine Ceva theorem.

b)Proof by counting angles.

c)Let P be a point of a given circle. Then measure of angle APB is fixed. Hence the sum of measures of angles CAP and CBP is fixed. Therefore, measure of angle $AP'B$ is fixed, hence P' lies on a fixed circle passing through A and B.

d)By P_A, P_B, P_C denote points symmetric to P in BC, CA, AB, respectively. Since $P_AC = PC = P_BC$, the perpendicular bisector of the segment P_AP_B passes through C, and hence it is the bisector CP' of angle P_ACP_B. Hence P' is the circumcenter of triangle $P_AP_BP_C$. By homothety with center P and ratio $1/2$, the midpoint T of PP' is the center of the circle passing through projections of P to the sidelines. Similarly, T is the center of the circle passing through projections of P'. These two circles coincide since T is equidistant from projections of P and P' to a centain line.

In the case when P' is a point at infinity we obtain the Theorem on Simson line. for a point lying on the circumcircle, its projections of to the sidelines are collinear.

e)Let the common point of XY' and YX' be P and the common point of XY and $X'Y'$ be Q. Let $XA \cap YX' = X_A$, and $Y'A \cap YX' = Y_A$. Also call as Q_A intersection point of QA and YX'. Prove that lines AQ and AP are symmetric with respect to the bisector of angle A. Since this is true for all angles points P and Q are isogonaly conjugated.

Consider cross-ratio (Y, Q_A, X_A, X'). Project these points from Q to line AX'. The map of Y is X', the map of Q_A is A, X_A transforms to itself and X' transforms to the common point of lines $X'Y'$ and XA. Now project obtained ratio from Y' to line YX'. The map of A is Y_A, the map of X is P, X_A transforms to itself and the map of X' transforms back to X'. Thus $(Y, Q_A, X_A, X') = (P, Y_A, X_A, X') = (Y_A, P, X', X_A)$. It means that AQ is the reflection of AP in the bisector of angle A.

Another proof see also in the book of A.Akopyan and A.Zaslavsky "Geometry of conics" AMS, 2007.

f)By P' denote the point of intersection. Let AB and CD meet at K. For one of triangles KBC and KDA, P' is isogonal conjugate to P. Hence, P lies on the line symmetric to KP in the bisector of angle K. For the other of two triangles KBC and KDA, P' is also isogonal conjugate to P. We obtain that P lies on all four lines symmetric to AP, BP, CP, and DP, in the bisectors of corresponding angles.

g)Suppose that projections of P are concyclic. Then, similarly to p.d), obtain that point P' symmetric to P in the center of the circle is isogonal conjugate to P. The converse statement is proved anagolously.

h)See the book mentioned above.

i)By the statement converse to the Theorem on Simson line, it is sufficient to prove that the projections of parabola focus to the sidelines are collinear. Since a point symmetric to the focus in any tangent lies on a directrix, the required statement follows.

2.

a)Let the circumcircles of ABF and CDF meet for the second time at M. We have $\angle(AM, MD) = \angle(AM, MF) + \angle(MF, MD) = \angle(BA, BF) + \angle(CF, CD) = \angle(BA, CD) =$

$\angle(AE, ED)$, hence M lies on the circumcircle of ADE. Similarly for the other triangle.
b)It is sufficient to prove that triangles MBE and MFD are similar. We have $\angle(EB, BM) = \angle(CE, CM) = \angle(CD, CM) = \angle(FD, FM)$. Angles MEB and MDF are equal by the same reasoning.
c)Let B' and C' are the positions of bugs at some moment. The intersection point P of the circumcircles of triangles ABC and $AB'C'$ is the center of spiral similarity taking line BB' to CC' (ratio of this similarity is equal to the ratio of the speeds of the bugs). Hence P is Miquel point for quadruple of lines $B'B$, BC, CC', $C'B'$.
d)Suppose E moves from B to C at some constant speed, while F moves from D to A at the same speed. The condition holds at any moment. We show that each of points Q and R moves at a constant speed, so statement of the problem follows from the previous problem. Note that $\angle EQB = \angle FQD$, $EB = FD$. Hence the circumcircles of triangles EBQ FDQ are equal, and ratio EQ/QF is constant. Therefore Q moves at a constant speed. Similarly, for R.
e)Miquel point is the focus of a parabola inscribed ro the quadrilateral.
f)By Therem on Simson line, eah three of four projections are collinear. Hence all four projections are collinear. It is known that homothety with center at some point of the circumcircle and ratio 2 takes Simson line to the line passing through the orthocenter. Therefore, by homothety with center at Miquel point and ratio 2 we obtain the line l passing through orthocenters of four triangles. The orthocenters have equal powers with respect to three circles constructed on the diagonals of quadrilateral, hence l is the radical axis for these three circles that is perpendicular to Gauss line (passing through the centers of the circles). For the inscribed parabola, the focus lies on the circumcircles of the triangles. Hence, Miquel point is the focus. The projections of Miquel point lie on a tangent to the parabola at its vertex. The line passing through the orthocenters is the directrix of the parabola. Hence, Gauss line is parallel to the axis od parabola.

3.
a)By M denote the midpoint of AC, N — midpoint of BD, T — midpoint of EF. Let F', A' and B' are the midpoints of the sides of ABF. Note that M lies on $F'B'$, N lies on $F'A'$, T lies on $A'B'$. By homotheties with centers at vertices of ABF and ratio 2, $\frac{\overrightarrow{F'M}}{\overrightarrow{MB'}} = \frac{\overrightarrow{BC}}{\overrightarrow{CF}}$, $\frac{\overrightarrow{B'T}}{\overrightarrow{TA'}} = \frac{\overrightarrow{AE}}{\overrightarrow{EC}}$, $\frac{\overrightarrow{A'N}}{\overrightarrow{NF'}} = \frac{\overrightarrow{FD}}{\overrightarrow{DA}}$. Multiplying these three equalities, and applying Menelaus theorem, we obtain the required.
b)Hint. Gauss line of quadrilateral $ABCD$ is the locus of points X such that $S_{XAB} + S_{XCD} = S_{XBC} + S_{XDA}$ (here the areas are oriented).
c)Follows from 2f.
d)This is reformulation of b).

2 Three Miquels for a Quartet.

4. From Problems 13, 15 it follows that AX, BY, CZ pass through D' (isogonal conjugate to D with respect to triangle XYZ).

5. Follows directly from Problem 13.

6. Follows directly from Problems 13, 15.

7. By problem 11, $\triangle XDZ \sim \triangle XYA$, $\triangle XD'Z \sim \triangle XYA'$. Hence $XA : XD' = (XA : XZ)(XZ : XD') = (XY : XD)(XA' : XY) = XA' : XD$, that equivalent to the statement of the problem.

8. Follows from the previous problem and Theorem on three homotheties applied to the segments AA', DD', and $B'B$. Indeed, Z is the center of homothety that takes A to B', A' to B, etc. Alternative solution could be easily derived from Problem 1e).

9.

a)X is a center of a spiral similarity that takes C to D, A to B. Since K_Y is the image of L_Y under this spiral similarity, angle K_YXL_Y equals to $\angle(AC, BD)$. Hence X lies on the circle $P_YK_YL_Y$. Similarly obtain that Z lies on the same circle.

b)Since X lies on the circle AP_YB, we have $\angle XP_YB = \angle XAB$. Similarly, $\angle BP_YZ = \angle BCZ$. From these and four analogous equalities it follows that $\angle XP_YZ + \angle ZP_XY + \angle YP_ZX = \pi$, that implies the statement of the problem.

c)From problem 15 it is clear that $\psi_X(P_Y) = P_Z$, etc. This means that P_X, P_Y, P_Z belong to a certain quartet, so $\psi_X(P_X) = \psi_Y(P_Y) = \psi_Z(P_Z)$. We need to show that $\psi_Z(P_Z)$ is isogonal conjugate to T, or equivalently, $\psi_Z(T)$ is isogonal conjugate to P_Z. Note that ψ_Z takes circles ZXP_Y and ZYP_X (passing through T) to lines YP_X and XP_Y, hence $\psi_Z(T) = YP_X \cap XP_Y$. From equalities $\angle P_XYX = \angle ZYP_Z$, $\angle P_YXY = \angle ZXP_Z$ obtain the required statement.

3 Quartets for three Miquels.

10. a)-b) Follows directly from the definition.

c) Each of the transformations ψ_X, ψ_Y, ψ_Z is *circular* (i.e. takes a circle either to a circle or to a line) and preserves the orientation. Hence any product of ψ_Z, ψ_Y, ψ_X is a circular transformation preserving the orientation. Moreover, from a)-b) follows X, Y, Z are invariant points for $\psi_Z\psi_Y\psi_X$. Note that a circular transformation preserving the orientation is uniquely defined by the images of three points. The statement of the problems is independent of the order of ψ_X, ψ_Y, ψ_Z in its product. Hence ψ_X, ψ_Y, ψ_Z commute to each other.

11. By definition of ψ_X, $\angle ZXD = \angle AXY$ and $XD \cdot XA = XY \cdot XZ$, that implies the first similarity. The second similarity is proved analogously.

12. From Problem 10 it follows, in particular, that $\psi_X \circ \psi_Y = \psi_Z^{-1} = \psi_Z$. Therefore, $\psi_Y(A) = \psi_Y(\psi_X(D)) = \psi_Z(D) = C$, so ψ_Y interchanges the points in the pairs (A, C), (B, D). Similarly we obtain that ψ_X interchanges the points in the pairs (A, D), (B, C), while ψ_Z interchanges the points in the pairs (A, B), (C, D).

13. Let D', A' be isogonal conjugates to D, A respectively. Then A' lies on XD, D' lies on XA. Moreover, $\angle XD'Z = \pi - \angle ZXD' - \angle D'ZX = \pi - \angle DXY - \angle YZD = \angle ZDX + \angle XYZ - \pi$. By Problem 11 we have $\angle ZDX = \angle AYX$, i.e. $\angle XD'Z = \angle XYA'$. Hence $\triangle XD'Z \sim \triangle XA'Y$, and $A' = \psi_X(D')$.

14.

a)The incenter and the excenters of XYZ.

b)Point O and three points symmetric to X, Y, Z in the opposite sidelines of XYZ.

c)From the definition of ψ_Z it follows that its invariant point lie on bisector of angle XZY, and on the circle with center Z and radius $\sqrt{ZX \cdot ZY}$. There are two such points U and V. From Problem 10 we obtain that $\psi_X(U) = \psi_Y(\psi_Z(U)) = \psi_Y(U) = \psi_Z(\psi_X(U))$, i.e. $\psi_X(U)$ is also an invariant point for ψ_Z. It is clear that $\psi_X(U) \neq U$, therefore, $\psi_X(U) = \psi_Y(U) = V$, and a required quartet (both for U and for V) is U, U, V, V. Further, the isogonal conjugates to U, V are also invariant under ψ_Z. Hence U and V are isogonal conjugates to each other, and quartet U, U, V, V coincides to conjugate quartet.

15.

a)From the definition of ψ_X and the Problem 12 it follows that X is the center of spiral similarity taking A to B, and C to D. By Problem 2b) this center is Miquel point.

b)Y is Miquel point for the quadruple of lines AB, BC, AD, CD; Z is Miquel point for the quadruple AD, AC, BD, BC.

c)X is Miquel point, hence (in particular) $\triangle XBD \sim \triangle XAC$. Let $P_X = AD \cap BC$, $P_Y = AC \cap BD$, $P_Z = AB \cap CD$. We have $XA \cdot XD = XB \cdot XC = XP_Y \cdot XP_Z = R_a^2$, and the angles AXD, BXC, P_YXP_Z have a common bisector l. The inversion with center X and radius R_a followed by the symmetry in l takes triangles ADP_Y and BCP_Y to DAP_Z and CBP_Z, respectively. Therefore it tales intersection point Y of the circles ADP_Y and BCP_Y to Z. Hence the product of an inversion and a symmetry defined above is ψ_X.

16. From Problem 13 it follows that the product of ψ_X and the isogonal conjugacy (in any order) takes D to A'.

1

Раскраски и кластеры

А. Белов-Канель,
И. Иванов-Погодаев, А. Малистов, М. Харитонов

1. Плоскость раскрашена **a)** в два цвета **b)** в три цвета. Докажите, что найдутся две точки одного цвета, расстояние между которыми 1.

2. Тот же вопрос для пространства, раскрашенного в 4 цвета.

3. Тот же вопрос для n-мерного пространства, раскрашенного в $n+1$ цвет.

4. Плоскость раскрашена в два цвета. Докажите, что в одном из цветов укладываются все расстояния.

5. Тот же вопрос для n-мерного пространства, раскрашенного в n цветов.

6.* Тот же вопрос для плоскости, раскрашенной в 3 цвета.

7.* Тот же вопрос для n-мерного пространства раскрашенного в n+1 цвет.

8. Раскрасьте плоскость в возможно меньшее число цветов, чтобы не было единичного отрезка с одноцветными концами.

К настоящему времени неизвестно минимальное число x цветов, такое что при некоторой раскраске плоскости в x цветов нет отрезка единичной длины с одноцветными вершинами. Известно только, что $4 \leqslant x \leqslant 7$.

Задача упрощается, если искать «почти единичные» отрезки.

9. Плоскость раскрашена **a)** в четыре цвета **b)** в пять цветов. Докажите, что найдутся две точки одного цвета, расстояние между которыми отличается от единицы не более чем на 0,001.

Решение пункта b предыдущей задачи основывается на следующем факте:

10. Клетки плоскости раскрашены в два цвета (общая граница клеток разных цветов считается пестрой, каждая клетка раскрашена полностью в один цвет). Докажите, что найдется одноцветная ломаная (все точки покрашены в один цвет), концы которой находятся на расстоянии больше 1000.

11. Пространственное обобщение предыдущей задачи на пространственную трехмерную решетку, клетки которой покрашены в три цвета

12. То же для n-мерной решетки, покрашенной в n цветов.

13. Куб $k \times k \times k$ разбит на k^3 единичных кубиков, каждый из которых покрашен в красный, синий или зеленый цвет. Докажите, что найдется одноцветная ломаная, соединяющая противоположные грани. Сформулируйте и докажите n-мерное обобщение. Докажите также, что при увеличении количества цветов на единицу утверждение становится неверным.

14. Трехмерное пространство покрашено в 9 цветов. Докажите, что найдутся две одноцветные точки, расстояние между которыми отличается от единицы меньше, чем на 0,001. Обобщите задачу на n-мерный случай.

Определение. Назовем *кластером* множество связанных клеток. Две клетки, имеющие общую точку, считаются *связанными.*

Результаты задачи 12 поддаются дальнейшему обобщению.

15. Пусть все клетки единичной n-мерной решетки покрашены в k цветов ($k < n+1$). Тогда в кубе с ребром $10M$ найдется связный одноцветный кластер объема M^{n+1-k}.

a)* Решите задачу для $k = 2$.

b)* Решите задачу для $k = n$.

c)** Попробуйте решить задачу в других случаях.

16. Покажите, что утверждение задачи 12 вытекает из следующего факта, который лежит в основе топологического определения размерности: если n-мерное пространство покрыто открытыми множествами ограниченного диаметра, то есть точка, покрытая $n+1$ раз.

17.* Докажите этот топологический факт.

3

Раскраски и кластеры

А. Белов-Канель,
И. Иванов-Погодаев, А. Малистов, М. Харитонов

18. Слой между двумя параллельными прямыми раскрашен в 2 цвета. Докажите, что в нём найдутся 2 одноцветные точки на единичном расстоянии.

19*. Слой между двумя параллельными плоскостями раскрашен в 4 цвета. Докажите, что в нём найдутся 2 одноцветные точки на единичном расстоянии.

20. a) **Лемма Шпернера.** Треугольник, вершины ABC которого раскрашены в цвета 1, 2, 3 соответственно, разбит на треугольники. Вершины этих треугольников раскрашены в цвета 1, 2, 3 так, чтобы точки, лежащие на $[AB]$, были раскрашены в цвета 1, 2; на $[BC]$ — 2, 3; $[CA]$ — 3, 1. Тогда найдётся треугольник, вершины которого раскрашены в разные цвета.

b) Сформулируйте и докажите это утверждение для n-мерного пространства.

21. a) Докажите, что не существует непрерывного отображения диска на свою границу, тождественного на этой границе (т. е. любая точка границы переходит сама в себя). Такие отображения называются *ретрактами*.

b) Докажите **теорему Брауэра о неподвижной точке**: Непрерывное отображение диска в себя имеет неподвижную точку.

c) Сформулируйте и докажите n-мерные аналоги этих фактов.

22. Индуктивное определение размерности. 0-мерное множество: если все его точки лежат в разных компонентах связности. 1-мерное: не 0-мерное, любые 2 точки которого разделяются 0-мерным множеством, и т.д. Докажите, что $\mathbb{R}^n$ согласно этому определению n-мерно.

4

Раскраски и кластеры

М. Матдинов

Вариации на задачи про кластеры

23.* Пусть k-мерный куб со ребром n разбит на n^k маленьких k-мерных кубиков со стороной 1, раскрашенных в ℓ цветов. Рассмотрим все тройки цветов a, b, c. Рассмотрим множество точек, каждая из которых покрашена во все эти цвета. Окружим все эти точки окрестностью радиуса 2 и рассмотрим объединения этих окрестностей и связные компоненты такого объединения. Пусть каждая такая связная компонента для любого цвета имеет диаметр не больше d.

Тогда найдётся константа $C(k, d, \ell) > 0$ такая, что существует кластер объёма $C(k, d, \ell) \cdot n^{k-1}$.

a) Докажите это для $k = 3$,

b) для произвольного k.

24.** Сформулируйте условие, аналогичное условию предыдущей задачи, для наборов из m цветов. Общая гипотеза: существует константа $C(k, m, d, \ell) > 0$ такая, что при раскраске k-мерного куба с ребром n в l цветов найдётся кластер объёма $C(k, m, d, \ell) \cdot n^{k+2-m}$.

Данная задача является своего рода обобщением задачи 15. Её решение нам не известно.

1

Colorings and clusters

A. Belov-Kanel,
I. Ivanov-Pogodaev, A. Malistov, M. Kharitonov

1. Each point of the plane is colored by one of **a)** two colors; **b)** three colors. Prove that there exist two points separated by the distance 1 which have the same color.

2. Prove the same for the space colored by 4 colors.

3. Prove the same for the n-dimensional space colored by $n+1$ colors.

4. The plane is colored by two colors. Prove that one can choose the color such that for any distance d there exist the points separated by the distance d and colored by this color.

5. Prove the same for the n-dimensional space colored by n colors.

6.* Prove the same for the plane colored by 3 colors.

7.* Prove the same for the n-dimensional space colored by $n+1$ colors.

8. Color the plane such that there are no unit segment with the edges colored by the same color. Try to use the minimal number of colors.

So far, it remains an open question to find the minimal number of colors x such for some coloring there are no unit segment with the single-colored edges. It is known that $4 \leq x \leq 7$.

If we look for "almost unit" segments then the problem became more simple.

9. The plane is colored by **a)** four colors; **b)** five colors. Prove that there exist two points of the same color such that the distance between them is different of 1 by less then 0.001.

The solution of the b-part of the previous problem is based on the following fact:

10. Consider the cell-like plane. Suppose that any cell is fully colored by one of two colors. Let the common edge of the two cells be colored by both cell-colors. Prove that there exists a single-colored polygonal path such that the edges of the path are separated for the distance more then 1000.

11. Prove the generalization of the previous problem for 3-dimensional cube lattice colored by 3 colors.

12. Prove the generalization of the previous problem for n-dimensional cube lattice colored by n colors.

13. Suppose that a cube $k \times k \times k$ consists of k^3 unit cubes which are colored by red, blue and green colors. Prove that there exists a single-colored polygonal path which connects opposite faces of the big cube. Formulate and prove the n-dimensional generalization. Prove that if we use $n+1$ colors then the statement is not true.

14. 3-dimensional space is colored by 9 colors. Prove that there exist two single-colored points which are separated by the distance d such that $|d-1| < 0.001$. Generalize this problem for n-dimensional case.

Definition. Connected set of cells called *claster* (Two cells with common point are *connected*).

The achievements of the problem 12 can be further generalized.

15. Suppose that the cells of the n-dimensional lattice are colored by k colors and $k < n+1$. Then in any cube with edge $10M$ there exist connected single-colored cluster with volume M^{n+1-k}.

a)* Solve the problem for $k = 2$.

b)* Solve the problem for $k = n$.

c)** Try to solve the problem for other cases

16. Show that the statement of the problem 12 can be obtained from the following fact, which is base for topological definition of dimension: if n-dimensional space is covered by some sets of bounded diameter then there exists a point covered $n+1$ times.

17.* Prove this topological fact.

3

Colorings and clusters

A. Belov-Kanel,

I. Ivanov-Pogodaev, A. Malistov, M. Kharitonov

There is **a mistake** in the first version of the problem 16. Here is the right version.

16. Show that the statement of the problem 12 can be obtained from the following fact, which is base for topological definition of dimension: if n-dimensional space is covered by some **open** sets of bounded diameter then there exists a point covered $n+1$ times.

18. Suppose the layer between two parallel lines is colored by two colors. Prove that there exist 2 points of the same color inside the layer such that $distance(A, B) = 1$.

19.* The same problem for the layer between two parallel planes, colored by 4 colors.

20. a) **Sperner Lemma.** Triangle ABC is divided by small triangles; vertices of these triangles are colored by 3 colors such that A is colored by color 1, B by color 2, C by color 3. Vertices on $[AB]$ are colored by colors 1 or 2; vertices on $[BC]$ – by colors 2 or 3, vertices on $[CA]$ — by colors 3 or 1. Prove that there exists a small triangle, which vertices are colored by different colors.

b) Generalize this lemma for n-dimensional space (consider also $n = 1$).

21. a) Prove that there is no continuous mapping of disc onto its boundary R such that it is identical on the border. This mapping is called *Retract*.

b) Prove the **Brauer theorem:** Continuous mapping F of the disc into itself has a stable point, i.e. there exists a point x_0 such that $F(x_0) = x_0$.

c) Generalize this for n-dimensional space.

22. Inductive definition of dimension. 0-dimensional space is a space such that any 2 points are situated in different connected components, 1-dimensional space is a space such that any 2 points can be separated by 0-dimensional space (and it is not 0-dimensional). n-dimensional space is a space, such that any two points are separated by $n-1$-dimensional space (and it is not $(n-1)$-dimensional). Prove that $\mathbb{R}^n$ is n-dimensional space according this definition.

4

Colorings and clusters

M. Matdinov

Additional problems

23.* k-dimensional cube $n \times n \times \cdots \times n$ is divided by n^k k-dimensional subcubes $1 \times 1 \times \cdots \times 1$ colored by ℓ colors. Consider all triples of colors (a, b, c). Consider the set of points colored by these 3 colors simultaneously. Surround each of them by circle of radius 2. Consider connected components of the union of these circles. Suppose that all such components for all triples (a, b, c) have diameter less than d.

Then there exists a positive constant $C(k, d, \ell) > 0$ such that for any coloring there exists a cluster of volume $C(k, d, \ell) \cdot n^{k-1}$.

a) Prove that for $k = 3$.

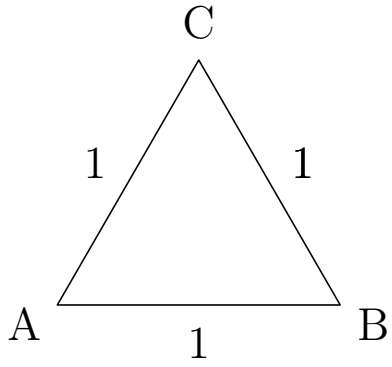
b) Prove that for all k.

24.** Generalize condition of the previous problem for m-tuples of colours. General hypothesis: there exists a constant $C(k, m, d, \ell) > 0$ such that for each coloring of k-dimensional cube $n \times n \times \cdots \times n$ by l colors there exists a cluster of volume $C(k, m, d, \ell) \cdot n^{k+2-m}$.

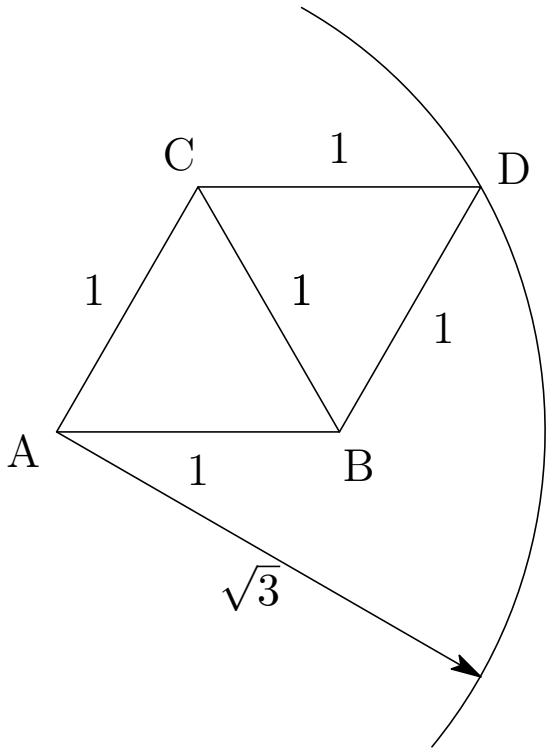
This hypothesis can be considered as generalization of the problem 15. We don't know how to solve it.

Colorings and clusters
Solutions

1a. В равностороннем треугольнике со стороной 1 по принципу Дирихле найдутся 2 вершины одного цвета.



1a. Consider the regular triangle with the side 1. Using Dirichlet principle we can find 2 vertices colored by the same color.



1b. Предположим обратное. Тогда рассмотрим точку A цвета 1. Если такой не найдётся, задача сводится к 1а. Докажем, что на расстоянии $\sqrt{3}$ от A все точки цвета 1. Рассмотрим произвольный равносторонний треугольник ABC со стороной 1. Точки B и C раскрашены в цвета 2 и 3. Рассмотрим равносторонний треугольник BCD. Точка D покрашена в первый цвет. Расстояние между A и D равно $\sqrt{3}$. Повторяя такую опе-

рацию для произвольных точек получим, что все точки на расстоянии $\sqrt{3}$ от A покрашены в первый цвет, а т.к. среди них найдутся 2 точки на расстоянии 1, имеем требуемое утверждение.

2. Аналогично 1b, только берём не равносторонние треугольники, а равносторонние тетраэдры.

3. Аналогично 2, только берём не равностороние тетраэдры, а n-мерные симплексы.

4. Предположим обратное. Тогда у цветов 1 и 2 найдутся такие расстояния x и y соответственно, без ограничения общности $x \geqslant y$, что x не укладывается в цвете 1, а y — в 2. Тогда рассмотрим точку A цвета 1. Вокруг неё опишем окружность радиуса x. Все точки этой окружности покрашены в цвет 2. Тогда на ней найдутся две точки цвета 2, расстояние между которыми — y.

5. Предположим противное. Аналогично задаче 4 рассмотрим расстояния $x_1 \geqslant x_2 \geqslant \ldots \geqslant x_n$ такие, что x_1 не укладывается в цвете 1, x_2 — в цвете 2, …, x_n — в цвете n. Далее доказательство проводим по индукции. Если точки цвета 1 нет, то применяем индукционное предположение. Рассматриваем точку 1-го цвета и описываем вокруг неё n-мерную сферу радиуса $r_1 = x_1$. На ней только точки 2-го, …, n-го цветов. Теперь рассмотрим на ней точку 2-го цвета. Если её нет, то задача опять сводится к случаю меньшей размерности. Описываем вокруг этой точки сферу радиуса x_2. В пересечении этих двух сфер получаем сферу меньшей размерности и радиуса r_2, раскрашенной в цвета 3, 4, …, n. Продолжая этот процесс далее получаем точку, которая не может быть покрашена ни в какой из цветов. Несложно доказать, что $d_i > r_i \geqslant \frac{\sqrt{2}}{2} d_i$, откуда следует требуемое утверждение.

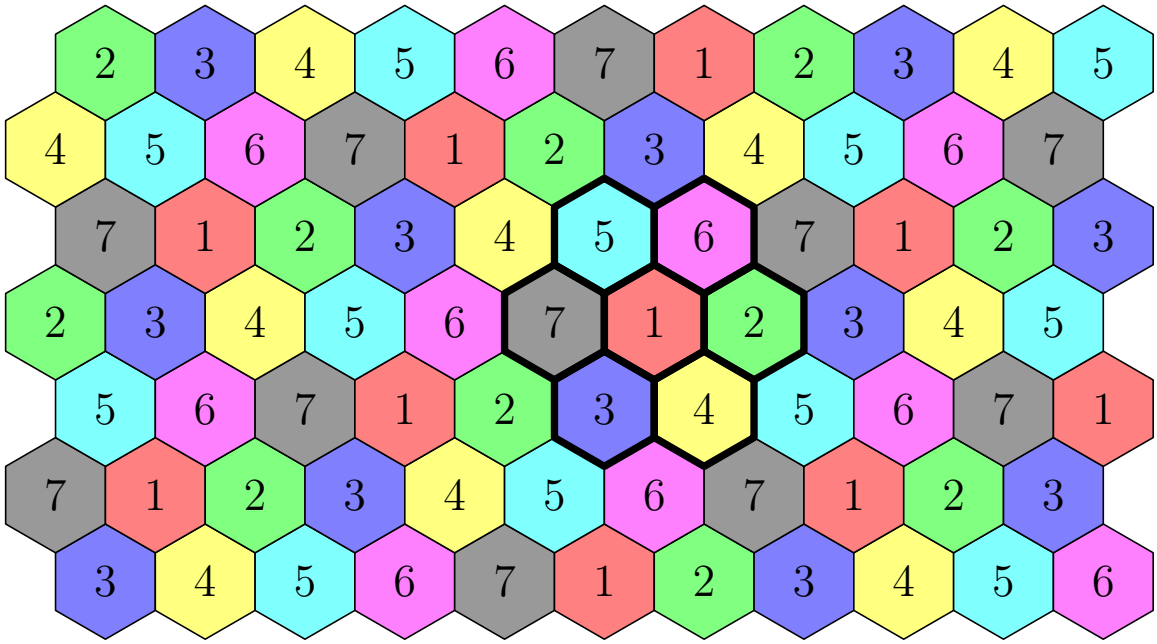
5. We shall act similarly to the problem 4. Assume the contrary. Suppose that there are no two points of color 1 on the distance x_1 from each other, there are no two points of color 2 on the distance x_2 from each other, etc. We can also suppose that $x_1 \geqslant x_2 \geqslant \ldots \geqslant x_n$.

Let us consider a point A_1 of the color 1 and sphere S_1 with center A_1 with the radius $r_1 = x_1$. Next we consider a point $A_2 \in S_1$ of color 2 and sphere $S_2 \subset S_1$ obtained by intersection of S_1 with sphere of radius x_2 centered in the point of A_2. Similarly, we construct point A_3 and $(n-4)$-dimensional sphere S_3 and so on.

If we can not find point on the sphere S_k with color $k+1$, we proceed with next color and distance. Finally we get a point, which can not be colored in any color and get a contradiction.

The only thing we have to take care that process can be continued on the each step, i.e. all spheres will be not empty. This can be guaranteed by proving that $2 \cdot r_i > d_{i+1}$.

8. Раскраска в 7 цветов. Границы шестиугольников покрашены в любой цвет, их диаметр — 0,99: (рисунок)



The coloring of the plane in 7 colors is based on the hexagonal lattice. diameter of each hexagon is 0.99 (see picture).

9b. Предположим противное. Разобьём всю плоскость на квадратики со стороной $\varepsilon/1000$. Если в квадратике найдутся точки 3 цветов, то все точки, расположенные от него на расстоянии от $1 - \varepsilon/2$ до $1 + \varepsilon/2$ раскрашены в 2 цвета, значит среди них найдутся 2 точки одного цвета, расстояние между которыми отличается от единицы не больше чем на ε.

Если каждый квадратик раскрашен не более чем в 2 цвета, то будем считать его квадратиком одного из этих цветов. Без ограничения общности найдётся квадратик 1-го цвета. Рассмотрим кластер максимальной площади из квадратиков 1го цвета. Если с ним по внешней границе граничат кластеры двух цветов, то найдётся квадрат со стороной $\varepsilon/10$, содержащий точки 3 цветов. Далее доказываем аналогично первой части доказательства.

Если этот кластер 1-го цвета граничит только с кластером одного цвета, без ограничения общности 2го, то рассматриваем этот кластер 2-го цвета. Он кроме кластера первоначального цвета граничит только с одним кластером. Продолжая эту операцию, получаем кластер диаметра не меньше 2, а в таком кластере найдутся 2 точки на расстоянии, отличающемся от единицы не больше чем на ε. Противоречие.

9b. Suppose contrary. Let us divide the plane onto squares with the side $\varepsilon/1000$.

Suppose we can find 3 points of 3 different colors a, bc in one of these squares. Let A be center of such square. Consider a circle C of radii 1 with the center A. It is clear that all its point are colored with remaining two colors. If both of them are present (otherwise we can find 2 points of the same color of distance 1), then there are two points in C of different colors arbitrary close to each other. By considering intersection point of C and circle C' of radii 1 centered by one of these point, we get a contradiction.

Suppose there are no such square. Let us color each square in color of an arbitrary point in it. Then no three squares of pairwise different colors meet and we have cluster with arbitrary big diameter, greater than 1. Then we have pair of points which we need.

14. The proof is similar to the proof of problem 9b. We divide a space on small cubes and color each of them in the color of its arbitrary point. If there is no pairwise contacted cubes of 4 different colors, then there exist a cluster diameter grater than 1 and we are done. (This fact can be obtained from dimension theorem quite similar as Lebeg theorem). Next we consider a sphere S of radii 1 centered by center of one of these cubes and continue.

In order to proceed next step we divide equator of S onto 4 equal parts P_1, P_2, P_3, P_4 and consider north hemisphere of S as a "square" with edges P_i. Next we proceed similarly to the problem 9b.

14. Разбиваем пространство на кубики с ребром $\varepsilon/1000$. Раскрашиваем их в один из цветов, которые в них содержатся. Из задачи 16 находим кубик, в котором найдутся точки 4-ёх цветов. Тогда точки на расстоянии от $1-\varepsilon/2$ до $1+\varepsilon/2$ раскрашены в 5 цветов. Далее действуем аналогично задаче 9*b*.

17. *Если n-мерное пространство покрыто открытыми множествами ограниченного диаметра, то есть точка, покрытая $n+1$ раз.*

Указание. Пусть все открытые множества ограничены диаметром d. Рассмотрим n-мерный правильный симплекс со стороной $1000d$.

Присвоим всем вершинам симплекса цвета от 1 до $n+1$. Также присвоим всем открытым множествам цвета от 1 до $n+1$ так, чтобы множества покрывающие вершины имели цвета этих вершин, множества, покрывающие ребра имели цвета одной из двух вершин этих рёбер, множества, покрывающие k-мерные грани имели цвет одной из $k+1$ вершин, ограничивающих грань.

Рассмотрим непрерывное отображение внутренних точек симплекса на его границу. Каждая точка, покрытая один раз переходит в вершину

соответствующего цвета. Точка, покрытая k раз переходит в точку на k-мерной грани, вершины которой раскрашены в соответствующие цвета, при этом точка выбирается как средневзвешенный центр масс с весами, равными расстояниям до границы соответствующего множества.

Такое отображение является ретрактом, а непрерывного ретракта не существует.

17. Note. Suppose that all open sets are bounded by diameter d. Consider n-dimensional simplex with edges equal to $1000d$.

Let us assign the colors from 1 up to $n+1$ for the simplex vertices. Also we assign these colors to the open sets such that the following conditions hold: the sets covering vertices colored by its colors; the sets covering edge colored by one of its vertices colors; the sets covering k-dimensional face colored by one of this face vertices colors.

Consider the continuous mapping from simplex to its boundary. Each point covered by one set is transformed to the vertex colored by the color of this set. Each point covered by k sets is transformed to a point on the k-dimensional face which vertices are colored by the corresponding colors. This point is situated at weighted mass center of these vertices according to the distances to the boundaries of the corresponding sets.

This mapping is a retract. But continuous retract does not exist.

10,11,12, 16. Указание. Все кластеры ограничены, иначе мы можем найти длинный путь в неограниченном кластере. Каждому кластеру, состоящему из кубов, сопоставим открытое множество, состоящее из точек кластера и некоторой ε-окрестности ($\varepsilon = 10^{-3}$). Из топологического факта задачи 17 следует, что существует точка, покрытая $n+1$ множеством. Из принципа Дирихле следует, что существуют две пересекающиеся окрестности кластеров одного цвета. Тогда это должен быть один кластер.

10,11,12, 16. Note. All the clusters are bounded. If not, there exists a long path in some unbounded cluster. For each cluster consider an open set formed by the cubes of the cluster and its ε-neighborhood. Using the fact of problem 17 we obtain that there exists a point covered by $n+1$ sets. Hence this point is covered by two clusters of the same color. But it is impossible.

13. Указание. Пусть существует раскрашенный куб $k \times k \times k$ без сквозного пути. Отразим куб относительно каждой его грани в соответствующие этим граням стороны. Будем отражать получившиеся кубы относительно других граней, заполняя отражениями исходного куба все новые и новые области. Получим заполнение пространства отражениями

нашего куба. Для любого кластера в исходном кубе существуют три соприкасающиеся по вершине грани, которых этот кластер не касается. В виду построенных отражений этот кластер не касается всех граней некоторого куба $2k \times 2k \times 2k$, внутри которого он находится. Таким образом все кластеры ограничены, что противоречит факту задачи 12.

13. Note. Suppose that there exists a colored cube $k \times k \times k$ having no path from some face to the opposite one. Let us reflect the cube using each its face. Then we reflect these reflected cubes again using other faces. So we can fill the space by reflections of our initial cube. For any cluster in the cube there exist three faces having some common vertex such that the cluster does not intersect them. It is clear that our cluster is bounded by some $2k \times 2k \times 2k$ cube. Hence all cluster are bounded. Using the fact from problem 12 we obtain a contradiction.

Задачи о покрытиях и функции роста

А.Толпыго, Б.Френкин, М.Прасолов, И.Богданов

Предисловие.

Тема данного цикла — покрытия фигур однотипными фигурами (как правило — кругами или шарами). Требуется оценить их количество, их суммарную площадь и т. п.

Наиболее трудными здесь, естественно, являются задачи, где требуется дать точную оценку. К счастью, оказывается, что основные приложения таких задач как раз не требуют точного ответа, достаточно знать порядок соответствующей величины. Точно соответствующие определения будут сформулированы в цикле В.

Цикл А.

А1

Нетрудно покрыть единичный квадрат кругом площади $\pi/2$. А можно ли покрыть квадрат несколькими кругами, суммарная площадь которых меньше $\pi/2$? Круги могут пересекаться и выходить за пределы квадрата.

А2

Требуется покрыть единичный квадрат несколькими кругами, радиус каждого из которых равен r. Пусть $N(r)$ — минимальное число кругов, которыми это можно сделать.

Очевидно, если $r \to 0$, то $N(r) \to \infty$.

Найти характер стремления к бесконечности этой функции (как быстро она растет)?

А3

Та же задача, если требуется покрыть единичный куб несколькими (пересекающимися) шарами радиуса r.

Цикл В.

В предыдущих задачах подразумевалось, что всякому интуитивно понятно, что такое «скорость роста». Но дальше нам потребуется точное определения: что такое «рост функции»?

Дать такое определение не очень легко, и оно будет дано чуть ниже. Вначале же мы, еще не давая точного определения «роста функции», сформулируем свойства этого «роста». А именно:

Далее рассматриваются только функции $f(x)$ такие, что
$(*)$ $f(x)$ определена для всех x, больше некоторого a (как правило, $a = 0$, но удобнее разрешить более общий случай — напр. функции типа $\ln(x-1)$). Кроме того, предполагается, что $f(x)$ всюду положительна (во всяком случае, при $x > a$), нестрого возрастает и стремится к бесконечности.

Обозначим рост функции $f(x)$ через $[f]$. В частности, для простоты будем далее обозначать рост функции x^n просто n. Таким образом, $n = [x^n]$.

Мы хотим, чтобы рост функции обладал следующими свойствами:

(1) Если для всех x, больших некоторого b (b не обязательно совпадает с a) выполняется $f(x) > g(x)$, то $[f] \geqslant [g]$.

(2) Пусть A, B, C — произвольные положительные числа, и $g(x) = Cf(Ax + B)$. Тогда $[g] = [f]$.

(3) Отсюда понятно, что вполне возможна ситуация, когда одновременно $[f] \geqslant [g]$ и $[g] \geqslant [f]$. В этом случае мы также полагаем, что $[f] = [g]$.

Если же $[f] \geqslant [g]$, но неверно, что $[g] \geqslant [f]$, то мы пишем: $[f] > [g]$.

Однако до сих пор мы не дали определения: что же такое рост? Правильный ответ состоит в том, что рост как раз и есть нечто, удовлетворяющее всем перечисленным свойствам.

Если говорить более формально, то следует рассмотреть класс функций, удовлетворяющих $(*)$, и разбить его на классы эквивалентности: $f \sim g$, если $[f] = [g]$ (в соответствии с пп. (2), (3)).

Каждый из этих классов и называется ростом всех входящих в него функций. При этом некоторые классы больше других, так что мы можем говорить о том, что такая-то функция растет быстрее (ее рост больше), чем другая.

B1

Докажите, что $1 < 2$.

(Напоминание: 1 и 2 — не числа, а рост функций).

Однако если для обычных чисел всегда верно одно из трех: либо $a > b$, либо $a = b$, либо $a < b$, то для функций это неверно. Существуют «несравнимые функции», т.е. такие функции, f и g, что неверно ни то, что $[f] \geqslant [g]$, ни то, что $[g] \geqslant [f]$.

B2

Найти две несравнимые функции.

B3

Найдите функцию $f(x)$ такую, что $1 < [f] < 3$, но при этом f несравнима с 2.

B4

Найдите функцию $f(x)$ такую, что для любого числа $a > 0$ выполняется неравенство $[f] < a$.

B5

Найдите функцию $f(x)$ такую, что $[f] = [f^2]$.

B6

Существует ли функция $f(x)$ такая, что $[f] = [\ln f]$?

Цикл C.

Этот цикл посвящен задачам на покрытия фигур. При этом основной вопрос будет состоять в следующем: найти порядок роста числа кругов (или шаров), которыми можно покрыть данную фигуру.

Определение. Эпсилон-сетью (ε-сетью) называется набор точек внутри данной фигуры такой, что любая точка фигуры находится на расстоянии не больше ε от одной из выбранных точек. (Иными словами, круги или шары радиуса ε полностью покрывают данную фигуру).

Минимальной ε-сетью называется ε-сеть с минимально возможным числом точек. Обозначим $M(\varepsilon)$ количество точек минимальной ε-сети.

Определение. Дельта-решеткой (δ-решеткой) называется набор точек внутри данной фигуры такой, что любые две точки набора находятся на расстоянии не меньше δ. (Разумеется, числа эпсилон и дельта могут совпадать).

δ-решетка называется максимальной, если она имеет максимально возможное число точек. Обозначим $N(\delta)$ количество точек максимальной δ-решетки.

Обдумайте сами, почему в первом случае надо рассматривать минимальную сеть, а во втором — максимальную решетку.

C1

Найдите $M(\varepsilon)$ и $N(\delta)$ для отрезка длины a (и произвольных ε, δ).

C2

Дайте по возможности точную оценку для $M(\varepsilon)$ и $N(\delta)$, если сеть и решетку следует строить в единичном круге; в единичном шаре. Оценку следует дать сверху и снизу, т.е. она должна иметь вид (условно говоря):

$$2/\varepsilon < M(\varepsilon) < 5/\varepsilon.$$

Определение. Размерностью фигуры называется рост $M(\varepsilon)$, как функции аргумента $1/\varepsilon$.

C3

Докажите, что размерность равна росту $N(\varepsilon)$, как функции аргумента $1/\varepsilon$.

Таким образом, **размерность можно определить как с помощью минимальных сетей, так и с помощью максимальных решеток**.

C4

Приведите пример фигуры размерности $3/2$.

C5

Какие еще размерности могут иметь различные фигуры? (Достаточно привести несколько примеров).

C6

Спираль — это линия, целиком лежащая внутри некоторого круга и сходящаяся к центру; ее можно задать в полярных координатах уравнением $r = f(\varphi)$, где f определена при $\varphi \geqslant 0$ и убывает, стремясь к нулю на бесконечности.

Найдите размерность спирали. Зависит ли эта размерность от конкретной функции f, или она всегда одна и та же?

Цикл D.

Уточнение задачи A1.

Какова точная оценка в задаче A1? Иначе говоря: требуется найти число α такое, что (1) квадрат нельзя покрыть кругами суммарной площади меньше α и (2) для любого $\beta > \alpha$ квадрат можно покрыть кругами площади β. (Постарайтесь также выяснить, можно ли его покрыть кругами площади ровно α).

Эта задача подразделяется на три. Для формулировки первой укажем, что одно из решений задачи A1 состоит в следующем: квадрат покрывается сеткой из правильных шестиугольников, и каждый шестиугольник покрывается кругом. Отношение площади круга к площади шестиугольника равно $\gamma = \dfrac{2\pi}{3\sqrt{3}}$. Поскольку площадь сетки немного больше площади квадрата, то суммарная площадь кругов (обозначим её β) будет больше γ, но разность $\beta - \gamma$ можно сделать сколь угодно малой, взяв достаточно малые шестиугольники.

D1

Докажите, что если все радиусы кругов должны быть одинаковыми, то приведённая выше конструкция оптимальна, то есть $\alpha = \gamma$.

D2

Найдите α, если разрешается брать круги двух произвольных радиусов.

D3

Найдите α, если разрешается брать круги произвольных радиусов без всяких ограничений.

(**Предупреждение.** Не думайте, что задачи упорядочены по возрастанию сложности!)

D4–D6

Та же задача, но требуется покрыть куб со стороной 1 шарами (шары, разумеется, могут пересекаться). (Здесь достаточно дать хорошие оценки; точный ответ неизвестен).

D7

Требуется покрыть квадрат несколькими кругами равного радиуса так, чтобы каждая точка была покрыта не менее, чем N кругами. Докажите, что при некотором N суммарную площадь кругов можно сделать меньше, чем $N\gamma$.

D8

Верно ли утверждение задачи D7 при $N = 2$?

Цикл E.

E1

На столе лежит листок бумаги в клеточку. Поверх него положен еще один лист бумаги в клеточку; клетки на обоих листах квадратные и одного размера, но второй лист положен наискось, так что его линии не параллельны линиям первого. Верхний листок прозрачный, и видно, как его линии делят один из квадратов нижнего листка.

На какое максимальное число частей может быть разделен нижний квадрат?

А на какое минимальное?

E2

На столе лежит листок бумаги в клеточку размером 10000×10000. Поверх него положен еще один лист бумаги в клеточку размером 1000×2000; клетки на обоих листах квадратные и одного размера. Верхний листок прозрачный, и видно, как линии нижнего листка делят квадраты верхнего листка на части; таким образом, мы видим, что число частей верхнего листка больше двух миллионов.

Докажите, что это число меньше десяти миллионов.

Цикл F.

Возвращаемся к задачам того же типа, что в цикле C. Однако теперь рассмотрим несколько иную конструкцию. Число ε мы будем считать фиксированным (напр. $\varepsilon = 1$), но будем рассматривать неограниченную фигуру Ф — например, всю плоскость, полуплоскость, и т.п.

Рассмотрим произвольную точку O данной фигуры и фигуру $\Phi(R)$, состоящую из всех точек, принадлежащих данной фигуре и отстоящих от O на расстояние не более R (пересечение данной фигуры с кругом или, может быть, шаром с центром в данной точке).

Обозначим через $N(R)$ число точек минимальной ε-сети в фигуре $\Phi(R)$. Пусть χ — скорость роста функции $N(R)$. Если эта функция не стремится к бесконечности при $R \to \infty$, положим условно $\chi = 0$.

χ называется *объёмной характеристикой* данной фигуры.

F1

У каких фигур объемная характеристика равна нулю?

F2

Найти объемную характеристику следующих фигур: (а) плоскости, (б) полуплоскости, (в) полосы, заключенной между двумя параллельными прямыми.

F3

Докажите, что объемная характеристика фигуры не зависит от того, как выбрана точка O.

F4

Даны два положительных числа ε и δ. Верно ли, что для любой фигуры Φ её объёмная характеристика, вычисленная с помощью кругов радиуса ε и радиуса δ, одна и та же?

F5

Найти объемную характеристику внутренности параболы и внешней части параболы.

F6

Найти объемную характеристику каждой из трех частей, на которые гипербола $xy = 1$ делит плоскость.

F7

Какие числовые значения может принимать объемная характеристика фигуры (фигуры на плоскости, фигуры в пространстве)?

F8

Может ли объемная характеристика фигуры не быть числом (т.е. быть функцией, которая не сравнима с числами); например, может ли она, подобно функции из задачи B4, быть больше 1, меньше 2 и притом несравнима с числом 3/2?

Решения

A1

Ответ. Можно.

Один из способов состоит в том, чтобы покрыть квадрат одним кругом радиуса $\frac{\sqrt{2}}{2}$, затем немного уменьшить этот радиус. Останется 4 не покрытых уголка, которые можно покрыть четырьмя маленькими кругами.

Другой способ приведен в комментариях к циклу задач D.

A2, A3

В первом случае функция $N(r)$ растет как квадрат (пользуясь терминологией цикла B, $[N(r)] = 2$), во втором случае — как куб.

B1

Каковы бы ни были A, B, C, при больших x всегда $x^2 > C(Ax + B)$. Таким образом, $[x^2] \geqslant [x]$, но неверно, что $[x] \geqslant [x^2]$. Это и значит, что $[2] > [1]$.

B2, B3

Искомую функцию можно построить, например, так. Построим последовательность быстро увеличивающихся интервалов (нам подойдут, например, интервалы: $[2, 4]$, $[4, 16]$, $[16, 256]$, $\ldots$, $[2^{2^n}, 2^{2^{n+1}}]$, $\ldots$), и рассмотрим вначале функцию $g(x)$, которая равна $x^{\frac{3}{2}}$ на нечетных по счету отрезках, и равна $x^{\frac{5}{2}}$ на четных.

Достаточно очевидно, что такая функция растет быстрее, чем x, но медленнее, чем x^3. Главный ее недостаток в том, что она разрывна (в точках $4, 16, \ldots$) и беда не в том, что разрывна, а в том, что из-за этого она не монотонна.

Однако это легко поправимо. Положим функцию $f(x)$ на r-м по счету выделенном отрезке равной $g(x) + A_r$, где числа A_r подбираются так, чтобы функция стала непрерывной. Конкретно положим $A_1 = 0, A_2 = 4^{\frac{3}{2}} - 4^{\frac{5}{2}} = -56$, и т. д.

Получающаяся функция удовлетворяет требованию задачи. Действительно, её график по-прежнему лежит между графиками функций $x^{3/2}$ и $x^{5/2}$. Покажем, что она несравнима с x^2. Рассмотрим отрезок $[d, d^2]$, где $d = 2^{2^{2n}}$. тогда $f(d) \leqslant d^{5/2}$, откуда $f(d^2) \leqslant d^{5/2} + \left((d^2)^{5/2} - d^{3/2}\right) \leqslant d^3 + d^{5/2} \leqslant 2(d^2)^{3/2}$. Значит, если $[f] \geqslant a$, то $a \geqslant 3/2$. Аналогично, рассматривая отрезок $[d, d^2]$, где $d = 2^{2^{2n+1}}$, получаем $f(d^2) \geqslant \frac{1}{2}(d^2)^{5/2}$, откуда $[f] \leqslant a$ лишь при $a \leqslant 3/2$.

Соответственно, построенная функция и функция x^2 дают решение задачи B2.

B4

Годится, например, функция $f(x) = \ln x$.

B5

Годится, например, функция $f(x) = 2^x$. В самом деле, тогда $f^2 = 4^x = f(2x)$, и согласно определению, это значит, что рост функций одинаков.

B6

Ответ. Да, существует.

Достаточно, например, чтобы выполнялось соотношение $\ln f(x) = f(x/2)$, или, что то же самое, $f(2x) = \exp(f(x))$. Для этого рассмотрим произвольную возрастающую функцию $f(x)$ на отрезке $[1, 2]$ так, чтобы $f(1) = 1 < f(2) = e = \exp(f(1))$. Тогда наше соотношение однозначно задаёт функцию на отрезках $[2, 4]$, $[4, 8]$, $\ldots$. Ясно, что полученная функция — требуемая.

C1

Ответ. $M(\varepsilon)$ равно целому числу, ближайшему к $a/2\varepsilon$, которое не меньше его. $N(\delta) = [a/\delta + 1]$.

C2

В задаче не требуется дать наилучшую оценку, поэтому она не имеет определенного ответа.

Вот одна из оценок.

(a) Для $M(\varepsilon)$.

Поскольку круги радиуса ε должны покрыть весь единичный круг, то их суммарная площадь должна быть больше, чем площадь круга. Соответственно, $M(\varepsilon) > \left(\frac{1}{\varepsilon}\right)^2$ (соответственно, $M(\varepsilon) > \left(\frac{1}{\varepsilon}\right)^3$ для куба).

С другой стороны, используя конструкцию из правильных 6-угольников, приведённую перед условием D1, мы легко убеждаемся в том, что можно покрыть весь круг (с небольшим «заходом наружу») правильными 6-угольниками. Покрыв эти 6-угольники кругами, мы получаем (при достаточно малых ε) оценку $M(\varepsilon) < A \cdot \left(\frac{1}{\varepsilon}\right)^2 \cdot \frac{2\pi}{3\sqrt{3}}$, где A — произвольное число, большее 1.

Для шара первая оценка совершенно аналогична, с тем изменением, что надо брать не квадрат, а куб. Вторую получить так просто не удается. Однако можно, например, покрыть весь шар маленькими кубиками, и затем каждый кубик поместить в шар. Это дает оценку сверху.

(b) Для $N(\delta)$.

Пусть в круге размещено N точек; опишем вокруг каждой круг радиуса δ. Если эти круги не полностью покрывают единичный круг, то заведомо можно поместить еще одну точку. Следовательно, $N(\delta) > \left(\frac{1}{\delta}\right)^2$. Остальные оценки получаются аналогично.

C3

Очевидно, что если δ-решетка максимальна, то круги радиуса δ с центрами в этих точках полностью покрывают фигуру, так что такая решетка является также и δ-сетью, хотя не обязательно максимальной. Во всяком случае, это значит, что $N(\delta) \geqslant M(\delta)$.

С другой стороны, в круге радиуса $\frac{\delta}{2}$ может находиться не более одной точки δ-решетки. Отсюда сразу следует, что если $\varepsilon < \frac{\delta}{2}$, то $M(\varepsilon) \geqslant N(\delta)$.

Таким образом, $M(\frac{\varepsilon}{3}) \geqslant N(\delta) \geqslant M(\delta)$, откуда и следует утверждение задачи.

C4

Искомую фигуру можно построить, в частности, как пересечение бесконечного числа фигур. А именно, пусть Φ_1 — круг радиуса 2 с центром в начале координат.

Далее, Φ_2 есть фигура, вписанная в Φ_1. А именно, Φ_2 есть объединение восьми кругов радиуса $\frac{1}{2}$ (т.е. в 4 раза меньше) с центрами в точках $(0, -\frac{3}{4})$; $(0, -\frac{1}{4})$; $(0, \frac{1}{4})(0, \frac{3}{4})$, и аналогично еще 4 круга с центрами на оси ординат. (Эти круги частично пересекаются.)

Теперь в каждый из этих кругов мы вписываем 8 кругов радиуса $1/8$, расположенных точно так же. Это значит, что их центры лежат на прямых, проходящих через центр очередного круга параллельно одной из осей. Эти 64 круга образуют фигуру Φ_3 и т. д. Обозначим через Φ пересечение всех фигур Φ_i.

Очевидно, получившаяся фигура как раз и может быть покрыта либо одним кругом радиуса 2, либо 8 кругами радиуса $\frac{1}{2}$, и т. д. С другой стороны, все центры кругов радиуса

2^{1-2n} (их количество равно 2^{3n}) образуют 2^{-2n}-решётку и лежат в нашей фигуре. Отсюда и находим её размерность.

C5

Ответ. Реализуются все размерности между 1 и 2. Также размерность может не быть числом. Размерность может быть меньше 1 только для несвязной фигуры.

C6

Ответ. Размерность спирали может быть различной.

В самом деле, допустим сначала, что длина спирали конечна. Так будет, например, если $f(\varphi) = e^{-\varphi}$: в этом случае длина каждого следующего витка спирали пропорциональна длине первого витка, так что длина всей спирали равна сумме убывающей геометрической прогрессии.

Пусть L — длина всей спирали.

Поскольку кусок спирали длины $2r$ заведомо можно покрыть кругом радиуса r, очевидно, что всю спираль можно будет покрыть кругами в количестве $\frac{L}{2r}$, а это означает, что размерность равна 1.

С другой стороны, если спираль плотно покрывает круг, то ее размерность будет больше 1. Но что значит «плотно»?

Для примера проведем внутри круга, в котором лежит спираль, концентрические окружности радиусов $\frac{1}{2}, \frac{1}{3}, \dots \frac{1}{n}, \dots$. Они разделяют круг на концентрические кольца.

Допустим, что в k-м кольце лежит 10^k витков спирали, расположенных равномерно (в этом кольце). Это будет «достаточно плотное» расположение витков, и легко убедиться, что размерность такой спирали больше 1.

Для доказательства сформулируем общее утверждение:

Утверждение Т. Пусть Ф — некоторая фигура, размер которой значительно больше ширины витка спирали. Тогда площадь Ф приблизительно равна произведению ширины витка на длину той части спирали, которая содержится внутри Ф.

Строго говоря, это утверждение весьма неточно, и легко привести к нему контрпримеры. Однако для тех двух фигур, которые нам только и понадобятся, оно верно и достаточно очевидно. Поэтому пока оставим вопрос открытым, и перейдем собственно к доказательству.

Пусть дано некоторое, достаточно малое ε. Выберем k таким образом, чтобы, одновременно выполнялись два неравенства: с одной стороны, $1/k^2 \gg \varepsilon \gg 10^{-k}$, с другой — чтобы ширина витка в k-том по счету концентрическом кольце (обозначим ее δ) была много меньше ε. Поскольку, по предположению, ширина витка меньше чем $1/10^{-k}$, то очевидно, что эти условия вполне совместимы.

Пусть Ф — k-тое кольцо (оно задается неравенством $1/k < r < 1/(k+1)$), и пусть L — длина той части спирали, которая лежит в Ф. Согласно утверждению Т (убедитесь сами, что оно справедливо для любого кольца, если радиус кольца не очень мал), $\pi\left(\frac{1}{k^2} - \frac{1}{(k+1)^2}\right) \approx L\delta$. Согласно тому же утверждению, длина участка спирали, лежащая в круге радиуса ε, приближенно равна $\pi\varepsilon^2/\delta$.

Следовательно, число кругов, лежащих внутри Ф, не может быть меньше (по порядку), чем частное этих двух величин, то есть отношения площади Ф к площади круга.

Это значит, что кругов должно быть столько же (по крайней мере, по порядку), сколько их было бы, если бы они полностью покрывали Ф.

Отсюда нельзя еще сделать вывод, что размерность спирали равна 2: хотя Ф, несомненно, фигура размерности 2, но наше рассуждение проводилось для определенного ε, и с уменьшением ε фигура Ф также уменьшается. Но поскольку k уменьшается намного медленнее, чем ε, во всяком случае, легко убедиться в том, что рост такой спирали больше 1, а только это мы и стремились доказать.

D1–D3

Ответ к задаче D3. Если разрешается брать круги любого радиуса, то ответ: 1. То есть квадрат можно покрыть несколькими кругами, если разрешается, чтобы их суммарная площадь равнялась $1 + \alpha$, как бы мало ни было α.

Задача **D2** (случай двух радиусов) сложнее.

Перейдем к доказательствам. Начнем с задачи **D3**, как более простой.

Очевидно, достаточно доказать следующую лемму:

Лемма. Если можно покрыть квадрат площади 1 кругами суммарной площади $1 + \alpha$, то существует также способ покрыть его кругами суммарной площади $1 + (\alpha/2)$.

Для **доказательства** сначала заметим, что если единичный квадрат можно покрыть кругами суммарной площади меньше $1 + \alpha$, то любую фигуру площади S можно покрыть кругами суммарной площади меньше $S(1 + \alpha)$.

Для доказательства этого вспомогательного утверждения достаточно заметить, что любую фигуру можно «почти точно» покрыть сеткой из мелких квадратиков. Покрыв каждый квадрат нужным образом, мы получим покрытие произвольной фигуры.

Теперь докажем лемму. Для этого мы впишем в единичный квадрат круг, а затем оставшуюся фигуру (ее площадь равна $S = 1 - (\pi/4)$ покроем мелкими квадратиками, с тем, чтобы покрыть ее кругами суммарной площади меньше $S(1 + \alpha)$.

Этого достаточно для доказательства леммы, а с тем доказано и утверждение задачи.

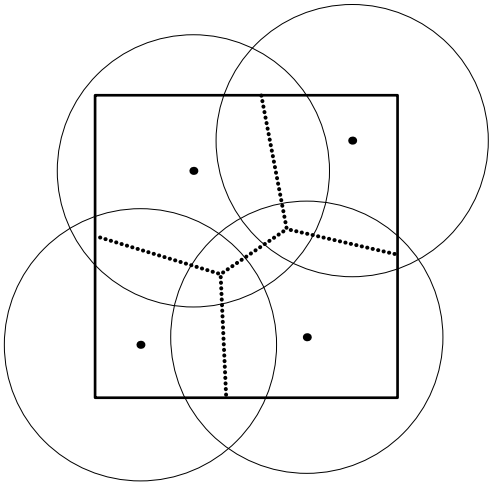
Задача D1.

Лемма. Пусть даны N кругов одинакового радиуса (можно считать, что радиус равен 1), и в каждом круге расположен выпуклый многоугольник, содержащий центр круга, причем количество углов всех многоугольников не превышает $6N$. Тогда суммарная площадь многоугольников не превышает суммарной площади вписанных в те же круги правильных 6-угольников.

Для **доказательства леммы** соединим каждую вершину многоугольника с центром O соответствующего круга; тем самым многоугольник разбит на треугольники. Удвоенная площадь треугольника не больше $\sin\alpha$, где α — угол при вершине O; при этом сумма всех таких углов равна $2\pi N$, а их количество равно $n \leqslant 6N$. Таким образом, требуется оценить сверху выражение $\sin\alpha_1 + \sin\alpha_2 + \ldots + \sin\alpha_n$ при условии $\alpha_1 + \alpha_2 + \cdots + \alpha_n = \pi N$. Можно считать, что $n = 6N$ (если это не так, добавим несколько нулевых углов). Теперь можно, например, воспользоваться тем, что график функции $\sin x$ на отрезке $[0, \pi]$ — выпуклый вверх, и потому максимум достигается, если все слагаемые равны между собой; в этом случае все углы будут равны по $\pi/3$, поэтому у нас и получится сумма площадей правильных шестиугольников.

Вернемся к нашей задаче, причем будем решать ее для произвольного многоугольника T площади 1 с углами, не превосходящими $2\pi/3$. Пусть T покрыт несколькими кругами одного радиуса ε. Тогда T можно разбить на многоугольники по следующему принципу: берем точки, для которых данный центр круга — ближайший (см. рисунок). Поскольку круги одного радиуса, то сторонами получающихся многоугольников являются общие хорды двух пересекающихся кругов (или части этих хорд), и каждый многоугольник R_i целиком лежит внутри соответствующего круга (иначе некоторые точки не были бы покрыты). Более того, центр круга, естественно, лежит в R_i.

Среднее значение внутреннего угла при каждой вершине разбиения не превосходит $2\pi/3$ (это проверяется отдельно для вершин внутри T, точек на границе и углов — в послед-



нем случае как раз и важно, что углы T не превышают $2\pi/3$), откуда легко следует, что число углов не превосходит $6N$. Значит, по лемме суммарная площадь многоугольников (которая равна 1) не больше, чем $N\dfrac{3\sqrt{3}}{2}\varepsilon^2$, то есть суммарная площадь кругов не меньше $N\cdot\pi\varepsilon^2 \geqslant \dfrac{2}{3\sqrt{3}\,\varepsilon^2}\cdot\pi\varepsilon^2 = \dfrac{2\pi}{3\sqrt{3}}$, что и требовалось.

Замечание. Отсюда, в частности, следует, что если T — правильный 6-угольник, то наилучший способ его покрытия — покрыть его одним кругом; все прочие способы дают результат строго хуже (суммарная площадь будет больше).

Заметим ещё, что приведённое доказательство с минимальными изменениями проходит для любого многоугольника с числом сторон, не большим 6.

Задача D2.

Здесь оптимальная конструкция такова.

Во-первых, ясно, что часть квадрата надо заполнить кругами большего радиуса (как именно — будет сказано ниже), а оставшуюся часть — по методу, описанному перед условием задачи D1, т.е. мелкой 6-угольной сеткой.

Во-вторых, из соображений, высказанных выше, ясно, что меньший радиус должен быть как можно меньше. Но и больший радиус тоже должен быть малым; иначе говоря, требуется, чтобы было $1 \gg r_1 \gg r_2$ (чем сильнее они уменьшаются, тем лучше; оптимальное соотношение не достигается, но говоря условно, требуется, чтобы отношения $r_1/1$ и r_2/r_1 оба равнялись нулю).

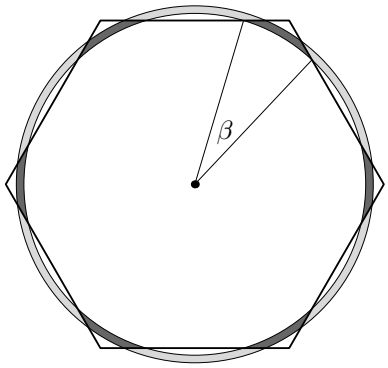
Заполним квадрат мелкой (относительно мелкой; применительно к радиусу r_2 она будет, напротив, очень крупной) 6-угольной сеткой. Затем каждому 6-угольнику сопоставим круг радиуса r_1 с тем же центром.

Таким образом, суммарная площадь всех покрывающих кругов (если пренебречь эффектами, связанными с границей квадрата — а, как мы знаем, это вполне корректно) равна площади кругов радиуса r_1 (их столько же, сколько 6-угольников), плюс площадь оставшихся «уголков», умноженная на γ. Будем называть второе слагаемое **полной** площадью уголков; она в γ раз больше их «настоящей» площади.

Отсюда понятно, что нам достаточно рассматривать покрытие одного 6-угольника, которому соответствует 1 «большой» круг (радиуса r_1) и 6 «уголков».

Пусть боковая сторона каждого 6-угольника равна a (число a можно выбрать произвольно, лишь бы оно было достаточно малым). Должны выполняться неравенства $r_1 = a\frac{\sqrt{3}}{2}$. Это значит, что круг не полностью покрывает соответствующий ему 6-угольник, но притом вылезает за его границу.

Остается найти соотношение между a и r_1, при котором достигается экстремум. Примем вначале $r_1 = a\frac{\sqrt{3}}{2}$, и будем медленно увеличивать этот радиус. Если он увеличивается на δ,



то площадь большого круга увеличилась на площадь кольца радиуса r_1 и ширины δ, т.е. приблизительно на $2\pi r_1\delta$. С другой стороны, площадь «уголков» уменьшилась на $6\beta r_1\delta$, (см. рисунок), соответственно, их полная площадь уменьшилась на $6\gamma\beta r_1\delta$.

Очевидно, суммарная площадь уменьшается, пока первое выражение меньше второго, и начинает расти после того, как они сравняются. Минимум, стало быть, достигается, если они равны, то есть требуется, чтобы выполнялось равенство $2\pi r_1\delta = 6\gamma\beta r_1\delta$. Сокращая, получаем $\beta = \dfrac{2\pi}{6\gamma} = \dfrac{\sqrt{3}}{2}$.

При этом $r_1 = a \cdot \dfrac{\sqrt{3}}{2} \cdot \dfrac{1}{\cos\left(\frac{\pi}{6} - \frac{\beta}{2}\right)}$. Коэффициент, с которым покрыт весь 6-угольник (и, тем самым, также и весь квадрат) нетрудно вычислить, но он имеет несколько «зубодробительный» вид.

Стоит заметить, что тем же способом можно найти оптимальное покрытие, если разрешается брать круги трех разных радиусов, и вообще, любого фиксированного числа k разных радиусов.

В заключение отметим, что приведенное доказательство имеет «лакуну»: не доказано, что центры «больших» кругов следует размещать именно в форме 6-угольной решетки. Жюри в данный момент не имеет четкого доказательства этого факта. Возможно, участники сумеют восполнить этот пробел.

D4–D6

Ответ в задаче D6, фактически, тот же, что в задаче D3: куб можно покрыть несколькими шарами, если разрешается, чтобы их суммарный объем равнялся $1 + a$, как бы мало ни было a.

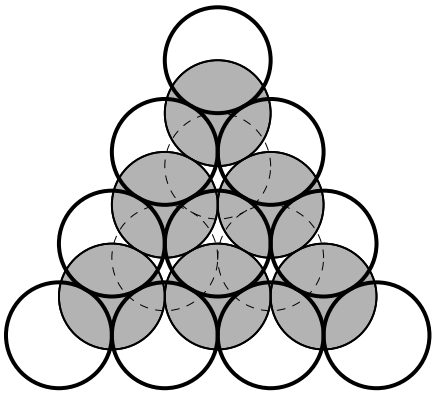
Более того, по сути и решение ее полностью аналогично. Заметим, что тут надо будет воспользоваться тем обстоятельством, что объем шара, вписанного в куб, больше, хоть и ненамного, чем половина объема куба. Если бы это было не так (к примеру, если бы мы занимались 4-мерной геометрией), то доказательство все равно прошло бы, но его пришлось бы немного модифицировать.

В задаче D4, так же, как и в D1, понятно, что радиус шаров должен быть мал, но суть вопроса заключается в том, как должны размещаться центры равных шаров, покрывающих куб.

По аналогии с D1 («плотная упаковка кругов») следует, по всей видимости, разместить эти центры так, чтобы получить «плотную упаковку шаров». Мы сначала предъявим «наиболее плотную» упаковку непересекающихся шаров; после этого останется увеличить все радиусы так, чтобы полученные шары покрыли всё.

Эта плотная упаковка имеет следующий вид: будем размещать шары горизонтальными слоями. Шары нижнего слоя размещаются так же, как в D1, т.е. их центры образуют правильную треугольную решетку. А центры шаров второго слоя располагаются так, чтобы каждый из них образовывал вместе с тремя шарами нижнего слоя правильный тетраэдр

(заметим в скобках, что расположить их так можно двумя разными способами, поскольку «ячеек», в которые можно поместить очередной шар, вдвое больше, чем шаров для этих ячеек). Третий слой помещается поверх второго по тому же принципу, и т. д. (Для наглядности скажем, что если, к примеру, шары первого слоя выложены правильным треугольником по n шаров вдоль стороны, шары второго — треугольником со стороной на единицу меньше и т. д., то в итоге шары будут выложены пирамидой в форме правильного тетраэдра; в нашем случае, впрочем, шары должны образовать не пирамиду, а «приблизительно» куб).



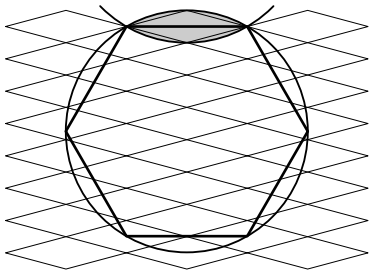
После этого остаётся вычислить радиус увеличенных шаров. Скажем без доказательства, что коэффициент увеличения будет равен отношению диаметра описанной сферы правильного октаэдра к его ребру, т.е. $\sqrt{2}$.

Наконец, в задаче D5 нужно, по образцу задачи D2, разместить центры шаров большего радиуса точно так же, как в D4, подобрать их радиусы так, чтобы заполнить куб не полностью, а оставшуюся часть заполнить шарами малого радиуса по образцу D4.

Однако мы хотим подчеркнуть, что сказанное по поводу задач D4, D5 — не решение, а только правдоподобные рассуждения о том, каким оно должно быть. Напротив, сказанное о задаче D6 — есть исчерпывающее решение, или, точнее, его конспект.

D7

Согласно задаче D1, при оптимальном покрытии квадрата кругами в один слой часть квадрата покрыта дважды. Эта часть состоит из «лунок». Заметим, что несколькими копиями «лунки» можно покрыть весь шестиугольник — например, так, как на рисунке: шестиугольник покрыт ромбами, каждый из которых, в свою очередь, покрывается одной лункой.



Пусть N — число ромбов в этом покрытии; каждый из этих ромбов получается из изначального сдвигом. Соответственно, если N раз сдвинуть наше исходное покрытие, то наш шестиугольник будет полностью покрыт лунками. Точно так же будет покрыт и любой другой 6-угольник покрытия, а значит, и весь квадрат. (Читателю предоставляется разобраться с граничным эффектом самостоятельно.)

Таким образом, N покрытиями мы в действительности покрыли квадрат не N, а $N+1$ раз.

D8

Решение авторам пока неизвестно.

E1

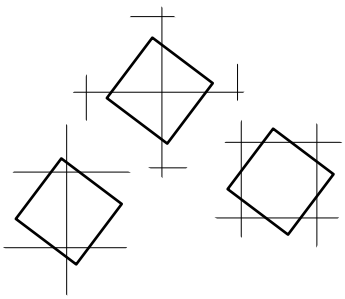
Лемма 1. Число частей равно $1 + a + b + c$, где a, b, c — соответственно число горизонтальных линий, пересекающих данный квадрат, вертикальных линий и узлов сетки.

Доказательство проще всего провести, сначала стерев все линии, а затем восстанавливая их одну за другой.

Лемма 2. В квадрате со стороной 1 нельзя поместить треугольник, у которого основание и высота, на него опущенная, оба не меньше 1 и не параллельны сторонам квадрата.

Доказательство легко следует из того, также элементарного, факта, что в квадрат со стороной 1 не может поместиться треугольник площади больше $1/2$, причем равенство возможно только если основание треугольника совпадает со стороной квадрата.

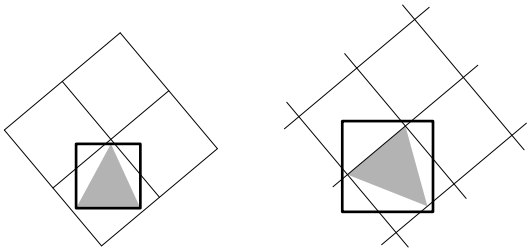
Ответ. число частей не меньше 4 и не больше 6. Примеры для 4, 5 и 6 частей легко нарисовать (см. рис.).



Решение. Для того, чтобы доказать, что частей не может быть меньше или больше, выясним, чему могут быть равны числа a, b, c. Поскольку «ширина» наклонно лежащего квадрата в горизонтальном или вертикальном направлении больше 1, но заведомо меньше 2, ясно, что первые два числа не могут быть меньше 1 или больше 2. Легко также убедиться в том, что c может быть равно 0, 1 или 2.

Допустим, что частей всего 3; из сказанного следует, что это возможно только в случае $a = b = 1$, $c = 0$. Но тогда наш квадрат целиком помещается в трех квадратах, именно так, как показано на рисунке справа, и мы видим, что эта картинка противоречит лемме 2.

Случай семи частей, как ни странно, аналогичен в том смысле, что сводится к той же лемме (см. рис. слева).



E2

Попытаемся приближенно вычислить число частей. Для этого будем считать, что верхний прямоугольник не положен на нижний, а нарисован на нем. Сотрём теперь все его линии и будем их восстанавливать одну за другой. Мы будем считать, что верхний прямоугольник — тот, число частей в котором надо оценить — расположен параллельно осям (его линии вертикальны и горизонтальны), а линии нижнего, по всей вероятности, наклонны.

Вначале нарисуем только верхний прямоугольник размером 1000×2000 без внутренних линий. Он разделен на два миллиона с небольшим частей, поскольку в данный момент части — это квадратики нижнего листка, а на границе — части квадратиков.

Теперь будем проводить одну за другой линии верхнего прямоугольника, сначала горизонтальные, потом вертикальные. Каждая линия дает столько новых частей, на сколько частей она разбивается точками пересечения. Имеется около двух миллионов точек пересечения горизонталей с вертикалями, и кроме того, надо сосчитать, сколько есть точек пересечения новых линий, которые мы рисуем, с линиями нижней сетки.

Пусть наименьший угол между линиями верхней и линиями нижней решеток равен α. Каждая верхняя линия имеет длину 1000 или 2000. Рассмотрим, например, линии длины 1000. Число пересечений такой линии с линиями нижней приближенно равно $1000(\sin\alpha + \cos\alpha)$. Эта величина максимальна, если $\alpha = \pi/4$, и в этом случае она составляет $1000\sqrt{2}$. Случай линий длины 2000 полностью аналогичен, и легко убедиться, что число частей приближенно равно $2 \cdot 10^6 + 2 \cdot 10^6 + 2 \cdot 2 \cdot 10^6\sqrt{2} \approx 2 \cdot 4{,}83 \cdot 10^6$.

Остается проверить, что эффекты, связанные с границей прямоугольника, незначительны, так что коэффициент при миллионе остается меньше 10.

F1

Ответ. У ограниченных фигур и только у них.

F2

Ответ. Для плоскости и полуплоскости объемная характеристика равна 2. Для полосы — 1.

F3

Надо воспользоваться тем, что если расстояние между точками O и O' равно A, то круг радиуса R с центром в точке O целиком лежит в круга радиуса $R + A$ с центром в O', и наоборот.

F4

Ответ. Да, верно.

В самом деле, если $\delta < \varepsilon$, то любой круг радиуса δ покрывается одним кругом радиуса ε, откуда следует, что $N_\delta(R) \geqslant N_\varepsilon(R)$. (Здесь $N_\varepsilon(R)$ и $N_\delta(R)$ как раз и обозначают две функции, определённые с помощью различных радиусов.)

С другой стороны, любой круг радиуса ϵ покрывается определенным количеством A кругов радиуса δ, откуда следует, что $N_\delta(R) \leqslant A \cdot N_\varepsilon(R)$, что и требуется.

F5

Ответ. $\frac{3}{2}$, 2.

F6

Ответ. Все три имеют характеристику 2.

F7, F8

Если фигура неограничена и связна, то ее характеристика не может быть меньше 1. Если она находится в трехмерном пространстве, то ее характеристика не может быть больше характеристики всего пространства, т. е. не больше 3.

В остальном возможно всё.

Problems on coverings and growth functions

A.Tolpygo, B.Frenkin, M.Prasolov, I.Bogdanov

Preface

This cycle is devoted to coverings of arbitrary figures by figures of a prescribed type (usually by disks or balls). The task is to estimate their number, their total area etc.

Naturally, the most difficult are the problems where the precise estimate is required. Fortunately it turns out that basic applications of these problems just don't require an exact answer, and it suffices to know the order of the respective value. The related definitions will be formulated accurately in Part B.

Part A.

A1.

It is not difficult to cover the unit square by a disk of area $\pi/2$. However, is it possible to cover the square by several disks of total area less than $\pi/2$? The disks may intersect and get out of the square.

A2.

Cover the unit square by several disks of the same radius r. Let $N(r)$ be the minimal number of disks sufficient for this.

Clearly $r \to 0$ implies $N(r) \to \infty$.

Determine the type of tending to infinity for this function. In other words: how fast does it grow?

A3.

The same task but it is required to cover the unit cube by several (possibly intersecting) balls of radius r.

Part B.

In the above problems we assumed that the notion of the "growth rate" is intuitively clear. But below we shall need the precise definition for the "growth of a function".

This definition is not quite obvious, and we will give it a bit later. Before that, having no precise definition for the "growth of a function", we will formulate certain properties of this "growth". Specifically:

In the sequel, we consider only functions $f(x)$ such that

(*) $f(x)$ is defined for all x exceeding some a (usually a=0, but for convenience we allow a more general case, for example, functions like $\ln(x-1)$). Moreover we assume that $f(x)$ is positive everywhere (at any rate, for $x>a$), non-strictly increases and tends to infinity.

Denote the growth of $f(x)$ by $[f]$. In the sequel, for simplicity we denote the growth of x^n by n. Thus, $n = [x^n]$.

We require that the growth of a function has the following features:

(1) If for all x exceeding some b (b not necessarily equals a) we have $f(x) > g(x)$ then $[f] \geq [g]$.

(2) Let A, B, C be arbitrary positive numbers, and $g(x) = Cf(Ax+B)$. Then $[g] = [f]$.

(3) This easily implies that a situation is possible when both relations $[f] \geq [g]$ and $[g] \geq [f]$ hold. In this case we set $[f] = [g]$ as well.

If $[f] \geq [g]$ but not $[g] \geq [f]$ then we write $[f] > [g]$.

However we still have not given a definition for growth. The correct answer is that growth is just something subject to the above conditions.

Speaking more formally, we have to consider the class of functions subject to (*) and split it into equivalence classes: $f \sim g$ iff $[f] = [g]$ (according to items (2), (3)).

Each of these classes is called the growth of all functions contained in it. Some of these classes occur to be greater than others, so that we may say that a certain function grows faster (its growth is greater) than another one.

B1.

Prove that $1 < 2$.

(A reminder: 1 and 2 are not numbers but the growth of some functions.)

For ordinary numbers there are three alternatives: either $a>b$, or $a=b$, or $a<b$. However for functions this doesn't hold. There exist "incomparable functions", that is, functions f and g such that both assertions $[f] \geq [g]$ and $[g] \geq [f]$ fail.

B2.

Find two incomparable functions.

B3.

Find a function $f(x)$ such that $1 < [f] < 3$ but f is incomparable with 2.

B4.

Find a function $f(x)$ such that for any $a>0$ we have $[f] < a$.

B5.

Find a function $f(x)$ such that $[f] = [f^2]$.

B6.

Does there exist a function $f(x)$ such that $[f] = [\ln f]$?

Part C.

This part is devoted to coverings of figures. The main task is as follows: determine the growth for the number of disks (or balls) sufficient to cover a given figure.

Definition. An epsilon-net (ε-net) is a set of points inside a given figure, such that the distance from any point of the figure to at least one of the chosen points doesn't exceed ε.
(In other words, disks or balls of radius ε cover the whole figure.)

Definition. A delta-lattice (a δ-lattice) is a set of points inside a given figure, such that the distance between any two points in the set is not less than δ. (Of course, the numbers epsilon and delta may coincide).

An ε-net is called minimal if it contains the minimal possible number of points. Denote the number of points in a minimal ε-net by $M(\varepsilon)$.

A δ-lattice is called maximal if it contains the maximal possible number of points. Denote the number of points in a maximal δ-lattice by $N(\delta)$.

Think on your own why in the first case we have to consider a minimal net, and in the second case a maximal lattice.

C1

Determine $M(\varepsilon)$ and $N(\delta)$ for a segment of length a (and arbitrary ε, δ).

C2

Give an estimate (as sharp as possible one) for $M(\varepsilon)$ and $N(\delta)$ when a net and a lattice are constructed in the unit disk, or in the unit ball. The estimate must be two-sided, that is of the form (for instance)

$2/\varepsilon < M(\varepsilon) < 5/\varepsilon$

Definition. The dimension of a figure is the growth of $M(\varepsilon)$ as of the function in $1/\varepsilon$.

C3

Prove that the dimension equals the growth of $N(\varepsilon)$ as of the function in $1/\varepsilon$.

Thus the dimension can be defined both in terms of minimal nets and of maximal lattices.

C4

Give an example of a figure of dimension 3/2 .

C5

What are other possible values for the dimension of various figures? (It suffices to provide several examples.)

C6

A *helix* is a curve which is located inside some disk and tends to its center; in polar coordinated, it can be defined by the equation $r = f(\varphi)$ where f is defined for $\varphi \geq 0$ and decreases tending to 0 at infinity.

Determine the dimension of a helix. Does this dimension depend on the specific function f, or it is always the same?

After the intermediate final

Part D.

Refinement of Problem A1.

Now, we are interested in a sharp estimate in Problem A1. In other words: we look for a number α such that (1) the unit square cannot be covered by disks of total area less than α, and (2) for any $\beta > \alpha$, this square can be covered by disks of total area β. (Try also to determine whether the square can be covered by disks of total area equal to α.)

We split this problem into three different ones. To formulate the first problem, we recall one of the solutions for problem A1: tile the square by the regular hexagons, and cover each hexagon by a disk. The ratio of areas of a hexagon and a disk is $\gamma = 2\pi / 3\sqrt{3}$. Since the total area of hexagons is larger than the area of the square, the total area of disks β will be greater than γ, but the difference $\beta - \gamma$ can be made arbitrarily small by taking sufficiently small hexagons.

D1.

Suppose that all the radii of the disks should be equal. Prove that the above construction is optimal, that is, $\alpha = \gamma$.

D2.

Find α if it is allowed to use the disks of two distinct radii.

D3.

Find α if it is allowed to use the disks of arbitrary radii without any restrictions.

(**Warning.** The problems are NOT necessarily ordered by the ascending difficulty!)

D4 – D6.

The same problem but the task is to cover the unit cube by balls which of course may intersect. (Here it suffices to give some good estimates; the precise answer is not known.)

D7.

We need to cover the square with some disks of equal radii so that each point is covered by at least N disks. Prove that there exists some N such that it is possible to make this with total area of the disks smaller than $N\gamma$.

D8.

Does the statement from D7 hold for $N = 2$?

Part E.

E1.

A checked sheet of paper lies on a table. Another checked sheet of paper is put on it; the squares on both sheets have the same size but the lines on the second list aren't parallel to those on the first sheet. The upper sheet is transparent, so that we see how its lines dissect some fixed square of the lower sheet.

Determine the maximum and the minimal possible number of parts in a dissection of a lower square.

E2.

A checked sheet of paper of size 10000*10000 lies on a table. Above it, another checked sheet of paper of size 1000*2000 is put; the squares on both sheets have the same size. The upper sheet is transparent, so that we see how the lines of the lower sheet dissect the squares of the upper one. So, the number of parts of the upper list occurs to be greater than two millions.

Prove that this number is less than ten millions.

Part F.

Now we return to the problems of the same type as in cycle B. However, we consider a somewhat different construction. We assume the number ε to be fixed (for instance ε=1) but consider an unbounded figure Φ, for instance, the whole plane, a half-plane etc.

Consider an arbitrary point O of the given figure, and construct the figure $\Phi(R)$ consisting of all points of the given figure whose distance from O does not exceed R (in other words, $\Phi(R)$ is the intersection of the given figure with a disk or a ball with center O and radius R).

Denote by $N(R)$ the number of points in a minimal ε-net in the figure $\Phi(R)$. Let χ be the growth rate of the function $N(R)$; if this function does not tend to infinity as $R \to \infty$, we conventionally say that $\chi = 0$.

The number χ is called *the volume characteristics* of the given figure.

F1.

Which figures have a zero volume characteristics?

F2.

Determine the volume characteristic of the following figures: (a) the plane, (b) a half-plane, (c) a strip bounded by parallel lines.

F3.

Prove that the volume characteristic of a figure does not depend on the choice of point O.

F4.

Given two positive numbers ε and δ. Is it true, that the volume characteristics of a figure Φ calculated using the disks of radius ε is the same as that calculated using the disks of radius ε?

F5.

Determine the volume characteristic of the interior and of the exterior of a parabola.

F6.

Determine the volume characteristic for each of three parts of the plane obtained by its dissection by the hyperbola $xy=1$.

F7.

What are possible numerical values of the volume characteristic of a figure (in the plane or in the space)?

F8.

Can the volume characteristic of a figure be not a number (in other words, can $N(R)$ be a function incomparable with some numbers)? For instance, can it be greater than 1, less than 2 and incomparable with 3/2, like the function from Problem B4?

Problems on coverings and growth functions

A. Tolpygo, B. Frenkin, M. Prasolov, I. Bogdanov

Solutions

A1

Answer. Yes, it is possible.

One of the possible ways is to cover a square by a disk of radius $\frac{\sqrt{2}}{2}$, then decrease this radius a bit, and cover the four regions which are uncovered by four small disks.

A different way is presented in the problem statements before problem D1.

A2, A3

In the first case the growth of $N(r)$ is quadratic (in terms of part B, $[N(r)] = 2$), while in the second case it is cubic.

B1

For any A, B and a sufficiently large x we have $x^2 > Ax + B$. Thus $[x^2] \geq [x]$ but it is false that $[x^2] \leq [x]$. This just means that $2 > 1$.

B2, B3

For instance, the required function can be constructed as follows. Construct a sequence of rapidly increasing intervals (for example, the intervals $[2, 4], [4, 16], [16, 256], \ldots, [2^{2^n}, 2^{2^{n+1}}] \ldots$ would fit), and consider first the function $g(x)$ equal to $x^{\frac{3}{2}}$ on odd intervals, and equal to $x^{\frac{5}{2}}$ on even intervals.

Obviously, this function grows faster than x but slower than x^3. The main its defect is its discontinuity (at points $4, 16, \ldots$) which implies a non-monotonicity. But this can be easily corrected. Define $f(x)$ at the rth interval as $g(x) + A_r$, where the constants A_r are chosen so that the function becomes continuous. Namely, let $A_1 = 0, A_2 = 4^{\frac{3}{2}} - 4^{\frac{5}{2}} = -56$, and so on.

The resulting function satisfies the conditions of the problem. Actually, its graph lies between the graphs of x and x^3 as before. Now we show that it is incomparable with x^2. Consider an interval $[d, d^2]$, where $d = 2^{2^{2n}}$. Then we have $f(d) \leq d^{5/2}$, and hence $f(d^2) \leq d^{5/2} + \left((d^2)^{5/2} - d^{3/2}\right) \leq d^3 + d^{5/2} \leq 2(d^2)^{3/2}$. Therefore, if $[f] \geq a$, then $a \geq 5/2$. Analogously, considering the interval $[d, d^2]$ for $d = 2^{2^{2n+1}}$, we get $f(d^2) \geq \frac{1}{2}(d^2)^{5/2}$, wherefore the relation $[f] \leq a$ implies $a \leq 3/2$.

Correspondingly, the constructed function and x^2 provide the solution for Problem B2.

B4

For instance, the function $f(x) = \ln x$ fits.

B5

For instance, the function $f(x) = 2^x$ fits. Indeed, we have $f^2 = 4^x = f(2x)$, and the growth of the functions is the same by the definition.

B6

Answer. Yes, it exists.

It suffices to provide for instance the relation $\ln f(x) = f(x/2)$, or equivalently $f(2x) = \exp(f(x))$. To get this, take any increasing function on the interval $[1, 2]$ such that $f(1) = 1$¡ $f(2) = e = \exp(f(1))$; then our relation determines uniquely the function f n the intervals $[2, 4]$, $[4, 8]$, and so on. Clearly, we get a desired example.

C1

Answer. $M(\varepsilon) = \left\lceil \frac{a}{2\varepsilon} \right\rceil$, $N(\delta) = \left\lfloor \frac{a}{\delta} + 1 \right\rfloor$.

C2

The best possible estimate is not required in this problem, so the answer is not unique.

We present one of the possible estimates.

(a) For $M(\varepsilon)$.

Since the disks of radius ε must cover the whole unit disk, their total area must exceed the area of the unit disk. Hence $M(\varepsilon) > \left(\frac{1}{\varepsilon}\right)^2$ (respectively, $M(\varepsilon) > \left(\frac{1}{\varepsilon}\right)^3$ for the cube).

On the other hand, using the construction shown in the Problems section above D1, we can easily see that the whole disk (with a minor "overcoming") can be covered by regular hexagons. Covering these hexagons by disks, we arrive (for a sufficiently small ε) to the estimate $M(\varepsilon) < A \cdot \left(\frac{1}{\varepsilon}\right)^2 \cdot \frac{2\pi}{3\sqrt{3}}$, where A is an arbitrary constant greater than 1.

For the ball, the first estimate is quite similar up to replacement of the square by the cube. The second one can't be obtained in this way. But we can cover the ball by small cubes and insert each cube into a ball. This leads to the upper estimate.

(b) For $N(\delta)$.

Suppose that N points form a lattice; for each point, consider a disk of radius δ with the center in this point. If all these disks do not cover the unit disk, then we can add one more point to the lattice, and it is not maximal. Hence $N(\delta) > \left(\frac{1}{\delta}\right)^2$. The other estimates are obtained similarly to the ones above.

C3

Clearly, a maximal δ-lattice is also a δ-net (the respective disks cover the whole figure). On the other hand, a disk of radius $\frac{\delta}{2}$ can contain not more than one point of a δ-lattice. This immediately implies that if $\varepsilon < \frac{\delta}{2}$ then $M(\varepsilon) \geq N(\delta)$.

Thus $M(\frac{\varepsilon}{3}) \geq N(\delta) \geq M(\delta)$, and this implies the statement of the problem.

C4

We construct the required figure as an intersection of an infinite number of figures. Actually, let Φ_1 be the disk of radius 2 with the center at the origin. Furthermore, inscribe a figure Φ_2 into Φ_1 as follows. Let Φ_2 be the union of 8 disks of radius $\frac{1}{2}$ (that is, 4 times less) with centers at $(0, -\frac{3}{4})$; $(0, -\frac{1}{4})$; $(0, \frac{1}{4})$; $(0, \frac{3}{4})$, and of 4 similar disks with centers on the y-axis. (Some of these disks do intersect.)

Next, into each of these disks we inscribe 8 disks of radius $\frac{1}{8}$ arranged similarly. Namely, their centers lie on lines passing through the center of the disk in question and parallel to one of the coordinate axes. These 64 disks form the figure Φ_3. The figures $\Phi_4, \Phi_5, \ldots$ are constructed similarly; let Φ be the intersection of all these figures.

Obviously the resulting figure can be covered either by a single disk of radius 2, or by 8 disks of radius $\frac{1}{2}$, or so on. On the other hand, all the centers of constructed disks of radius 2^{1-2n} form a 2^{-2n}-lattice in Φ, and there are 2^{3n} such centers. This determines the dimension of Φ.

C5

Answer. The dimension of a (connected) figure can attain each value from $[1, 2]$ (on the plane). The dimension can be incomparable with some numbers as well. It can be less than 1 only for a non-connected figure.

C6

Answer. The dimensions of different helices may be different.

Let us assume first that a helix has finite length. (This is true, for example, when $f(\varphi) = e^{-\varphi}$, since in this case the length of each turn is proportional to the length of the first one, thus the length of the whole helix is the sum of a descending geometrical progression.)

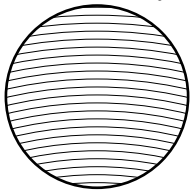
Let L be the length of the whole helix. Since a piece of the helix having length $2r$ obviously can be covered by a disk of radius r, the whole helix can be covered by $\frac{L}{2r}$ disks, which means that the dimension is equal to 1.

On the other hand, if a helix covers a disk densely, then its dimension is greater than 1. We are left to define what is "densely".

To give an example, let us draw concentrical circles of radii $\frac{1}{2}, \frac{1}{3}, \ldots \frac{1}{n}, \ldots$ in the unit disk. They dissect the disk into concentrical rings. Suppose that the kth ring contains 10^k turns of the helix arranged uniformly (in this ring). We claim that this is a "sufficiently dense" arrangement of turns, that is, the dimension of this helix is greater than 1.

For the proof we formulate a general

Proposition T. Let Φ be some figure in a ring, the size of Φ being much larger than the width of a helix turn. Then the area of Φ is approximately equal to the product of the turn width and the total length of the parts of the helix lying inside Φ.



Strictly speaking, this proposition is incorrect; one can easily find some counterexamples. But we will apply it only to the disks and the rings, for which it holds, and can be easily proved.

So, we turn to the dimension of our helix. Consider some small ε. Choose integer k such that $1/k^2 \gg \varepsilon \gg 10^{-k}$, that is, ε much smaller that the width of kth ring, but the turn width δ in this ring is much smaller than ε. Note that such k can be found if ε is small enough. Denote by Φ the $k$the ring.

Now, apply the Proposition T to Φ and to each disk of radius ε in the covering. Let L be the length of the piece of the helix lying inside Φ; then the area of Φ is $\pi\left(\frac{1}{k^2} - \frac{1}{(k+1)^2}\right) \approx L\delta$. On the other hand, each disk will contain the pieces of helix with total length $\approx \pi\varepsilon^2/\delta$. This means that the number of disks covering the part of a helix inside Φ is (almost) at least the ratio of these two values. In other words, these disks have almost the area sufficient to cover the whole Φ.

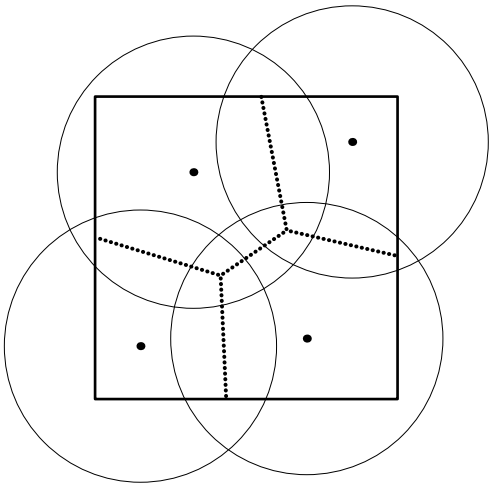
From this, we cannot make a conclusion that the dimension of a helix is 2, though Φ has a dimension 2: actually, we choose different figures Φ for different ε. But, since k grows much slower than ε, one can easily see that the dimension of Φ is definitely greater than 1, and this was exactly our goal.

D1

Lemma. Given N polygons lying inside unit disks such that each polygon contains the center of its disk, and the total amount of vertices of polygons does not exceed $6N$. Then the total area of polygons is at most $N \cdot \frac{3\sqrt{3}}{2}$ (so, this estimate is sharp when the polygons are the regular hexagons).

Proof. Dissect each polygon into triangles by the radii of its disk. The doubled area of each triangle is not greater than $\sin\alpha$, where α is the angle at the central vertex. So, the doubled total area of given polygons equals to $\sin\alpha_1 + \sin\alpha_2 + \ldots + \sin\alpha_n$, where α_i are the corresponding angles; we have $\alpha_1 + \cdots + \alpha_n = \pi N$, and $n \leq 6N$. Moreover, adding some zero angles one can assume that $n = 6N$. Finally, one can notice that the graph of the function $\sin x$ on $[0, \pi]$ is concave, so the function attains its maximal value when $\alpha_i = \frac{\pi}{3}$. So we get the maximal area if all polygons are regular hexagons.

Solution of problem D1. We will prove the statement for an arbitrary polygon T with angles not exceeding $2\pi/3$. Suppose that T is covered by N equal disks. Divide this unit square into convex polygons by the following rule: for each disk center, its polygon contains all the points such that this center is the closest center to them (see Figure). The discs are equal, therefore the sides of polygons are the parts of common chords of our disks. Since all points are covered, each polygon lies in a corresponding disk. Moreover, each polygon obviously contains the center of its disk.



Now we estimate the average value of the angle of our polygon. The average angle in each vertex of our dissection is not greater than $2\pi/3$ (it can be easily seen separately for the vertices inside T, on the sides of T, and for the vertices of T — exactly here we use the estimates for the angles of T). It follows easily that the total number of vertices of the polygons does not exceed $6N$. Hence by the Lemma the total area of the polygons (which is 1) is at most $N\dfrac{3\sqrt{3}}{2}$, and the total area of the disks is $N \cdot \pi\varepsilon^2 \geq \dfrac{2}{3\sqrt{3}\,\varepsilon^2} = \gamma$, as desired.

Remark. If T is a regular hexagon, then the best way to cover it is to use exactly one disk; all other ways are strictly worse.

With some minimal changes, the proof above is valid for any polygon with at most six sides.

D2

The optimal construction is the following one.

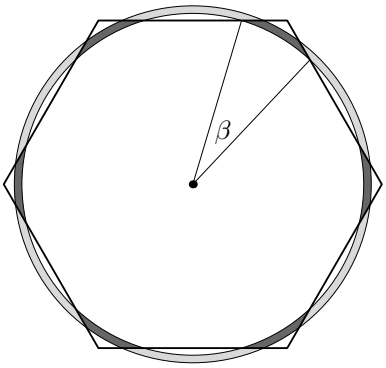
We cover a part of the square by the disks of a larger radius, and the remaining part will be covered by the smaller disks using the method from D1 (that is, using a covering by the small hexagons). Naturally, to reach an (almost) optimal configuration by this method, we should take the radii r_1 and r_2 such that $1 \gg r_1 \gg r_2$.

Cover a unit square by regular hexagon lattice. Denote a side of a hexagon by a and put on each hexagon a disk of radius r_1. We wish these disks to intersect but not to cover the whole square; these conditions rewrite as $\frac{\sqrt{3}}{4}a < r_1 < a$. It remains to cover the rest by the smaller disks.

Now we find the radius r_1 for which this configuration is optimal. We see that the total area of the disks is the sum of the total areas of large and of small disks; the latter is γ times larger than the area of the parts ("corners") uncovered by the large disks. We call this latter summand the *full* area of the corners (to distinguish it from their *total* area).

Thus, neglecting the boungary effect, we can consider only the disks covering of one hexagon; the ratio of their total area to the hexagon area is exactly the total area of all disks.

First, let $r_1 = a\frac{\sqrt{3}}{2}$, and then let us increase this radius. When it increases by δ, the area of the large disks increases approximately by $2\pi r_1\delta$, while the area of six "corners" decreases by



$6\beta r_1\delta$, hence their *full* area decreases by $6\beta r_1\delta\gamma$. Obviously, the total area is minimal when the increment and the decrement become equal, that is, $2\pi r_1\delta = 6\beta r_1\delta\gamma$, wherefore $\beta = \dfrac{2\pi}{6\gamma} = \dfrac{\sqrt{3}}{2}$.

In this case $r_1 = a \cdot \dfrac{\sqrt{3}}{2} \cdot \dfrac{1}{\cos\left(\frac{\pi}{6} - \frac{\beta}{2}\right)}$. The total area of the disks can be computed but it has a combersome form.

One can notice that the (presumably) optimal arrangement of the disks with three (or more) different radii can be found in a similar manner.

Unfortunately, the jury does not know the proof for the optimality of this configuration. Perhaps, the participants can fill this gap?

D3

Answer. 1.

The solution follows from the following

Lemma. If a unit square is covered by disks of total area $1 + \alpha$ then it is possible to cover this square by disks ot total area $1 + \frac{\alpha}{2}$.

Proof. Assume that a unit square can be covered by the disks of total area $< 1 + \alpha$. We claim first that each figure F of area S can also be covered by the disks of total area $< S(1+\alpha)$. To prove this, one can cover F "almost sharp" by some small squares, and then cover these squares with the "non-efficiency" $< 1 + \alpha$.

Now we are ready to prove the Lemma. Inscribe a disk of area $\frac{\pi}{4}$ into a unit square; then we can cover the remaining figure by the disks of total area $< (1 + \alpha)\left(1 - \frac{\pi}{4}\right)$. Since $\pi > 2$, we get the desired covering.

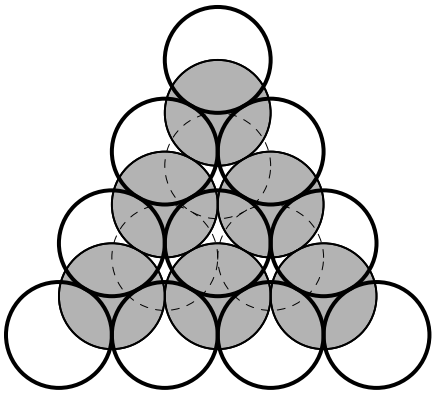
D4

Naturally, as in D1, the radii should be small enough, and the main queation is about the arrangement of the ball centers. By analogy with D1, it is presumably better to take some "dense" packing of balls and try to expand it. We present a dense packing of non-intersecting balls; then it remains to increase their radii to obtain the covering of the whole cube.

Let the centers of the balls in a "first layer" lie in the vertices of a triangular lattice in one plane. The next layer will contain the balls also forming the triangular lattice; moreover, each center in the second layer will form a regular tetrahedron with three centers from the first layer.

The third layer is constructed in a same manner from the second one, and so on. Notice here that, having made two first layers, one can construct the third one in two (essentially different) ways; but the density of both coverings will be the same. To visualize his, we remark that if one starts with the first layer in a form of the regular triangle with side n, then we can obtain a pyramid with n layers.

It remains to calculate the ratio of the radii of the original and the expanded balls. We mention (without a proof) that this ratio is equal to the ratio of the circumdiameter and the side of a regular octahedron, or $\sqrt{2}$.



D5

It seems that the optimal configuration has a form similar to that in D2. That is, we take the configuration from the previous problem, decrease the radii a bit, and cover the remainng spaceby the balls of a smaller radius.

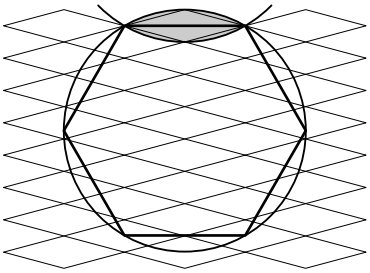
D6

The answer for D6 is the same as for D3. The solution is similar due to the fact that the volume of the ball is greater than half of the volume of the corresponding cube (in 4-dimensional case this is not true, hence one should upgrade the proof a bit).

Remark. Notice that in D4–D5 we provide only some plausible reasonings on an optimal example; conversely, in D6 we show an outline of the full solution.

D7

Cover the square as in D1. Part of the square which is covered twice consists of equal figures; we call such a figure *a lens*. Cover a hexagon by lenses as in the figure: a hexagon is covered by rhombs, and each rhomb can be covered by a lens.



Denote by N a number of rhombs in this covering. Each rhomb can be obtain from the initial one by a shift. So, let us shift the covering N times to obtain a square covered $N+1$ times. The correcting of the boundary effect is left to the reader.

D8

Jury does not know the solution.

E1

We start with two elementary lemmas.

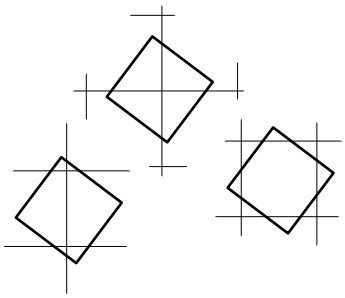
Lemma 1. The number of dissection parts equals to $1+a+b+c$ where a, b respectively are numbers of horizontal and vertical lines which intersect the lower square and c is a number of vertices of the upper sheet which lies in the lower square.

Proof by induction is easy: delete all the horizontal and vertical lines, and then draw them one by one.

Lemma 2. If a base and the corresponding altitude of triangle are at least 1 and this triangle lies in the square then its base coincide with the side of the square.

Proof follows from the fact that if a unit square contains a triangle with area $\geq 1/2$, then this area is $1/2$; moreover, in this case the triangle and the square have a common side.

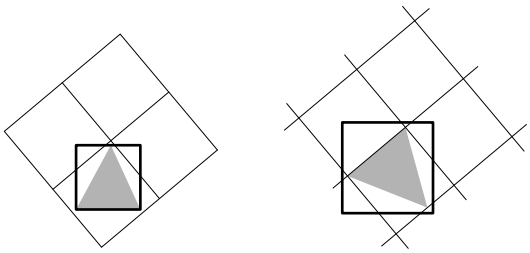
Answer. The number of parts equals to 4, 5 or 6. Examples are shown in the figure.



Solution. First, notice that each projection of a unit square onto some line has a length between 1 and $\sqrt{2}$; this means that a and b can be only 1 or 2. Moreover, it is easy to see than $0 \leq c \leq 2$. This means, by Lemma 1, that the desired number lies between 3 and 7. We are left to show that the border cases are impossible.

Suppose that the number of parts equales to 3, thus $a = b = 1$ and $c = 0$. This means that the lower square is covered by three upper squares (see the left figure below); this contradicts Lemma 2.

The case of 7 parts is similar in the sense that it follows from the same Lemma 2 (see the right figure below).



E2

We will find an approximate estimate for the number of pieces; the further details are left to the reader. We will assume that the upper rectangle is drawn on the plane; so we erase it and then we reconstruct it in several steps. Now we assume that the sides of top (small) rectangle are oriented vertically and horizontally, while the lower rectangle is sloped.

First, we draw the boundary of the upper rectangle; then it will be split into a bit more than two millions parts (almost all of them — except those on the border — are the unit squares). Next, we draw the vertical and horizontal lines (of the upper rectangle) one by one. Each line increases the number of parts by the number of its intersections with other lined (drawn up to this moment). Hence, the total increment will equal to the total number of the points of intersection (where the 3- and 4-fold points are considered in an appropriate manner). So, we are to estimate this number. There are not more than two millions points of intersection of horizontal and vertical lines with each other. The remaining points are the points of intersection of lines from different sheets.

Let α be the angle between a horizontal and (some) sloped line. Then a horizontal line (of length 1000) intersects approximately $1000 \sin\alpha$ lower lines of this type, and approximately $1000 \cos\alpha$ sloped lines of another type, so all horizontal lines add approximately $2000 \cdot 1000(\sin\alpha + \cos\alpha)$ points. Similarly, the vertical lines add approximately $1000 \cdot 2000(\sin\alpha + \cos\alpha)$ points. Note that the expression $\sin\alpha + \cos\alpha$ reaches its maximum value $\sqrt{2}$ when $\alpha = \pi/4$, so the

obtained bound for the number of parts is approximately $2 \cdot 10^6 + 2 \cdot 10^6 + 2 \cdot 2 \cdot 10^6 \sqrt{2} \approx 2 \cdot 4.83 \cdot 10^6$ parts. It is left to see that the errors in our calculations sum up at less than one million.

F1

Answer. Figure has a zero volume characteristics if and only if it is bounded.

F2

Answer. A volume characteristics of the plane and the halfplane equals to 2. A volume characteristics of the strip equals to 1.

F3

Denote by $N_1(R)$ and $N_2(R)$ functions corresponding to points O_1 and O_2. Then $N_1(R + O_1O_2) \geq N_2(R)$ and $N_1(R) \leq N_2(R + O_1O_2)$. Thus $[N_1] = [N_2]$.

F4

Answer. Yes.

Assume that $\varepsilon > \delta$. Then each disk of radius δ can be covered by a disk of radius ε. So $N_\delta(R) \geq N_\varepsilon(R)$. (Here, N_ε and N_δ denote the two functions defined by the disks of the corresponding radii.)

Conversely, each disk of radius ϵ can be covered by A disks of radius δ (for some constant A). Then $N_\delta(R) \leq A \cdot N_\varepsilon(R)$, quod erat demonstrandum.

F5

Answer. $\frac{3}{2}, 2$.

F6

Answer. A volume characteristics of all of them equals to 2.

F7, F8

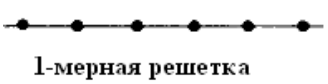
If a figure is unbounded and connected then its volume characteristics is at least 1. Volume characteristics of a figure is at most volume characteristics of the whole plane/space, i.e. 2 or 3. These are the only restrictions.

Случайные блуждания и электрические цепи

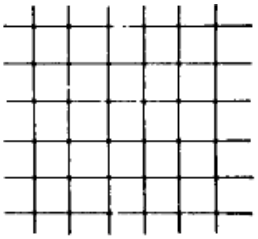
Дмитрий Баранов, Михаил Скопенков, Алексей Устинов

Цель нашего проекта — доказать следующую теорему и изучить смежные вопросы.
Теорема Пойа. **(a)** *Если человек случайным образом перемещается по* 2*-мерной решетке, то он вернется в начальную точку с вероятностью* 1.
(b) *Если же он перемещается по* 3*-мерной решетке, то вероятность его возврата в начальную точку строго меньше* 1.

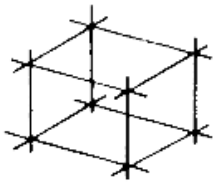
Точные формулировки даны ниже. Предлагаемый подход к доказательству основан на физической интерпретации. Никаких специальных знаний физики не требуется.



1-мерная решетка



2-мерная решетка



3-мерная решетка

---

1. Одномерные блуждания

Мы сначала сформулируем задачу и лишь потом дадим необходимые определения.

1.1. Человек ходит по отрезку улицы, состоящему из 5 кварталов. Начав свой путь на границе кварталов в точке x, он с вероятностью $1/2$ перемещается на один квартал влево и с вероятностью $1/2$ — на один квартал вправо. Подойдя к границе кварталов, он опять выбирает направление движения случайным образом. Если он оказывается в точке 5 (его дом) или в точке 0 (бар), то он прекращает движение: см Рис 1.

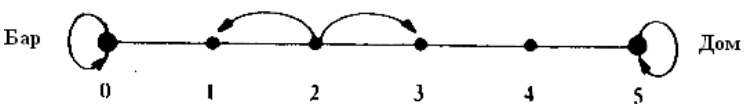


Рис. 1: Случайное движение по улице; см. задачу 1.1.

(A)* Напишите компьютерную программу, моделирующую движение этого человека. Запустите ее много раз и определите процент числа случаев, в которых он приходит домой. Вы можете использовать этот способ для угадывания ответов в последующих задачах.

E-mail address: dimbaranov@mail.ru, skopenkov@rambler.ru, ustinov.alexey@gmail.com

(B) Пусть $P_T(x)$ — вероятность того, что человек, начавший свое движение в точке x и сделавший не более T ходов, оказался дома. Заполните следующую таблицу десятичными дробями с точностью до сотых.

Таблица 1: Вероятности $P_T(x)$ для малых T

T \ x	0	1	2	3	4	5
1	0.00	0.00	0.00	0.00	0.50	1.00
2						
3						
4						

(C) Найдите вероятность $P(x)$ того, что человек дойдет до дома через какое-то количество ходов.

Определение. (A) Предположим, что у некоторого эксперимента имеется n *равновероятных* исходов, и событие X происходит ровно в m из них. Тогда *вероятностью* события X называется число $P_1(X) := m/n$.

Например, вероятность выпадения орла при бросании монеты — $1/2$; вероятность выпадения 6 очков на кубике — $1/6$; вероятность пойти направо по нашей улице — $1/2$.

(B) Теперь предположим, что событие X зависит от последовательности таких экспериментов. Последовательность из T экспериментов имеет n^T возможных исходов. Предположим, что событие X происходит ровно при m_T исходов из них. Тогда *вероятностью* события X называется число $P_T(X) := m_T/n^T$.

Например, есть ровно 4 возможных исхода при бросании монеты 2 раза:

1-ый бросок	орел	орел	решка	решка
2-ой бросок	орел	решка	орел	решка

Пусть событие X состоит в появлении решки хотя бы один раз. Событие X происходит в 3 случаях из 4 возможных. Поэтому вероятность события X есть $P_2(X) = 3/4$.

Вероятность получения более 10 очков при бросании двух кубиков — $1/12$, так как это событие происходит в ровно 3 случаях ($5+6$, $6+5$ или $6+6$) из 36 возможных. Вероятность того, что человек, начавший с точки 3, сдвинется вправо два раза подряд составляет $1/4$.

(C) Пусть теперь событие X зависит от бесконечной последовательности таких экспериментов. Мы будем называть число $P(X)$ *вероятностью* события X, если вероятности $P_T(X)$ стремятся к числу $P(X)$ при стремлении T к бесконечности[1].

Например, вероятность выпадения решки хотя бы один раз в бесконечной серии бросков составляет $P(X) = 1$, так как $P_T(X) = 1 - 1/2^T$ стремится к 1 при стремлении T к бесконечности.

То, что вероятности $P(X)$ *существуют* для всех событий X, рассматриваемых в проекте, можно использовать без доказательства.

[1] Формально это означает, что для каждого $\varepsilon > 0$ существует число T_0 такое, что для каждого $T > T_0$ выполнено $|P(X) - P_T(X)| < \varepsilon$.

1.2. Петя и Паша играют на монетки; всего у них есть 5 монеток; в каждом раунде Петя выигрывает у Паши одну монетку с вероятностью $1/2$ и проигрывает с вероятностью $1/2$; они играют до тех пор, пока у Пети не станет 0 монеток (он проиграл) или 5 (он выиграл все монеты Паши). Найдите вероятность $P(x)$ того, что Петя выиграет, начав игру с x монетками.

1.3. Предположим, нашего "путешественника" сносит в одну сторону; точнее, пусть он каждый раз перемещается вправо с вероятностью p и влево с вероятностью $q = 1 - p$. Найдите вероятности $P(x)$ в этом случае.

1.4. Предположим, что вы играете на деньги; сначала у вас 20 монет, а у вашего соперника — 50 монет; в каждой игре вы выигрываете одну монету с веротностью 0.45 и проигрываете с вероятностью 0.55; игра продолжается до тех пор, пока у кого-либо не закончатся деньги. Найдите вероятность своего разорения.

Определение. *Электрическая цепь* — это связный конечный граф, у которого каждому ребру xy приписано положительное вещественное число, называемое его *проводимостью*[2] $C(xy)$, и задано два непересекающихся выделенных множества вершин (P и N). Вершины из множества N соединены с отрицательным полюсом батарейки и землей, а вершины из множества P — с положительным; см рисунок 2.

Потенциалы вершин $v(x)$ определяются следующими аксиомами:

1. *Граничные условия.* Если $x \in N$, то $v(x) = 0$. Если $x \in P$, то $v(x) = 1$.
2. *Закон Кирхгофа.* Если $x \notin P \cup N$, то $\sum_{xy} C(xy)\,(v(x) - v(y)) = 0$, где суммирование ведется по всем ребрам xy, содержащим вершину x.

Число $i(xy) := C(xy)\,(v(x) - v(y))$ называется *током*, идущим по ребру xy; $i(x) := \sum_{xy} i(xy)$ — *током*, втекающим в цепь через вершину x (так, $i(x) = 0$ для каждого $x \notin P \cup N$ по аксиоме 2); $C := \sum_{x \in P} i(x)$ называется *эффективной проводимостью* цепи между множествами P и N; $Q := \sum_{xy} C(xy)\,(v(x) - v(y))^2$, где суммирование ведется по всем ребрам цепи, называется *тепловой мощностью* цепи.

1.5. Одинаковые резисторы соединены последовательно и подключены к батарейке в 1 вольт как показано на рисунке 2. Найдите потенциалы $v(x)$ в точках $x = 0, 1, 2, 3, 4, 5$. Вы можете использовать программы, эмулирующие электрические цепи, для угадывания ответа.

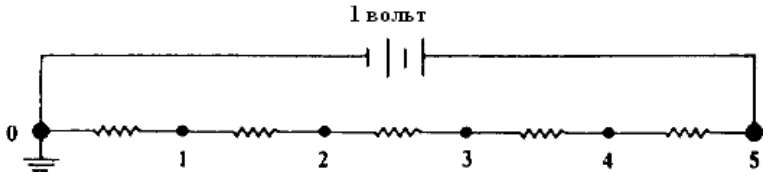


Рис. 2: Электрическая цепь; см. задачу 1.3.

[2]Величина, обратная проводимости, называется *сопротивлением*.

1.6. Рассмотрим цепь с вершинами $0, 1, \ldots, n$, ребрами $01, 12, \ldots, (n-1)n$ единичной проводимости, и выделенными множествами $N = \{0\}$, $P = \{n\}$.
(A) *Принцип максимума.* Функция $v(x)$, удовлетворяющая аксиоме 2 достигает своего максимума и минимума в вершинах из множества $P \cup N$.
(B) *Единственность.* Если $v(x)$ и $u(x)$ — две функции, удовлетворяющие аксиомам 1–2, то $v(x) = u(x)$ для всех x.
(C) Найдите потенциалы $v(x)$ и эффективную проводимость данной цепи. К чему они стремятся при стремлении числа n к бесконечности?

1.7. Сформулируйте и докажите 1-мерную теорему Пойа.

2. Двумерные блуждания

2.1. Рассмотрим город, схема которого приведена на рисунке 3 слева. Отрезки обозначают улицы. Пути отхода помечены буквой E, а буквой P помечены точки, занятые полицией. Найдите с точностью до сотых вероятность $P(x)$ того, что начав свой путь в точке x, человек убежит, а не попадет в руки полиции. Из точки $x = (a, b)$ он перемещается в каждую из точек $(a+1, b), (a-1, b), (a, b+1), (a, b-1)$ с вероятностью $1/4$. Если он достигает одной из точек E или P, то его передвижения заканчиваются.

2.2. Найдите потенциалы $v(x)$ в цепи из единичных резисторов на рисунке 3 справа.

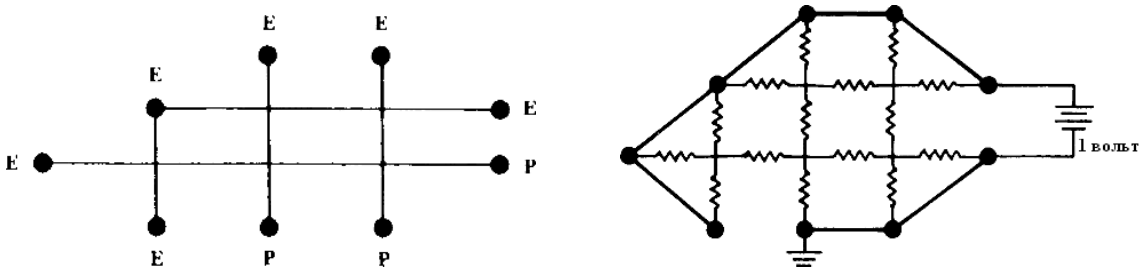


Рис. 3: Случайное движение по городу и электрическая цепь; см. задачи 2.1 и 2.2.

2.3. Паук перемещается случайным образом по ребрам
(A) куба; **(B)** октаэдра; **(C)** додекаэдра; **(D)** икосаэдра;
если он начинает движение в точке a, то какова вероятность того, что он достигнет противоположной вершины h быстрее, чем вернется в начальную вершину a; см. рисунок 4 слева?

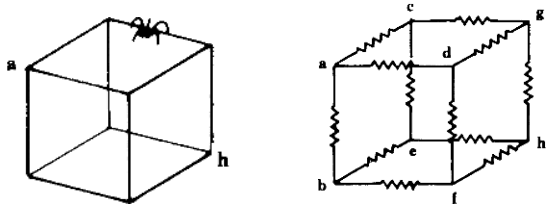


Рис. 4: Случайное блуждание по кубу и электрическая цепь; см. задачи 2.3(A) и 2.5(A).

2.4. Следующие преобразования сохраняют эффективную проводимость цепи:
(A) замена двух резисторов, соединенных последовательно, на один резистор проводимости $1/\left(\frac{1}{C_1}+\frac{1}{C_2}\right)$; см. рисунок 5 слева;
(B) замена двух параллельно соединенных резисторов на один резистор с проводимостью $C_1 + C_2$; см. рисунок 5 справа;
(C) объединение двух вершин с одинаковыми потенциалами в одну новую вершину.

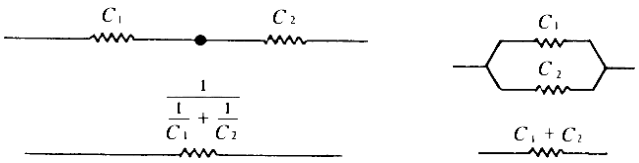


Рис. 5: Последовательное и параллельное соединение; см. задачу 2.4.

2.5. Найдите эффективную проводимость между
(1) противоположными вершинами; **(2)*** смежными вершинами;
(A) куба; **(B)** октаэдра; **(C)** додекаэдра; **(D)** икосаэдра;
с ребрами единичной проводимости; см рисунок 4 справа.

2.6. Пьяный турист выходит из отеля и перемещается случайным образом по улицам Парижа, схема центра которого приведена на рисунке 6 слева. Найдите вероятность того, что он дойдет до Триумфальной арки до того, как доберется до окраины города.

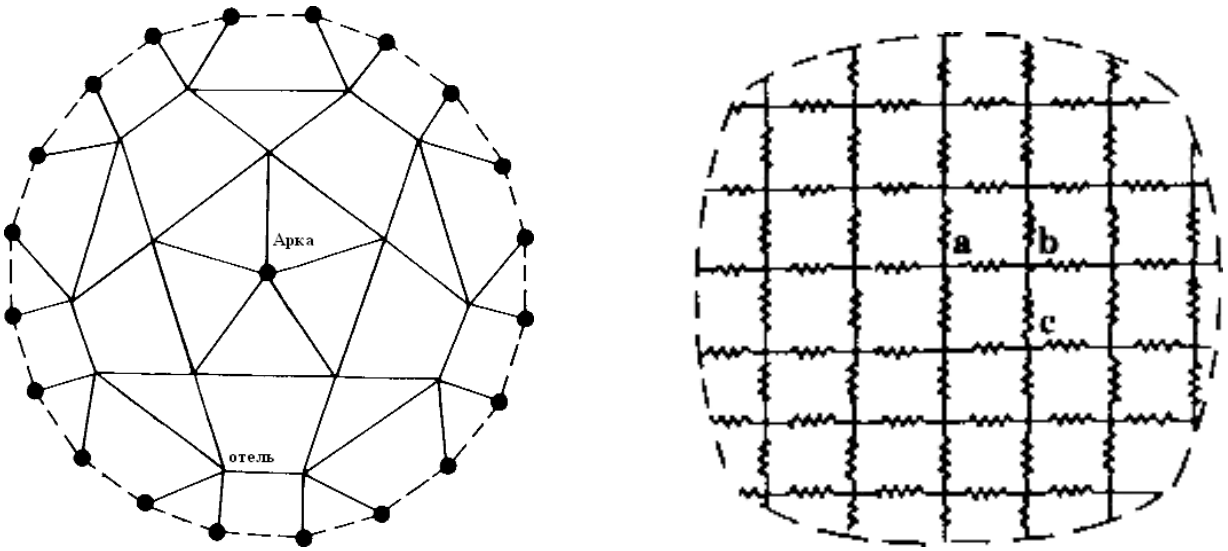


Рис. 6: Туристическая карта Парижа и решеточная цепь; см. задачи 2.6 и 2.7.

2.7. Проводимость между вершинами **(A)*** a и b; **(B)**** a и c; двумерной решетки из единичных резисторов равна 2 и $\pi/2$, соответственно; см. Рис. 6 справа.

Для 2-мерной решетки в определении потенциала мы добавляем еще одну аксиому:

3. $v(x)$ стремится к $1/2$ при стремлении расстояния между x и некоторой фиксированной вершиной к бесконечности.

Разрешается пользоваться без доказательства существованием функции $v(x)$, удовлетворяющей аксиомам 1–3.

2.8. *Закон монотонности.* Разрезание каких-либо ребер цепи может только уменьшить эффективную проводимость между данными вершинами; см. рисунок 7 слева. Объединение каких-либо вершин в одну вершину может только увеличить эффективную проводимость между данными вершинами; см. рисунок 7 в центре.

2.9. **(A)** Докажите, что эффективная проводимость между центром и границей квадратной решетки 4×4 из единичных резисторов меньше 3; см рисунок 7 справа.
(B) К чему стремится проводимость между центром и границей квадратной решетки $2n \times 2n$ из единичных резисторов при стремлении n к бесконечности?

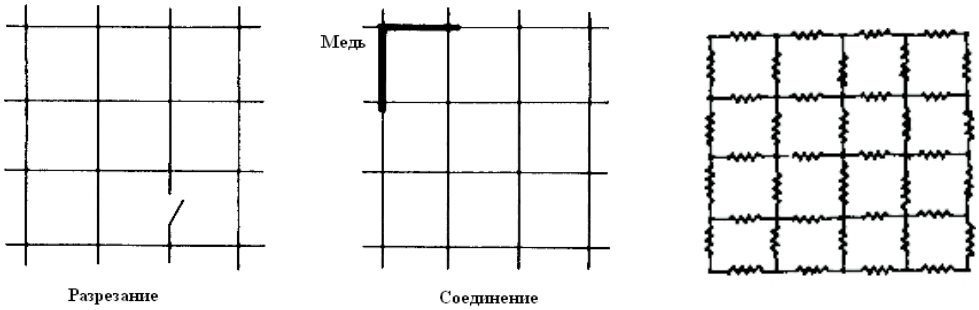


Рис. 7: Разрезание, соединение, и квадратная решетка 4×4; см. задачи 2.8 и 2.9.

2.10. Докажите теорему Пойа для 2-мерной решетки.

3. Сопротивление правильных графов

Задачи данного раздела могут помочь при решении задач предыдущего.

Функция $v(x)$ на вершинах электрической цепи называется *гармонической*, если она удовлетворяет аксиоме 2 из определения электрической цепи. Вершины из множества $P \cup N$ назовем *граничными*, остальные — *внутренними*.

3.1. **(A)** *Принцип суперпозиции.* Если функции $u(x)$ и $v(x)$ — гармонические и $a, b \in \mathbb{R}$, то и функция $au(x) + bv(x)$ — гармоническая.
(B) *Принцип максимума.* Гармоническая на конечной электрической цепи функция принимает свое наибольшее и наименьшее значения на граничных вершинах.
(C) *Единственность.* Если $u(x)$ и $v(x)$ — две гармонические функции, совпадающие во всех граничных вершинах конечной электрической цепи, то $u(x) = v(x)$ во всех вершинах x графа.
(D) Человек, случайно блуждающий по конечному городу, с вероятностью 1 побывает на всех перекрестках его улиц.

3.2. **(A)*** *Альтернатива Фредгольма.* Пусть имеется система линейных уравнений

$$\begin{cases} a_{1,1}x_1 + \ldots + a_{1,n}x_n = b_1, \\ \quad \ldots\ldots\ldots\ldots\ldots \\ a_{n,1}x_1 + \ldots + a_{n,n}x_n = b_n \end{cases},$$

в которой число уравнений равно числу неизвестных. Докажите, что всегда выполняется одна из следующих возможностей:

1. для любых $b_1, \ldots, b_n$ система имеет ровно одно решение (в частности, при $b_1 = \ldots = b_n = 0$ существует только нулевое решение);
2. для некоторых $b_1, \ldots, b_n$ система неразрешима, а для некоторых (в том числе нулевых) имеет бесконечно много решений.

(B)* *Задача Дирихле.* Докажите, что для любой конечной электрической цепи существует функция $v(x)$, удовлетворяющая аксиомам 1–2.

3.3. (A) *Вариационный принцип.* Пусть $v(x)$ — произвольная функция на вершинах конечной электрической цепи, удовлетворяющая аксиоме 1, но не обязательно аксиоме 2. Занумеруем вершины графа числами $1, \ldots, n$ и пусть $1, \ldots, k$ — внутренние вершины. Обозначим $v_1 := v(1)$, $v_2 := v(2)$, $\ldots$, $v_n = v(n)$. Будем рассматривать $v_1, \ldots, v_k$ как переменные. Рассмотрим тепловую мощность $Q(v_1, \ldots, v_k) := \sum_{xy} C(xy)\,(v_x - v_y)^2$ как функцию переменных $v_1, \ldots, v_k$. Докажите, что функция $Q(v_1, \ldots, v_k)$ принимает наименьшее значение, когда функция $v(x)$ — гармоническая.
(B)* Докажите, что для функции $Q(v_1, \ldots, v_k)$ существует ровно один набор значений $v_1, \ldots, v_k$, при котором она достигает своего минимального значения. Используя это, дайте второе доказательство того факта, что для конечной электрической цепи функция $v(x)$, заданная аксиомами 1–2, существует и единственна.
(C) *Закон сохранения энергии.* Докажите, что минимальное значение величины $Q(v_1, \ldots, v_k$ численно равно эффективной проводимости C.
(D) *Принцип монотонности.* Докажите, что если в цепи одну проводимость увеличить, то эффективная проводимость не уменьшится.
(E) Из 2-мерной решетки выбросили произвольное множество ребер. Докажите, что по-прежнему случайное блуждание с вероятностью 1 вернется в исходную вершину.

Предположим, что имеется граф Γ, у которого сопротивление каждого ребра равно 1. Возьмем в графе Γ два смежных ребра AB и AC. Эти ребра назовем *эквивалентными*, если существует перестановка вершин графа, переводящая соединенные ребром вершины в соединенные ребром вершины, при которой A переходит в A и B — в C. Вершину графа назовем *центром симметрии* графа Γ, если все ребра, содержащие ее, эквивалентны. Граф Γ называется *правильным*, если все его вершины — центры симметрии графа.

Примеры правильных графов: правильные многогранники любой размерности; правильные решетки на евклидовой плоскости, плоскости Лобачевского и их многомерных аналогах; симметричные решетки на торе и т. п. Нетривиальный пример: граф ромбододекаэдра. Это многогранник, который получается, если к каждой грани куба приставить по четырехугольной пирамиде так, что все треугольники, граничащие по ребрам куба, сольются в ромбы. Поверхность ромбододекаэдра состоит из 12 ромбов. Он нетривиален тем, что его вершины имеют разную степень (3 и 4).

3.4. (A) Пусть правильный граф содержит n вершин, A_1 и A_2 — соседние вершины степеней k_1 и k_2, соответственно. Докажите, что сопротивление между ними равно

$$\left(\frac{1}{k_1} + \frac{1}{k_2}\right)\left(1 - \frac{1}{n}\right).$$

(B) Если же взять 2-мерную решетку, то $1/n$ в последней формуле нужно заменить нулем.

3.5. К двум соседним вершинам проволочного **(A)** икосаэдра; **(B)** додекаэдра; подвели напряжение так, что по соединяющему их ребру потек ток I. Какой при этом будет течь ток по диаметрально противоположному ребру?

3.6. * Докажите, что суммы $D_n = \sum_{k=0}^{n} \frac{1}{C_n^k}$ и $F_n = \frac{n+1}{2^{n+1}} \sum_{k=1}^{n+1} \frac{2^k}{k}$ равны друг другу. Как эти суммы связаны с сопротивлением многомерного куба? В качестве следствия получите, что порядок вхождения двойки в число $\sum_{k=1}^{n} \frac{2^k}{k}$ стремится к бесконечности с ростом n.

4. Трехмерные блуждания

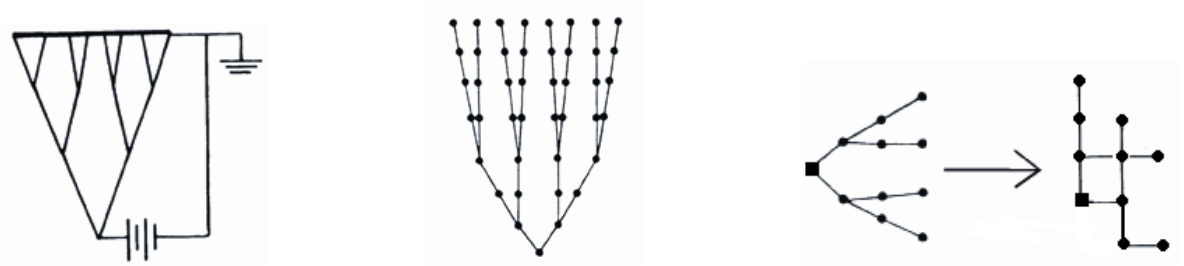


Рис. 8: (Слева) бинарное дерево глубины 3; (в центре) модифицированное бинарное дерево глубины 3; (справа) разрешенные пересечения ребер в этом дереве; см. задачи 4.1, 4.2 и 4.4.

4.1. Найдите сопротивление бинарного дерева глубины **(A)** 3; **(B)** 2010, составленного из единичных резисторов (см. рис. 8 слева).

4.2. Найдите сопротивление *модифицированного* бинарного и троичного деревьев глубины 2010, в которых каждый резистор на k-ом уровне заменяется на 2^k последовательно соединенных единичных резисторов (см. рис. 8 в центре).

4.3. Какие из деревьев, упомянутых **(A)** в задаче 4.1; **(B)** в задаче 4.2; можно вырезать из трехмерной решетки?

4.4. А если разрешаются пересечения (см. рис. 8 справа) ребер на равном расстоянии от корня?

4.5. Докажите теорему Пойа для 3-мерной решетки.

5. Сопротивление кольца*

Во всех задачах этого раздела будет фигурировать квадратная металлическая сетка. Будет предполагаться, что все ее узлы — это точки двумерной целочисленной решетки. Соединены друг с другом только соседние узлы (расстояние между которыми равно единице). Сопротивление ребра между любыми соседними узлами также считается равным единице.

5.1. Источник тока подключается к узлам сетки с координатами $(0,0)$ и $(1,0)$. Докажите, что в узлах $(2,2)$ и $(3,2)$ будут одинаковые потенциалы (см. замечание в задаче 3.4(В)).

5.2. Из сетки вырезан квадрат размерами $n \times n$, граничные точки которого соединены шиной с нулевым сопротивлением. Докажите, что сопротивление между любым узлом квадрата и его границей не превосходит $\sqrt{n}$.

5.3. В условиях задачи 5.2 источник тока подключен к внутреннему узлу квадрата и к границе. Потенциал на границе равен нулю. Докажите, что если источник подает ток ε, то в каждой из точек квадрата потенциал не превосходит $\varepsilon\sqrt{n}$.

5.4. Для функций, заданных в узлах целочисленной решетки определим

$$\Delta f(x,y) := f(x-1,y) + f(x+1,y) + f(x,y-1) + f(x,y+1) - 4f(x,y).$$

Пусть $r(x,y) = \sqrt{x^2+y^2}$. Докажите, что для функции $f(x,y) := \ln r(x,y)$ (при $r(x,y) \geq 2$) выполнено $\Delta f(x,y) = O\left(\frac{1}{r^4(x,y)}\right)$.

Здесь и далее для двух функций A и B запись $A = O(B)$ означает, что для некоторой положительной константы c всегда выполняется неравенство $|A| \leq cB$.

5.5. Из металлической сетки вырезано кольцо с внутренним радиусом $r_1 n$ и внешним — $r_2 n$ (центры обоих кругов — в начале координат). Если некоторое ребро разрезано, то сопротивление оставшегося куска пропорционально его длине. На внутренний контур кольца подается напряжение $\ln r_1$, а на внешний — $\ln r_2$. Докажите, что в каждой точке кольца (x,y) потенциал имеет вид $U_n(x,y) = \ln r(x,y) + O\left(\frac{1}{n^{3/2}}\right)$.

5.6. С помощью равенства $\operatorname{arctg} x = x + O(x^3)$ докажите, что при $0 \leq y < R$ выполнено $\frac{R}{R^2+y^2} = \operatorname{arctg}\frac{y+1}{R} - \operatorname{arctg}\frac{y}{R} + O\left(\frac{1}{R^2}\right)$.

5.7. Докажите, что $\sum_{y=0}^{R-1} \frac{R}{R^2+y^2} = \frac{\pi}{4} + O\left(\frac{1}{R}\right)$.

5.8. Пусть в условиях задачи 5.5 имеется дополнительное ограничение $r_2 > 3r_1/2$ (достаточное для того, чтобы квадрат, описанный около внутреннего круга, целиком содержался бы во внешнем). Докажите, что между внутренним и внешним контурами течет ток $2\pi + O\left(\frac{1}{\sqrt{n}}\right)$. Выведите отсюда формулу для сопротивления кольца

$$R(r_1 n, r_2 n) = \frac{1}{2\pi} \ln \frac{r_2}{r_1} + O\left(\frac{1}{\sqrt{n}}\right). \tag{1}$$

5.9. Докажите, что формула (1) выполняется и без дополнительного ограничения $r_2 > 3r_1/2$.

5.10. С помощью равенства (1) уточните оценки в задачах 5.2, 5.3 и докажите формулы с более точными остаточными членами в задачах 5.5 и 5.8:

$$U(x,y) = \ln r(x,y) + O\left(\frac{\ln n}{n^2}\right), \qquad R(r_1 n, r_2 n) = \frac{1}{2\pi} \ln \frac{r_2}{r_1} + O\left(\frac{\ln n}{n}\right).$$

6. Сложные задачи*

6.1. *Теорема Лиувилля.*Пусть заданная на $\mathbb{Z}^2$ функция $f(m,n)$ удовлетворяет неравенству $0 \le f(m,n) \le 1$ и условию

$$f(m,n) = \frac{1}{4}\left(f(m-1,n) + f(m+1,n) + f(m,n-1) + f(m,n+1)\right) \tag{2}$$

для всех $m, n \in \mathbb{Z}$. Докажите, что $f(m,n)$ является константой.

6.2. *Существование потенциалов.* Докажите, что существует такая функция $f(m,n)$ на $\mathbb{Z}^2$, что $f(0,0) = 0$, $f(0,1) = 1$, для каждого $(m,n) \ne (0,0), (0,1)$ условие (2) выполняется, и $f(m,n)$ стремится к $1/2$ при стремлении $r(m,n) := \sqrt{m^2+n^2}$ к бесконечности.

6.3. *Функция Грина.* Пусть $f(m,n)$ — сопротивление 2-мерной решетки между началом координат и точкой (m,n).
(A) Докажите, что для каждого $(m,n) \ne (0,0)$ условие (2) выполнено.
(B) Докажите, что $f(m,n) = g(r(m,n)) + O(1)$ для некоторой функции $g(x)$.
(C) Докажите, что сопротивление между и центром и границей диска радиуса r, вырезанного из 2-мерной решетки равно $\frac{1}{2\pi}\ln r + O(1)$.
(D) Докажите, что $f(m,n) = \frac{1}{2\pi}\ln r(m,n) + O(1)$.

6.4. Найдите с точностью до сотых вероятность того, что при случайном блуждании по 3-мерной решетке мы вернемся в начальную точку.

6.5. Робот ходит по вершинам 3-мерной решетки, переходя каждый раз в одну из соседних вершин. В одной из вершин находится клад; робот находит его, когда оказывается в вершине с кладом. Существует ли программа, управляющая движением робота и использующая конечный объем памяти и генератор случайных чисел, такая, что робот найдет клад с вероятностью 1?

7. Указания и решения

1.1 **(A)** Правильность работы программы проверяется следующим образом: разность между "настоящей" и посчитанной вероятностями должна быть пропорциональна числу $\frac{1}{\sqrt{n}}$, где n — число экспериментов.
(B) Ответ смотрите в таблице.

Таблица 2: Вероятности $P_T(x)$ и $P(x)$

T \ x	0	1	2	3	4	5
1	0.00	0.00	0.00	0.00	0.50	1.00
2	0.00	0.00	0.00	0.25	0.50	1.00
3	0.00	0.00	0.13	0.25	0.63	1.00
4	0.00	0.06	0.13	0.38	0.63	1.00
$P(x)$	0.00	0.20	0.40	0.60	0.80	1.00

(C) *Ответ*: $P(x) = x/5$; см последнюю строку в таблице выше.
Доказательство. Рассмотрим случайное блуждание по точкам $0, 1, 2, \ldots, n$. Обозначим за $P(x)$ вероятность дойти из x до N раньше, чем до 0. Рассмотрим получившуюся функцию $P(x)$, определенную в точках $x = 0, 1, 2, \ldots, n$. Она обладает следующими свойствами:

1. $P(0) = 0$ и $P(n) = 1$.
2. $P(x) = \frac{1}{2}P(x-1) + \frac{1}{2}P(x+1)$ для каждого $x = 1, 2, ..., n-1$.

Свойство 1 следует из того, что при достижении точек 0 и n перемещения заканчиваются; для игры на монетки это означает конец игры. Свойство 2 заключается в том, что вероятность попасть домой из внутренней точки x равна среднему арифметическому вероятностей попадания домой из соседних точек. Свойство 2 выводится из следующего утверждения:

Базовое Утверждение. *Пусть E — некоторое событие, F и G — два события, из которых всегда случается ровно одно. Тогда*

$$P(E) = P(F) \cdot P(E \text{ последует за } F) + P(G) \cdot P(E \text{ последует за } G).$$

В нашем случае E="человек дойдет до бара", F="первый раз он пойдет налево" и G="первый раз он пойдет направо". Тогда получаем $P(E) = P(x)$, $P(F) = P(G) = 1/2$, $P(E$ последует за $F) = P(x-1)$, $P(E$ последует за $G) = P(x+1)$ и свойство 2 доказано.

Из этих двух свойств вытекает, что $P(x)$ представляет собой арифметическую прогрессию $P(x) = x/n$.

1.2 *Ответ*: $P(x) = x/5$; эта задача эквивалентна задаче 1.1(C).

1.3 *Ответ*: $P(x) = \frac{(q/p)^x - 1}{(q/p)^5 - 1}$.
Указание: Рассуждайте так же, как в решении задачи 1.1(C). Покажите, что свойства 1–2 надо заменить на такие:

1. $P(0) = 0$ и $P(n) = 1$.
2. $P(x) = qP(x-1) + pP(x+1)$ для каждого $x = 1, 2, ..., n-1$.

Выберите A и B такими, чтобы функция $f(x) = A(q/p)^x + B$ удовлетворяла новым свойствам 1–2.

1.4 *Ответ*: $\approx 99.995\%$. Точное значение: $1 - \frac{(0.55/0.45)^{20} - 1}{(0.55/0.45)^{70} - 1}$; смотрите решение задачи 1.3.

1.5 *Ответ*: $v(x) = x/5$. *Указание.* Из аксиом 1–2 следует, что функция $v(x)$ будет линейной для этой цепи.

1.6 **(A)** Пусть M — максимум функции $v(x)$. Тогда если $v(x) = M$ для $x \notin P \cup N$, то это же равенство должно быть выполнено для $v(x-1)$ и $v(x+1)$ так как $v(x)$ — среднее арифметическое $v(x-1)$ и $v(x+1)$. Если $x-1$ оказалась внутренней точкой, применяем то же самое рассуждение и получаем $f(x-2) = M$; продолжая рассуждение, получаем $f(0) = M$. Для минимального значения аналогично.

(B) Положим $h(x) = v(x) - u(x)$. Тогда для любой внутренней точки x имеем:

$$\frac{h(x-1)+h(x+1)}{2} = \frac{v(x-1)+v(x+1)}{2} - \frac{u(x-1)+u(x+1)}{2}$$

и поэтому функция $h(x)$ также удовлетворяет аксиоме 2. Но $h(x) = 0$ при x из $P \cup N$; из принципа максимума получаем, что максимальное и минимальное значения h равны 0. Значит, $h(x) = 0$ и $v(x) = u(x)$.

(C) *Ответ*: $v(x) = x/n$, $C = 1/n$; $C \to 0$ и $v(x) \to 0$ для каждого фиксированного x при $n \to \infty$.

Указание: Легко проверить, что функция $f(x) := x/n$ удовлетворяет аксиомам 1–2. Из единственности (см 1.6(B)) следует, что $v(x) = x/n$.

1.7 Теорема. *При случайном блуждании по* 1*-мерной решетке вероятность вернуться когда-либо в начальную точку равна* 1.

Доказательство. Пусть P — вероятность вернуться когда-либо в начальную точку. Обозначим P_n вероятность вернуться в начальную точку до попадания в n или $-n$. Предположим, что все эти вероятности существуют. Тогда $P_n \le P \le 1$ для любого n.

Сейчас мы докажем, что $P_n = 1 - 1/n$. После первого "хода" человек попадает в одну из точек 1 и -1 с вероятностью $1/2$. Если он оказался в точке 1, то из задачи 1.1(C) получаем, что вероятность вернуться в начало до попадания в точку n равна $1 - 1/n$. Если он оказался в точке -1, рассуждаем аналогично. Применяя Базовое Утверждение из решения задачи 1.1(C), получаем $P_n = \frac{1}{2}\left(1 - \frac{1}{n}\right) + \frac{1}{2}\left(1 - \frac{1}{n}\right) = 1 - \frac{1}{n}$. (Еще можно было заметить, что $P_n = 1 - C$, где $C = 1/n$ — проводимость цепи из задачи 1.6.)

Так как $1 - 1/n \le P \le 1$ для каждого n, то P равно 1. □

2.1 *Ответ*: см. рисунок 9 слева.

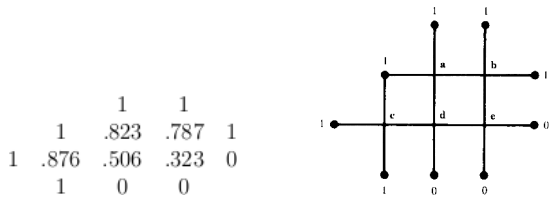


Рис. 9: Вероятности $P(x)$ или потенциалы $v(x)$; см. задачи 2.1 и 2.2.

Указание. Схема города представлена на рисунке 9 справа. Вероятности $P(x)$ обозначены за a, b, c, d, и e. Как и в 1-мерном случае, функция $P(x)$ удовлетворяет аксиомам 1–2 из определения электрической цепи. Отсюда мы получаем систему линейных уравнений:

$$\begin{aligned} a &= (b + d + 2)/4; \\ b &= (a + c + 2)/4; \\ c &= (d + 3)/4; \\ d &= (a + c + e)/4; \\ e &= (b + d)/4. \end{aligned}$$

Ответ получаем, решая эту систему.

Замечание. Нахождение точного решения для двухмерной "задачи Дирихле" — дело сложное; поэтому мы рассмотрим два метода нахождения приближенных решений.

Первый метод использует случайные блуждания. Он называется *методом Монте-Карло*, так как случайные блуждания связаны с вероятностями, а в Монте-Карло находится известные игорные дома, азартные игры в которых тоже связаны с вероятностями. Мы моделируем много случайных блужданий из точки x и находим долю путей, закончившихся в точках E. Из закона больших чисел следует, что полученная оценка будет приближением для "настоящей" вероятности $P(x)$. Этот яркий и простой метод позволяет находить решения, но он не очень эффективен.

Теперь опишем более эффективный *метод релаксации*. Напомним, что мы ищем функцию с заданными значениями на границе у которое значение в любой внутренней точке равно среднему

арифметическому значений ее соседей. Возьмем какую-нибудь функцию с подходящими граничными значениями и возьмем некоторую внутреннюю точку. В общем случае значение функции не будет равно среднему арифметическому значений в соседних точках. Тогда попробуем "подогнать": положим новое значение функции в этой точке равным среднему арифметическому значений в соседних точках. Теперь будем по очереди брать остальные внутренние точки и делать с ними ту же операцию. Когда мы пройдем по всем внутренним точкам, функция не будет удовлетворять аксиоме 2, так как после изменения значения функции в одной точке мы могли изменить значения в соседних с ней точках, нарушив равенство. Тем не менее, полученная функция будет "лучше" удовлетворять аксиоме 2, чем та функция, с которой мы начали; повторяя этот процесс (проходя каждый раз по всем внутренним точкам) мы будем получать приближения к решению лучше и лучше.

2.2 *Ответ*: см. рисунок 9 слева; эта задача эквивалентна задаче 2.1.

2.3 *Ответ*: **(A)** 2/5; **(B)** 1/2; **(C)** 2/7; **(D)** 2/5.
Указание. Сведем задачу к задаче 2.5 при помощи следующего утверждения:

Физическая интерпретация вероятности. *Вероятность того, что случайное блуждание по графу G из вершины a достигнет вершины h до возврата в a, равна*

$$P = C/\deg a,$$

где C — проводимость графа G (все резисторы единичные) между a и h, а $\deg a$ — число ребер, выходящих из вершины a.

2.4 **(C)** *Указание.* Функция $v(x)$ однозначно определена на вершинах получившейся цепи. Проверьте, что она удовлетворяет аксиомам 1–2.

2.5 *Ответ*: **(1A)** 6/5; **(1B)** 2; **(1C)** 6/7; **(1D)** 2.
(2A) 12/7; **(2B)** 12/5; **(2C)** 30/19; **(2D)** 30/11.

Короткое решение смотрите в разделе 3.
(2A) *Указание.* Соединим точки a и b с батарейкой; см. рисунок 4 справа. Потенциалы в точках c и d равны из симметрии; аналогично в точках e и f. Таким образом, наша схема эквивалентна схеме, изображенной на рисунке 10 слева.

Используя формулы для параллельного и последовательного соединения резисторов, эта цепь сводится в к цепи из одного резистора сопротивлением 7/12 ом (см рисунок 10 справа). Таким образом, сопротивление равно 7/12.

2.6 *Ответ*: 1/7. Решение аналогично решению задачи 2.3.

2.7 **(A)**. Короткое решение смотрите в разделе 3.
(B) Авторам неизвестно элементарное решение задачи. Красивое решение, использующее дискретное преобразование Фурье, вы можете найти в книге [7].

2.8 Смотрите раздел 3.

2.9 **(B)** *Ответ*: $C \to 0$ при $n \to \infty$.
Указание. Применим закон монотонности: объединим вместе точки, расположенные на квадратах, как показано на рисунке 11 сверху. Полученная цепь эквивалентна цепи на рисунке 11 в центре. Так как можно заменить n параллельных резисторов в 1 ом на один резистор в $1/n$ ом, цепь эквивалентна цепи на рисунке 11 снизу. Проводимость этой цепи равна

$$\frac{1}{\sum_{k=1}^{n} \frac{1}{8k-4}}.$$

Это число стремится к нулю при стремлении n к бесконечности. Так как проводимость старой цепи не больше, она тоже стремится к нулю.

2.10 *Указание.* Пусть P — вероятность того, что при случайном блуждании по 2-мерной решетке мы вернемся в начальную точку. Обозначим за P_n вероятность того, что случайное блуждание вернется в начальную точку до достижения граничных точек квадрата $2n \times 2n$ с центром в начальной точке. Предположим, что все эти вероятности существуют. Ясно, что $P_n \le P \le 1$ для каждого n. Из физической интерпретации вероятности получаем, что $P_n = 1 - C/4$, где C — эффективное сопротивление

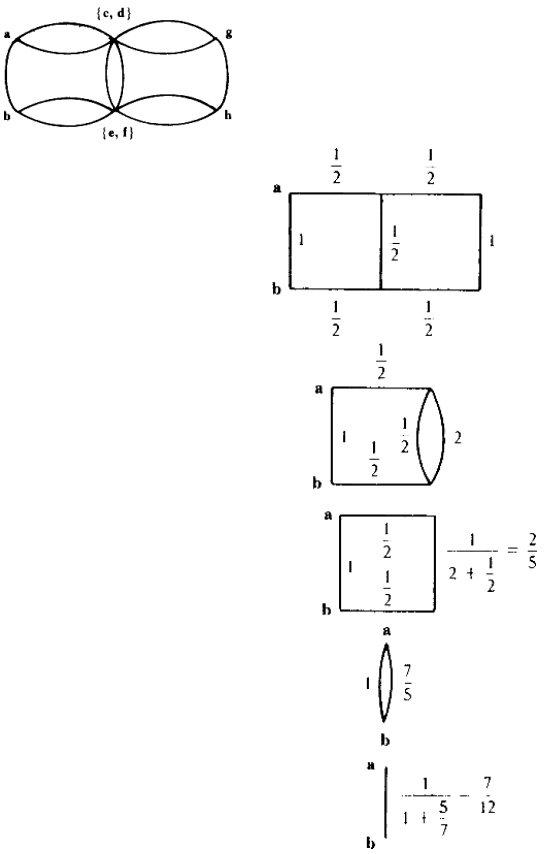


Рис. 10: Упрощение цепи; см решение задачи 2.5.

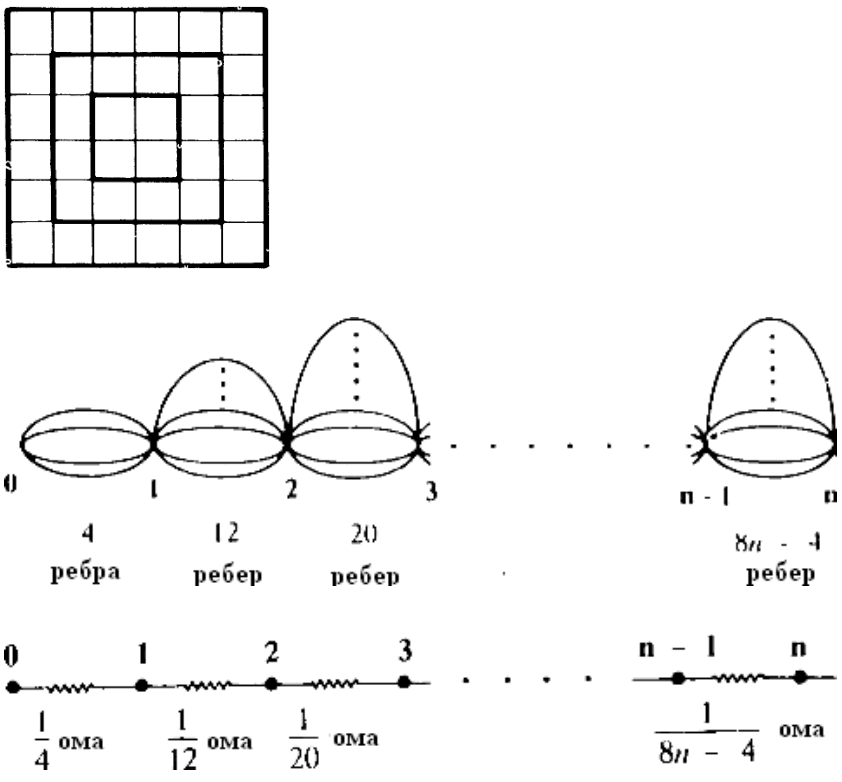


Рис. 11: Объединение в квадратной цепи и эквивалентная цепь; см решение задачи 2.9.

между центром и границей квадрата $2n \times 2n$. Из решения задачи 2.9(В) следует, что C стремится к нулю при стремлении n к бесконечности. Поэтому $P_n \to 1$ при $n \to \infty$ и P равно 1. $\square$

3.1 См., например, статью [6].

3.2 (В) Рассмотрим случайное блуждание по электрической цепи. Пусть $P_T(x)$ — вероятность того, что, стартуя из вершины x и делая T шагов, мы достигнем положительного полюса батареи раньше, чем отрицательного. Ясно, что при фиксированном x последовательность $P_T(x)$ возрастает, значит, имеет предел $P(x)$. Функция $P(x)$ удовлетворяет аксиомам 1–2.

Замечание. Существование и единственность решения системы Кирхгофа - факты действительно фундаментальные. Например, из единственности решения следует теорема Дена о том, что прямоугольник с иррациональным отношением сторон нельзя разрезать на квадраты. Из существования решения (в непрерывном случае) следует теорема Римана о конформном отображении [2].

3.3 См., например, статью [6].

3.4 (А) Пусть правильный граф содержит n вершин и A_1, A_2 — соседние вершины степеней k_1 и k_2 соответственно. Рассмотрим сначала ситуацию, когда в вершину A_1 подается ток $\frac{n-1}{n}$, а из всех остальных вершин вытекает ток $\frac{1}{n}$. В силу правильности графа по ребру A_1A_2 будет течь ток $\frac{1}{k_1}\left(1-\frac{1}{n}\right)$. Если же ток $\frac{n-1}{n}$ подается в вершину A_2, а из остальных вершин вытекает ток $\frac{1}{n}$, то по ребру A_1A_2 будет течь ток $\frac{1}{k_2}\left(1-\frac{1}{n}\right)$. Объединим обе ситуации, заменив во втором случае все токи на противоположные. Тогда получится, что в вершину A_1 подается единичный ток, который вытекает из вершины A_2. При этом по ребру A_1A_2 течет ток $\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)$. Значит, сопротивление между вершинами A_1A_2 равно $\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)$.

Если граф бесконечный, то $1/n$ нужно заменить нулем. Корректность обосновывается с помощью предельного перехода.

Формула для сопротивления между соседними вершинами правильного графа принадлежит А.Б. Ходулч
Вместе с определением правильного графа она взята из статьи [4].

(В) Решим сначала задачу на физическом уровне строгости. Наложим шину с нулевым сопротилением на периметр прямоугольника $[-N, N+1] \times [-N, N]$. Поскольку при подключении батарейки к узлам $(0,0)$ и $(1,0)$ потенциалы на бесконечности стремятся к нулю, то наложение шины мало повлияет на значение искомого сопротивления. (Дальнейшие рассуждения также проводятся с точностью до погрешности, которая стремится к нулю с ростом N.) Если источник единичного тока подключен к точке $(0,0)$ и шине, то из точки $(0,0)$ в каждый из четырех соседних узлов течет ток равный $1/4$. Если же источник единичного тока подключен к шине и точке $(1,0)$, то из четырех соседних узлов в точку $(1,0)$ втекает ток равный $1/4$. Значит, при подключении обоих источников тока по ребру $(0,0)-(1,0)$ будет течь ток $1/2$, и разность потенциалов в этих точках тоже будет равна $1/2$. Но между ними течет общий ток равный единице, поэтому эквивалентное сопротивление решетки между соседними узлами равно $1/2$.

Придадим этим рассуждениям математическую строгость. Снова предположим, что источник единичного тока подключен к точке $(0,0)$ и шине с нулевым сопротилением, наложенной на периметр прямоугольника $[-N, N+1] \times [-N, N]$. Согласно задаче 5.2 сопротивление R такого графа не превосходит $N^{1/2}$, поэтому при нулевом потенциале на шине потенциал V в точке $(0,0)$ не превосходит $N^{1/2}$:

$$V = IR = R \le N^{1/2}. \tag{3}$$

Покажем, что потенциал в точках, близких к наложенной шине, мало отличается от нуля. Обозначим через u_j наибольший потенциал на периметре прямоугольника $[-j, j+1] \times [-j, j]$. Тогда из гармоничности распределения потенциалов следует, что $u_{j-1} \ge 2u_j - u_{j+1}$ $(1 \le j \le N-1)$. Поэтому, если $u_{N-1} = \varepsilon$ (по предположению $u_N = 0$), то для всех j в пределах $0 \le j \le N$ будет выполняться неравенство $u_j \ge (N-j)\varepsilon$. В частности, $u_0 = V \ge N\varepsilon$. Учитывая неравенство (3), получаем, что $\varepsilon \le N^{-1/2}$.

Если теперь наложить шину по периметру квадрата $[-N, N] \times [-N, N]$, то получится новое (симметричное) распределение потенциалов, которое, согласно принципу максимума, отличается от исходного не более чем на $N^{-1/2}$. Значит, до переноса правого края шины четыре тока, выходящие из точки $(0,0)$ отличались от $1/4$ не более чем на $N^{-1/2}$. Аналогично, если источник тока подключен к

шине и к точке $(1,0)$, то из четырех соседних узлов в точку $(1,0)$ втекают четыре тока, отличающиеся от $1/4$ не более чем на $N^{-1/2}$.

Объединяя обе ситуации, получаем, что в прямоугольнике $[-N,N+1]\times[-N,N]$ с закороченным периметром источник единичного тока подключен к узлам $(0,0)$ и $(1,0)$, а по ребру $(0,0)-(1,0)$ течет ток $1/2+O(N^{-1/2})$. Значит, разность потенциалов, а соответственно и сопротивление, равны $1/2+O(N^{-1/2})$.

Для завершения доказательства применим аксиому 3. Рассмотрим батарейку, подключенную к точкам $(0,0)$ и $(1,0)$ так, что потенциалы в этих точках равны $1/4$ и $-1/4$ соответственно. Ток, протекающий между $(0,0)$ и $(1,0)$ обозначим через I. Выберем прямоугольник $[-N,N+1]\times[-N,N]$ так, чтобы потенциалы на его периметре по модулю не превосходили некоторого $\varepsilon>0$. При замене всех потенциалов на периметре прямоугольника нулями, согласно принципу максимума, все потенциалы внутри так же изменятся не более чем на $\varepsilon>0$. Тогда получится, что ток I течет между узлами с разностью потенциалов $1/2+O(\varepsilon)$, а сопротивление между которыми равно $1/2+O(N^{-1/2})$. Следовательно $I=1+O(\varepsilon)+O(N^{-1/2})$. Так как ε может быть выбрано сколь угодно малым, а N растет при уменьшении ε, то ток I в точности равен единице. Он течет между узлами с разностью потенциалов $1/2$, значит, эквивалентное сопротивление решетки между соседнми узлами в точности равно $1/2$.

3.5 Пусть теперь B_1 и B_2 — вершины графа, диаметрально противоположные A_1 и A_2 соответственно. Мы доказали, что если в A_1 подается единичный ток, который вытекает из вершины A_2, то по ребру A_1A_2 течет ток $I=\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)$. Ток текущий при этом по ребру B_2B_1 обозначим через x. Дополнительно подключим источник единичного тока к вершинам B_1 и B_2 (в B_2 подается, из B_1 — вытекает). Тогда по каждому из ребер A_1A_2 и B_2B_1 будет течь ток $I+x$. Но то же распределение токов получится, если единичный ток подается в вершину A_1 и вытекает из B_1, а дополнительный ток подается в B_2 и вытекает из A_2. В такой ситуации в силу правильности графа первый источник по ребру A_1A_2 дает ток $\frac{1}{k_1}$, а второй — $\frac{1}{k_2}$. Отсюда

$$I+x=\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)+x=\frac{1}{k_1}+\frac{1}{k_2},$$
$$x=\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\frac{1}{n},\qquad \frac{x}{I}=\frac{1}{n-1}.$$

Поэтому для икосаэдра получается ток $\frac{I}{11}$, додекаэдра — $\frac{I}{19}$, ромбододекаэдра — $\frac{I}{13}$, куба — $\frac{I}{7}$.

3.6 *Указание.* Суммы D_n и F_n удовлетворяют одному и тому же рекуррентному соотношению. Например, $D_n=1+\dfrac{n+1}{2n}D_{n-1}$. Кроме того, $D_0=F_0=1$. Следовательно, они равны друг другу. Значит, при $n\geq 1$

$$\sum_{k=1}^{n}\frac{2^k}{k}=\frac{2^n}{n}F_{n-1}=\frac{2^n}{n}D_{n-1}=\frac{2^n}{n}\sum_{k=0}^{n-1}\frac{1}{C_{n-1}^k}.$$

Для завершения доказательства нужно оценить степень вхождения числа 2 в общий знаменатель дробей из полученной суммы с помощью формулы Лежандра для показателя, с которым простое число входит в разложение факториала.

Сопротивление n-мерного проволочного куба (у которого каждое ребро имеет единичное сопротивление) между противоположными вершинами R_n связанно с данными суммами равенствами (более подробно см. в статье [5]).

$$D_n=F_n=(n+1)R_{n+1}.$$

4.1 *Указание.* Докажите по индукции, что сопротивление бинарного дерева глубины n из единичных резисторов равно $1-\frac{1}{2^n}$.

4.2 *Указание.* Потенциалы в точках, расположенных на одинаковом расстоянии от корня дерева, равны из симметрии. Объединив такие точки в бинарном дереве, получим цепь, изображенную на рисунке 12. Ее сопротивление равно $\frac{1}{2}\cdot n=\frac{n}{2}$. Для троичного дерева аналогично получаем $R=\frac{1}{3}+\frac{2}{9}+\cdots+\frac{2^{n-1}}{3^n}$. Отсюда $R=1-\frac{2^n}{3^n}$.

4.3 *Указание.* Двоичное дерево глубины 3 вырезать не сложно. Покажем, что двоичное дерево глубины 2010 вырезать нельзя. Если его удалось вырезать, то все его вершины расположены на расстоянии

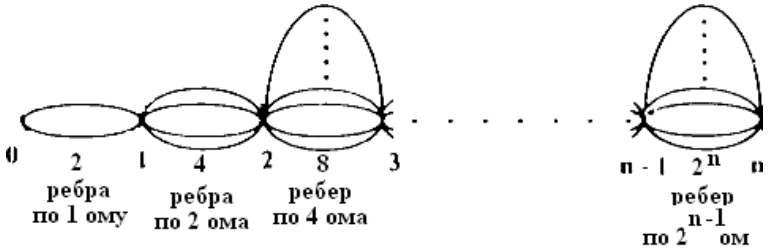


Рис. 12: Подсчет сопротивления дерева; см решение задачи 4.2.

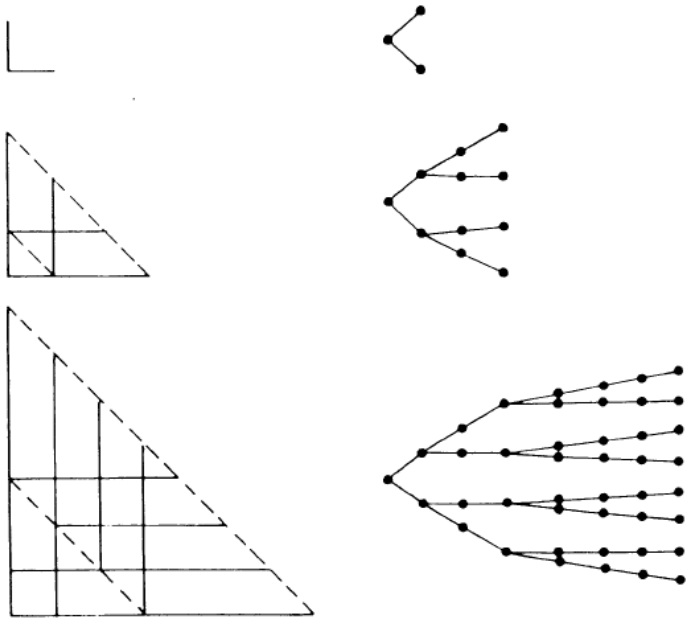


Рис. 13: Вырезание двоичного дерева с пересечениями из плоскости; см решение задачи 4.4.

не более 2010 от корня; отсюда получаем, что дерево находится в кубе со стороной $2\cdot 2010+1$. Поэтому число его вершин не превосходит $4021^3 \le 2^{36}$. С другой стороны, число его вершин равно $2^{2011}-1$. Полученное противоречие завершает доказательство. Задача о вырезании модифицированного дерева решается не просто.

4.4 *Указание.* Двоичное дерево вырезать нельзя; рассуждайте аналогично решению задачи 4.3, пользуясь тем, что более двух вершин склеиться не могут. Модифицированное двоичное дерево можно вырезать из плоскости (см рисунок 13), а троичное из пространства аналогичным образом (см рисунок 14). Доказательство проводится индукцией по глубине дерева.

4.5 *Указание.* Для любого $n = 2^i - 1$ рассмотрим множество вершин (x, y, z), где $|x| + |y| + |z| \le n$. Пусть R_i — сопротивление между началом координат и границей такой фигуры. Как известно из задачи 4.4, из такой части решетки можно вырезать модифицированное троичное дерево глубины i с пересечениями ребер на равном расстоянии от корня. Легко заметить, что сопротивление дерева с такими пересечениями равно сопротивлению такого же дерева без пересечений. Как известно из задачи 4.2, сопротивления модифицированных троичных деревьев не превосходят 1. Поэтому не превосходят 1 и сопротивления вырезаемых деревьев с пересечениями. Из закона монотонности получаем, что $R_i \le 1$. Значит, при подключении батарейки в 1 вольт ток будет не меньше 1. Следовательно, потенциалы в вершинах, соседних с началом координат будут не больше $1 - \frac{1}{6} = \frac{5}{6}$. Они равны вероятности возврата в начало координат до попадания на границу. Переходя к пределу, получаем требуемое.

5.1 Рассмотрим точки $A(2,3)$, $B(3,3)$, $C(1,2)$, $D(2,2)$, $E(3,2)$, $F(4,2)$, $G(2,1)$, $H(3,1)$. Сначала подсоединим источник тока, который подает единичный ток в начало координат, а вторым концом подсоединен к контуру с нулевым сопротивлением, наложенному на периметр квадрата $[-R, R]^2$. Из-за сим-

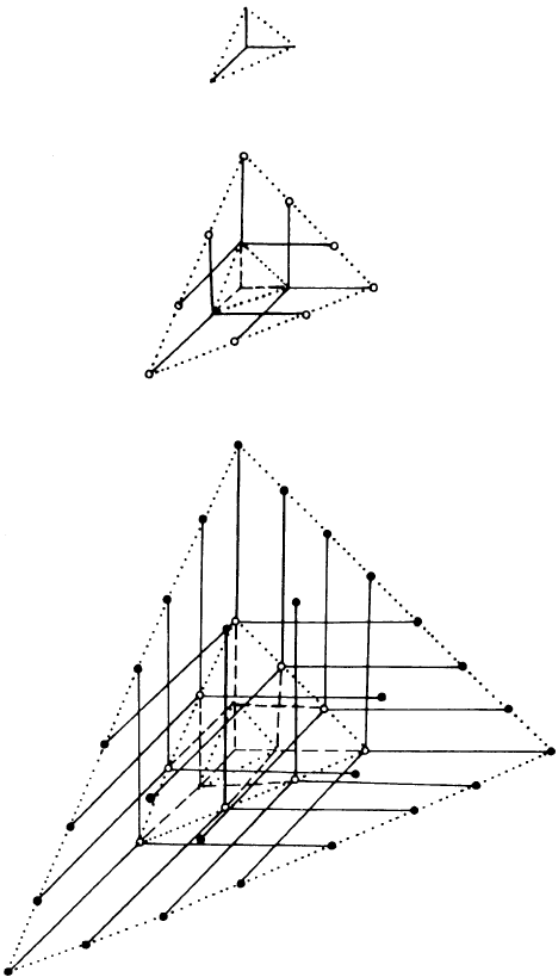


Рис. 14: Вырезание троичного дерева с пересечениями из пространства; см решение задачи 4.4.

метрии для некоторого числа I будут выполняться равенства $i(CD) = i(GD) = i(DA) = i(DE) = I$. Рассмотрим теперь другую ситуацию, когда единичный ток подается на периметр квадрата $[-R, R]^2$ и выходит из точки $(1, 0)$. Тогда $i(DE) \approx i(HE) \approx i(EB) \approx i(EF) \approx -I$. Здесь равенство понимается с точностью до малого ε, которое стремится к нулю с ростом R (как и в задаче 3.4, это следует из аксиомы 3). Комбинируя обе ситуации получаем, что когда источник тока подключен к узлам $(0, 0)$, $(1, 0)$, а на периметр квадрата $[-R, R]^2$ наложен контур с нулевым сопротивлением, по ребру DE течет ток меньший, чем ε. Устремляя R в бесконечность (и снова применяя аксиому 3), приходим к утверждению задачи.

Отметим, что при решении мы без доказательства пользовались сложным утверждением о существовании и единственности.

5.2 Рассмотрите какое-нибудь дерево, соединяющее данную точку с периметром квадрата.

5.3 Примените принцип максимума.

5.4 Запишем оператор Лапласа в виде

$$\Delta f(x, y) = f(x - 1, y) + f(x + 1, y) - 2f(x, y) + f(x, y - 1) + f(x, y + 1) - 2f(x, y).$$

Тогда для функции $f(x, y) = \ln r(x, y)$

$$f(x - 1, y) + f(x + 1, y) - 2f(x, y) = \frac{1}{2} \ln \frac{((x + 1)^2 + y^2)((x - 1)^2 + y^2)}{(x^2 + y^2)^2} =$$
$$= \frac{1}{2} \ln \left(1 + \frac{2x + 1}{r^2}\right) \left(1 + \frac{-2x + 1}{r^2}\right) = \frac{1}{2} \ln \left(\left(1 + \frac{1}{r^2}\right)^2 - \frac{4x^2}{r^4}\right).$$

Аналогично

$$f(x, y - 1) + f(x, y + 1) - 2f(x, y) = \frac{1}{2} \ln \left(\left(1 + \frac{1}{r^2}\right)^2 - \frac{4y^2}{r^4}\right).$$

Поэтому

$$\Delta f(x, y) = \frac{1}{2} \ln \left(\left(1 + \frac{1}{r^2}\right)^2 - \frac{4x^2}{r^4}\right) \left(\left(1 + \frac{1}{r^2}\right)^2 - \frac{4y^2}{r^4}\right) =$$
$$= \frac{1}{2} \ln \left(1 - \frac{1}{r^4} + \frac{16x^2y^2}{r^8}\right) = \frac{1}{2} \ln \left(1 + O\left(\frac{1}{r^4}\right)\right) = O\left(\frac{1}{r^4}\right).$$

5.5 Если точка (x, y) лежит вблизи границы кольца, то в ней имеет смысл изменить определение оператора Лапласа, чтобы оно согласовывалось с правилами Кирхгофа. Например, если для некоторых a, $b \in [0, 1)$ точки $(x - a, y)$ и $(x, y - b)$ попадают на границу, то будем считать, что

$$\Delta f(x, y) = \frac{f(x - a, y) - f(x, y)}{a} + \frac{f(x, y - b) - f(x, y)}{b} + f(x + 1, y) + f(x, y + 1) - 2f(x, y).$$

Тогда для функции $f(x, y) = \ln r(x, y)$ в такой точке

$$\frac{f(x - a, y) - f(x, y)}{a} + f(x + 1, y) - f(x, y) = \frac{1}{2a} \ln \left(1 + \frac{-2ax + a^2}{r^2}\right) + \frac{1}{2} \ln \left(1 + \frac{2x + 1}{r^2}\right) = O\left(\frac{1}{r^2}\right).$$

Аналогично

$$\frac{f(x, y - b) - f(x, y)}{b} + f(x, y + 1) - f(x, y) = O\left(\frac{1}{r^2}\right).$$

Таким образом $\Delta f(x, y) = O\left(r^{-2}\right)$, причем эта оценка остается справедливой, если из четырех соседних с (x, y) точек лишь одна лежит за пределами кольца.

Рассмотрим теперь функцию $f(x, y) = U_n(x, y) - \ln r(x, y)$. Она равна нулю на границе кольца, а во всех внутренних точках удовлетворяет уравнению $\Delta f(x, y) = \varphi(x, y)$, где $\varphi(x, y) = O(n^{-2})$ в точках вблизи границы, и $\varphi(x, y) = O(n^{-4})$ в остальных точках кольца.

Значения функции $\varphi(x, y)$ — это токи, которые подаются в соотвествующие узлы кольца. Потенциалы, индуцированные токами в точках вблизи границы оцениваются как $O(n^{-2})$. Действительно, если, считать, что во все внутренние точки области имеют равный потенциал U, то в точки границы текут токи не меньше чем U. Поэтому, при замене всех токов на U потенциалы внутри области не увеличатся.

Оценим теперь потенциал порождаемый токами в остальных точках (отделенных от границы). Число таких точек есть $O(n^2)$, и ток в каждой из них (согласно задаче 5.3) приводит к потенциалам не превосходящим $O(n^{-7/2})$. Поэтому общий потенциал, порожденный внутренними токами есть $O(n^{-3/2})$. То есть $f(x, y) = O(n^{-3/2})$.

5.6 Воспользуемся равенством

$$\operatorname{arctg} x - \operatorname{arctg} y = \operatorname{arctg} \frac{x - y}{1 + xy},$$

которое справедливо при $|xy| < 1$. Тогда

$$\operatorname{arctg} \frac{y+1}{R} - \operatorname{arctg} \frac{y}{R} = \operatorname{arctg} \frac{1/R}{1 + y(y+1)/R^2} =$$
$$= \operatorname{arctg} \left(\frac{R}{y^2 + R^2} + O\left(\frac{1}{R^2}\right)\right) = \frac{R}{R^2 + y^2} + O\left(\frac{1}{R^2}\right).$$

5.7 Просуммируйте формулу из задачи 5.6.

5.8 Для нахождения сопротивления кольца найдем ток, который будет через него протекать при условии, что на внутренний контур кольца подается напряжение $\ln nr_1$, а на внешний — $\ln nr_2$. Будем искать ток протекающий через периметр квадрата $[-R - 1/2, R + 1/2]^2$, где $R = [r_1 n] + 1$. Посчитаем его приближенно, заменяя потенциалы в узлах на значения функции $\ln r(x, y)$. Всего будет просуммировано $O(n)$ токов, каждый с погрешностью $O(n^{-3/2})$. Поэтому итоговая погрешность будет равна $O(n^{-1/2})$.

В силу симеетрии квадрата, протекающий через его периметр ток может быть записан в виде

$$I = 8 \sum_{y=0}^{R} (\ln r(R+1, y) - \ln r(R, y)) + O\left(\frac{1}{n^{1/2}}\right).$$

Так как

$$\ln r(R+1, y) - \ln r(R, y) = \frac{R}{R^2 + y^2} + O\left(\frac{1}{R^2}\right),$$

то, применяя формулу из задачи 5.7, получаем нужное равенство $I = 2\pi + O(n^{-1/2})$.

5.9 Как и в предыдущей задаче, для вычисления сопротивления нужно найти ток, протекающий через замкнутую ломаную, опоясывающую внутреннюю окружность. Снова для приближенного нахождения тока потенциалы в узлах решетки можно заменить на значения функции $\ln r(x, y)$. Если теперь ломаную заменить на описанный около нее квадрат, то внутри полученного контура появится $O(n^2)$ новых источников тока, в каждом из которых втекает или вытекает ток равный $O(n^{-4})$. Значит, искомое значение для суммарного тока отличается от уже найденного тока через периметр квадрата $I = 2\pi + O(n^{-1/2})$ не более чем на $O(n^{-2})$.

5.10 Докажем, что в задаче 5.2 сопротивление между любым узлом квадрата и его границей есть $O(\ln n)$. Тогда и во всех следующих задачах при буквальном повторении доказательств все остаточные члены умножаться на $\frac{\ln n}{\sqrt{n}}$.

Вместо квадрата рассмотрим треугольник, вырезанный из квадратной сетки прямыми $x = 0$, $y = 0$, $x + y = n$, и оценим его сопротивление между началом координат и гипотенузой. Будем предполагать, что в целых точках на отрезка $x + y = k$, $x, y \geq 0$ потенциалы равны $V_k = \sum_{j=2}^{k+1} \frac{1}{j}$ $(0 \leq k \leq n)$. В частности, в начале координат потенциал нулевой. Будем также предполагать, что в каждую точку на прямой $x + y = k$ втекает ток $1/k$, то есть через каждый уровень протекает единичный ток. Чтобы ситуация не противоречила закону Ома, сопротивления внутри треугольника

нужно будет увеличить. (Сопротивления, соединяющие точки вида $(0,k)$, $(k,0)$ с точками $(0,k+1)$, $(k+1,0)$ остаются единичными.) Чтобы выполнялся закон Кирхгофа, из точки $(j,k-j)$ должны течь токи $\frac{k-j}{k(k+1)}$ и $\frac{j+1}{k(k+1)}$ в точки $(j,k-j+1)$ и $(j+1,k-j)$ соответственно. Посколку разность потенциалов равна $1/(n+1)$, то единичные сопротивления нужно будет заменить на сопротивления $\frac{k}{k-j}$ и $\frac{k}{j+1}$ соответственно. Полученная схема имеет сопротивление $V_n \le \ln n$. Значит, сопротивление исходной схемы также не превосходит $\ln n$.

8. Благодарности

Большинство задач частей 1, 2 и 4 данного проекта заимствованы из статьи П. Дойля и Дж. Снелл [6]. Авторы благодарны И. Богданову, В. Бугаенко и М. Прасолову за помощь при переводе данногог проекта.

[1] I. Benjamini and O. Schramm, Random walks and harmonic functions on infinite planar graphs using square tilings, Ann. Prob. **24:3** (1996), 1219–1238.

[2] J. Cannon, W. Floyd, W. Parry, Squaring rectangles: the finite Riemann mapping theorem, Contemp. Math. **169** (1994), 133–211.

[3] P. G. Doyle and J. L. Snell, Random walks and electric networks, Mathematical Association of America, 1984, http://arxiv.org/abs/math.PR/0001057.

[4] Г.А. Гальперин "Мой друг Андрей Ходулчв" (Математическое просвещение, сер. 3, вып. 4 (2000), 8–32, http://www.mccme.ru/free-books/matpros/i5008032.pdf.zip.

[5] Ф. Недемейер и Я. А. Смородинский, *Сопротивление ребер многомерного куба*, Квант, № 6, 1986.

[6] M. Prasolov and M. Skopenkov, *Tiling by rerctangles and alternating current*, submitted (2010). http://arxiv.org/abs/1002.1356.

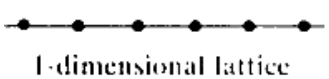
[7] F. Spitzer, Principles of random walks, Springer–Verlag, 1976.

Random walks through electrical networks

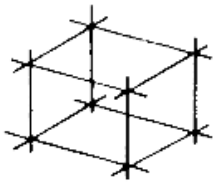
Dmitry Baranov, Mikhail Skopenkov, Alexey Ustinov

The aim of the project is to prove the following result and investigate related problems.
The Polya Theorem. (a) *A man which is randomly walking in a* 2*-dimensional lattice will return to the initial point with probability* 1.
(b) *A man which is randomly walking in a* 3*-dimensional lattice will return to the initial point with probability strictly less than* 1.

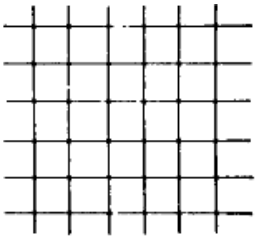
Accurate statements are given in the project. The suggested approach to the result is based on a physical interpretation. However, no physical background is assumed.



1-dimensional lattice



3-dimensional lattice



2-dimensional lattice

---

1. Walking in one dimension

Let us first state a problem and then give all the necessary definitions.

1.1. A man walks along a 5-block stretch of Madison Avenue. He starts at corner x and, with probability 1/2, walks one block to the right and, with probability 1/2, walks one block to the left; when he comes to the next corner he again randomly chooses his direction along Madison Avenue. He continues until he reaches corner 5, which is home, or corner 0, which is a bar. If he reaches either home or the bar, he stays there; see Figure 1.

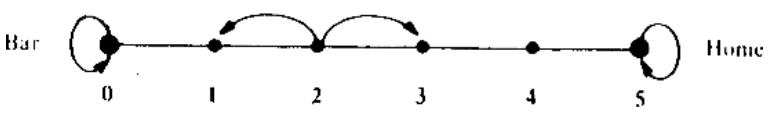


Figure 1: Random walk along Madison Avenue; see Problem 1.1.

(A)* Write a computer program which models the motion of the man. Run the program a large number of times, and find the percentage of cases in which the man returns home. You may use this to guess the answers in further problems.

E-mail address: dimbaranov@mail.ru, skopenkov@rambler.ru, ustinov.alexey@gmail.com

(B) Let $P_T(x)$ be the probability that the man, starting at corner x and making at most T "moves", will reach home. Complete the following table by 2-digit decimals.

Table 1: The probabilities $P_T(x)$ for small T

T \ x	0	1	2	3	4	5
1	0.00	0.00	0.00	0.00	0.50	1.00
2						
3						
4						

(C) Find the probability $P(x)$ that the man will reach home eventually.

Definition. (A) Suppose that an experiment has n *equally possible* outcomes, and an event X occurs in exactly m of the outcomes. Then the *probability* of the event X is by definition the number $P_1(X) := m/n$.

For instance, the probability of getting tails in a coin throw is $1/2$; the probability of getting 6 points in a die roll is $1/6$; the probability that our walker moves one block to the right is $1/2$.

(B) Now suppose that the event X depends on a sequence of such experiments. A sequence of T experiments has n^T possible outcomes. Assume that X occurs for exactly m_T outcomes among them. Then the *probability* of X is the number $P_T(X) := m_T/n^T$.

For instance, there are 4 possible outcomes of throwing a coin 2 times:

1st throw	heads	heads	tails	tails
2nd throw	heads	tails	heads	tails

Let the event X be getting tails in at least one throw. The event X occurs for 3 cases among the 4 possible ones. Thus the probability of the event X is $P_2(X) = 3/4$.

The probability of getting more than 10 points in two die rolls is $1/12$ because this event occurs for 3 cases (5+6, 6+5 or 6+6) among the 36 possible ones. The probability that our walker, starting at corner 3, consequently moves right twice is $1/4$.

(C) Finally, suppose that the event X depends on an infinite sequence of such experiments. We say that the *probability* of the event X is $P(X)$, if the probabilities $P_T(X)$ tend to a number $P(X)$ as T tends to infinity[1].

For instance, the probability of getting tails at least once in an infinite sequence of coin throws is $P(X) = 1$, because $P_T(X) = 1 - 1/2^T$ tends to 1 as T tends to infinity.

You may use without proof that the probability $P(X)$ *exists* for all the events X considered in the project.

1.2. Peter and Paul match pennies; they have a total of 5 pennies; on each match, Peter wins one penny from Paul with probability $1/2$ and loses one with probability $1/2$; they play until Peter's fortune reaches 0 (he is ruined) or reaches 5 (he wins all Paul's money). Find the probability $P(x)$ that Peter wins if he starts with x pennies.

[1] Formally, this means that for each $\varepsilon > 0$ there is a number T_0 such that for each $T > T_0$ we have $|P(X) - P_T(X)| < \varepsilon$.

1.3. Assume that our walker has a tendency to drift in one direction: more specifically, assume that each step is to the right with probability p or to the left with probability $q = 1 - p$. Find the probability $P(x)$ in this case.

1.4. You are gambling against a professional gambler; you start with 20 dollars and the gambler with 50 dollars; you play a game in which you win one dollar with probability 0.45 and lose one dollar with probability 0.55; play continues until you or the gambler runs out of money. Find the probability of being ruined.

Definition. An *electrical network* is a connected finite graph with a positive real number (*conductance*[2] $C(xy)$) assigned to each edge xy, and two disjoint marked sets of vertices (P and N). The vertices of the set N are joined with the ground and the negative pole of a battery, and the vertices of P are joined with the positive pole; see Figure 2.

The *voltages* $v(x)$ of the vertices in the network are defined by the following axioms:

1. *Boundary condition.* If $x \in N$ then $v(x) = 0$; if $x \in P$ then $v(x) = 1$.
2. *Kirchhoff current law.* If $x \notin P \cup N$ then $\sum_{xy} C(xy)\,(v(x) - v(y)) = 0$, where the summation is over all the edges xy containing the vertex x.

The number $i(xy) := C(xy)\,(v(x) - v(y))$ is the *current* through edge xy; $i(x) := \sum_{xy} i(xy)$ is the *current* flowing inside the network through vertex x (thus $i(x) = 0$ for each $x \notin P \cup N$ by axiom 2); $C := \sum_{x \in P} i(x)$ is the *effective conductance* of the network between the subsets P and N; $Q := \sum_{xy} C(xy)\,(v(x) - v(y))^2$, where the summation is over all the edges of the network, is the *heat power* of the network.

1.5. We connect equal resistors in series and put a unit voltage across the ends as in Figure 2. Find the voltages $v(x)$ established at the points $x = 0, 1, 2, 3, 4, 5$. Hereafter you may use network-simulation software to guess the answer.

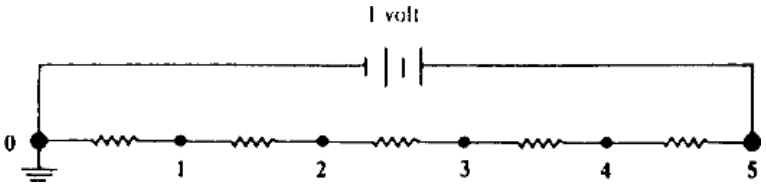


Figure 2: An electrical network; see Problem 1.5.

1.6. Consider the network with the vertices $0, 1, \ldots, n$, the edges $01, 12, \ldots, (n-1)n$ of unit conductance, and marked sets $N = \{0\}$, $P = \{n\}$.
(A) *Maximum Principle.* A function $v(x)$ satisfying the above axiom 2 takes on its maximum value and its minimum value on the set $P \cup N$.
(B) *Uniqueness Principle.* If $v(x)$ and $u(x)$ are two functions satisfying the above axioms 1–2 then $v(x) = u(x)$ for all x.
(C) Find the voltages $v(x)$ and the effective conductance of the network. To which numbers tend these values as n tends to infinity?

1.7. State and prove an analogue of the Polya theorem for the 1-dimensional lattice.

[2]The reciprocal of conductance is called *resistance*.

2. Walking in two dimensions

2.1. Consider the town in Figure 3 to the left. Segments represent streets. Large dots marked E indicate escape routes and those marked P are police. Find with 2-digit precision the probability $P(x)$ that our walker, starting at an interior point x, will reach an escape route before he reaches a policeman. The walker moves from $x = (a, b)$ to each of the four neighboring points $(a + 1, b), (a - 1, b), (a, b + 1), (a, b - 1)$ with probability $1/4$. If he reaches one of the points E or P, he remains at this point.

2.2. Find the voltages $v(x)$ in the network (of unit resistors) in Figure 3 to the right.

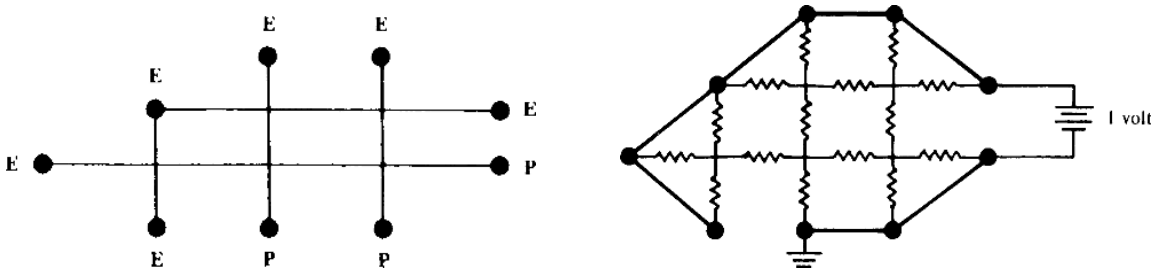


Figure 3: Random walk in a town and an electrical network; see Problems 2.1 and 2.2.

2.3. A bug walks randomly on the edges of
(A) a cube; **(B)** an octahedron; **(C)** a dodecahedron; **(D)** an icosahedron;
If the bug starts at a vertex a, what is the probability that it reaches food at the opposite vertex h before returning to a; see Figure 4 to the left?

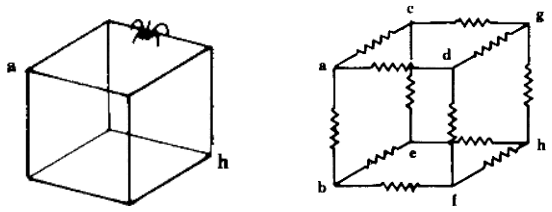


Figure 4: Random walk in a cube and an electrical network; see Problems 2.3(A) and 2.5(A).

2.4. The following transformations preserve the effective conductance of a network:
(A) replacing two resistors connected in series by a single resistor whose conductance is $1/\left(\frac{1}{C_1} + \frac{1}{C_2}\right)$; see Figure 5 to the left;
(B) replacing two resistors connected in parallel by a single resistor whose conductance is $C_1 + C_2$; see Figure 5 to the right;
(C) shortening together two vertices having the same voltage.

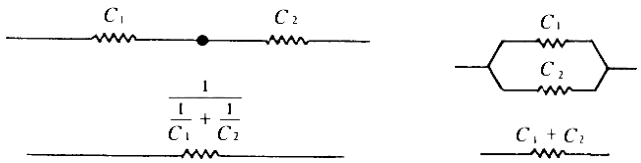


Figure 5: Series and parallel connections; see Problem 2.4.

2.5. Find the effective conductance between **(1)** opposite; **(2)*** adjacent; vertices of **(A)** a cube; **(B)** an octahedron; **(C)** a dodecahedron; **(D)** an icosahedron; with edges of unit conductance; see Figure 4 to the right.

2.6. A drunken tourist starts at her hotel and walks at random through the streets of the idealized Paris shown in Figure 6 to the left. Find the probability that she reaches the Arc de Triomphe before she reaches the outskirts of town.

2.7. The conductance between vertices **(A)*** a and b; **(B)**** a and c; of the 2-dimensional lattice (of unit resistors) equals to 2 and $\pi/2$, respectively; see Figure 6 to the right.

Notice that for the 2-dimensional lattice we need to modify the definition of the voltages $v(x)$ by adding one more axiom:

3. *Condition at infinity.* $v(x)$ tends to $1/2$ as the distance between x and a fixed vertex tends to infinity.

It is allowed to use without proof that there exists a function $v(x)$ satisfying axioms 1–3.

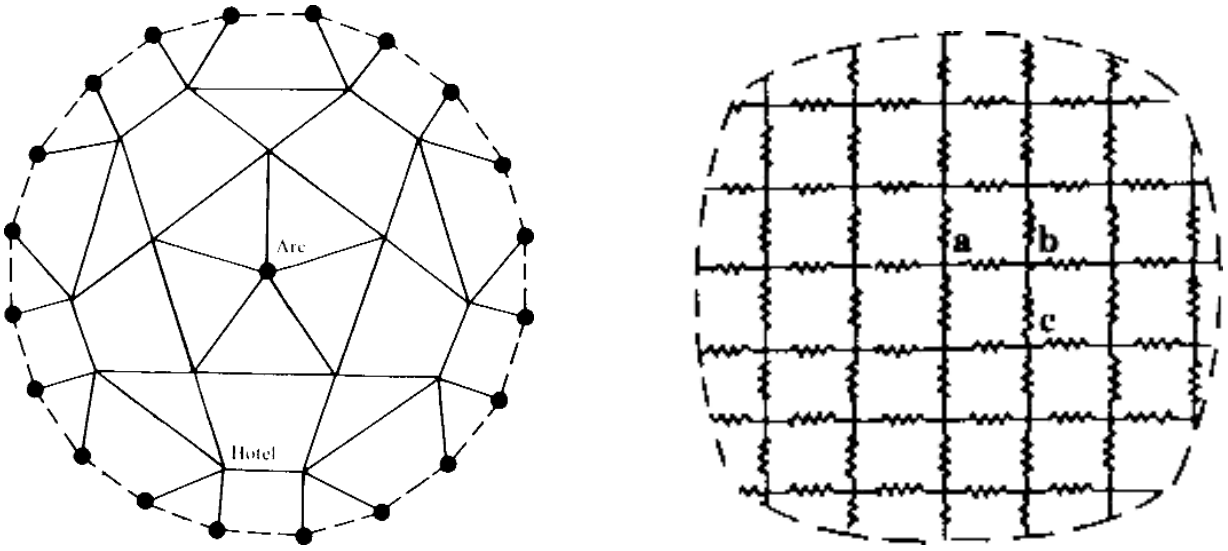


Figure 6: Paris tourist map and a lattice network; see Problems 2.6 and 2.7.

2.8. *Rayleigh's Monotonicity Law.* Cutting certain edges can only decrease the effective conductance between two given nodes; see Figure 7 to the left. Shorting certain sets of nodes together can only increase the effective conductance of the network between two given nodes; see Figure 7 in the middle.

2.9. **(A)** Prove that the conductance between the center and the boundary of a square 4×4 lattice of unit resistors is less than 3; see Figure 7 to the right.
(B) To which number tends the conductance between the center and the boundary of a square $2n \times 2n$ lattice of unit resistors as n tends to infinity?

2.10. Prove the Polya theorem for the 2-dimensional lattice.

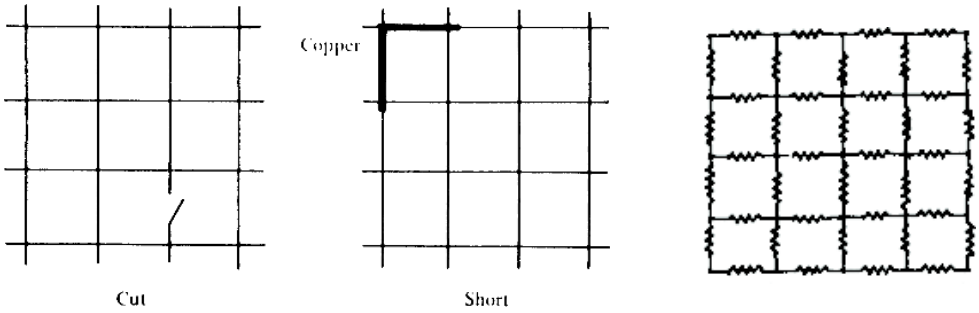


Figure 7: Cutting, shortening, and a square 4×4 lattice; see Problems 2.8 and 2.9.

3. Conductances of symmetric graphs

The problems of this sections may help to solve the problems from the previous one.

A function $v(x)$ on vertices of an electrical network is *harmonic*, if it satisfies axiom 2 from the definition of an electrical network. The vertices from the set $P \cup N$ are called *boundary* vertices, the remaining vertices are called *interior* ones.

3.1. (A) *Superposition principle.* If functions $u(x)$ and $v(x)$ are harmonic and $a, b \in \mathbb{R}$ then the function $au(x) + bv(x)$ is harmonic.
(B) *Maximum principle.* Prove that a harmonic function $u(x)$ defined on a finite electrical network takes its maximum and minimum values on boundary vertices.
(C) *Uniqueness principle.* If harmonic functions $u(x)$ and $v(x)$ coincide at each boundary vertex of a finite network then $u(x) = v(x)$ for each vertex x of the network.
(D) A man randomly walking in a finite town visits all street corners with probability 1.

3.2. (A)* *Fredholm's alternative.* Take a system of linear equations

$$\begin{cases} a_{1,1}x_1 + \ldots + a_{1,n}x_n = b_1, \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n,1}x_1 + \ldots + a_{n,n}x_n = b_n \end{cases},$$

in which the number of equations equals to the number of unknowns. Then exactly one of the following alternatives holds:

1. for any $b_1, \ldots, b_n$ the system has a unique solution (in particular, for $b_1 = \ldots = b_n = 0$ there is only zero solution);
2. for some $b_1, \ldots, b_n$ the system has no solutions, and for some other (in particular, for $b_1 = \ldots = b_n = 0$) it has infinitely many solutions.

(B)* *Dirichlet's problem.* Prove that for any finite electrical network there exist a function $v(x)$ satisfying axioms 1–2.

3.3. (A) *Variational principle.* Let $v(x)$ be an arbitrary function on vertices of a finite electrical network satisfying axiom 1, but *not* necessarily axiom 2. Enumerate the vertices by $1, \ldots, n$ and let $1, \ldots, k$ be the interior vertices. Denote by $v_1 := v(1)$, $v_2 := v(2)$, $\ldots$, $v_n = v(n)$. Consider $v_1, \ldots, v_k$ as independent variables. Consider the heat power $Q(v_1, \ldots, v_k) := \sum_{xy} C(xy)\,(v_x - v_y)^2$ as a function in variables $v_1, \ldots, v_k$. Prove that the function $Q(v_1, \ldots, v_k)$ takes its minimum when the function $v(x)$ is harmonic.

(B) Prove that for the function $Q(v_1, \dots, v_k)$ there exist a unique sequence $v_1, \dots, v_k$, for which the function takes its minimum. Apply this to obtain the second proof that for each finite electrical network there is a unique function $v(x)$ satisfying axioms 1–2.
(C) *Energy conservation law.* Prove that the minimal value of the heat power $Q(v_1, \dots, v_k)$ numerically equals to the effective conductance C.
(D) *Rayleigh's Monotonicity Law.* If the conductances of the edges of a network are increased, the effective conductance can only increase.
(E) One removed an arbitrary set of edges from the 2-dimensional lattice. Prove that a random walk still returns to the initial vertex with probability 1.

Let Γ be an electrical network of unit resistors. Take a pair of adjacent edges AB and AC of the network. These edges are called *equivalent*, if there is a permutation of vertices of the network, taking adjacent vertices to adjacent ones, and taking A to A and B to C. A vertex is a *symmetry center* of the network Γ, if all the edges containing the vertex are equivalent. The network Γ is *regular*, if all its vertices are symmetry centers.

Examples of regular graphs: regular polyhedra of arbitrary dimensions, regular lattices of arbitrary dimensions, regular lattices in a torus etc. A nontrivial example: graph of a rombododecahedron. The vertices of this polyhedron have different degrees (3 and 4).

3.4. (A) A regular network contains n vertices. Let A_1 and A_2 be two adjacent vertices of degrees k_1 and k_2, respectively. Prove that the resistance between them is

$$\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right).$$

(B) If the network is 2-dimensional lattice then $1/n$ should be replaced by 0.

3.5. The battery is joined with two adjacent vertices of **(A)** an icosahedron; **(B)** a dodecahedron; so that the current through the edge joining the two vertices is I. Find the current through the opposite edge.

3.6. * Prove that the sums $D_n = \sum_{k=0}^{n} \frac{1}{C_n^k}$ and $F_n = \frac{n+1}{2^{n+1}} \sum_{k=1}^{n+1} \frac{2^k}{k}$ are equal. How these sums are connected with the conductance of an n-dimensional cube? Apply this to prove that the order of 2 in the number $\sum_{k=1}^{n} \frac{2^k}{k}$ tends to infinity as n tends to infinity.

4. Walking in three dimensions

4.1. Find the conductance of a binary tree of depth **(A)** 3; **(B)** 2010;
made of unit resistors; see Figure 8 to the left.

4.2. Find the conductance of a *modified* binary and a trinary trees of depth 2010, in which each resistor at k-th level is replaced by 2^k unit resistors in series; see Figure 8 in the middle.

4.3. Which trees from **(A)** Problem 4.1; **(B)** Problem 4.2; can be cut out from the 3-dimensional lattice?

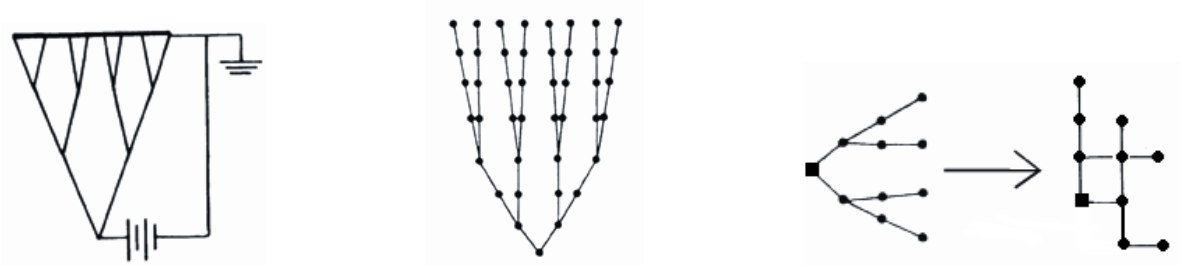


Figure 8: (Left) A binary tree of depth 3; (middle) a modified binary tree of depth 3; (right) allowed intersections of edges of the trees; see Problems 4.1, 4.2, and 4.4.

4.4. — And if one allows intersections of edges at equal distance from the root of the tree; see Figure 8 to the right?

4.5. Prove the Polya theorem for the 3-dimensional lattice.

5. Conductance of a ring*

Consider a metal 2-dimensional lattice with unit resistance of each edge.

5.1. The battery is joined with nodes $(0,0)$ and $(1,0)$. Prove that the voltages at the nodes $(2,2)$ and $(3,2)$ are the same; see the remark in Problem 3.4(B).

5.2. Prove that the resistance between any node of a square $n\times n$ lattice and the boundary is less than $\sqrt{n}$.

5.3. The battery is joined with an interior node A of the square $n \times n$ lattice and with the boundary. The voltage at the boundary is zero. Prove that if the current through node A is ε, then the voltage at each other node is less than $\varepsilon\sqrt{n}$.

5.4. Denote by $\Delta f(x,y) = f(x-1,y)+f(x+1,y)+f(x,y-1)+f(x,y+1)-4f(x,y)$ and $r(x,y) = \sqrt{x^2+y^2}$. If $f(x,y) = \ln r(x,y)$ for $r(x,y) \geq 2$ then $\Delta f(x,y) = O\left(\frac{1}{r^4(x,y)}\right)$.

Hereafter for two functions A and B we write $A = O(B)$, if there exist a positive constant c we have $|A| \leq cB$.

5.5. A ring with inner radius $r_1 n$, outer radius $r_2 n$, and center at the origin is cut out from the 2-dimensional lattice. If an edge is cut then the resistance of the part is proportional to its length. Assign the voltage $\ln nr_1$ to the inner boundary circle and $\ln nr_2$ — to the outer one. Let $U_n(x,y)$ be the voltage at node (x,y). Prove that for each (x,y) inside the ring $U_n(x,y) = \ln r(x,y) + O\left(\frac{1}{n^{3/2}}\right)$.

5.6. Using $\arctan x = x + O(x^3)$ prove that for each $0 \leq y < R$ we have $\frac{R}{R^2+y^2} = \arctan\frac{y+1}{R} - \arctan\frac{y}{R} + O\left(\frac{1}{R^2}\right)$.

5.7. Prove that $\sum_{y=0}^{R-1} \frac{R}{R^2+y^2} = \frac{\pi}{4} + O\left(\frac{1}{R}\right)$.

5.8. Assume that in Problem 5.5 $r_2 > 3r_1/2$ (so that our ring contains a square). Prove that the current incoming through the inner boundary is $2\pi + O\left(\frac{1}{\sqrt{n}}\right)$. Apply this to

get the following formula for the conductance of the ring:

$$R(r_1 n, r_2 n) = \frac{1}{2\pi} \ln \frac{r_2}{r_1} + O\left(\frac{1}{\sqrt{n}}\right). \tag{1}$$

5.9. Prove formula (1) without additional assumption $r_2 > 3r_1/2$.

5.10. Using (1) make estimations in Problems 5.2, 5.3 more precise and prove more precise formulas in Problems 5.5, 5.8:

$$U_n(x, y) = \ln r(x, y) + O\left(\frac{\ln n}{n^2}\right), \qquad R(r_1 n, r_2 n) = \frac{1}{2\pi} \ln \frac{r_2}{r_1} + O\left(\frac{\ln n}{n}\right).$$

6. Challenge*

6.1. *Liouville's theorem.* Suppose that a function $f(m, n)$ on $\mathbb{Z}^2$ satisfies the inequality $0 \leq f(m, n) \leq 1$ and the equality

$$f(m, n) = \frac{1}{4}\left(f(m-1, n) + f(m+1, n) + f(m, n-1) + f(m, n+1)\right) \tag{2}$$

for each $m, n \in \mathbb{Z}$. Prove that the function $f(m, n)$ is constant.

6.2. *Existence of a voltage.* Prove that there exists a function $f(m, n)$ on $\mathbb{Z}^2$ such that $f(0, 0) = 0$, $f(0, 1) = 1$, for each $(m, n) \neq (0, 0), (0, 1)$ equality (2) holds, and $f(m, n)$ tends to $1/2$ as $r(m, n) := \sqrt{m^2 + n^2}$ tends to infinity.

6.3. *Green's function.* Let $f(m, n)$ be the resistance of the 2-dimensional lattice between the origin and the point (m, n).
(A) Prove that for each $(m, n) \neq (0, 0)$ equality (2) holds.
(B) Prove that $f(m, n) = g(r(m, n)) + O(1)$ for some function $g(x)$.
(C) Prove that the resistance between the center and the boundary of a disc of radius r cut from the 2-dimensional lattice equals to $\frac{1}{2\pi} \ln r + O(1)$.
(D) Prove that $f(m, n) = \frac{1}{2\pi} \ln r(m, n) + O(1)$.

6.4. Find with 2 digit precision the probability that a random walk on a 3-dimensional lattice eventually returns to the initial point.

6.5. Robot walks on the vertices of the 3-dimensional lattice, each time moving from a vertex to one of the neighbors. One of the vertices contains a treasure, which is found when the robot reaches the vertex. Is there a program for a robot using a finite memory and a random number generator such that the robot finds the treasure with probability 1?

7. Hints to the solutions

1.1 (A) To check the correctness of the program we use the following criterion: the difference between the percentage and the probability should be approximately inverse proportional to the square root of the number of times the program is run.
(B) See the answer in the table. The proof is straightforward.

Table 2: The probabilities $P_T(x)$ and $P(x)$

x / T	0	1	2	3	4	5
1	0.00	0.00	0.00	0.00	0.50	1.00
2	0.00	0.00	0.00	0.25	0.50	1.00
3	0.00	0.00	0.13	0.25	0.63	1.00
4	0.00	0.06	0.13	0.38	0.63	1.00
$P(x)$	0.00	0.20	0.40	0.60	0.80	1.00

(C) *Answer*: $P(x) = x/5$; see the last row of the above table.
Proof. Consider a random walk on the integers $0, 1, 2, \ldots, n$. Let $P(x)$ be the probability, starting at x, of reaching N before 0. We regard $P(x)$ as a function defined on the points $x = 0, 1, 2, \ldots, n$. The function $P(x)$ has the following properties:

1. $P(0) = 0$ and $P(n) = 1$.
2. $P(x) = \frac{1}{2}P(x-1) + \frac{1}{2}P(x+1)$ for each $x = 1, 2, ..., n-1$.

Property 1 follows from our convention that 0 and n are traps; if the walker reaches one of these positions, he stops there; in the game interpretation, the game ends when one player has all of the pennies. Property 2 states that, for an interior point, the probability $P(x)$ of reaching home from x is the average of the probabilities $P(x-1)$ and $P(x+1)$ of reaching home from the points that the walker may go to from x. We can derive property 2 from the following basic fact about probability:

Basic Fact. *Let E be any event, and F and G be events such that one and only one of the events F or G will occur. Then*

$$P(E) = P(F) \cdot P(E \text{ given } F) + P(G) \cdot P(E \text{ given } G).$$

In this case, let E be the event "the walker ends at the bar", F the event "the first step is to the left", and G the event "the first step is to the right". Then, if the walker starts at x, $P(E) = P(x)$, $P(F) = P(G) = 1/2$, $P(E \text{ given } F) = P(x-1)$, $P(E \text{ given } G) = P(x+1)$, and property 2 follows.

Properties 1–2 imply together that $P(x)$ is the arithmetic progression $P(x) = x/n$.

1.2 *Answer*: $P(x) = x/5$; in fact this problem is equivalent to 1.1(C).

1.3 *Answer*: $P(x) = \frac{(q/p)^x - 1}{(q/p)^5 - 1}$.
Hint: Argue as in the solution of Problem 1.1(C). Show that properties 1–2 in the solution should be replaced by

1. $P(0) = 0$ and $P(n) = 1$.
2. $P(x) = qP(x-1) + pP(x+1)$ for each $x = 1, 2, ..., n-1$.

Show that you can choose A and B so that the function $f(x) = A(q/p)^x + B$ satisfies these modified properties.

1.4 *Answer*: $\approx 99.995\%$. The exact answer is $1 - \frac{(0.55/0.45)^{20} - 1}{(0.55/0.45)^{70} - 1}$; the problem is equivalent to 1.3.

1.5 *Answer*: $v(x) = x/5$. *Hint.* Axioms 1–2 imply that $v(x)$ is linear for this network.

1.6 (A) Let M be the largest value of $v(x)$. Then if $v(x) = M$ for $x \notin P \cup N$, the same must be true for $v(x-1)$ and $v(x+1)$ since $v(x)$ is the average of these two values. If $x - 1$ is still an interior point, the same argument implies that $f(x-2) = M$; continuing in this way, we eventually conclude that $f(0) = M$. That same argument works for the minimum value m.

(B) Let $h(x) = v(x) - u(x)$. Then if x is any interior point,

$$\frac{h(x-1)+h(x+1)}{2} = \frac{v(x-1)+v(x+1)}{2} - \frac{u(x-1)+u(x+1)}{2}$$

and $h(x)$ also satisfies axiom 2. But $h(x) = 0$ for x in $P \cup N$, and hence, by the Maximum Principle, the maximum and minimum values of h are 0. Thus $h(x) = 0$ for all x, and $v(x) = u(x)$ for all x.
(C) *Answer*: $v(x) = x/n$, $C = 1/n$; $C \to 0$ and $v(x) \to 0$ for each fixed x as $n \to \infty$.
Hint: It is easy to check that the function $f(x) := x/n$ satisfies axioms 1–2. By uniqueness principle it follows that $v(x) = x/n$.

1.7 Theorem. *A random walk, starting at the origin of the* 1*-dimensional lattice, eventually returns to the origin with probability* 1.
Proof. Let P be the probability that a random walk, starting at the origin, eventually returns to the origin. Let P_n be the probability that a random walk, starting at the origin, returns to the origin before reaching the points n and $-n$. Assume that all these probabilities exist. Clearly, then $P_n \le P \le 1$ for each n.

Let us prove that $P_n = 1 - 1/n$. After first "move" our walker comes to either point 1 or -1 with probability 1/2. Given that he comes to 1 by Problem 1.1(C) the probability that he returns to the origin before reaching the point n equals to $1-1/n$. Analogously, given that he comes to -1 the probability that he returns to the origin before reaching the point $-n$ equals to $1 - 1/n$. Applying Basic Fact from the solution of Problem 1.1(C) one gets $P_n = \frac{1}{2}\left(1 - \frac{1}{n}\right) + \frac{1}{2}\left(1 - \frac{1}{n}\right) = 1 - \frac{1}{n}$. (Alternatively, one can observe that $P_n = 1 - C$, where $C = 1/n$ is the conductance of the network from Problem 1.6.)

So $1 - 1/n \le P \le 1$ for each n, hence P must be 1. □

2.1 *Answer*: see Figure 9 to the left.

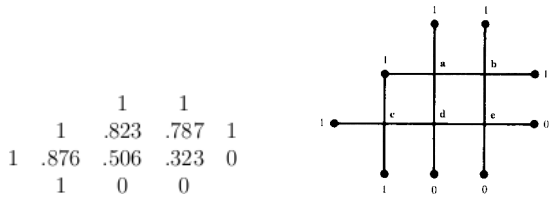


Figure 9: The probabilities $P(x)$ or, equivalently, the voltages $v(x)$; see Problems 2.1 and 2.2.

Hint. The town is shown again in Figure 9 to the right. The probabilities $P(x)$ are denoted by a, b, c, d, and e. Similarly to 1-dimensional case, the function $P(x)$ satisfies axioms 1–2 from the definition of an electrical network. Thus we get a system os linear equations:

$$\begin{aligned} a &= (b + d + 2)/4; \\ b &= (a + c + 2)/4; \\ c &= (d + 3)/4; \\ d &= (a + c + e)/4; \\ e &= (b + d)/4. \end{aligned}$$

Solving the system, we get the answer.

Remark. Finding the exact solution to a "Dirichlet problem" in two dimensions is not always a simple matter, so we will consider two methods for generating approximate solutions.

First let us present a method using random walks. This method is known as a *Monte Carlo method*, since random walks are random, and gambling involves randomness, and there is a famous gambling casino in Monte Carlo. We start many random walks at x and count the percentage of walks reaching the points marked by E. By the law of averages (the law of large numbers in probability theory), the estimate that we obtain this way will approach the true expected probability $P(x)$. This method is a colorful way to solve the problem, but quite inefficient.

Now let us present the more efficient *method of relaxations*. Recall that we are looking for a function that has specified boundary values, for which the value at any interior point is the average of the values at its neighbors. Begin with any function having the specified boundary values, pick an interior point, and

see what is happening there. In general, the value of the function at the point we are looking at will not be equal to the average of the values at its neighbors. So adjust the value of the function to be equal to the average of the values at its neighbors. Now run through the rest of the interior points, repeating this process. When you have adjusted the values at all of the interior points, the function that results will not satisfy axiom 2, because most of the time after adjusting the value at a point to be the average value at its neighbors, we afterwards came along and adjusted the values at one or more of those neighbors, thus destroying the harmony. However, the function that results after running through all the interior points is more nearly to satisfy axiom 2 than the function we started with; if we keep repeating this averaging process, running through all of the interior points again and again, the function will approximate more and more closely the solution to our problem.

2.2 *Answer*: see Figure 9 to the left; this problem is equivalent to 2.1.

2.3 *Answer*: **(A)** 2/5; **(B)** 1/2; **(C)** 2/7; **(D)** 2/5.
Hint. Reduce to Problem 2.5 using the following simple result:

Physical interpretation of probability. *The probability that a random walk in a graph G, starting at a vertex a, reaches a vertex h before returning to the initial point a, equals to*

$$P = C/\deg a,$$

where C is the conductance of the graph G (of unit resistors) between a and h, and $\deg a$ is the number of edges containing the vertex a.

2.4 (C) *Hint.* The function $v(x)$ is well-defined on the vertices of the network obtained by the shortening. Check that $v(x)$ still satisfies axioms 1–2.

2.5 *Answer*: **(1A)** 6/5; **(1B)** 2; **(1C)** 6/7; **(1D)** 2.
(2A) 12/7; **(2B)** 12/5; **(2C)** 30/19; **(2D)** 30/11.

For a short solution refer to section 3.

(2A) *Hint.* We put a unit battery between a and b; see Figure 4 to the right. Then, by symmetry, the voltages at c and d will be the same as will those at e and f. Thus our circuit is equivalent to the circuit shown in Figure 10 to the left.

Using the laws for the effective resistance of resistors in series and parallel, this network can be successively reduced to a single resistor of resistance 7/12 ohms, as shown in Figure 10 to the right. Thus the effective resistance is 7/12.

2.6 *Answer*: 1/7. Argue analogously to the solution of Problem 2.3.

2.7 (A). For a short solution refer to section 3.
(B) The authors do not know elementary solution of the problem. A nice solution based on discrete Fourier transformation can be found in the book [7].

2.8 Refer to section 3.

2.9 (B) *Answer*: $C \to 0$ as $n \to \infty$.
Hint. We apply Monotonicity Law as follows: short together nodes on squares about the origin, as shown in Figure 11 in the top. The network we obtain is equivalent to the network shown in Figure 11 in the middle. Now as n 1-ohm resistors in parallel are equivalent to a single resistor of resistance $1/n$ ohms, the modified network is equivalent to the network shown in Figure 11 in the bottom. The conductance of this network is

$$\frac{1}{\sum_{k=1}^{n} \frac{1}{8k-4}}.$$

This number tends to zero as n tends to infinity. As the conductance of the old network can only be smaller, we conclude that it too must tend to zero.

2.10 *Hint.* Let P be the probability that a random walk on the 2-dimensional lattice, starting at the origin, eventually returns to the origin. Let P_n be the probability that a random walk, starting at the origin, returns to the origin before reaching the boundary points of the $2n \times 2n$ square centered at the origin. Assume that all these probabilities exist. Clearly, then $P_n \le P \le 1$ for each n. By the physical interpretation of the probability one gets $P_n = 1 - C/4$, where C is the effective conductance between the

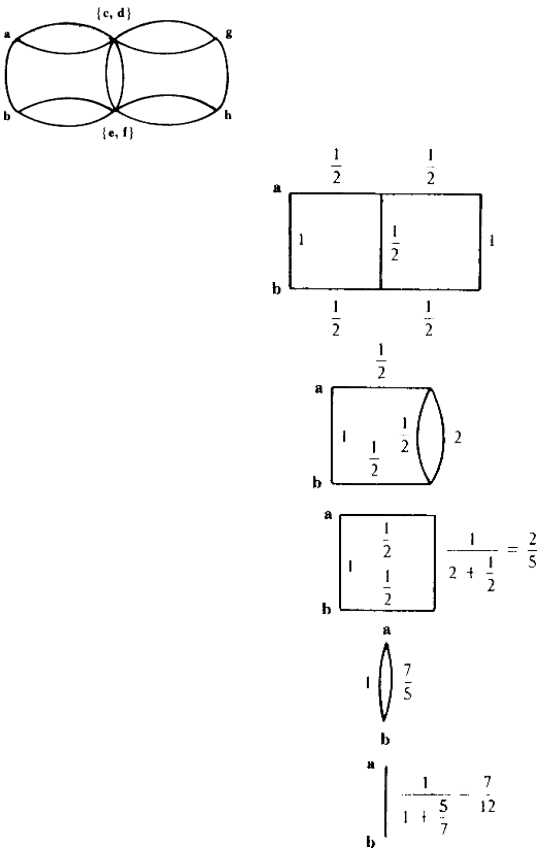


Figure 10: Simplification of a network; see the solution of Problem 2.5.

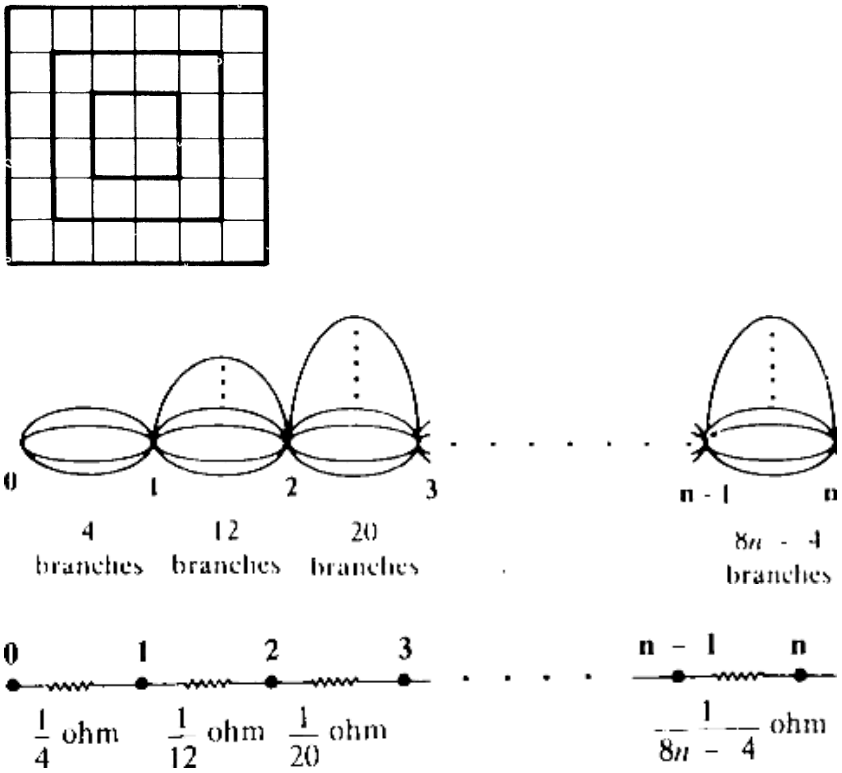


Figure 11: Shortening a square network and an equivalent network; see the solution of Problem 2.9.

center and the boundary of the $2n \times 2n$ square. By Problem 2.9(B) C tends to zero as n tends to infinity. Thus $P_n \to 1$ as $n \to \infty$, hence P must be 1. $\square$

3.1–3.3 E. g., see [6].

3.4 **(A)** First consider the following network: a current flowing through the vertex A_1 equals to $\frac{n-1}{n}$ and to $-\frac{1}{n}$ for all other vertices. Since graph network is regular the current flowing through the edge A_1A_2 equals to $\frac{1}{k_1}\left(1-\frac{1}{n}\right)$. Now consider a similar network: a current flowing through the vertex A_2 equals to $-\frac{n-1}{n}$ and to $\frac{1}{n}$ for all other vertices. Sum two networks and obtain by superposition principle the network with unit current source connected to A_1 and A_2. The current flowing through the edge A_1A_2 in obtained network equals to $\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)$. Therefore, the resistance between A_1 and A_2 equals to $\left(\frac{1}{k_1}+\frac{1}{k_2}\right)\left(1-\frac{1}{n}\right)$.

If the graph is infinite substitute $\frac{1}{n}$ by zero. An explanation of this answer bases on limiting process.

The formula of the resistance between two adjacent vertices of the regular graph is due to A.B.Hodulev. The definition of a regular graph and this formula are taken from [4].

(B) First we introduce some physical explanations. Apply a bus of zero resistance to the boundary of rectangle $[-N, N+1] \times [-N, N]$. Since the potential tends to zero at infinity, applying a bus changes the resistance a little bit. (Next we give our explanations with error which tends to zero as N grows to infinity). If a current source of unit current is connected with point $(0,0)$ and the bus then the current flowing through the edges outgoing from $(0,0)$ equals to $\frac{1}{4}$. If current source of unit current is connected to the bus and the point $(1,0)$ then the current flowing through the edges ingoing to $(1,0)$ equals to $\frac{1}{4}$. Therefore if both current sources are connected then the current flowing through the edge $(0,0)-(1,0)$ equals to $\frac{1}{2}$. So a potential difference between these points equals to $\frac{1}{2}$ also. But the current flowing through source connecting these points equals to 1. Therefore the resistance between these points equals to $\frac{1}{2}$.

Now we make our arguments more strict. Again suppose that a current source is connected to the point $(0,0)$ and a bus of zero resistance applied to the boundary of rectangle $[-N, N+1] \times [-N, N]$. According to Problem 5.2 the resistance of such graph is at most $N^{1/2}$. Hence if the bus is joined with the ground, the potential at the point $(0,0)$ is at most $N^{\frac{1}{2}}$:

$$V = IR = R \leq N^{1/2}. \tag{3}$$

We will show that the potential at the points which are close to the bus differs from zero a little. Denote the maximal potential at the points of the boundary of the rectangle $[-j, j+1] \times [-j, j]$ by u_j. From harmonicity of potential distribution it follows that $u_{j-1} \geq 2u_j - u_{j+1}$ $(1 \leq j \leq N-1)$. So if $u_{N-1} = \varepsilon$ (by assumption $u_n = 0$) then for all j such that $0 \leq j \leq N$ an inequality $u_j \geq (N-j)\varepsilon$ holds. In particular, $u_0 = V \geq N\varepsilon$. Using inequality (3), we obtain that $\varepsilon \leq N^{-1/2}$.

This time apply a bus to the boundary of the square $[-N, N] \times [-N, N]$. The obtained potential distribution is symmetric. Due to the maximum principle it differs from the initial one at most by $N^{-1/2}$. Therefore, before the moving of the right side of the bus four currents outcoming from the point $(0,0)$ differed from $\frac{1}{4}$ at most by $N^{-\frac{1}{2}}$. Similarly, if the source is connected to a bus and the point $(1,0)$ then ingoing currents to the point $(1,0)$ differ from $\frac{1}{4}$ at most by $N^{-\frac{1}{2}}$.

Unite these two situations. So by superposition principle we obtain that in the network obtained from the rectangle by short-circuiting its boundary and connecting the unit current source to the points $(0,0)$ and $(1,0)$ the current flowing through the edge $(0,0)-(1,0)$ equals to $1/2+O(N^{-1/2})$. Hence the potential difference and, therefore, resistance equal to $1/2 + O(N^{-1/2})$.

Consider the initial network. To finish the proof apply the axiom 3. Assume that potentials at the points $(0,0)$ and $(1,0)$ equal to $\frac{1}{4}$ and $-\frac{1}{4}$ respectively. Denote the current flowing through the source by I. Choose the rectangle $[-N, N+1] \times [-N, N]$ such that a potential on its boundary is smaller than some $\varepsilon > 0$. By maximum principle if one substitute all potentials at the boundary of the rectangle by zero (fixing the current through the battery) then potentials at all points will change at most by ε. In particular the potential difference between $(0,0)$ and $(1,0)$ will equal to $1/2 + O(\varepsilon)$. But resistance between them equals to $1/2 + O(N^{-1/2})$. Therefore $I = 1 + O(\varepsilon) + O(N^{-1/2})$. As ε tends to zero, N grows to infinity and we obtain that $I = 1$. Therefore the resistance of the lattice between adjacent vertices equals to $\frac{1}{2}$.

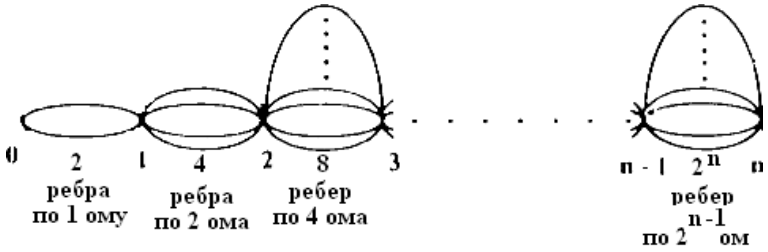


Figure 12: Calculation of the resistance of a tree; see solution of Problem 4.2.

3.5 Let B_1 and B_2 be the vertices of the graph which are opposite to A_1 and A_2 respectively. By Problem 3.4(A) the current flowing through the edge A_1A_2 equals to $I = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)\left(1 - \frac{1}{n}\right)$ if the source of given network is of unit current. Denote the current flowing through the edge B_2B_1 by x. Additionally connect the unit current source to vertices B_1 and B_2 such that the current flows in the vertex B_2. Then the current flowing through edges A_1A_2 and B_2B_1 equals to $I + x$. But if one connect two unit current sources to A_1, B_1 and B_2, A_2 (current flows in A_1 and in B_2) then the current distribution will be the same. By Problem 3.4 the first source gives a current $\frac{1}{k_1}$ through A_1A_2 and the second — $\frac{1}{k_2}$. Therefore

$$I + x = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)\left(1 - \frac{1}{n}\right) + x = \frac{1}{k_1} + \frac{1}{k_2},$$
$$x = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)\frac{1}{n}, \qquad \frac{x}{I} = \frac{1}{n-1}.$$

So for icosahedron the answer is $\frac{I}{11}$, for a dodecahedron — $\frac{I}{19}$, for a rhombdodecahedron — $\frac{I}{13}$, for a cube — $\frac{I}{7}$.

3.6 *Hint.* Sums D_n and F_n fit the same recurrence relation. For instance, $D_n = 1 + \frac{n+1}{2n}D_{n-1}$. Also $D_0 = F_0 = 1$. Therefore $D_n = F_n$. So if $n \geq 1$ then

$$\sum_{k=1}^{n} \frac{2^k}{k} = \frac{2^n}{n}F_{n-1} = \frac{2^n}{n}D_{n-1} = \frac{2^n}{n}\sum_{k=0}^{n-1}\frac{1}{C_{n-1}^k}.$$

To finish the proof bound the power of the number 2 in the factorization of the common denominator of the fractions from the obtained sum using Legendre formula for the power of the prime number in the factorization of a factorial.

The resistance R_n between two opposite vertices of an n-dimensional cube (with unit resistance on each its edge) is related with the sums D_n and F_n by the relations

$$D_n = F_n = (n+1)R_{n+1}$$

(see details in the paper [5]).

4.1 *Hint.* Prove by induction that the resistance of a binary tree of depth n constructed of unit resistors equals to $1 - \frac{1}{2^n}$.

4.2 *Hint.* Voltages at points situated at the same distance from the tree root are equal in virtue of symmetry. Shortening such points in a binary tree we receive a series from Figure 12. It's resistance equals to $\frac{1}{2} \cdot n = \frac{n}{2}$. Similarly for a trinary tree we receive $R = \frac{1}{3} + \frac{2}{9} + \cdots + \frac{2^{n-1}}{3^n}$. Hence $R = 1 - \frac{2^n}{3^n}$.

4.3 *Hint.* It is not difficult to cut a tree of depth 3. Let us show that it is impossible to cut a tree of depth 2010. Suppose we cut it, thus all it's vertices are situated at the distance not more than 2010 from the root; hence the tree is contained in a cube with the side $2 \cdot 2010 + 1$. So it has no more than $4021^3 \leq 2^{36}$ vertices. On the other hand the number of it's vertices equals to $2^{2011} - 1$. The contradiction completes the proof. The problem of cutting of a modified tree couldn't be solved easily.

4.4 *Hint.* A binary tree could not be cut; the arguments are similar to the solution 4.3, taking in account that more than two vertices could not be glued. A modified binary tree could be cut from the plane(see

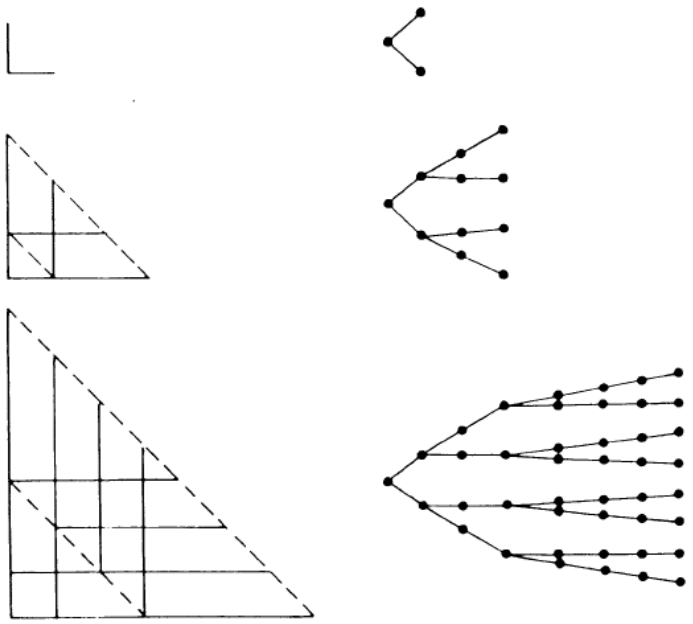


Figure 13: Cutting of a binary tree with intersections from the plain; see solution 4.4.

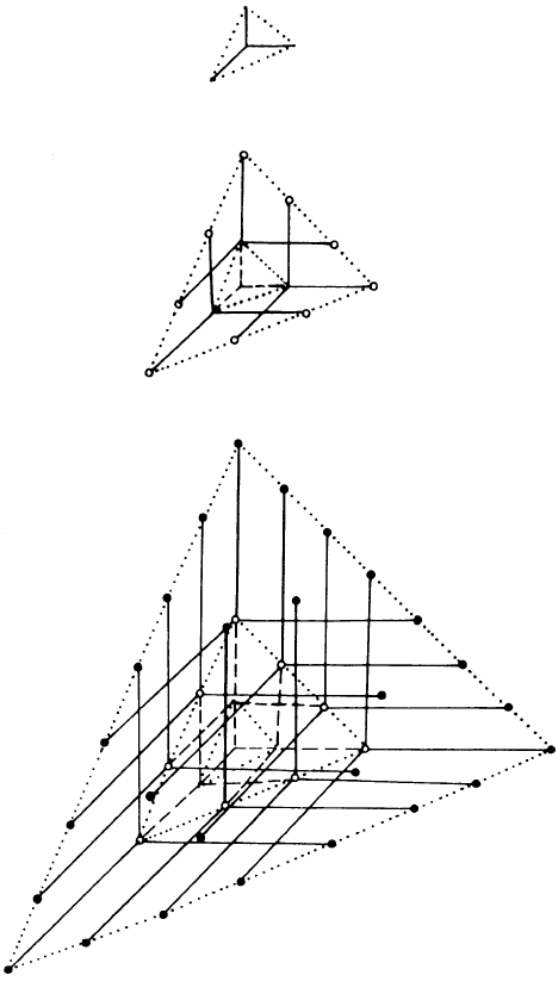


Figure 14: Cutting of a trinary tree with intersections from the space; see solution 4.4.

Figure 13), and similarly a trinary tree could be cut from the space (see Figure 14). Proofs could be done using induction by the tree depth.

4.5 *Hint.* For any $n = 2^i - 1$ let us consider the set of vertices (x, y, z), where $|x| + |y| + |z| \le n$. Let R_i be the resistance between the origin and the border of the figure. As we already know from Problem 4.4, it is possible to cut from such a part of the lattice a modified trinary tree of depth i with edges intersections at the same distance from the root. As we know from Problem 4.2, the resistances of modified trinary trees not more than 1. So the resistances of modified trinary trees also not more than 1. From the monotonicity law we receive that $R_i \le 1$. Hence switching on a battery of 1 Volt the current will be not more than 1. Hence the voltages at the vertices joint to the origin will be not more than $1 - \frac{1}{6} = \frac{5}{6}$. They equal to the probability or returning to the origin before reaching of the border. Taking the limit we obtain the statement we need.

5.1 Consider the points $A(2,3)$, $B(3,3)$, $C(1,2)$, $D(2,2)$, $E(3,2)$, $F(4,2)$, $G(2,1)$, $H(3,1)$. First let the power supply deliver a unit current to the origin, the second clip being connected to the perimeter of the square $[-R, R]^2$ (we set the zero resistance to this perimeter). From the symmetry reasons we get the equalities $i(CD) = i(GD) = i(DA) = i(DE) = I$ for some number I.

Now, consider the second situation, when the same current is delivered to the perimeter of the same square (with zero resistance as well), the second clip being connected to the point $(1, 0)$. Then we get $i(DE) \approx i(HE) \approx i(EB) \approx i(EF) \approx -I$, where the equalties are accurate within some small ε which tends to zero as $R \to \infty$ (similarly to Problem 3.4, it follows from axiom 3). Combining both situations, we get the following. Suppose that the power supply is connected to $(0, 0)$ and $(1, 0)$, while the perimeter of square $[-R, R]^2$ is replaced by a zero resistance loop; then the current flowing through edge DE is less than ε. Taking the limit as $R \to \infty$ (and applying axiom 3 again), we obtain the desired statement.

Note that we used without proof a difficult fact on the existence and uniqueness.

5.2 Consider an arbitrary tree connecting the given point with the perimeter of the square.

5.3 Apply the maximum principle.

5.4 Write the Laplace operator in the form

$$\Delta f(x,y) = f(x-1,y) + f(x+1,y) - 2f(x,y) + f(x,y-1) + f(x,y+1) - 2f(x,y).$$

Then for the function $f(x,y) = \ln r(x,y)$ we get

$$f(x-1,y) + f(x+1,y) - 2f(x,y) = \frac{1}{2}\ln\frac{((x+1)^2+y^2)((x-1)^2+y^2)}{(x^2+y^2)^2} =$$
$$= \frac{1}{2}\ln\left(1+\frac{2x+1}{r^2}\right)\left(1+\frac{-2x+1}{r^2}\right) = \frac{1}{2}\ln\left(\left(1+\frac{1}{r^2}\right)^2 - \frac{4x^2}{r^4}\right).$$

Similarly,

$$f(x,y-1) + f(x,y+1) - 2f(x,y) = \frac{1}{2}\ln\left(\left(1+\frac{1}{r^2}\right)^2 - \frac{4y^2}{r^4}\right).$$

Hence

$$\Delta f(x,y) = \frac{1}{2}\ln\left(\left(1+\frac{1}{r^2}\right)^2 - \frac{4x^2}{r^4}\right)\left(\left(1+\frac{1}{r^2}\right)^2 - \frac{4y^2}{r^4}\right) =$$
$$= \frac{1}{2}\ln\left(1 - \frac{1}{r^4} + \frac{16x^2y^2}{r^8}\right) = \frac{1}{2}\ln\left(1 + O\left(\frac{1}{r^4}\right)\right) = O\left(\frac{1}{r^4}\right).$$

5.5 For a point (x, y) nearby the boundary of the ring, it makes sense to change the definition of the Laplace operator, in order to agree with the Kirchhoff rules. For instance, if the points $(x - a, y)$ and $(x, y - b)$ for some $a, b \in [0, 1)$ lie on the boundary, then we set

$$\Delta f(x,y) = \frac{f(x-a,y) - f(x,y)}{a} + \frac{f(x,y-b) - f(x,y)}{b} + f(x+1,y) + f(x,y+1) - 2f(x,y).$$

Thus, considering the function the function $f(x,y)=\ln r(x,y)$ in such a point we have

$$\frac{f(x-a,y)-f(x,y)}{a}+f(x+1,y)-f(x,y)=\frac{1}{2a}\ln\left(1+\frac{-2ax+a^2}{r^2}\right)+\frac{1}{2}\ln\left(1+\frac{2x+1}{r^2}\right)=O\left(\frac{1}{r^2}\right).$$

Analogously,

$$\frac{f(x,y-b)-f(x,y)}{b}+f(x,y+1)-f(x,y)=O\left(\frac{1}{r^2}\right).$$

So, we get $\Delta f(x,y)=O\left(r^{-2}\right)$; moreover, this estimate remains valid if only one of the points neighboring to (x,y) lies outside the ring.

Consider now the function $f(x,y)=U_n(x,y)-\ln r(x,y)$. It vanishes on the boundary of the ring, while in each interior point it satisfies the equation $\Delta f(x,y)=\varphi(x,y)$, where $\varphi(x,y)=O(n^{-2})$ at the points nearby the boundary, and $\varphi(x,y)=O(n^{-4})$ at all other points of the ring.

Interpret the values $\varphi(x,y)$ as the currents delivered to the corresponding points of the ring. The potentials at the points nearby the boundary can be estimated by the maximum principle (cf. Problem 3.1). Actually, the potential on the boundary is zero, the current between our point (x,y) and the closest point of the boundary is $O(n^{-2})$, therefore the potential at (x,y) is $O(n^{-2})$ as well. By the maximum principle, all such currents induce the potentials not exceeding $O(n^{-2})$ in all points of our region.

Now, we are left to estimate the potentials generated by the currents at other points of the ring (that are those far from the boundary). The number of those points is $O(n^2)$, and the current in each of them induces the potentials not exceeding $O(n^{-7/2})$ (according to Problem 5.3). Thus, the total potential generated by our points is $O(n^{-3/2})$. Hence, we finally get $f(x,y)=O(n^{-3/2})$.

5.6 We involve the relation

$$\arctan x-\arctan y=\arctan\frac{x-y}{1+xy},$$

which holds when $|xy|<1$. So, we get

$$\arctan\frac{y+1}{R}-\arctan\frac{y}{R}=\arctan\frac{1/R}{1+y(y+1)/R^2}=$$
$$=\arctan\left(\frac{R}{y^2+R^2}+O\left(\frac{1}{R^2}\right)\right)=\frac{R}{R^2+y^2}+O\left(\frac{1}{R^2}\right).$$

5.7 Sum up the formula from problem 5.6.

5.8 To find the resistance of the ring, we find the current flowing through it when the inner and the outer loops of the ring are under the voltages $\ln nr_1$ and $\ln nr_2$ respectively. Let us find the current through the perimeter of a square $[-R-1/2,R+1/2]^2$, where $R=[r_1n]+1$. We will calcuat it approximately, changing the potentials at all the points by the corresponding values of the function $\ln r(x,y)$. Thus, we sum up $O(n)$ of currents, each with the error of $O(n^{-3/2})$. Hence the total error is $O(n^{-1/2})$.

By the symmetry of the square, the current flowing trough its perimeter can be found as

$$I=8\sum_{y=0}^{R}\left(\ln r(R+1,y)-\ln r(R,y)\right)+O\left(\frac{1}{n^{1/2}}\right).$$

Since

$$\ln r(R+1,y)-\ln r(R,y)=\frac{R}{R^2+y^2}+O\left(\frac{1}{R^2}\right),$$

we can apply the formula from problem 5.7 obtaining the desired relation $I=2\pi+O(n^{-1/2})$.

5.9 Similarly to the previous problem, to calculate the resistance we will find the current flowing through some closed broken line. Again, for the approximate calculation we replace the potentials by the values of the function $\ln r(x,y)$. Now let us replace our broken line by the square circumscribed around it; there will be $O(n^2)$ new current sources inside this contour, in each of them there appears an (incoming or outcoming) current of order $O(n^{-4})$. Therefore, the desired value of the total current differs from the

current $I = 2\pi + O(n^{-1/2})$ through the perimeter of the square (which was found above) by at most $O(n^{-2})$.

5.10 Let us prove that in Problem 5.2 the resistance between any vertex of the square and its boundary equals to $O(\ln n)$. This allows us to multiply residual terms by $\frac{\ln n}{\sqrt{n}}$ in solutions of all next problems.

Instead of the square consider a triangle cut from the square lattice by lines $x = 0$, $y = 0$, $x + y = n$, and bound its resistance between its origin and its hypotenuse. Assume that a potential at integer points of the segment $x + y = k$, $x, y \geq 0$ equals to $V_k = \sum_{j=2}^{k+1} \frac{1}{j}$ $(0 \leq k \leq n)$. In particular the potential at the origin equals zero. Also assume that the current flowing into each vertex at the line $x + y = k$ equals to $\frac{1}{k}$, i.e. through each level flows unit current. Increase resistances inside the triangle to fit Ohm law. (Resistances connecting points of the form $(0, k), (k, 0)$ with points $(0, k + 1)$, $(k + 1, 0)$ stay unit.) To fit Kirchhoff law currents flowing from the point $(j, k - j)$ should equal to $\frac{k-j}{k(k+1)}$ and $\frac{j+1}{k(k+1)}$ and flow to the points $(j, k - j + 1)$ and $(j + 1, k - j)$ respectively. Since potential difference equals to $\frac{1}{n+1}$ then unit resistances must be substituted by $\frac{k}{k-j}$ and $\frac{k}{j+1}$ respectively. The resistance of the obtained network equals to $V_n \leq \ln n$. So the resistance of the initial network is smaller than $\ln n$ also.

8. Acknowledgements

Most of the problems from sections 1, 2 and 4 are taken from the paper of P. Doyle and J. Snell [3]. The authors are grateful to I. Bogdanov, V. Bugaenko and M. Prasolov for help in translation.

[1] I. Benjamini and O. Schramm, Random walks and harmonic functions on infinite planar graphs using square tilings, Ann. Prob. **24:3** (1996), 1219–1238.

[2] J. Cannon, W. Floyd, W. Parry, Squaring rectangles: the finite Riemann mapping theorem, Contemp. Math. **169** (1994), 133–211.

[3] P. G. Doyle and J. L. Snell, Random walks and electric networks, Mathematical Association of America, 1984, http://arxiv.org/abs/math.PR/0001057.

[4] G. A. Galperin, My friend Andrey Khodulev, Mat. Prosv. 3rd series, **4** (2000), 8–32, http://www.mccme.ru/free-books/matpros/i5008032.pdf.zip.

[5] F. Nedemeyer, A. Smorodinskiy, *Resistance of edges of higher dimensional cube*, Kvant **6** (1986), in Russian.

[6] M. Prasolov and M. Skopenkov, *Tiling by rerctangles and alternating current*, submitted (2010). http://arxiv.org/abs/1002.1356.

[7] F. Spitzer, Principles of random walks, Springer–Verlag, 1976.

Tiling by rectangles and alternating current

M. Prasolov[a], M. Skopenkov[b,c,*]

[a] *Moscow State University, Faculty of Mechanics and mathematics, Leninskie Gory, 1, GSP-1, Moscow, 119991, Russian Federation*
[b] *Institute for information transmission problems of the Russian Academy of Sciences, Bolshoy Karetny per. 19, bld. 1, Moscow, 127994, Russian Federation*
[c] *King Abdullah University of Science and Technology, P.O. Box 2187, 4700 KAUST, 23955-6900 Thuwal, Kingdom of Saudi Arabia*

Abstract

This paper is on tilings of polygons by rectangles. A celebrated physical interpretation of such tilings due to R.L. Brooks, C.A.B. Smith, A.H. Stone and W.T. Tutte uses direct-current circuits. The new approach of the paper is an application of alternating-current circuits. The following results are obtained:

- a necessary condition for a rectangle to be tilable by rectangles of given shapes;
- a criterion for a rectangle to be tilable by rectangles similar to it but not all homothetic to it;
- a criterion for a generic polygon to be tilable by squares.

These results generalize the ones of C. Freiling, R. Kenyon, M. Laczkovich, D. Rinne and G. Szekeres.

Keywords:
Tiling, rectangle, orthogonal polygon, alternating current
2010 MSC: 52C20, 94C05, 31C20, 30C15, 60J10

1. Introduction

A rectangle $a \times b$, where a and b are integers, can be tiled by $a \cdot b$ squares. Thus a rectangle with rational side ratio can be tiled by squares. In 1903 M. Dehn proved the converse assertion:

Theorem 1.1. [10] *A rectangle can be tiled by squares (not necessarily equal) if and only if the ratio of two orthogonal sides of the rectangle is rational.*

Although this assertion is expectable, the proof is complicated. After original proof, many improvements have been made [2, 3, 18, 25, 32].

The most interesting for us is the approach of R.L. Brooks, C.A.B. Smith, A.H. Stone and W.T. Tutte [3]. To a tiling of a rectangle they assign a direct-current circuit, and then deduce Theorem 1.1 from certain properties of the circuit. They also apply the technique to find a tiling of a square by squares of distinct sizes, see the figure in the front cover of the journal [13].

*Corresponding author
Email address: skopenkov@rambler.ru, Tel.: +966 531 557 960, Fax: +966 2 802 0064 (M. Skopenkov)

We study finite tilings by arbitrary nondegenerate rectangles. The sides of rectangles are assumed to be parallel to coordinate axes, i.e., either vertical or horizontal. By *the ratio* of a rectangle we mean the length of the horizontal side divided by the length of the vertical one. We study the following problem posed in [16, p. 218] and [19, p. 3]:

Problem 1.2. Which rectangles can be tiled by rectangles of given ratios $c_1, \dots, c_n$?

A related problem of *signed* tilings is solved in [19].

For $n = 1$ and $c_1 = 1$ the question of Problem 1.2 is answered by Theorem 1.1. A necessary condition for arbitrary n was actually proved by M. Dehn: if a rectangle of ratio c can be tiled by rectangles of ratios $c_1, \dots, c_n$ then c is (the value of) a rational function in $c_1, \dots, c_n$ with rational coefficients.

This function depends only on "combinatorial structure" of the tiling. For instance, if a rectangle of ratio c is dissected into 2 rectangles of ratios c_1 and c_2 by a vertical (respectively, horizontal) cut then $c(c_1, c_2) = c_1 + c_2$ (respectively, $c(c_1, c_2) = \frac{c_1 c_2}{c_1 + c_2}$). The problem reduces to description of possible functions $c(c_1, \dots, c_n)$. By the mentioned physical interpretation this is equivalent to a natural problem: *describe possible formulas $c(c_1, \dots, c_n)$ expressing the conductance of a planar direct-current circuit through the conductances $c_1, \dots, c_n$ of individual resistors.*

The main idea of the paper is to apply *alternating*-current circuits (equivalently, circuits with complex-valued conductances) to the above problems. Our first result is

Theorem 1.3. *Suppose that a rectangle of ratio c can be tiled by rectangles of ratios $c_1, \dots, c_n$. Then $c = C(c_1, \dots, c_n)$ for some rational function $C(z_1, \dots, z_n)$ such that*

(1) *$C(z_1, \dots, z_n)$ has rational coefficients, i.e., $C(z_1, \dots, z_n) \in \mathbb{Q}(z_1, \dots, z_n)$;*
(2) *$C(z_1, \dots, z_n)$ is degree 1 homogeneous, i.e., $C(tz_1, \dots, tz_n) = tC(z_1, \dots, z_n)$;*
(3) *if $\operatorname{Re} z_1, \dots, \operatorname{Re} z_n > 0$ then $\operatorname{Re} C(z_1, \dots, z_n) > 0$.*

Problem 1.4. Is the converse theorem true for $n \geq 3$?

Parts (1) and (2) of Theorem 1.3 were actually proved by Dehn, see also [17, Lemma 4]. Case $n = 1$ (respectively, $n = 2$) of both Theorem 1.3 and its converse is equivalent to Theorem 1.1 (respectively, to [16, Theorem 5], see also Theorem 3.1 below). For $n \geq 3$ the converse theorem cannot be proved by our method, see Example 3.2.

Theorem 1.3 has a clear physical meaning, see §2.4. But this theorem (even together with its converse) is not *algorithmic*, i.e., it does not give an algorithm to decide if there exists a required tiling. Thus it is interesting to get less general but algorithmic results.

A result of this kind was obtained independently by C. Freiling, D. Rinne in 1994 and M. Laczkovich, G. Szekeres in 1995. It uses the following notion. An *algebraic conjugate* of an algebraic number c is a complex root of the minimal integral polynomial of c.

Theorem 1.5. [17, 22] *For $c > 0$ the following 3 conditions are equivalent:*

(1) *a square can be tiled by rectangles of ratios c and $1/c$;*
(2) *the number c is algebraic and all its algebraic conjugates have positive real parts;*
(3) *for certain positive rational numbers $d_1, \dots, d_m$ we have*

$$d_1 c + \cfrac{1}{d_2 c + \dots + \cfrac{1}{d_m c}} = 1.$$

We present a new short self-contained proof of this result. This new proof is an example of a natural application of alternating-current circuits. We also get a new algorithmic result:

Theorem 1.6. *For a number $c > 0$ the following 3 conditions are equivalent:*

(1) *a rectangle of ratio c can be tiled by rectangles of ratios c and $1/c$ (in such a way that there is at least one rectangle of ratio $1/c$ in the tiling);*
(2) *the number c^2 is algebraic and all its algebraic conjugates distinct from c^2 are negative real numbers.*
(3) *for certain positive rational numbers $d_1, \dots, d_m$ we have*

$$\cfrac{1}{d_1 c + \cfrac{1}{d_2 c + \cdots + \cfrac{1}{d_m c}}} = c.$$

More algorithmic results can be found in [16, p. 224]. For similar results on tiling by triangles see [29]. For higher dimensional generalizations see [25].

We also consider tilings of arbitrary (not necessarily convex) polygons by rectangles. This generalization reveals new connections between tilings and electrical circuits.

We apply direct-current circuits with several terminals to get a criterion for a generic polygon to be tilable by squares (Theorem 4.2 below, again not algorithmic). This result generalizes Theorem 1.1 and [21, Theorems 9 and 12]. An easier related problem of *signed* tiling by squares is solved in [15, 20].

We apply alternating-current circuits with several terminals to get a short proof of a generalization of Theorem 1.5 to polygons with rational vertices [28] (Theorem 4.3 below). We also give basic results on *electrical impedance tomography* for alternating-current circuits, cf. [9, 6, 7, 23].

There is a close relationship among electrical circuits, discrete harmonic functions and random walks on graphs [11, 24, 1]. Our results have equivalent statements in the language of each of the theories, e.g., see Corollary 4.9 below.

The paper splits naturally into two formally independent parts: §§1–3 and §§4–6.

The first part contains the proof of Theorems 1.3, 1.5 and 1.6. In §2 the basics of electrical circuits and their connection with tilings are recalled. In §3 the results of §1 are proved.

The second part concerns some variations. In §4 the results on tilings of polygons, electrical impedance tomography and random walks are stated. In §5 the results of §2 are generalized to electrical circuits with several terminals. In §6 the results of §4 are proved.

2. Main ideas

2.1. Electrical circuits

Our approach is based on electrical circuits theory [26]. However, the reader is not assumed to be familiar with physics. In this section we recall all the required physical concepts (although the presentation is formal and physical meaning is explained very briefly). This section does not contain new results. For short proofs see §5.

An *electrical network* is a connected graph with a nonnegative real number (*conductance*) assigned to each edge, and two marked (*boundary*) vertices.

For simplicity assume that the graph does not have neither multiple edges nor loops. Although all the concepts below can be adopted easily for the graphs with multiple edges. We say that electrical network is *planar* if the graph is drawn in the unit disc in such a way that the boundary vertices are in the boundary of the disc and the edges do not intersect each other.

Fix an enumeration of the vertices $1, 2, \ldots, n$ of the graph such that 1 and 2 are the boundary ones. It is convenient to denote the number of boundary vertices by $b = 2$. Let m the number of edges. Denote by c_{kl} the conductance of the edge between the vertices k and l. Set $c_{kl} = 0$ if there is no edge between k and l in the graph.

An *electrical circuit* is an electrical network along with two real numbers U_1 and U_2 (*incoming voltages*) assigned to the boundary vertices.

Each electrical circuit gives rise to certain numbers U_k, where $1 \le k \le n$ (*voltages* at the vertices), and I_{kl}, where $1 \le k, l \le n$ (*currents* through the edges). These numbers are defined by the following axioms:

(C) *The Ohm law.* For each pair of vertices k, l we have $I_{kl} = c_{kl}(U_k - U_l)$.
(I) *The Kirchhoff current law.* For each vertex $k > b$ we have $\sum_{l=1}^{n} I_{kl} = 0$.

Informal meaning of law (I) is that electrical charge is not aggregated at the nonboundary vertices. In other words, these laws assert that U_k is a *discrete harmonic function*. The numbers U_k and I_{kl} are well-defined by these axioms by the following classical result.

Theorem 2.1. [31] *For any electrical circuit the system of linear equations* (C),(I) *in variables U_k, $b < k \le n$, and I_{kl}, $1 \le k, l \le n$, has a unique solution.*

Denote by $I_1 = \sum_{k=1}^{n} I_{1k}$ the current flowing inside the circuit through vertex 1. *The conductance* of an electrical circuit with $U_1 \ne U_2$ is the number $C = I_1/(U_1 - U_2)$. Clearly, the conductance does not depend on U_1 and U_2. Thus *the conductance* of an electrical network is well-defined. Basic examples of networks and their conductances are shown in figure 1.

$C(a,b) = \frac{ab}{a+b}$ $\qquad$ $C(a,b) = a + b$

Figure 1: Series and parallel electrical networks

2.2. Tilings and networks

There is a close relationship between electrical networks and tilings. We say that an edge kl of a circuit is *essential*, if $I_{kl} \ne 0$. Clearly, the property of an edge being essential does not depend on U_1 and U_2 if $U_1 \ne U_2$.

Lemma 2.2. [3, 4, Theorem 1.4.1] *The following two conditions are equivalent:*

(1) *a rectangle of ratio c can be tiled by m rectangles of ratios $c_1, \ldots, c_m$;*
(2) *there is a planar electrical network having conductance c and consisting of m essential edges of conductances $c_1, \ldots, c_m$.*

Let us sketch the proof of assertion (1) $\Longrightarrow$ (2). Given a tiling as in (1) construct an electrical network as follows (see figure 2). Take a point in each maximal horizontal cut of the tiling and in each horizontal side of the tiled rectangle. These points are vertices of the network. For each rectangle in the tiling draw an edge between the vertices in the cuts containing the horizontal sides of the rectangle. Set the conductance of the edge to be the ratio of the rectangle. The obtained network has conductance c, see §5.2 for the proof.

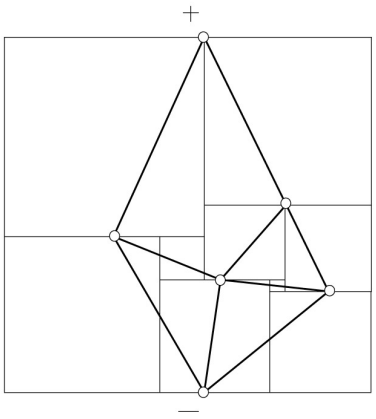


Figure 2: Correspondence between tilings and electrical networks

2.3. Formulas for conductance

Let us summarize some useful properties of formulas for conductance.

Lemma 2.3. *Suppose that an electrical network consists of m edges of conductances $c_1, \dots, c_m$. Then the conductance of the network $C(c_1, \dots, c_m)$ has the following properties:*

(1) [3] $C(c_1, \dots, c_m) \in \mathbb{Q}(c_1, \dots, c_m)$;
(2) [3] $C(c_1, \dots, c_m)$ *is degree* 1 *homogeneous;*
(3) [3] $\frac{\partial}{\partial c_j} C(c_1, \dots, c_m) = (\frac{U_k - U_l}{U_1 - U_2})^2$, *where k and l are the endpoints of the edge j;*
(4) [27] *if* $c_1, \dots, c_m > 0$ *then* $\frac{\partial}{\partial c_j} C(c_1, \dots, c_m) \geq 0$; *if the edge j is essential then the latter inequality is strict;*
(5) [5] *if* $\operatorname{Re} c_1, \dots, \operatorname{Re} c_m > 0$ *then* $\operatorname{Re} C(c_1, \dots, c_m) > 0$.

Remark 2.4. (A. Akopyan, private communication) Property (4) follows from (1), (2) and (5). Property (5) does not follow from (1), (2) and (4), e.g., the function $C(c_1, c_2) = (c_1 + c_2)\frac{c_1^2 + c_2^2}{c_1^2 + 2c_2^2}$ satisfies (1), (2), (4) but not (5).

Property (5) concerns the extension of the function $C(c_1, \dots, c_m)$ to the complex plane. This fundamental property does not seem to be payed attention for *direct*-current circuits. Certainly it is well-known for *alternating*-current circuits. Short proof of the lemma is given in §5.1.

2.4. Alternating-current circuits

Let us explain informal physical meaning of fundamental Lemma 2.3(5) and condition (3) of Theorem 1.3. This is not used elsewhere in the paper and the reader may easily skip this subsection.

Informally, *an alternating-current circuit* is a collection of conductors, condensers, inductors and a single alternating-voltage source connected with each other.

Formally, an *alternating-current circuit* is a graph with the following structure:

- two marked (*boundary*) vertices;
- two functions (*voltages*) $\tilde{U}_1(t) = U \cos \omega t$ and $\tilde{U}_2(t) = 0$ assigned to them;
- division the edges into three types (*conductors, condensers and inductors*);
- a positive number $\tilde{c}_{kl}$ assigned to each edge (called *conductance, capacitance or inductance*, depending on the type of the edge).

The *voltages* $\tilde U_k(t)$ and the *currents* $\tilde I_{kl}(t)$ are defined by the following axioms:

($\tilde{\text{C}}$) *The generalized Ohm law.* For each edge kl we have

$$\tilde I_{kl}(t) = \begin{cases} \tilde c_{kl}(\tilde U_k(t) - \tilde U_l(t)) & \text{if } kl \text{ is a conductor;} \\ \tilde c_{kl}\frac{d}{dt}(\tilde U_k(t) - \tilde U_l(t)) & \text{if } kl \text{ is a condenser;} \\ \tilde c_{kl}\int_{\pi/2\omega}^t (\tilde U_k(t) - \tilde U_l(t))dt & \text{if } kl \text{ is an inductor.} \end{cases}$$

($\tilde{\text{I}}$) *The Kirchhoff current law.* For each vertex $k \neq 1, 2$ we have $\sum_{l=1}^n \tilde I_{kl}(t) = 0$.

The voltages and the currents can be found using the following well-known algorithm. Denote by $i = \sqrt{-1}$. Put $U_1 = U$, $U_2 = 0$ and

$$c_{kl} = \begin{cases} \tilde c_{kl}, & \text{if } kl \text{ is a conductor;} \\ i\omega\tilde c_{kl}, & \text{if } kl \text{ is a condenser;} \\ \frac{1}{i\omega}\tilde c_{kl}, & \text{if } kl \text{ is an inductor.} \end{cases}$$

Define the complex numbers U_k, $3 \le k \le n$, and I_{kl}, $1 \le k, l \le n$, by *direct*-current laws (C), (I). Then $\tilde U_k(t) = Re(U_k e^{i\omega t})$, $\tilde I_{kl}(t) = Re(I_{kl}e^{i\omega t})$. In this sense alternating-current circuits are "equivalent" to direct-current circuits with complex-valued conductances (also called *admittances*).

Notice that always $Re\, c_{kl} \ge 0$. Physically this means nonnegative *energy dissipation* at the edge kl (which is $Re\, c_{kl}|U_k - U_l|^2$). Thus a physical meaning of Lemma 2.3(5) is: "a network consisting of elements dissipating energy also dissipates energy".

2.5. Positive real functions

This subsection is used in the proof of only assertions (2) $\Longrightarrow$ (3) in Theorems 1.5 and 1.6.

Consider electrical circuits, in which all the edges have conductances z and $1/z$, $Re\, z > 0$. (They have a natural physical meaning: circuits consisting of condensers and inductors with incoming voltage of complex frequency z/i.) Let us describe possible conductances $C(z)$ of such electrical circuits. By Lemma 2.3(1), (2) and (5) the functions $C(z)$ are *positive real*, i.e., satisfy condition (1) of the following lemma. Denote by $Re\, \infty = 0$, $C(\infty) = \lim_{z\to 0} C(1/z)$ and $C'(\infty) = \lim_{z\to 0}(C(1/z))'$.

Lemma 2.5. [5, 14, 16, Lemma 4] *For an odd function $C(z) \in \mathbb{R}(z)$ the following 5 conditions are equivalent:*

(1) *if $Re\, z > 0$ then $Re\, C(z) > 0$;*
(2) *if $C(z) = 1$ then $Re\, z > 0$;*
(3) *if $C(z) = 0$ then $Re\, z = 0$ and $C'(z) > 0$ (here $z \in \mathbb{C}$ or $z = \infty$);*
(4) *either $C(z)$ or $1/C(z)$ equals*

$$d_1 z \prod_{k=1}^n \frac{z^2 + a_k^2}{z^2 + b_k^2},$$

for some integer number $n \ge 0$ and real numbers $d_1 > 0$, $a_1 > b_1 > a_2 > \cdots > b_n \ge 0$;
(5) *either $C(z)$ or $1/C(z)$ equals*

$$d_1 z + \cfrac{1}{d_2 z + \cdots + \cfrac{1}{d_m z}},$$

for some integer number $m \ge 1$ and real numbers $d_1, \dots, d_m > 0$.

Parts of the lemma are proved in [5, 14] and in [16] using the results of [30]. A short proof is given in §5.3.

3. Proof of main results

3.1. Proof of Theorem 1.3

Hereafter in an *electrical circuit* or *network* we allow the conductances to be arbitrary complex numbers with positive real part. This generalization of the above notion is motivated by §2.4 (and describes both direct- and alternating-current circuits). Theorem 1.3 is an easy consequence of the results of §2:

Proof of Theorem 1.3. Suppose that a rectangle of ratio c can be tiled by rectangles of ratios $c_1, \dots, c_n$. By Lemma 2.2 there is an electrical network of conductance c consisting of edges of conductances $c_1, \dots, c_n$. For each $k = 1, \dots, n$ replace each edge of conductance c_k in the network by an edge of complex conductance z_k, $Re\, z_k > 0$. Let $C(z_1, \dots, z_n)$ be the conductance of the obtained network. The function $C(z_1, \dots, z_n)$ has the properties (1)–(3) of Theorem 1.3 by Lemma 2.3(1),(2) and (5). □

3.2. Proof of Theorem 1.5

Proof of Theorem 1.5. (3) $\Longrightarrow$ (1) [16] Suppose that condition (3) of Theorem 1.5 holds and, say, m is odd. Take a unit square. Cut off a rectangle of ratio $d_1 c$ from the square by a vertical cut. The remaining part is a rectangle of ratio

$$1 - d_1 c = \cfrac{1}{d_2 c + \cdots + \cfrac{1}{d_m c}}.$$

Now cut off a rectangle of ratio $1/d_2 c$ from the remaining part by a horizontal cut. We get a rectangle of ratio

$$d_3 c + \cfrac{1}{d_4 c + \cdots + \cfrac{1}{d_m c}}.$$

Continue this process alternating vertical and horizontal cuts. Condition (3) guaranties that after step $(m-1)$ we get a rectangle of ratio $d_m c$. We obtain a tiling of the square by rectangles of ratios $d_1 c$, $1/d_2 c$, $d_3 c$, $1/d_4 c$, $\dots$, $d_m c$. Since all $d_k \in \mathbb{Q}$ one can chop the tiling into rectangles of ratios c and $1/c$.

(1) $\Longrightarrow$ (2). Suppose that a square is tiled by rectangles of ratios c and $1/c$. By Lemma 2.2 there exists an electrical network of conductance 1 with edge conductances c and $1/c$. Replace each edge of conductance c (respectively, $1/c$) in this network by an edge of conductance $z \in \mathbb{C}$ (respectively, $1/z$). Let $C(z)$ the conductance of the obtained network. Then $C(z) \in \mathbb{Q}(z)$ by Lemma 2.3(1).

Since $C(c) = 1$ it follows that c is algebraic ($C(z)$ is nonconstant because $C(-c) = -C(c) = -1$ by Lemma 2.3(2)). Let z be an algebraic conjugate of c. Then still $C(z) = 1$.

Let us prove that $Re\, z > 0$. Indeed, first assume that $Re\, z < 0$. Then $Re(-z) > 0$ and $Re(-1/z) > 0$. Thus by Lemma 2.3(5) we have $0 < Re\, C(-z) = -Re\, C(z) = -1$, a contradiction. Now assume that $Re\, z = 0$. Let $z_k \to z$, where each $Re\, z_k < 0$. Still $0 < Re\, C(-z_k) = -Re\, C(z_k) \to -1$, a contradiction. Thus $Re\, z > 0$.

(2) $\Longrightarrow$ (3) [16] Let $p(z)$ be a minimal polynomial of c. Put $C(z) = \frac{p(-z)-p(z)}{p(-z)+p(z)}$. Then $C(c) = 1$, $C(z) \in \mathbb{Q}(z)$, $C(z)$ is odd and all the roots of the equation $C(z) = 1$ have positive real part. By Lemma 2.5(2) $\Longrightarrow$ (5) the function $C(z)$ satisfies condition (5) of Lemma 2.5. Since $C(z) \in \mathbb{Q}(z)$ it follows by Euclidean algorithm that all $d_k \in \mathbb{Q}$. Substituting $z = c$ we get the required condition. □

3.3. Proof of Theorem 1.6

The proof follows the ideas of §3.2 and §5.3.

Proof of Theorem 1.6. (3) $\Longrightarrow$ (1) Analogously to the proof of Theorem 1.5(3) $\Longrightarrow$ (1).

(1) $\Longrightarrow$ (2). Suppose that a rectangle of ratio c is tiled by rectangles of ratios c and $1/c$. Rotating through $\pi/2$ and stretching the figure we get a square tiled by squares and rectangles of ratio c^2. By Lemma 2.2 there exists an electrical circuit of conductance 1 with edge conductances 1 and c^2, in which all the edges are essential. Since there is at least one rectangle of ratio $1/c$ in the initial tiling, it follows that the network contains at least one edge of conductance c^2. Replace each edge of conductance c^2 (respectively, 1) in the network by an edge of conductance $z \in \mathbb{C}$ (respectively, $w \in \mathbb{C}$). Let $C(z,w)$ the conductance of the obtained network. Denote by $C(z) = C(z,1)$.

Let us prove that c^2 is algebraic. Indeed, by Lemma 2.3(4) we have $C'(c^2) > 0$ because there is at least one essential edge of conductance c^2 in the network. Thus $C(z)$ is nonconstant. By Lemma 2.3(1) it follows that $C(z) \in \mathbb{Q}(z)$. Since $C(c^2) = 1$ it follows that c^2 is algebraic.

Let z be an algebraic conjugate of c^2 distinct from c^2 itself. Then $C(z,1) = C(c^2) = 1$.

Let us prove that z is a negative real number. First assume $Im\, z < 0$. Then $Re\, iz > 0$. By Lemma 2.3(2) it follows that $Re\, C(iz,i) = Re(iC(z,1)) = Re\, i = 0$. Since $C(iz,i)$ is a rational function it follows that any neighborhood of iz contains a point z' such that $Re\, C(z',i) < 0$. Taking sufficiently small neighborhood we get $Re\, z' > 0$ because $Re\, iz > 0$. By continuity a neighborhood of i contains a point w' such that $Re\, w' > 0$ and still $Re\, C(z',w') < 0$. The obtained inequalities contradict to Lemma 2.3(5). Case $Im\, z > 0$ is violated similarly. Assume now $z > c^2$. Then by Lemma 2.3(4) we have $1 = C(z) > C(c^2) = 1$, a contradiction. Case $0 \le z < c^2$ is violated similarly. Thus $z < 0$.

(2) $\Longrightarrow$ (3) Let $p(z)$ be a minimal polynomial of c^2. Since the roots of a minimal polynomial are all simple it follows that $p(z^2) = (z^2 - c^2)\prod_{k=1}^{n}(z^2 + b_k^2)$ for some $b_1 > \cdots > b_n > 0$. Take a polynomial $q(z)$ with rational coefficients such that $q(z) = z\prod_{k=1}^{n}(z^2 + a_k^2)$, where $a_1 > b_1 > a_2 \cdots > b_n > 0$. Consider the odd rational function $C(z) = q(z)/(zq(z) - p(z^2))$. We have $C(c) = 1/c$.

Let us check that the function $C(z)$ satisfies condition (3) of Lemma 2.5. The roots of $C(z)$ are the numbers $0, \pm ia_1, \ldots, \pm ia_n$. A direct evaluation shows that for each $l = 1, \ldots, n$

$$C'(\pm ia_l) = -\frac{q'(\pm ia_l)}{p(-a_l^2)} = \frac{2a_l^2}{(c^2 + a_l^2)(a_l^2 - b_l^2)} \prod_{k \ne l} \frac{a_k^2 - a_l^2}{b_k^2 - a_l^2} > 0$$

by the assumption $a_1 > b_1 > a_2 > \cdots > b_n > 0$. Analogously $C'(0) = -q'(0)/p(0) > 0$.

Then by Lemma 2.5(3) $\Longrightarrow$ (5) the function $C(z)$ satisfies condition (5) of Lemma 2.5. Since $C(z) \in \mathbb{Q}(z)$ it follows by Euclidean algorithm that all $d_k \in \mathbb{Q}$. Substituting $z = c$ we get the required condition. □

3.4. Remarks to main results

Let us define inductively a *series-parallel* electrical network. By definition, a network consisting of a single edge is series-parallel. If a and b are two series-parallel networks then both their series and parallel "unions" (see figure 1) are series-parallel.

Theorem 3.1. [5] *If a function $C(c_1, c_2)$ satisfies conditions* (1)–(3) *of Theorem 1.3 then $C(c_1, c_2)$ is the conductance of a series-parallel electrical network with edge conductances c_1 and c_2.*

Proof. By conditions (1)–(3) of Theorem 1.3 we have $C(c_1, c_2) = \sqrt{c_1 c_2}\, C(z, 1/z)$, where $z = \sqrt{c_1/c_2}$, and the function $C(z) = C(z, 1/z)$ satisfies condition (1) of Lemma 2.5. By Lemma 2.5(1) $\Longrightarrow$ (5) it satisfies condition (5). Therefore, say, for m even and $C(0) = 0$,

$$C(c_1, c_2) = d_1 c_1 + \cfrac{1}{\cfrac{d_2}{c_2} + \cfrac{1}{d_3 c_1 + \cdots + \cfrac{d_m}{c_2}}}.$$

All the numbers $d_k \in \mathbb{Q}$ by the Euclidean algorithm. Now the required series-parallel network is constructed analogously to the proof of Theorem 1.5(3) $\Longrightarrow$ (1). $\square$

Example 3.2. A generalization of Theorem 3.1 to the case of 3 variables c_1, c_2, c_3 is not true. E.g., consider the network with 4 vertices and edge conductances $c_{13} = c_1$, $c_{23} = c_2$, $c_{24} = c_1$, $c_{14} = c_2$, $c_{34} = c_3$. By Lemma 2.3(4) and a symmetry argument it follows that $\partial C(c_1, c_2, c_3)/\partial c_3 = 0$ if $c_1 = c_2$. So $C(c_1, c_2, c_3)$ cannot be the conductance of a series-parallel network, because all the edges of such networks are essential.

4. Variations

4.1. Tilings of polygons by rectangles

In this subsection we study the following problem.

Problem 4.1. Which polygons can be tiled by rectangles of given ratios $c_1, \ldots, c_n$?

Case $n = 1$, $c_1 = 1$ of the problem is a description of polygons which can be tiled by squares, a problem posed in [15]. In case of hexagons such a description was obtained by R. Kenyon [21]. We give such description for a wide class of polygons.

Hereafter P is an *orthogonal* polygon, i.e., a polygon with sides parallel to coordinate axes. Assume that P is *simple*, i.e., the boundary ∂P has one connected component. Enumerate the sides parallel to the x-axis counterclockwise in ∂P. Let I_u be the *signed length* of the side u, where the sign of I_u is "+" ("−") if the P locally lies below (above) the side u. Let U_u be the y-coordinate of the side u. Assume that P is *generic*, i.e., the numbers $U_1, \ldots, U_b$ are pairwise distinct.

We need the following notion [9]. A sequence of boundary vertices $(p_1, \ldots, p_k, q_1, \ldots, q_k)$ of a planar network is *circular*, if the sequence $(p_1, \ldots, p_k, q_k, \ldots, q_1)$ is in counterclockwise order in the boundary of the unit disc. Denote by Ω_b the set of real $b \times b$ matrices C_{uv} satisfying the following properties:

- C_{uv} is symmetric;
- the sum of the entries of C_{uv} in each row is zero;
- if $(p_1, \ldots, p_k, q_1, \ldots, q_k)$ is a circular sequence then $(-1)^k \det\{C_{p_i q_j}\}_{i,j=1}^{k} \geq 0$.

Theorem 4.2. *Let P be a generic orthogonal polygon with b horizontal sides having signed lengths $I_1, \ldots, I_b$ and y-coordinates $U_1, \ldots, U_b$. Then the following two conditions are equivalent:*

(1) *the polygon P can be tiled by squares;*
(2) *there is a matrix $C_{uv} \in \Omega_b$ with rational entries such that $I_v = \sum_{u=1}^{b} C_{uv} U_u$ for each $v = 1, \ldots, b$.*

Cases $b = 2$ and $b = 3$ of this theorem are equivalent to Theorem 1.1 and [21, Theorem 9], respectively. Theorem 4.2 is algorithmic in the particular case when $U_1, \dots, U_b$ are linearly independent over $\mathbb{Q}$. Proof of the theorem is constructive, i.e., gives an algorithm to construct the required tiling if the latter exists. Theorem 4.2 does not necessarily hold for *nongeneric* polygons, e.g., for an orthogonal polygon with

$$U_1 = U_3 = 0, \quad U_2 = 2, \quad U_4 = -4, \quad I_1 = \sqrt{2}, \quad I_2 = 2, \quad I_3 = 2 - \sqrt{2}, \quad I_4 = -4.$$

We also give a short proof of the following result:

Theorem 4.3. [28] *A generic orthogonal polygon with rational vertices can be tiled by rectangles of ratios c and $1/c$ if and only if a square can be tiled by rectangles of ratios c and $1/c$.*

4.2. Electrical impedance tomography

Our approach to Problem 4.1 follows the idea of [21, 7] and uses electrical networks with several terminals.

Hereafter we allow electrical circuits to have several boundary vertices $1, \dots, b$ with prescribed voltages $U_1, \dots, U_b$. If an electrical circuit is planar, we assume that the boundary vertices are enumerated counterclockwise along the boundary of the unit disc. We do *not* assume that an electrical circuit is connected but require that each connected component contains a boundary vertex. The voltages and currents in such circuits are defined by the Ohm and the Kirchhoff current laws (C) and (I) from §2.

Consider the linear map $\mathbb{C}^b \to \mathbb{C}^b$ which takes the vector of voltages $(U_1, \dots, U_b)$ to the vector of *incoming currents* $(I_1, \dots, I_b) = (\sum_{k=1}^n I_{1k}, \dots, \sum_{k=1}^n I_{bk})$ flowing inside the network through the vertices $1, \dots, b$, respectively. The matrix C_{uv} of this linear map is called the *response* of the network. This matrix is symmetric [9].

We reduce the results of §4.1 to the following problems even more interesting in themselves:

- *Direct problem.* Describe possible responses of electrical networks.
- *Inverse problem.* Describe possible networks having a given response.

These problems are solved for planar direct-current networks [9, 6, 7, 23]. Let us state certain deep results of Y. Colin de Verdière, E.B. Curtis and J.A. Morrow.

Theorem 4.4. [9, 8, 7, Theorem 5] *The set of all possible responses of planar electrical networks with b boundary vertices and positive edge conductances is the set Ω_b.*

An electrical network is *minimal* (or *critical*) if it has minimal number of edges among all planar electrical networks with positive edge conductances and with the same response. The minimality of a network depends only on its graph [7]. In [9, 8, §9] an algorithm for finding edge conductances in a minimal network with given response is presented. This algorithm implies the following result.

Theorem 4.5. [9, §6.4] *Conductances of the edges in a minimal electrical network are uniquely determined by the response of the network. Each edge conductance is a rational function with rational coefficients in the entries of the response.*

For alternating-current circuits the direct problem is probably open. Let us state some basic results. The rest of §4 is not used in the proof of the above results.

Theorem 4.6. *For $b = 2$ or $b = 3$ the following 2 conditions are equivalent:*

(1) *C_{uv} is the response of a connected electrical network with b boundary vertices and with edge conductances having positive real parts;*

(2) *C_{uv} is a complex $b \times b$ matrix has the following 4 properties:*

- *C_{uv} is symmetric;*
- *the sum of the entries of C_{uv} in each row is zero;*
- *$\operatorname{Re} C_{uv}$ is non-negatively definite;*
- *if $\sum_{1 \le u,v \le b} \operatorname{Re} C_{uv} U_u U_v = 0$ then $U_1 = \cdots = U_b$.*

Problem 4.7. Does this result remain true for arbitrary $b \ge 4$?

Unlike direct-current networks *nonboundary vertices in alternating-current networks can be detected by the response*. For instance, by Theorem 4.6 there are electrical networks with response $\left(\begin{smallmatrix} 2 & 1 & -3 \\ 1 & 2 & -3 \\ -3 & -3 & 6 \end{smallmatrix}\right)$; any such network necessarily has nonboundary vertices.

4.3. Random walks

A *random work* on an electrical network (or on a *weighted graph*) is the Markov chain with the transition matrix $P_{kl} = c_{kl} / \sum_{j=1}^{n} c_{jk}$. Such Markov chain is ergodic and reversible. Denote by $k_1 l_1, \dots, k_m l_m$ all the edges of the Markov chain. The following theorem allows to translate the results of §1–§2 to the language of random walks.

Theorem 4.8. [11, page 42] *Let $P(c_{k_1 l_1}, \dots, c_{k_m l_m})$ be the probability that a random walk starting at vertex 1 reaches vertex 2 before returning to 1. Let $C(c_{k_1 l_1}, \dots, c_{k_m l_m})$ be the conductance of the network (with boundary vertices 1 and 2). Then $P(c_{k_1 l_1}, \dots, c_{k_m l_m}) = C(c_{k_1 l_1}, \dots, c_{k_m l_m}) / (c_{12} + \cdots + c_{1n})$.*

For instance, a translation of Lemmas 2.3(1) and (5) is:

Corollary 4.9. *The probability $P(c_{k_1 l_1}, \dots, c_{k_m l_m})$ is a rational function in $c_{k_1 l_1}, \dots, c_{k_m l_m}$. If $\operatorname{Re} c_{k_1 l_1}, \dots, \operatorname{Re} c_{k_m l_m} > 0$ then $\operatorname{Re}\left((c_{12} + \cdots + c_{1n}) P(c_{k_1 l_1}, \dots, c_{k_m l_m})\right) > 0$.*

The latter result does not necessarily hold for *nonreversible* Markov chains, e.g., for a Markov chain with vertices 1, 2, 3, 4 and oriented edges 14, 42, 43.

Nonreversible planar Markov chains have a geometric interpretation as tilings of trapezoids by trapezoids [21]. Here a *trapezoid* is a 4-gon with two sides parallel to the x-axis. The *ratio* of the trapezoid is the length of the horizontal middle edge divided by the hight. Natural problems are: *generalize the results of the paper to tilings by trapezoids; infinite tilings; signed tilings.*

5. Generalization of main ideas

5.1. Electrical circuits

Our approach is based on a generalization of the results of §2 to electrical circuits with b terminals. Short proofs of the results of §2 are obtained in this section as particular case $b = 2$. Our proof of Lemma 5.2(3), generalizing Lemma 2.3(3), is probably new. All the proofs are based on the following fundamental *energy conservation law*.

Claim 5.1. *Let $E(U,I)$ be a bilinear function. Consider an electrical network with the vertices $1,\dots,n$ such that $1,\dots,b$ are the boundary ones. Suppose that the numbers U_k, $1\le k\le n$, and I_{kl}, $1\le k,l\le n$, satisfy laws* (C),(I) *from §2. Set $I_u=\sum_{k=1}^n I_{uk}$. Then*

$$\sum_{1\le k<l\le n} E(U_k-U_l,I_{kl})=\sum_{1\le u\le b} E(U_u,I_u).$$

We usually apply this claim for the *energy dissipation* function $E(U,I)=Re(U\bar I)$.

Proof of Claim 5.1. By law (C) we have $I_{lk}=-I_{kl}$. Hence by law (I) we have

$$\sum_{1\le k<l\le n} E(U_k-U_l,I_{kl})=\sum_{k=1}^n E(U_k,\sum_{l=1}^n I_{kl})=\sum_{1\le u\le b} E(U_u,I_u).$$

□

Let us prove Theorem 2.1 for electrical circuits with b boundary vertices and with complex edge conductances having positive real part.

Proof of Theorem 2.1. Uniqueness. Suppose there are two collections of currents $I_{kl}^{I,II}$ and voltages $U_k^{I,II}$ satisfying laws (C),(I). Then their difference $I_{kl}=I_{kl}^I-I_{kl}^{II}$, $U_k=U_k^I-U_k^{II}$ satisfies (C),(I) for zero incoming voltages $U_1=\dots=U_b=0$. Then by Claim 5.1 we have

$$\sum_{1\le k<l\le n} Re\,\bar c_{kl}|U_k-U_l|^2=\sum_{1\le k<l\le n} Re((U_k-U_l)\bar I_{kl})=\sum_{1\le u\le b} Re(U_u\bar I_u)=0.$$

Here for each k,l either $Re\,c_{kl}>0$ or $c_{kl}=0$. Thus each $Re\,\bar c_{kl}|U_k-U_l|^2=0$. Since all the connected components of the circuit contain boundary vertices it follows that all U_k are equal. Hence each $U_k=0$, $I_{kl}=0$ and thus each $I_{kl}^I=I_{kl}^{II}$, $U_k^I=U_k^{II}$.

Existence. The number of linear equations in the system (C),(I) equals the number of variables. By the previous paragraph the system has a unique solution for $U_1=\dots=U_b=0$. Thus by the finite-dimensional Fredholm alternative it has a solution for any $U_1,\dots,U_b$. □

The following result generalizes Lemma 2.3.

Lemma 5.2. *Suppose that an electrical network has b boundary vertices and m edges of conductances $c_1,\dots,c_m$. Then the response of the network $C_{uv}(c_1,\dots,c_m)$ has the following properties:*

(1) $C_{uv}(c_1,\dots,c_m)\in\mathbb{Q}(c_1,\dots,c_m)^{b\times b}$;
(2) $C_{uv}(c_1,\dots,c_m)$ *is degree* 1 *homogeneous;*
(3) $\frac{\partial}{\partial c_j}C_{uv}(c_1,\dots,c_m)=(V_{ku}-V_{lu})(V_{kv}-V_{lv})$, *where k and l are the endpoints of the edge j and V_{pq} is the matrix of the linear map $(U_1,\dots,U_b)\mapsto(U_1,\dots,U_n)$;*
(4) *if $c_1,\dots,c_m>0$ then $\frac{\partial}{\partial c_j}C_{uv}(c_1,\dots,c_m)$ is non-negatively definite;*
(5) *if $Re\,c_1,\dots,Re\,c_m>0$ then $Re\,C_{uv}(c_1,\dots,c_m)$ is non-negatively definite.*

Proof of Lemma 5.2. (1) By Theorem 2.1 and the Crammer rule the solution $\{I_{kl}(U_1,\dots,U_b)\}$ of the system of linear equations (C), (I) consists of linear functions in $U_1,\dots,U_b$ with coefficients being rational functions in $c_1,\dots,c_m$. So the entries of the matrix of the linear map $(U_1,\dots,U_b)\mapsto\sum_{k=1}^n I_{uk}(U_1,\dots,U_b)$ are rational functions in $c_1,\dots,c_m$.

(2) Consider the system of linear equations obtained from laws (C), (I) by substituting $tc_1, \dots, tc_m$ for $c_1, \dots, c_m$. It defines the same voltages as the initial one and the currents are scaled by t. So $C(tc_1, \dots, tc_m) = tC(c_1, \dots, c_m)$.

(3) Set $E(U, I) = \frac{\partial U}{\partial c_{kl}} I - U \frac{\partial I}{\partial c_{kl}}$. Then $E(U_k - U_l, I_{kl}) = (U_k - U_l)^2$ and $E(U_p - U_q, I_{pq}) = 0$ for $pq \neq kl$. Thus by Claim 5.1 we have

$$\sum_{1 \le u,v \le b} \frac{\partial C_{uv}}{\partial c_{kl}} U_u U_v = \sum_{1 \le u \le b} E(U_u, I_u) = \sum_{1 \le p < q \le n} E(U_p - U_q, I_{pq}) =$$
$$= (U_k - U_l)^2 = \sum_{1 \le u,v \le b} (V_{ku} - V_{lu})(V_{kv} - V_{lv}) U_u U_v.$$

(4) This follows directly from the latter formula.

(5) Assume that for each k, l either $Re\, c_{kl} > 0$ or $c_{kl} = 0$. Take $U_1, \dots, U_b \in \mathbb{R}$. By Claim 5.1 we have

$$\sum_{1 \le u,v \le b} Re\, C_{uv} U_u U_v = \sum_{1 \le u \le b} Re(U_u \bar{I}_u) =$$
$$= \sum_{1 \le k < l \le n} Re((U_k - U_l)\bar{I}_{kl}) = \sum_{1 \le k < l \le n} Re\, c_{kl} |U_k - U_l|^2 \ge 0.$$

□

Remark 5.3. If the network is connected then the latter inequality is strict unless $U_1 = \dots = U_b$.

5.2. Tilings and networks

Part (2) $\Longrightarrow$ (1) of the following result is probably new, cf. [1, 21].

Lemma 5.4. *Let P be a generic orthogonal polygon with horizontal sides of signed lengths $I_1, \dots, I_b$ and y-coordinates $U_1, \dots, U_b$. Then the following 2 conditions are equivalent:*

(1) *the polygon P can be tiled by m rectangles of ratios $c_1, \dots, c_m$;*
(2) *there is a planar electrical circuit with b boundary vertices, m essential edges of conductances $c_1, \dots, c_m > 0$, incoming voltages $U_1, \dots, U_b$ and incoming currents $I_1, \dots, I_b$.*

Remark 5.5. Condition (2) itself does not guarantee the existence of a rectangular polygon with horizontal sides of signed lengths $I_1, \dots, I_b$ and y-coordinates $U_1, \dots, U_b$. Lemma 5.4(1) $\Longrightarrow$ (2) is not necessarily true for *nongeneric* polygons.

Proof of Lemma 5.4. (1) $\Longrightarrow$ (2). Take a generic polygon P tiled by rectangles.

Let us construct the graph of the required network, see figure 3. Consider the union of the horizontal sides of all rectangles of the tiling. This union splits into several disjoint segments called *horizontal cuts*. Paint red (bold) all horizontal cuts except small neighborhoods of their endpoints. Paint blue (dashed) the vertical centerline of each rectangle in the tiling.

Contract all red segments. Then the blue set "becomes" a graph G and the polygon P "becomes" a topological disc D (since the y-coordinates of the horizontal sides of P are distinct it follows that each read segment has not more than one common point with ∂P). Denote by $1, \dots, b$ the vertices of the graph G obtained from the red segments in the horizontal cuts containing the sides of P and by $b+1, \dots, n$ — the other vertices.

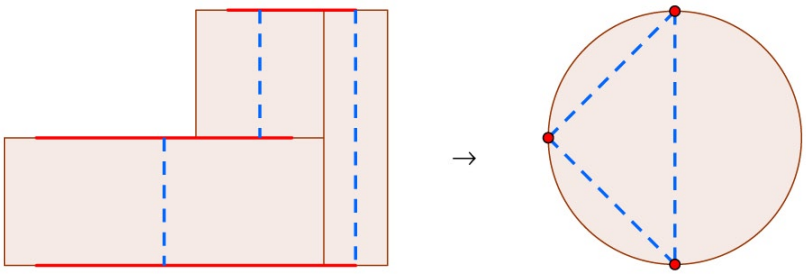


Figure 3: Construction of an electrical network

Clearly, $G \subset D$, $G \cap \partial D = \{1, \dots, b\}$ and each connected component of G contains a boundary vertex. Thus G is a graph of a planar network.

Let us define the voltages, currents and conductances in the network. For each vertex $k = 1, \dots, n$ of the graph G set U_k to be the y-coordinate of the horizontal red segment contracted to the vertex. For each edge kl of the graph G, obtained from the vertical centerline of a rectangle in the tiling, set I_{kl} and c_{kl} to be the horizontal side (with an appropriate sign) and the ratio of the rectangle, respectively. The laws (C), (I) are now checked directly. The constructed network is the required.

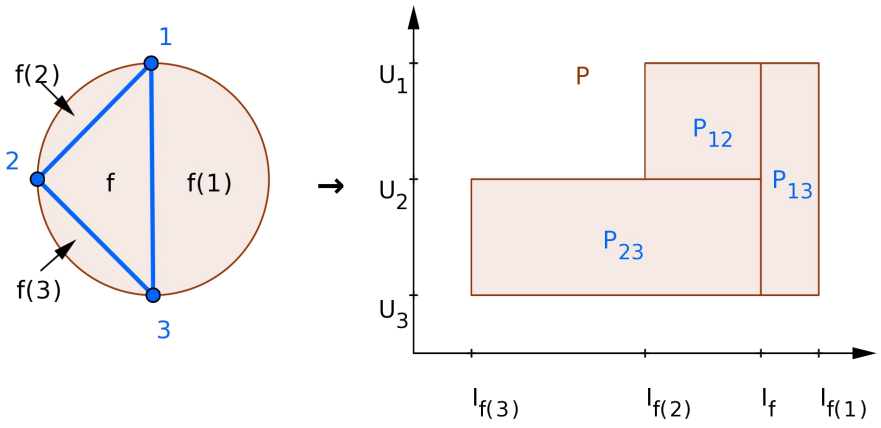


Figure 4: Construction of a tiling

(2) $\Longrightarrow$ (1). Take an electrical network as in (2). Construct a tiling of P as follows.

Let e be an edge of the network. Denote by $e\uparrow$ ($e\downarrow$) the endpoint of e with higher (lower) voltage (it is well-defined by the assumption that all the edges are essential). By a *face* we mean a connected component of the complement to the network in the unit disc D. Denote by $e\leftarrow$ ($e\rightarrow$) the face that borders the edge e from the left-hand (right-hand) side while one moves along the edge e from $e\uparrow$ to $e\downarrow$.

By law (I) it follows that to each face f one can assign a number I_f in such a way that $I_{kl\leftarrow} - I_{kl\rightarrow} = I_{kl}$. Without loss of generality assume $\min_f I_f = \min_{(x,y)\in P} x$, where the minimum in the left-hand side is over all the faces f meeting ∂D.

Let P_e be the rectangle with the vertices $(I_{e\rightarrow}, U_{e\uparrow}), (I_{e\rightarrow}, U_{e\downarrow}), (I_{e\leftarrow}, U_{e\uparrow}), (I_{e\leftarrow}, U_{e\downarrow})$. The rectangles P_e, where e runs through all the edges of the network, tile the polygon P by the following two claims (P_e-s cover P by Claim 5.6 and do not overlap by Claim 5.7). $\square$

Claim 5.6. $\bigcup_e P_e = P$.

Proof. It suffices to prove that $\partial \bigcup_e P_e \subset \partial P$. Since ∂P is a simple closed curve in the plane and $\bigcup_e P_e$ is bounded, the claim will follow.

We need the following description of the boundary ∂P, see figure 4. Boundary vertices split ∂D into b arcs. Start from vertex b and move along the circle ∂D counterclockwise. Enumerate the arcs in the order they appear in the motion. Denote by $f(v)$ the face containing the arc v. Denote by H_v the segment joining the points $(I_{f(v)}, U_v)$ and $(I_{f(v+1)}, U_v)$. Denote by V_v the segment joining the points $(I_{f(v)}, U_{v-1})$ and $(I_{f(v)}, U_v)$, where we set $U_0 = U_b$. Clearly, $\partial P = \bigcup_{v=1}^{b} (H_v \cup V_v)$.

Take a "generic" point $p \in \partial \bigcup_e P_e$, say, in a horizontal side of the "polygon" $\bigcup_e P_e$. The point p necessarily belongs to a horizontal side of a rectangle in the tiling, say, to the top side of a rectangle P_e. Denote by $v = e\uparrow$ the vertex of e of higher voltage.

Draw a horizontal line H through the top side of the rectangle P_e. We say that a rectangle P_d is *adjacent* if the vertex v is an endpoint of the edge d. Adjacent rectangles border upon the line H either from above or from below.

First assume that v is nonboundary. A simple induction shows that each point of H (except a finite set) is bordered by the same number of adjacent rectangles P_d from above and from below. Since the rectangle P_e borders upon the point p from below and p is "generic" it follows that some adjacent rectangle P_d borders upon it from above. Thus p belongs to $Int\, P_e \cup P_f \subset Int \bigcup_e P_e$, a contradiction.

So v is a boundary vertex. Analogously to the above each point of $H - H_v$ (except a finite set) is bordered by the same number of adjacent rectangles P_d from above and from below. Hence $p \in H_v$ and thus $p \in \partial P$. □

Claim 5.7. $\sum_e Area(P_e) = Area(P)$.

Proof. This follows immediately from Claim 5.1 because $Area(P_{kl}) = (U_k - U_l)I_{kl}$ and $Area(P) = \sum_{1 \le u \le b} U_u I_u$. □

5.3. Positive real functions

Let us prove Lemma 2.5. For a generalization to the case $b > 2$ see [12].

Proof of Lemma 2.5. (1) $\Longrightarrow$ (2). Indeed, if $Re\, z \le 0$ then $Re\, C(z) = -Re\, C(-z) \le 0$ and thus $C(z) \ne 1$.

(2) $\Longrightarrow$ (1). Consider the equation $C(z) = w$. Move w continuously in the half-plane $Re\, w > 0$. The roots cannot cross the line $Re\, z = 0$ (because $Re\, z = 0$ implies $Re\, C(z) = 0$ for an odd function $C(z) \in \mathbb{R}(z)$). Thus for each w in the half-plane $Re\, w > 0$ all roots of $C(z) = w$ are in the half-plane $Re\, z > 0$. Since $C(z)$ is odd it follows that the same is true for the half-planes $Re\, w < 0$, $Re\, z < 0$. So (1) holds.

(1) $\Longrightarrow$ (3). Suppose that $C(z) = 0$, where $z \in \mathbb{C}$. Then $Re\, z = 0$ because $Re\, z > 0 \implies Re\, C(z) > 0$ and $Re\, z < 0 \implies Re\, C(z) = -Re\, C(-z) < 0$. Since condition (1) and its converse hold in a neighborhood of the point z it follows that $C'(z) > 0$. A simple limiting argument proves the same for $z = \infty$.

(3) $\Longrightarrow$ (4) Assume for simplicity that $C(\infty) \ne 0$. Let $z_1, \dots, z_m$ be the roots of $C(z)$. Since $C'(z_k) > 0$ it follows that the roots are simple. Thus $C(z)$ has not more than m poles. The roots split the projective line $Re\, z = 0$ into m "segments". Since $C'(z_k) > 0$ it follows that for sufficiently small $\epsilon > 0$ we have $C(z_k - i\epsilon) < 0$ and $C(z_k + i\epsilon) > 0$. By intermediate value theorem it follows that each of the segments contains a pole of $C(z)$. Thus all the m poles of $C(z)$ belong to the line $Re\, z = 0$ and alternate with the roots. So (4) holds.

(4) $\Longrightarrow$ (5). Denote by $ht\,C(z)$ the sum of the degrees of the nominator and the denominator of $C(z)$. The proof is by induction over $ht\,C(z)$. If $ht\,C(z) = 1$ then there is nothing to prove. Assume that, say, $C(z)$ equals the expression from condition (4), where $n \geq 1$ and $b_n \neq 0$.

Denote by $r(z) = 1/(C(z) - d_1 z)$ and $q(z) = 1/C(z)$. Let us prove that $r(z)$ satisfies condition (3). Indeed, the roots of $r(z)$ are the numbers $\pm ib_1, \ldots, \pm ib_n$. For each $l = 1, \ldots, n$

$$r'(\pm ib_l) = q'(\pm ib_l) = \frac{2}{d_1(a_l^2 - b_l^2)} \prod_{k \neq l} \frac{b_k^2 - b_l^2}{a_k^2 - b_l^2} > 0$$

by the condition $a_1 > b_1 > a_2 > \cdots > b_n \geq 0$.

Hence by Lemma 2.5(3) $\Longrightarrow$ (4) it follows that $r(z)$ satisfies condition (4) as well. On the other hand $ht\,r(z) < ht\,C(z)$. By inductive hypothesis, $r(z)$ satisfies condition (5). Thus $C(z) = 1/(d_1 z + r(z))$ also satisfies condition (5).

(5) $\Longrightarrow$ (1). This follows by a simple induction over m. □

6. Proof of variations

6.1. Proof of Theorem 4.2

Proof of Theorem 4.2. (1) $\Longrightarrow$ (2). Let the polygon P be tiled by squares. By Lemma 5.4 there is a planar electrical circuit with edge conductances 1, incoming voltages $U_1, \ldots, U_b$ and incoming currents $I_1, \ldots, I_b$. Let C_{uv} be the response of the circuit. Then $I_v = \sum C_{uv} U_u$. By Lemma 5.2(1) all the entries of C_{uv} are rational. By Theorem 4.4 we have $C_{uv} \in \Omega_b$.

(2) $\Longrightarrow$ (1). Let $C_{uv} \in \Omega_b$ be a matrix with rational entries such that $I_v = \sum C_{uv} U_u$. By Theorem 4.4 there are planar electrical networks with the response C_{uv}. Take a minimal network with this property. By Theorem 4.5 the conductances of all the edges of the network are rational. Set the incoming voltages to be $U_1, \ldots, U_b$. Then the incoming currents are $I_1, \ldots, I_b$. Delete all unessential edges from the circuit. By Lemma 5.4 it follows that the polygon P can be tiled by rectangles of rational ratio, and hence by squares. □

Corollary 6.1. (of Lemmas 5.2, 5.4 and Theorem 4.4) *If a generic orthogonal polygon P can be tiled by rectangles of ratios $c_1, \ldots, c_n$ then there is a function $C_{uv}(z_1, \ldots, z_n)$ satisfying conditions (1), (2) and (5) of Lemma 5.2 such that $C(c_1, \ldots, c_n) \in \Omega_b$ and $I_v = \sum_{1 \leq u \leq b} C_{uv}(c_1, \ldots, c_n) U_u$ for each $v = 1, \ldots, b$.*

6.2. Proof of Theorem 4.3

Proof of Theorem 4.3. $\Longleftarrow$. This holds because a polygon with rational vertices can be tiled by squares.

$\Longrightarrow$. Suppose that P can be tiled by rectangles of ratios c and $1/c$. Let us prove analogously to the proof of Theorem 1.5(1) $\Longrightarrow$ (2) that all algebraic conjugates of c have positive real parts. Then Theorem 4.3 will follow from Theorem 1.5(2) $\Longrightarrow$ (1).

Consider the circuit given by Lemma 5.4. Replace each edge of conductance c (respectively, $1/c$) in the circuit by an edge of conductance $z \in \mathbb{C}$ (respectively, $1/z$). Let $C_{uv}(z)$ be the response of the obtained circuit. Consider the *energy dissipation function* $E(z) = \sum_{1 \leq u,v \leq b} C_{uv}(z) U_u U_v$. Since each $U_u \in \mathbb{Q}$ it follows by Lemma 5.2(1) that $E(z) \in \mathbb{Q}(z)$. Clearly, $E(c) = \sum_{1 \leq u \leq b} I_u U_u = Area(P)$. Thus $E(c) \in \mathbb{Q}$ and $E(c) > 0$.

Since $E(z) \in \mathbb{Q}(z)$ and $E(c) \in \overline{\mathbb{Q}}$ it follows that c is algebraic ($E(z)$ is nonconstant because $E(-c) = -E(c) < 0$ by Lemma 5.2(2)). Let z be an algebraic conjugate of c. Then $E(z) = E(c) > 0$.

Let us prove that $Re\, z > 0$. Indeed, first assume that $Re\, z < 0$. Then by Lemma 5.2(5) we have $0 \le Re\, E(-z) = -Re\, E(z) < 0$, a contradiction. A simple limiting argument shows that assumption $Re\, z = 0$ also leads to a contradiction. Thus $Re\, z > 0$. □

6.3. Proof of Theorem 4.6

Proof of Theorem 4.6. (1) $\Longrightarrow$ (2). This follows from Lemma 5.2(5) and Remark 5.3.

(2) $\Longrightarrow$ (1). For $b = 2$ there is nothing to prove. Assume that $b = 3$. Let $\delta > 0$ be a small number, $r_{uv} = -Re\, C_{uv} - \delta$, $m_{uv} = -Im\, C_{uv}$. By the assumption of the theorem it follows that $\left(\begin{smallmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{smallmatrix}\right)$ is positively definite. Thus $\left(\begin{smallmatrix} r_{31}+r_{12} & -r_{12} \\ -r_{12} & r_{12}+r_{23} \end{smallmatrix}\right)$ is positively definite for sufficiently small δ. Hence $r_{12}+r_{23}, r_{31}+r_{12}, r_{12}r_{23}+r_{23}r_{31}+r_{31}r_{12} > 0$. Analogously $r_{23}+r_{31} > 0$. Thus at least two of the numbers r_{12}, r_{23}, r_{31} are positive.

If $r_{12}, r_{23}, r_{31} > 0$ then the required network is a triangle 123 with edge conductances $c_{kl} = r_{kl} + im_{kl} + \delta$.

Now assume that exactly one of the numbers r_{12}, r_{23}, r_{31}, say, r_{31} is nonpositive. Take a large number M and denote by $\Delta_M = r_{12}r_{23} + r_{23}r_{31} + r_{31}r_{12} + iM(r_{23}+r_{12})$. The required network is a complete graph on the vertices $1, 2, 3, 4$ with edge conductances

$$\begin{array}{llll} c_{12} = & im_{12}+\delta, & c_{14} = & \Delta_M/r_{23}, \\ c_{23} = & im_{23}+\delta, & c_{34} = & \Delta_M/r_{12}, \\ c_{31} = & im_{31}+\delta-iM, & c_{24} = & \Delta_M/(r_{31}+iM). \end{array}$$

Clearly, for $M^2 > (r_{12}r_{23}+r_{23}r_{31}+r_{31}r_{12})|r_{31}|/(r_{23}+r_{12})$ we have all $Re\, c_{kl} > 0$.

Let us show by electrical transformations that the network has response C_{uv}. Indeed, replace the "letter Y" formed by the edges 14, 24 and 34 by a "triangle Δ" formed by 3 new edges of conductances $c'_{12} = r_{12}$, $c'_{23} = r_{23}$ and $c'_{31} = r_{31} + iM$. This *$Y\Delta$-transformation* does not change the response [21, page 12]. The obtained network has 3 pairs of multiple edges. Thus it has the same response as a triangle with edge conductances $r_{12}+im_{12}+\delta$, $r_{23}+im_{23}+\delta$, $r_{31}+im_{31}+\delta$. So the network has the response C_{uv}. □

Acknowledgements

The authors are grateful to A. Akopyan for many useful discussions and suggestions over the years. M. Skopenkov was supported in part by INTAS grant 06-1000014-6277, Russian Foundation of Basic Research grants 06-01-72551-NCNIL-a, Moebius Contest Foundation for Young Scientists and Euler Foundation.

References

[1] I. Benjamini and O. Schramm, Random walks and harmonic functions on infinite planar graphs using square tilings, Ann. Prob. **24:3** (1996), 1219–1238.

[2] V.G. Boltianskii, Hilbert's Third Problem (trans. by R. Silverman), V. H. Winston & Sons, Washington D.C., 1978.

[3] R.L. Brooks, C.A.B. Smith, A.H. Stone and W.T. Tutte, The Dissection of Rectangles into Squares, Duke Math. J. **7** (1940), 312–340.

[4] J. Cannon, W. Floyd, W. Parry, Squaring rectangles: the finite Riemann mapping theorem, Contemp. Math. **169** (1994), 133–211.

[5] W. Cauer, Die Verwirklichung der Wechselstromwiderst ände vorgeschriebener Frequenzabh ängigkeit, Archiv für Elektrotechnik **17** (1926), 355–388 (in German).

[6] Y. Colin de Verdière, Réseaux électriques planaires I, Comm. Math. Helv. **69:1** (1994), 351–374 (in French).

[7] Y. Colin de Verdière, I. Gitler and D. Vertigan, Réseaux électriques planaires II, Comm. Math. Helv. **71:1** (1996), 144–167 (in French).

[8] E.B. Curtis, D. Ingerman and J.A. Morrow, Circular planar graphs and resistor networks, Lin. Alg. Appl. **283** (1998), 115–150.

[9] E.B. Curtis and J.A. Morrow, Inverse problems for electrical networks, Series on Applied Mathematics **13**, World Scientific, Singapore, 2000.

[10] M. Dehn, Über die Zerlegung von Rechtecken in Rechtecke, Math. Ann. **57** (1903), 314–332 (in German).

[11] P.G. Doyle and J.L. Snell, Random walks and electric networks, Mathematical Association of America, 1984, http://arxiv.org/abs/math.PR/0001057.

[12] R.J. Duffin, Elementary operations which generate network matrices, Proc. AMS **6:3** (1955), 335–339.

[13] A.J.W. Duijvestijn, Simple perfect squared square of lowest order, J. Comb. Theory B **25** (1978), 240–243.

[14] R.M. Foster, A reactance theorem, Bell System Techn. J. **3** (1924), 259–267.

[15] C. Freiling, R. Hunter, C. Turner and R. Wheeler, Tiling with Squares and Anti-Squares, Amer. Math. Monthly **107:3** (2000), 195–204.

[16] C. Freiling, M. Laczkovich, D. Rinne, Rectangling a rectangle, Discr. Comp. Geometry **17** (1997), 217-225.

[17] C. Freiling, D. Rinne, Tiling a square with similar rectangles, Math. Res. Lett. **1** (1994), 547–558.

[18] H. Hadwiger, Vorlesungen Uber Inhalt, Oberflähe und Isoperimetrie, Springer-Verlag, 1957 (in German).

[19] K. Keating and J.L. King, Shape tiling, Elect. J. Comb. **4:2** (1997), R12.

[20] K. Keating and J.L. King, Signed tilings with squares, J. Comb. Theory A **85:1** (1999), 83–91.

[21] R. Kenyon, Tilings and discrete Dirichlet problems, Israel J. Math. **105:1** (1998), 61–84.

[22] M. Laczkovich, G. Szekeres, Tiling of the square with similar rectangles, Discr. Comp. Geometry **13** (1995), 569–572.

[23] G.F. Lawler and J. Sylvester, Determining resistances from boundary measurements in finite networks, SIAM J. Discr. Math., **2** (1989), 211–239.

[24] L. Lovasz, Random walks on graphs: a survey, In: Combinatorics, Paul Erdos is Eighty, D. Milos, V.T. Sos, and T. Szony, Eds., Budapest, Hungary: Janos Bolyai Math. Soc., 1996, 353–398.

[25] V.G. Pokrovskii, Slicings of n-dimensional parallelepipeds, Math. Notes **33:2** (1983), 137–140.

[26] M. Prasolov, M. Skopenkov, Dissections of a metal rectangle, Kvant (2008) (in Russian), submitted.

[27] J.W.S. Rayleigh, On the theory of resonance, In: Collected scientific papers **1** (1899), 33–75.

[28] Z. Su and R. Ding, Tilings of orthogonal polygons with similar rectangles or triangles, J. Appl. Math. Comp. **17:1** (2005), 343–350.

[29] B. Szegedy, Tilings of the square with similar right triangles, Combinatorica **21:1** (2001), 139–144.

[30] H.S. Wall, Analytic theory of continued fractions, Chelsea Pub. Co., Bronx, N.Y., 1967, 433 p.

[31] H. Weyl, Repartición de corriente en uno red conductora, Rev. Mat. Hisp. Amer. **5** (1923), 153–164 (in Spanish).

[32] I.M. Yaglom, How to dissect a square? Mathematicheskaya bibliotechka, Nauka, Moscow, 1968, 112 p. (in Russian), http://ilib.mirror1.mccme.ru/djvu/yaglom/square.htm.

Random walks and electric networks

Peter G. Doyle J. Laurie Snell

Version 3.02, 5 January 2000

Preface

Probability theory, like much of mathematics, is indebted to physics as a source of problems and intuition for solving these problems. Unfortunately, the level of abstraction of current mathematics often makes it difficult for anyone but an expert to appreciate this fact. In this work we will look at the interplay of physics and mathematics in terms of an example where the mathematics involved is at the college level. The example is the relation between elementary electric network theory and random walks.

Central to the work will be Polya's beautiful theorem that a random walker on an infinite street network in d-dimensional space is bound to return to the starting point when $d = 2$, but has a positive probability of escaping to infinity without returning to the starting point when $d \geq 3$. Our goal will be to interpret this theorem as a statement about electric networks, and then to prove the theorem using techniques from classical electrical theory. The techniques referred to go back to Lord Rayleigh, who introduced them in connection with an investigation of musical instruments. The analog of Polya's theorem in this connection is that wind instruments are possible in our three-dimensional world, but are not possible in Flatland (Abbott [1]).

The connection between random walks and electric networks has been recognized for some time (see Kakutani [12], Kemeny, Snell, and

Knapp [14], and Kelly [13]). As for Rayleigh's method, the authors first learned it from Peter's father Bill Doyle, who used it to explain a mysterious comment in Feller ([5], p. 425, Problem 14). This comment suggested that a random walk in two dimensions remains recurrent when some of the streets are blocked, and while this is ticklish to prove probabilistically, it is an easy consequence of Rayleigh's method. The first person to apply Rayleigh's method to random walks seems to have been Nash-Williams [24]. Earlier, Royden [30] had applied Rayleigh's method to an equivalent problem. However, the true importance of Rayleigh's method for probability theory is only now becoming appreciated. See, for example, Griffeath and Liggett [9], Lyons [20], and Kesten [16].

Here's the plan of the work: In Section 1 we will restrict ourselves to the study of random walks on finite networks. Here we will establish the connection between the electrical concepts of current and voltage and corresponding descriptive quantities of random walks regarded as finite state Markov chains. In Section 2 we will consider random walks on infinite networks. Polya's theorem will be proved using Rayleigh's method, and the proof will be compared with the classical proof using probabilistic methods. We will then discuss walks on more general infinite graphs, and use Rayleigh's method to derive certain extensions of Polya's theorem. Certain of the results in Section 2 were obtained by Peter Doyle in work on his Ph.D. thesis.

To read this work, you should have a knowledge of the basic concepts of probability theory as well as a little electric network theory and linear algebra. An elementary introduction to finite Markov chains as presented by Kemeny, Snell, and Thompson [15] would be helpful.

The work of Snell was carried out while enjoying the hospitality of Churchill College and the Cambridge Statistical Laboratory supported by an NSF Faculty Development Fellowship. He thanks Professors Kendall and Whittle for making this such an enjoyable and rewarding visit. Peter Doyle thanks his father for teaching him how to think like a physicist. We both thank Peter Ney for assigning the problem in Feller that started all this, David Griffeath for suggesting the example to be used in our first proof that 3-dimensional random walk is recurrent (Section 2.2.9), and Reese Prosser for keeping us going by his friendly and helpful hectoring. Special thanks are due Marie Slack, our secretary extraordinaire, for typing the original and the excessive number of revisions one is led to by computer formatting.

1 Random walks on finite networks

1.1 Random walks in one dimension

1.1.1 A random walk along Madison Avenue

A *random walk*, or *drunkard's walk*, was one of the first chance processes studied in probability; this chance process continues to play an important role in probability theory and its applications. An example of a random walk may be described as follows:

A man walks along a 5-block stretch of Madison Avenue. He starts at corner x and, with probability $1/2$, walks one block to the right and, with probability $1/2$, walks one block to the left; when he comes to the next corner he again randomly chooses his direction along Madison Avenue. He continues until he reaches corner 5, which is home, or corner 0, which is a bar. If he reaches either home or the bar, he stays there. (See Figure 1.)

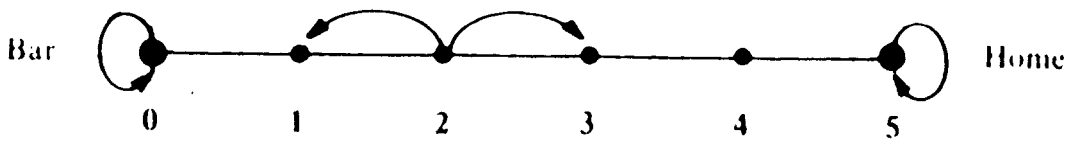


Figure 1: ♣

The problem we pose is to find the probability $p(x)$ that the man, starting at corner x, will reach home before reaching the bar. In looking at this problem, we will not be so much concerned with the particular form of the solution, which turns out to be $p(x) = x/5$, as with its general properties, which we will eventually describe by saying "$p(x)$ is the unique solution to a certain Dirichlet problem."

1.1.2 The same problem as a penny matching game

In another form, the problem is posed in terms of the following game: Peter and Paul match pennies; they have a total of 5 pennies; on each match, Peter wins one penny from Paul with probability $1/2$ and loses one with probability $1/2$; they play until Peter's fortune reaches 0 (he

is ruined) or reaches 5 (he wins all Paul's money). Now $p(x)$ is the probability that Peter wins if he starts with x pennies.

1.1.3 The probability of winning: basic properties

Consider a random walk on the integers $0, 1, 2, \ldots, N$. Let $p(x)$ be the probability, starting at x, of reaching N before 0. We regard $p(x)$ as a function defined on the points $x = 0, 1, 2, \ldots, N$. The function $p(x)$ has the following properties:

(a) $p(0) = 0$.

(b) $p(N) = 1$.

(c) $p(x) = \frac{1}{2}p(x-1) + \frac{1}{2}p(x+1)$ for $x = 1, 2, \ldots, N-1$.

Properties (a) and (b) follow from our convention that 0 and N are traps; if the walker reaches one of these positions, he stops there; in the game interpretation, the game ends when one player has all of the pennies. Property (c) states that, for an interior point, the probability $p(x)$ of reaching home from x is the average of the probabilities $p(x-1)$ and $p(x+1)$ of reaching home from the points that the walker may go to from x. We can derive (c) from the following basic fact about probability:

Basic Fact. Let E be any event, and F and G be events such that one and only one of the events F or G will occur. Then

$$\mathbf{P}(E) = \mathbf{P}(F) \cdot \mathbf{P}(E \text{ given } F) + \mathbf{P}(G) \cdot \mathbf{P}(E \text{ given } G).$$

In this case, let E be the event "the walker ends at the bar", F the event "the first step is to the left", and G the event "the first step is to the right". Then, if the walker starts at x, $\mathbf{P}(E) = p(x)$, $\mathbf{P}(F) = \mathbf{P}(G) = \frac{1}{2}$, $\mathbf{P}(E \text{ given } F) = p(x-1)$, $\mathbf{P}(E \text{ given } G) = p(x+1)$, and (c) follows.

1.1.4 An electric network problem: the same problem?

Let's consider a second apparently very different problem. We connect equal resistors in series and put a unit voltage across the ends as in Figure 2.

Voltages $v(x)$ will be established at the points $x = 0, 1, 2, 3, 4, 5$. We have grounded the point $x = 0$ so that $v(0) = 0$. We ask for the voltage $v(x)$ at the points x between the resistors. If we have N resistors, we

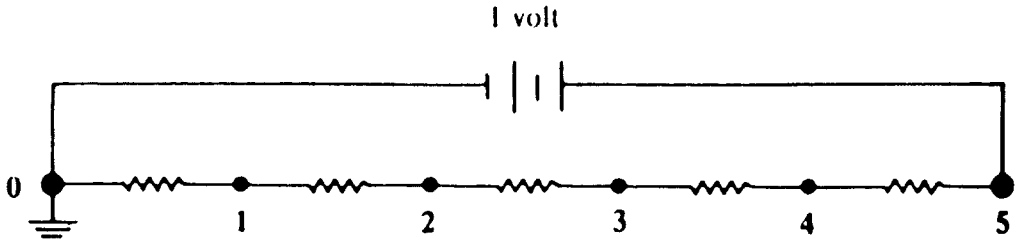


Figure 2: ♣

make $v(0) = 0$ and $v(N) = 1$, so $v(x)$ satisfies properties (a) and (b) of Section 1.1.3. We now show that $v(x)$ also satisfies (c).

By Kirchhoff's Laws, the current flowing into x must be equal to the current flowing out. By Ohm's Law, if points x and y are connected by a resistance of magnitude R, then the current i_{xy} that flows from x to y is equal to

$$i_{xy} = \frac{v(x) - v(y)}{R}.$$

Thus for $x = 1, 2, \ldots, N-1$,

$$\frac{v(x-1) - v(x)}{R} + \frac{v(x+1) - v(x)}{R} = 0.$$

Multiplying through by R and solving for $v(x)$ gives

$$v(x) = \frac{v(x+1) + v(x-1)}{2}$$

for $x = 1, 2, \ldots, N-1$. Therefore, $v(x)$ also satisfies property (c).

We have seen that $p(x)$ and $v(x)$ both satisfy properties (a), (b), and (c) of Section 1.1.3. This raises the question: are $p(x)$ and $v(x)$ equal? For this simple example, we can easily find $v(x)$ using Ohm's Law, find $p(x)$ using elementary probability, and see that they are the same. However, we want to illustrate a principle that will work for very general circuits. So instead we shall prove that these two functions are the same by showing that there is only one function that satisfies these properties, and we shall prove this by a method that will apply to more general situations than points connected together in a straight line.

Exercise 1.1.1 Referring to the random walk along Madison Avenue, let $X = p(1)$, $Y = p(2)$, $Z = p(3)$, and $W = p(4)$. Show that properties (a), (b), and (c) of Section 1.1.3 determine a set of four linear equations with variables X, Y, Z and W. Show that these equations have a unique solution. What does this say about $p(x)$ and $v(x)$ for this special case?

Exercise 1.1.2 Assume that our walker has a tendency to drift in one direction: more specifically, assume that each step is to the right with probability p or to the left with probability $q = 1 - p$. Show that properties (a), (b), and (c) of Section 1.1.3 should be replaced by

(a) $p(0) = 0$.

(b) $p(N) = 1$.

(c) $p(x) = q \cdot p(x-1) + p \cdot p(x+1)$.

Exercise 1.1.3 In our electric network problem, assume that the resistors are not necessarily equal. Let R_x be the resistance between x and $x+1$. Show that

$$v(x) = \frac{\frac{1}{R_{x-1}}}{\frac{1}{R_{x-1}} + \frac{1}{R_x}} v(x-1) + \frac{\frac{1}{R_x}}{\frac{1}{R_{x-1}} + \frac{1}{R_x}} v(x+1).$$

How should the resistors be chosen to correspond to the random walk of Exercise 1.1.2?

1.1.5 Harmonic functions in one dimension; the Uniqueness Principle

Let S be the set of points $S = \{0, 1, 2, \ldots, N\}$. We call the points of the set $D = \{1, 2, \ldots, N-1\}$ the *interior points* of S and those of $B = \{0, N\}$ the *boundary points* of S. A function $f(x)$ defined on S is *harmonic* if, at points of D, it satisfies the averaging property

$$f(x) = \frac{f(x-1) + f(x+1)}{2}.$$

As we have seen, $p(x)$ and $v(x)$ are harmonic functions on S having the same values on the boundary: $p(0) = v(0) = 0$; $p(N) = v(N) = 1$. Thus both $p(x)$ and $v(x)$ solve the problem of finding a harmonic

function having these boundary values. Now the problem of finding a harmonic function given its boundary values is called the *Dirichlet problem*, and the *Uniqueness Principle* for the Dirichlet problem asserts that there cannot be two different harmonic functions having the same boundary values. In particular, it follows that $p(x)$ and $v(x)$ are really the same function, and this is what we have been hoping to show. Thus the fact that $p(x) = v(x)$ is an aspect of a general fact about harmonic functions.

We will approach the Uniqueness Principle by way of the *Maximum Principle* for harmonic functions, which bears the same relation to the Uniqueness Principle as Rolle's Theorem does to the Mean Value Theorem of Calculus.

Maximum Principle . A harmonic function $f(x)$ defined on S takes on its maximum value M and its minimum value m on the boundary.

Proof. Let M be the largest value of f. Then if $f(x) = M$ for x in D, the same must be true for $f(x-1)$ and $f(x+1)$ since $f(x)$ is the average of these two values. If $x-1$ is still an interior point, the same argument implies that $f(x-2) = M$; continuing in this way, we eventually conclude that $f(0) = M$. That same argument works for the minimum value m. $\diamondsuit$

Uniqueness Principle. If $f(x)$ and $g(x)$ are harmonic functions on S such that $f(x) = g(x)$ on B, then $f(x) = g(x)$ for all x.

Proof. Let $h(x) = f(x) - g(x)$. Then if x is any interior point,

$$\frac{h(x-1)+h(x+1)}{2} = \frac{f(x-1)+f(x+1)}{2} - \frac{g(x-1)+g(x+1)}{2},$$

and h is harmonic. But $h(x) = 0$ for x in B, and hence, by the Maximum Principle, the maximum and mininium values of h are 0. Thus $h(x) = 0$ for all x, and $f(x) = g(x)$ for all x. $\diamondsuit$

Thus we finally prove that $p(x) = v(x)$; but what does $v(x)$ equal? The Uniqueness Principle shows us a way to find a concrete answer: just guess. For if we can find any harmonic function $f(x)$ having the right boundary values, the Uniqueness Principle guarantees that

$$p(x) = v(x) = f(x).$$

The simplest function to try for $f(x)$ would be a linear function; this leads to the solution $f(x) = x/N$. Note that $f(0) = 0$ and $f(N) = 1$ and

$$\frac{f(x-1)+f(x+1)}{2} = \frac{x-1+x+1}{2N} = \frac{x}{N} = f(x).$$

Therefore $f(x) = p(x) = v(x) = x/N$.

As another application of the Uniqueness Principle, we prove that our walker will eventually reach 0 or N. Choose a starting point x with $0 < x < N$. Let $h(x)$ be the probability that the walker never reaches the boundary $B = \{0, N\}$. Then

$$h(x) = \frac{1}{2}h(x+1) + \frac{1}{2}h(x-1)$$

and h is harmonic. Also $h(0) = h(N) = 0$; thus, by the Maximum Principle, $h(x) = 0$ for all x.

Exercise 1.1.4 Show that you can choose A and B so that the function $f(x) = A(q/p)^x + B$ satisfies the modified properties (a), (b) and (c) of Exercise 1.1.2. Does this show that $f(x) = p(x)$?

Exercise 1.1.5 Let $m(x)$ be the expected number of steps, starting at x, required to reach 0 or N for the first time. It can be proven that $m(x)$ is finite. Show that $m(x)$ satisfies the conditions

(a) $m(0) = 0$.

(b) $m(N) = 0$.

(c) $m(x) = \frac{1}{2}m(x+1) + \frac{1}{2}m(x-1) + 1$.

Exercise 1.1.6 Show that the conditions in Exercise 1.1.5 have a unique solution. Hint: show that if m and $\bar{m}$ are two solutions, then $f = m - \bar{m}$ is harmonic with $f(0) = f(N) = 0$ and hence $f(x) = 0$ for all x.

Exercise 1.1.7 Show that you can choose A, B, and C such that $f(x) = A + Bx + Cx^2$ satisfies all the conditions of Exercise 1.1.5. Does this show that $f(x) = m(x)$ for this choice of A, B, and C?

Exercise 1.1.8 Find the expected duration of the walk down Madison Avenue as a function of the walker's starting point (1, 2, 3, or 4).

1.1.6 The solution as a fair game (martingale)

Let us return to our interpretation of a random walk as Peter's fortune in a game of penny matching with Paul. On each match, Peter wins one penny with probability 1/2 and loses one penny with probability 1/2. Thus, when Peter has k pennies his expected fortune after the next play is

$$\frac{1}{2}(k-1) + \frac{1}{2}(k+1) = k,$$

so his expected fortune after the next play is equal to his present fortune. This says that he is playing a *fair game*; a chance process that can be interpreted as a player's fortune in a fair game is called a *martingale*.

Now assume that Peter and Paul have a total of N pennies. Let $p(x)$ be the probability that, when Peter has x pennies, he will end up with all N pennies. Then Peter's expected final fortune in this game is

$$(1 - p(x)) \cdot 0 + p(x) \cdot N = p(x) \cdot N.$$

If we could be sure that a fair game remains fair to the end of the game, then we could conclude that Peter's expected final fortune is equal to his starting fortune x, i.e., $x = p(x) \cdot N$. This would give $p(x) = x/N$ and we would have found the probability that Peter wins using the fact that a fair game remains fair to the end. Note that the time the game ends is a random time, namely, the time that the walk first reaches 0 or N for the first time. Thus the question is, is the fairness of a game preserved when we stop at a random time?

Unfortunately, this is not always the case. To begin with, if Peter somehow has knowledge of what the future holds in store for him, he can decide to quit when he gets to the end of a winning streak. But even if we restrict ourselves to stopping rules where the decision to stop or continue is independent of future events, fairness may not be preserved. For example, assume that Peter is allowed to go into debt and can play as long as he wants to. He starts with 0 pennies and decides to play until his fortune is 1 and then quit. We shall see that a random walk on the set of all integers, starting at 0, will reach the point 1 if we wait long enough. Hence, Peter will end up one penny ahead by this system of stopping.

However, there are certain conditions under which we can guarantee that a fair game remains fair when stopped at a random time. For our purposes, the following standard result of martingale theory will do:

Martingale Stopping Theorem. A fair game that is stopped at a random time will remain fair to the end of the game if it is assumed

that there is a finite amount of money in the world and a player must stop if he wins all this money or goes into debt by this amount.

This theorem would justify the above argument to obtain $p(x) = x/N$.

Let's step back and see how this martingale argument worked. We began with a harmonic function, the function $f(x) = x$, and interpreted it as the player's fortune in a fair game. We then considered the player's expected final fortune in this game. This was another harmonic function having the same boundary values and we appealed to the Martingale Stopping Theorem to argue that this function must be the same as the original function. This allowed us to write down an expression for the probability of winning, which was what we were looking for.

Lurking behind this argument is a general principle: If we are given boundary values of a function, we can come up with a harmonic function having these boundary values by assigning to each point the player's expected final fortune in a game where the player starts from the given point and carries out a random walk until he reaches a boundary point, where he receives the specified payoff. Furthermore, the Martingale Stopping Theorern allows us to conclude that there can be no other harmonic function with these boundary values. Thus martingale theory allows us to establish existence and uniqueness of solutions to a Dirichlet problem. All this isn't very exciting for the cases we've been considering, but the nice thing is that the same arguments carry through to the more general situations that we will be considering later on.

The study of martingales was originated by Levy [19] and Ville [34]. Kakutani [12] showed the connection between random walks and harmonic functions. Doob [4] developed martingale stopping theorems and showed how to exploit the preservation of fairness to solve a wide variety of problems in probability theory. An informal discussion of martingales may be found in Snell [32].

Exercise 1.1.9 Consider a random walk with a drift; that is, there is a probability $p \neq \frac{1}{2}$ of going one step to the right and a probability $q = 1 - p$ of going one step to the left. (See Exercise 1.1.2.) Let $w(x) = (q/p)^x$; show that, if you interpret $w(x)$ as your fortune when you are at x, the resulting game is fair. Then use the Martingale Stopping Theorem to argue that

$$w(x) = p(x)w(N) + (1 - p(x))w(0).$$

Solve for $p(x)$ to obtain

$$p(x) = \frac{\left(\frac{q}{p}\right)^x - 1}{\left(\frac{q}{p}\right)^N - 1}.$$

Exercise 1.1.10 You are gambling against a professional gambler; you start with A dollars and the gambler with B dollars; you play a game in which you win one dollar with probability $p < \frac{1}{2}$ and lose one dollar with probability $q = 1 - p$; play continues until you or the gambler runs out of money. Let R_A be the probability that you are ruined. Use the result of Exercise 1.1.9 to show that

$$R_A = \frac{1 - \left(\frac{p}{q}\right)^B}{1 - \left(\frac{p}{q}\right)^N}$$

with $N = A + B$. If you start with 20 dollars and the gambler with 50 dollars and $p = .45$, find the probability of being ruined.

Exercise 1.1.11 The gambler realizes that the probability of ruining you is at least $1 - (p/q)^B$ (Why?). The gambler wants to make the probability at least .999. For this, $(p/q)^B$ should be at most .001. If the gambler offers you a game with $p = .499$, how large a stake should she have?

1.2 Random walks in two dimensions

1.2.1 An example

We turn now to the more complicated problem of a random walk on a two-dimensional array. In Figure 3 we illustrate such a walk. The large dots represent boundary points; those marked E indicate escape routes and those marked P are police. We wish to find the probability $p(x)$ that our walker, starting at an interior point x, will reach an escape route before he reaches a policeman. The walker moves from $x = (a, b)$ to each of the four neighboring points $(a + 1, b)$, $(a - 1, b)$, $(a, b + 1)$, $(a, b-1)$ with probability $\frac{1}{4}$. If he reaches a boundary point, he remains at this point.

The corresponding voltage problem is shown in Figure 4. The boundary points P are grounded and points E are connected and fixed at one volt by a one-volt battery. We ask for the voltage $v(x)$ at the interior points.

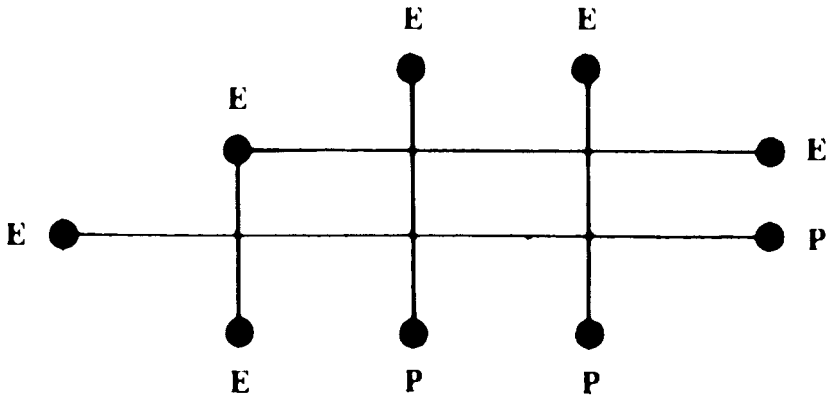


Figure 3: ♣

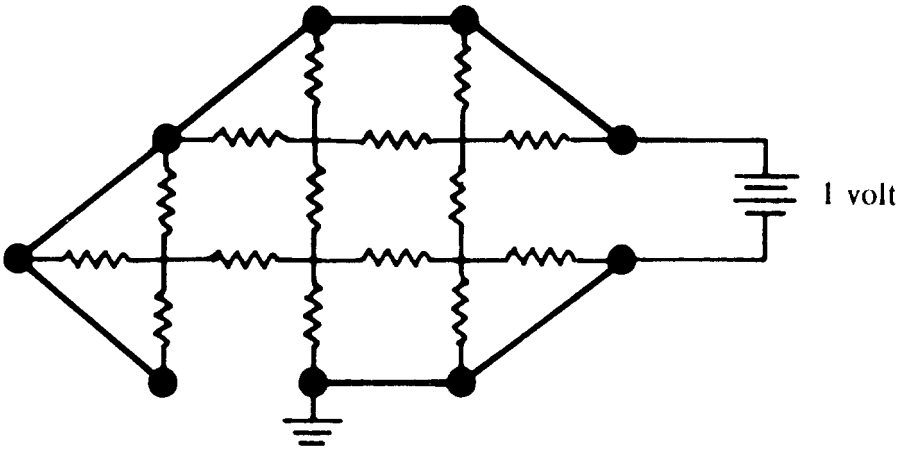


Figure 4: ♣

1.2.2 Harmonic functions in two dimensions

We now define harmonic functions for sets of lattice points in the plane (a lattice point is a point with integer coordinates). Let $S = D \cup B$ be a finite set of lattice points such that (a) D and B have no points in common, (b) every point of D has its four neighboring points in S, and (c) every point of B has at least one of its four neighboring points in D. We assume further that S hangs together in a nice way, namely, that for any two points P and Q in S, there is a sequence of points P_j in D such that $P, P_1, P_2, \ldots, P_n, Q$ forms a path from P to A. We call the points of D the *interior points* of S and the points of B the *boundary points* of S.

A function f defined on S is *harmonic* if, for points (a, b) in D, it has the averaging property

$$f(a,b) = \frac{f(a+1,b) + f(a-1,b) + f(a,b+1) + f(a,b-1)}{4}.$$

Note that there is no restriction on the values of f at the boundary points.

We would like to prove that $p(x) = v(x)$ as we did in the one-dimensional case. That $p(x)$ is harmonic follows again by considering all four possible first steps; that $v(x)$ is harmonic follows again by Kirchhoff's Laws since the current coming into $x = (a, b)$ is

$$\frac{v(a+1,b) - v(a,b)}{R} + \frac{v(a-1,b) - v(a,b)}{R} + \frac{v(a,b+1) - v(a,b)}{R} + \frac{v(a,b-1) - v(a,b)}{R} = 0.$$

Multiplying through by R and solving for $v(a, b)$ gives

$$v(a,b) = \frac{v(a+1,b) + v(a-1,b) + v(a,b+1) + v(a,b-1)}{4}.$$

Thus $p(x)$ and $v(x)$ are harmonic functions with the same boundary values. To show from this that they are the same, we must extend the Uniqueness Principle to two dimensions.

We first prove the Maximum Principle. If M is the maximum value of f and if $f(P) = M$ for P an interior point, then since $f(P)$ is the average of the values of f at its neighbors, these values must all equal M also. By working our way due south, say, repeating this argument at every step, we eventually reach a boundary point Q for which we can conclude that $f(Q) = M$. Thus a harmonic function always attains its maximum (or minimum) on the boundary; this is the Maximum

Principle. The proof of the Uniqueness Principle goes through as before since again the difference of two harmonic functions is harmonic.

The fair game argument, using the Martingale Stopping Theorem, holds equally well and again gives an alternative proof of the existence and uniqueness to the solution of the Dirichlet problem.

Exercise 1.2.1 Show that if f and g are harmonic functions so is $h = a \cdot f + b \cdot g$ for constants a and b. This is called the *superposition principle*.

Exercise 1.2.2 Let $B_1, B_2, \ldots, B_n$ be the boundary points for a region S. Let $e_j(a, b)$ be a function that is harmonic in S and has boundary value 1 at B_j and 0 at the other boundary points. Show that if arbitrary boundary values $v_1, v_2, \ldots, v_n$ are assigned, we can find the harmonic function v with these values from the solutions $e_1, e_2, \ldots, e_n$.

1.2.3 The Monte Carlo solution

Finding the exact solution to a Dirichlet problem in two dimensions is not always a simple matter, so before taking on this problem, we will consider two methods for generating approximate solutions. In this section we will present a method using random walks. This method is known as a *Monte Carlo method*, since random walks are random, and gambling involves randomness, and there is a famous gambling casino in Monte Carlo. In Section 1.2.4, we will describe a much more effective method for finding approximate solutions, called the *method of relaxations*.

We have seen that the solution to the Dirichlet problem can be found by finding the value of a player's final winning in the following game: Starting at x the player carries out a random walk until reaching a boundary point. He is then paid an amount $f(y)$ if y is the boundary point first reached. Thus to find $f(x)$, we can start many random walks at x and find the average final winnings for these walks. By the law of averages (the law of large numbers in probability theory), the estimate that we obtain this way will approach the true expected final winning $f(x)$.

Here are some estimates obtained this way by starting 10,000 random walks from each of the interior points and, for each x, estimating $f(x)$ by the average winning of the random walkers who started at this

point.

$$\begin{array}{ccccc} & & 1 & 1 & \\ & 1.824 & .785 & 1 & \\ 1 & .876 & .503 & .317 & 0 \\ & 1 & 0 & 0 & \end{array}$$

This method is a colorful way to solve the problem, but quite inefficient. We can use probability theory to estimate how inefficient it is. We consider the case with boundary values I or 0 as in our example. In this case, the expected final winning is just the probability that the walk ends up at a boundary point with value 1. For each point x, assume that we carry out n random walks; we regard each random walk to be an experiment and interpret the outcome of the ith experiment to be a "success" if the walker ends at a boundary point with a 1 and a "failure" otherwise. Let $p = p(x)$ be the unknown probability for success for a walker starting at x and $q = 1 - p$. How many walks should we carry out to get a reasonable estimate for p? We estimate p to be the fraction $\bar{p}$ of the walkers that end at a 1.

We are in the position of a pollster who wishes to estimate the proportion p of people in the country who favor candidate A over B. The pollster chooses a random sample of n people and estimates p as the proportion $\bar{p}$ of voters in his sample who favor A. (This is a gross oversimplification of what a pollster does, of course.) To estimate the number n required, we can use the central limit theorem. This theorem states that, if S_n, is the number of successes in n independent experiments, each having probability p for success, then for any $k > 0$

$$\mathbf{P}\left(-k < \frac{S_n - np}{\sqrt{npq}} < k\right) \approx A(k),$$

where $A(k)$ is the area under the normal curve between $-k$ and k. For $k = 2$ this area is approximately .95; what does this say about $\bar{p} = S_n/n$? Doing a little rearranging, we see that

$$\mathbf{P}\left(-2 < \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}} < 2\right) \approx .95$$

or

$$\mathbf{P}\left(-2\frac{\sqrt{pq}}{n} < \bar{p} - p < 2\frac{\sqrt{pq}}{n}\right) \approx .95.$$

Since $\sqrt{pq} \leq \frac{1}{2}$,

$$\mathbf{P}\left(-\frac{1}{\sqrt{n}} < \bar{p} - p < \frac{1}{\sqrt{n}}\right) \gtrapprox .95.$$

Thus, if we choose $\frac{1}{\sqrt{n}} = .01$, or $n = 10,000$, there is a 95 percent chance that our estimate $\bar{p} = S_n/n$ will not be off by more than .01. This is a large number for rather modest accuracy; in our example we carried out 10,000 walks from each point and this required about 5 seconds on the Dartmouth computer. We shall see later, when we obtain an exact solution, that we did obtain the accuracy predicted.

Exercise 1.2.3 You play a game in which you start a random walk at the center in the grid shown in Figure 5. When the walk reaches

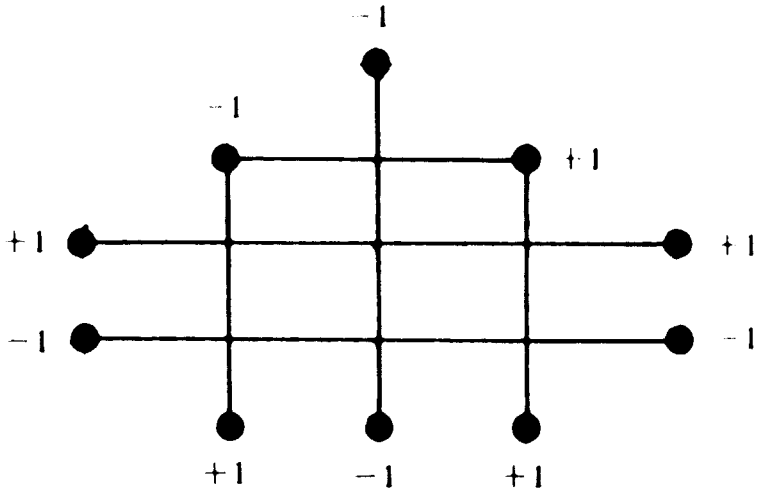


Figure 5: ♣

the boundary, you receive a payment of $+1$ or -1 as indicated at the boundary points. You wish to simulate this game to see if it is a favorable game to play; how many simulations would you need to be reasonably certain of the value of this game to an accuracy of .01? Carry out such a simulation and see if you feel that it is a favorable game.

1.2.4 The original Dirichlet problem; the method of relaxations

The Dirichlet problem we have been studying is not the original Dirichlet problem, but a discrete version of it. The original Dirichlet problem

concerns the distribution of temperature, say, in a continuous medium; the following is a representative example.

Suppose we have a thin sheet of metal gotten by cutting out a small square from the center of a large square. The inner boundary is kept at temperature 0 and the outer boundary is kept at temperature 1 as indicated in Figure 6. The problem is to find the temperature at

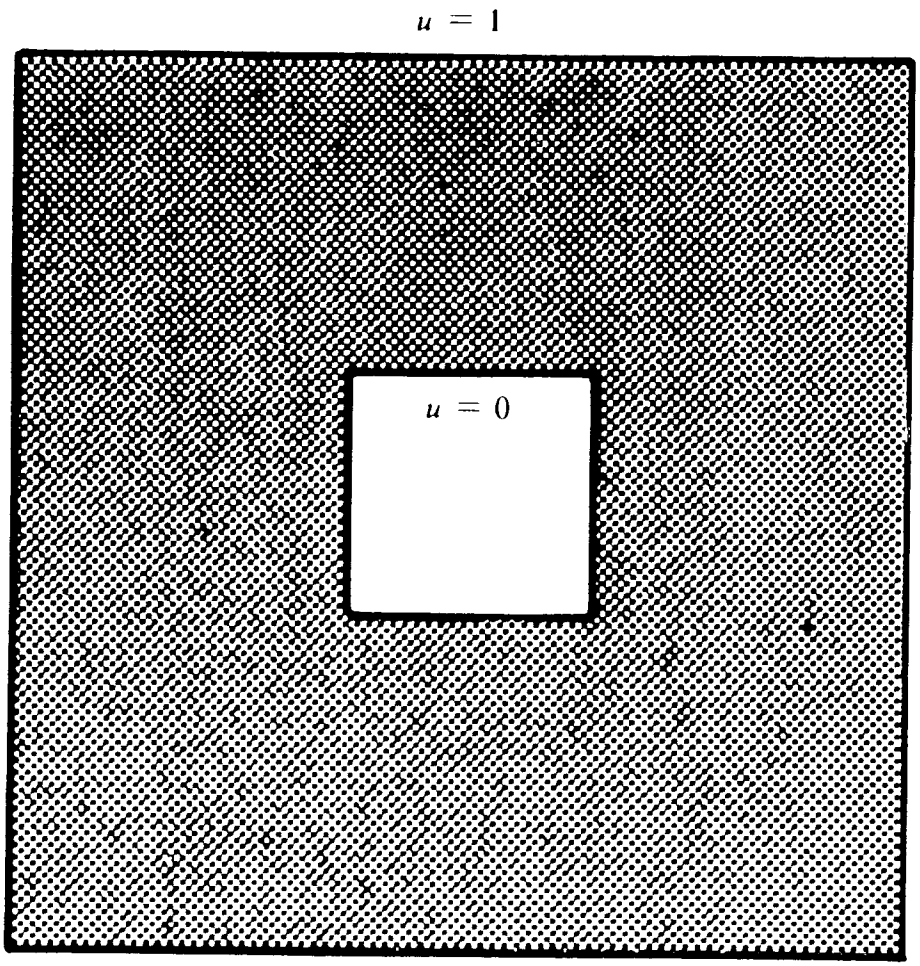


Figure 6: ♣

points in the rectangle's interior. If $u(x, y)$ is the temperature at (x, y), then u satisfies Laplace's differential equation

$$u_{xx} + u_{yy} = 0.$$

A function that satisfies this differential equation is called *harmonic*. It has the property that the value $u(x, y)$ is equal to the average of the values over any circle with center (x, y) lying inside the region. Thus to determine the temperature $u(x, y)$, we must find a harmonic function defined in the rectangle that takes on the prescribed boundary values. We have a problem entirely analogous to our discrete Dirichlet problem, but with continuous domain.

The *method of relaxations* was introduced as a way to get approximate solutions to the original Dirichlet problem. This method is actually more closely connected to the discrete Dirichlet problem than to the continuous problem. Why? Because, faced with the continuous problem just described, no physicist will hesitate to replace it with an analogous discrete problem, approximating the continuous medium by an array of lattice points such as that depicted in Figure 7, and searching for a function that is harmonic in our discrete sense and that takes on the appropriate boundary values. It is this approximating discrete problem to which the method of relaxations applies.

Here's how the method goes. Recall that we are looking for a function that has specified boundary values, for which the value at any interior point is the average of the values at its neighbors. Begin with any function having the specified boundary values, pick an interior point, and see what is happening there. In general, the value of the function at the point we are looking at will not be equal to the average of the values at its neighbors. So adjust the value of the function to be equal to the average of the values at its neighbors. Now run through the rest of the interior points, repeating this process. When you have adjusted the values at all of the interior points, the function that results will not be harmonic, because most of the time after adjusting the value at a point to be the average value at its neighbors, we afterwards came along and adjusted the values at one or more of those neighbors, thus destroying the harmony. However, the function that results after running through all the interior points, if not harmonic, is more nearly harmonic than the function we started with; if we keep repeating this averaging process, running through all of the interior points again and again, the function will approximate more and more closely the solution to our Dirichlet problem.

We do not yet have the tools to prove that this method works for a general initial guess; this will have to wait until later (see Exercise 1.3.12). We will start with a special choice of initial values for which we can prove that the method works (see Exercise 1.2.5).

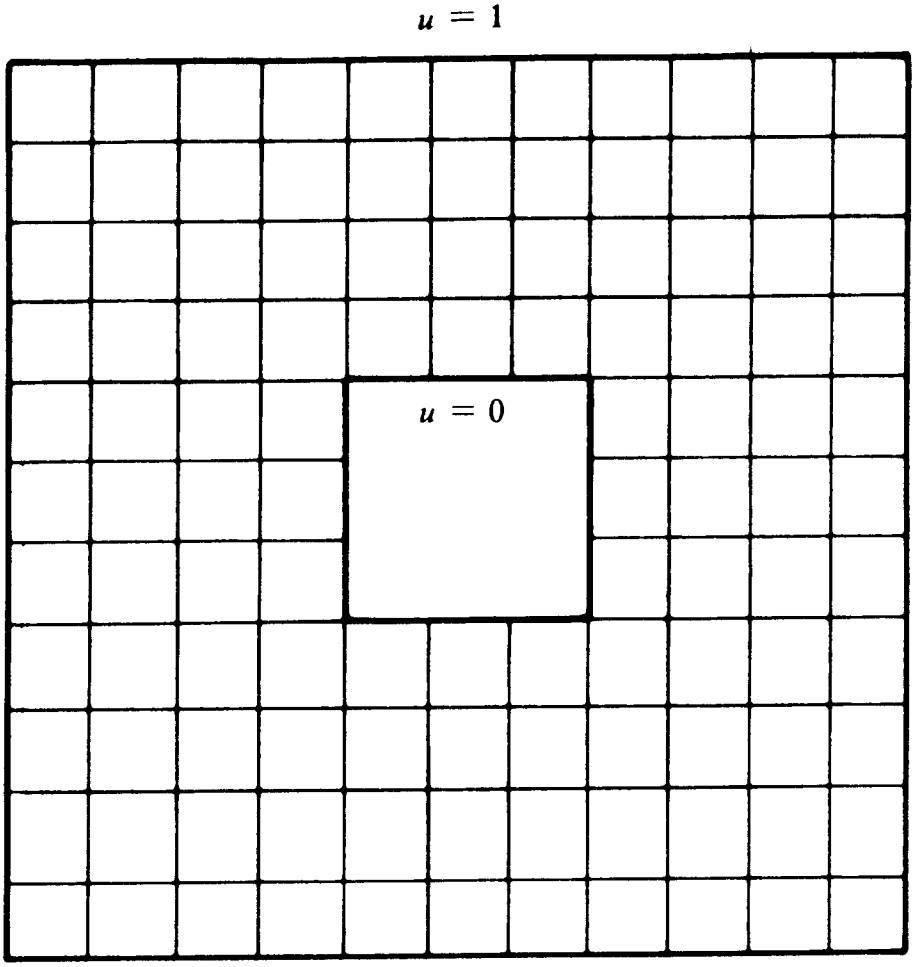


Figure 7: ♣

We start with all interior points 0 and keep the boundary points fixed.

$$\begin{array}{ccccc} & & 1 & 1 & \\ & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & \end{array}$$

After one iteration we have:

$$\begin{array}{ccccc} & & 1 & & \\ & 1 & .547 & .648 & 1 \\ 1 & .75 & .188 & .047 & 0 \\ & 1 & 0 & 0 & \end{array}$$

Note that we go from left to right moving up each column replacing each value by the average of the four neighboring values. The computations for this first iteration are

$$.75 = (1/4)(1+1+1+0)$$

$$.1875 = (1/4)(.75+0+0+0)$$

$$.5469 = (1/4)(.1875+1+1+0)$$

$$.0469 = (1/4)(.1875+0+0+0)$$

$$.64844 = (1/4)(.0469+.5769+1+1)$$

We have printed the results to three decimal places. We continue the iteration until we obtain the same results to three decimal places. This occurs at iterations 8 and 9. Here's what we get:

$$\begin{array}{ccccc} & & 1 & & \\ & 1 & .823 & .787 & 1 \\ 1 & .876 & .506 & .323 & 0 \\ & 1 & 0 & 0 & \end{array}$$

We see that we obtain the same result to three places after only nine iterations and this took only a fraction of a second of computing time. We shall see that these results are correct to three place accuracy. Our Monte Carlo method took several seconds of computing time and did not even give three place accuracy.

The classical reference for the method of relaxations as a means of finding approximate solutions to continuous problems is Courant, Friedrichs, and Lewy [3]. For more information on the relationship between the original Dirichlet problem and the discrete analog, see Hersh and Griego [10].

Exercise 1.2.4 Apply the method of relaxations to the discrete problem illustrated in Figure 7.

Exercise 1.2.5 Consider the method of relaxations started with an initial guess with the property that the value at each point is $\leq$ the average of the values at the neighbors of this point. Show that the successive values at a point u are monotone increasing with a limit $f(u)$ and that these limits provide a solution to the Dirichlet problem.

1.2.5 Solution by solving linear equations

In this section we will show how to find an exact solution to a two-dimensional Dirichlet problem by solving a system of linear equations. As usual, we will illustrate the method in the case of the example introduced in Section 1.2.1. This example is shown again in Figure 8; the interior points have been labelled a, b, c, d, and e. By our

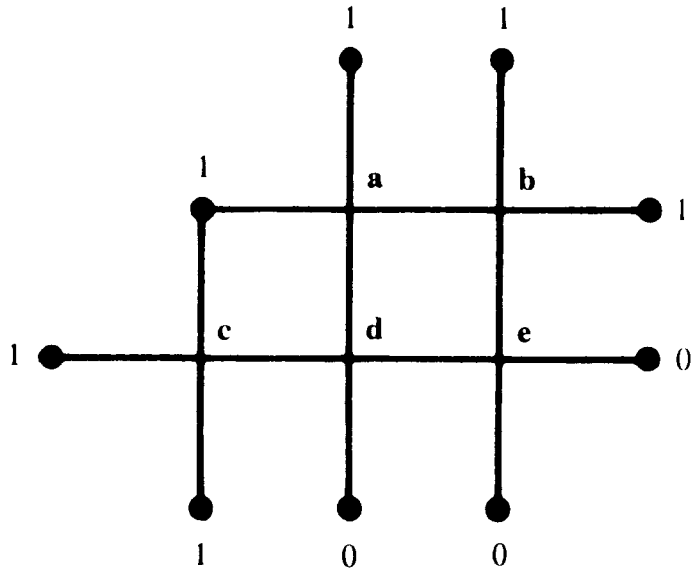


Figure 8: ♣

averaging property, we have

$$x_a = \frac{x_b + x_d + 2}{4}$$

$$x_b = \frac{x_a + x_c + 2}{4}$$
$$x_c = \frac{x_d + 3}{4}$$
$$x_d = \frac{x_a + x_c + x_e}{4}$$
$$x_e = \frac{x_b + x_d}{4}.$$

We can rewrite these equations in matrix form as

$$\begin{pmatrix} 1 & -1/4 & 0 & -1/4 & 0 \\ -1/4 & 1 & 0 & 0 & -1/4 \\ 0 & 0 & 1 & -1/4 & 0 \\ -1/4 & 0 & -1/4 & 1 & -1/4 \\ 0 & -1/4 & 0 & -1/4 & 1 \end{pmatrix} \begin{pmatrix} x_a \\ x_b \\ x_c \\ x_d \\ x_e \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \\ 3/4 \\ 0 \\ 0 \end{pmatrix}.$$

We can write this in symbols as

$$\mathbf{Ax} = \mathbf{u}.$$

Since we know there is a unique solution, $\mathbf{A}$ must have an inverse and

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{u}.$$

Carrying out this calculation we find

$$\text{Calculated } \mathbf{x} = \begin{pmatrix} .823 \\ .787 \\ .876 \\ .506 \\ .323 \end{pmatrix}.$$

Here, for comparison, are the approximate solutions found earlier:

$$\text{Monte Carlo } \mathbf{x} = \begin{pmatrix} .824 \\ .785 \\ .876 \\ .503 \\ .317 \end{pmatrix}.$$

$$\text{Relaxed } \mathbf{x} = \begin{pmatrix} .823 \\ .787 \\ .876 \\ .506 \\ .323 \end{pmatrix}.$$

We see that our Monte Carlo approximations were fairly good in that no error of the simulation is greater than .01, and our relaxed approximations were very good indeed, in that the error does not show up at all.

Exercise 1.2.6 Consider a random walker on the graph of Figure 9. Find the probability of reaching the point with a 1 before any of the

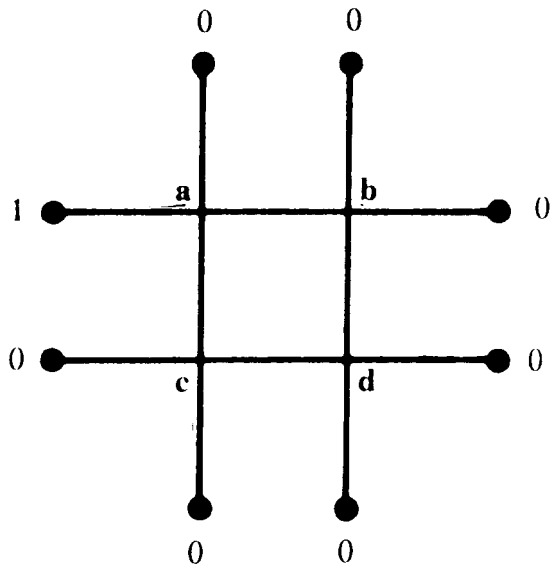


Figure 9: ♣

points with 0's for each starting point a, b, c, d.

Exercise 1.2.7 Solve the discrete Dirichlet problem for the graph shown in Figure 10. The interior points are a, b, c, d. (Hint: See Exercise 1.2.2.)

Exercise 1.2.8 Find the exact value, for each possible starting point, for the game described in Exercise 1.2.3. Is the game favorable starting in the center?

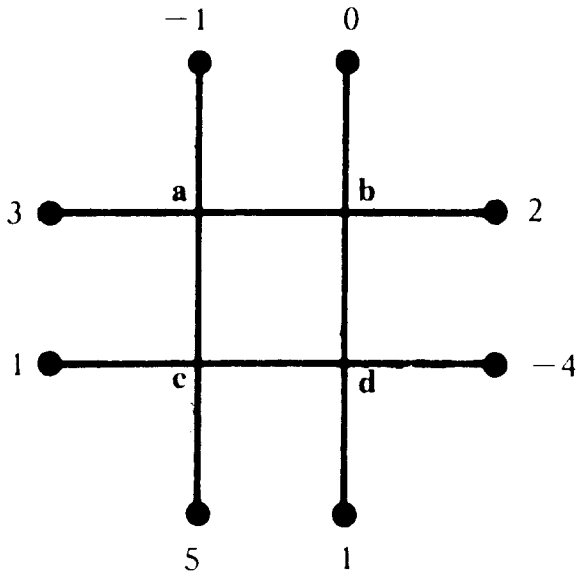


Figure 10: ♣

1.2.6 Solution by the method of Markov chains

In this section, we describe how the Dirichlet problem can be solved by the method of Markov chains. This method may be viewed as a more sophisticated version of the method of linear equations.

A *finite Markov chain* is a special type of chance process that may be described informally as follows: we have a set $S = \{s_1, s_2, \ldots, s_r\}$ of *states* and a chance process that moves around through these states. When the process is in state s_i, it moves with probability P_{ij} to the state s_j. The transition probabilities P_{ij} are represented by an r-by-r matrix $\mathbf{P}$ called the *transition matrix*. To specify the chance process completely we must give, in addition to the transition matrix, a method for starting the process. We do this by specifying a specific state in which the process starts.

According to Kemeny, Snell, and Thompson [15], in the Land of Oz, there are three kinds of weather: rain, nice, and snow. There are never two nice days in a row. When it rains or snows, half the time it is the same the next day. If the weather changes, the chances are equal for a change to each of the other two types of weather. We regard the

weather in the Land of Oz as a Markov chain with transition matrix:

$$\mathbf{P} = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix} \end{array}.$$

When we start in a particular state, it is natural to ask for the probability that the process is in each of the possible states after a specific number of steps. In the study of Markov chains, it is shown that this information is provided by the powers of the transition matrix. Specifically, if $\mathbf{P}^n$ is the matrix $\mathbf{P}$ raised to the nth power, the entries P_{ij}^n represent the probability that the chain, started in state s_i, will, after n steps, be in state s_j. For example, the fourth power of the transition matrix $\mathbf{P}$ for the weather in the Land of Oz is

$$\mathbf{P}^4 = \begin{array}{c} \\ \mathrm{R} \\ \mathrm{N} \\ \mathrm{S} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{R} & \mathrm{N} & \mathrm{S} \end{array} \\ \begin{pmatrix} .402 & .199 & .398 \\ .398 & .203 & .398 \\ .398 & .199 & .402 \end{pmatrix} \end{array}.$$

Thus, if it is raining today in the Land of Oz, the probability that the weather will be nice four days from now is .199. Note that the probability of a particular type of weather four days from today is essentially independent of the type of weather today. This Markov chain is an example of a type of chain called a regular chain. A Markov chain is a *regular* chain if some power of the transition matrix has no zeros. In the study of regular Markov chains, it is shown that the probability of being in a state after a large number of steps is independent of the starting state.

As a second example, we consider a random walk in one dimension. Let us assume that the walk is stopped when it reaches either state 0 or 4. (We could use 5 instead of 4, as before, but we want to keep the matrices small.) We can regard this random walk as a Markov chain with states 0, 1, 2, 3, 4 and transition matrix given by

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

The states 0 and 4 are *traps* or *absorbing states*. These are states that, once entered, cannot be left. A Markov chain is called *absorbing* if it has at least one absorbing state and if, from any state, it is possible (not necessarily in one step) to reach at least one absorbing state. Our Markov chain has this property and so is an absorbing Markov chain. The states of an absorbing chain that are not traps are called *non-absorbing*.

When an absorbing Markov chain is started in a non-absorbing state, it will eventually end up in an absorbing state. For non-absorbing state s_i and absorbing state s_j, we denote by B_{ij} the probability that the chain starting in s_i will end up in state s_j. We denote by $\mathbf{B}$ the matrix with entries B_{ij}. This matrix will have as many rows as non-absorbing states and as many columns as there are absorbing states. For our random walk example, the entries $B_{x,4}$ will give the probability that our random walker, starting at x, will reach 4 before reaching 0. Thus, if we can find the matrix $\mathbf{B}$ by Markov chain techniques, we will have a way to solve the Dirichlet problem.

We shall show, in fact, that the Dirichlet problem has a natural generalization in the context of absorbing Markov chains and can be solved by Markov chain methods.

Assume now that $\mathbf{P}$ is an absorbing Markov chain and that there are u absorbing states and v non-absorbing states. We reorder the states so that the absorbing states come first and the non-absorbing states come last. Then our transition matrix has the canonical form:

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}.$$

Here $\mathbf{I}$ is a u-by-u identity matrix; $\mathbf{0}$ is a matrix of dimension u-by-v with all entries 0.

For our random walk example this canonical form is:

$$\begin{array}{c} \\ 0 \\ 4 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 4 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \end{pmatrix} \end{array}.$$

The matrix $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$ is called the *fundamental matrix* for the absorbing chain $\mathbf{P}$. The entries N_{ij} of this matrix have the following probabilistic interpretation: N_{ij} is the expected number of times that

the chain will be in state s_j before absorption when it is started in s_i. (To see why this is true, think of how $(\mathbf{I}-\mathbf{Q})^{-1}$ would look if it were written as a geometric series.) Let $\mathbf{1}$ be a column vector of all 1's. Then the vector $\mathbf{t} = \mathbf{NI}$ gives the expected number of steps before absorption for each starting state.

The absorption probabilities $\mathbf{B}$ are obtained from $\mathbf{N}$ by the matrix formula

$$\mathbf{B} = (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{R}.$$

This simply says that to get the probability of ending up at a given absorbing state, we add up the probabilities of going there from all the non-absorbing states, weighted by the number of times we expect to be in those (non-absorbing) states.

For our random walk example

$$\mathbf{Q} = \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{pmatrix}$$

$$\mathbf{I}-\mathbf{Q} = \begin{pmatrix} 1 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ 0 & -\frac{1}{2} & 1 \end{pmatrix}$$

$$\mathbf{N} = (\mathbf{I}-\mathbf{Q})^{-1} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \end{array}$$

$$\mathbf{t} = \mathbf{N1} = \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 3 \end{pmatrix}$$

$$\mathbf{B} = \mathbf{NR} = \begin{pmatrix} \frac{3}{2} & 1 & \frac{1}{2} \\ 1 & 2 & 1 \\ \frac{1}{2} & 1 & \frac{3}{2} \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 4 \end{array} \\ \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}. \end{array}$$

Thus, starting in state 3, the probability is 3/4 of reaching 4 before 0; this is in agreement with our previous results. From $\mathbf{t}$ we see that the expected duration of the game, when we start in state 2, is 4.

For an absorbing chain $\mathbf{P}$, the nth power $\mathbf{P}^n$ of the transition probabilities will approach a matrix $\mathbf{P}^\infty$ of the form

$$\mathbf{P}^\infty = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{B} & \mathbf{Q} \end{pmatrix}.$$

We now give our Markov chain version of the Dirichlet problem. We interpret the absorbing states as boundary states and the non-absorbing states as interior states. Let B be the set of boundary states and D the set of interior states. Let f be a function with domain the state space of a Markov chain $\mathbf{P}$ such that for i in D

$$f(i) = \sum_j P_{ij} f(j).$$

Then f is a *harmonic function* for $\mathbf{P}$. Now f again has an averaging property and extends our previous definition. If we represent f as a column vector $\mathbf{f}$, f is harmonic if and only if

$$\mathbf{Pf} = \mathbf{f}.$$

This implies that

$$\mathbf{P}^2\mathbf{f} = \mathbf{P} \cdot \mathbf{Pf} = \mathbf{Pf} = \mathbf{f}$$

and in general

$$\mathbf{P}^n\mathbf{f} = \mathbf{f}.$$

Let us write the vector $\mathbf{f}$ as

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}_B \\ \mathbf{f}_D \end{pmatrix}$$

where $\mathbf{f}_B$ represents the values of f on the boundary and $\mathbf{f}_D$ values on the interior. Then we have

$$\begin{pmatrix} \mathbf{f}_B \\ \mathbf{f}_D \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{B} & \mathbf{Q} \end{pmatrix} \begin{pmatrix} \mathbf{f}_B \\ \mathbf{f}_D \end{pmatrix}$$

and

$$\mathbf{f}_D = \mathbf{B}\mathbf{f}_B.$$

We again see that the values of a harmonic function are determined by the values of the function at the boundary points.

Since the entries B_{ij} of $\mathbf{B}$ represent the probability, starting in i, that the process ends at j, our last equation states that if you play a game in which your fortune is f_j when you are in state j, then your expected final fortune is equal to your initial fortune; that is, fairness is preserved. As remarked above, from Markov chain theory $\mathbf{B} = \mathbf{NR}$ where $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$. Thus

$$\mathbf{f}_D = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{R}\mathbf{f}_B.$$

(To make the correspondence between this solution and the solution of Section 1.2.5, put $\mathbf{A} = \mathbf{I} - \mathbf{Q}$ and $\mathbf{u} = \mathbf{R}\mathbf{f}_B$.)

A general discussion of absorbing Markov chains may be found in Kemeny, Snell, and Thompson [15].

Exercise 1.2.9 Consider the game played on the grid in Figure 11. You start at an interior point and move randomly until a boundary

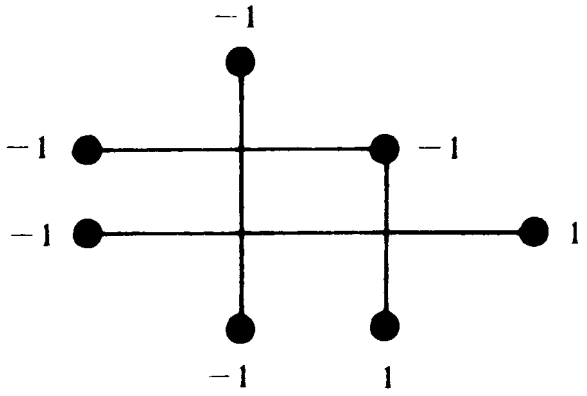


Figure 11: ♣

point is reached and obtain the payment indicated at this point. Using Markov chain methods find, for each starting state, the expected value of the game. Find also the expected duration of the game.

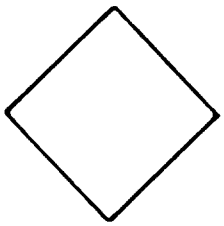
1.3 Random walks on more general networks

1.3.1 General resistor networks and reversible Markov chains

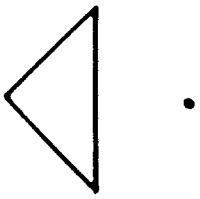
Our networks so far have been very special networks with unit resistors. We will now introduce general resistor networks, and consider what it means to carry out a random walk on such a network.

A *graph* is a finite collection of *points* (also called *vertices* or *nodes*) with certain pairs of points connected by *edges* (also called *branches*). The graph is *connected* if it is possible to go between any two points by moving along the edges. (See Figure 12.)

We assume that G is a connected graph and assign to each edge xy a resistance R_{xy}; an example is shown in Figure 13. The *conductance* of an edge xy is $C_{xy} = 1/R_{xy}$; conductances for our example are shown in Figure 14.

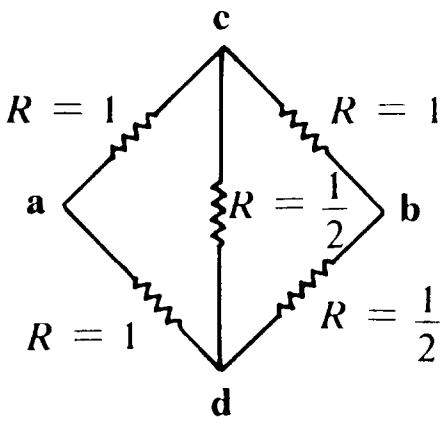


Connected graph



Disconnected graph

Figure 12: ♣



Resistances

Figure 13: ♣

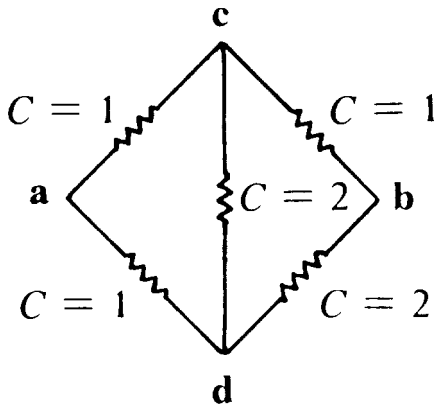


Figure 14: ♣

We define a *random walk* on G to be a Markov chain with transition matrix $\mathbf{P}$ given by

$$P_{xy} = \frac{C_{xy}}{C_x}$$

with $C_x = \sum_y C_{xy}$. For our example, $C_a = 2$, $C_b = 3$, $C_c = 4$, and $C_d = 5$, and the transition matrix $\mathbf{P}$ for the associated random walk is

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} & 0 \end{pmatrix} \end{array}$$

Its graphical representation is shown in Figure 15.

Since the graph is connected, it is possible for the walker to go between any two states. A Markov chain with this property is called an *ergodic* Markov chain. Regular chains, which were introduced in Section 1.2.6, are always ergodic, but ergodic chains are not always regular (see Exercise 1.3.1).

For an ergodic chain, there is a unique probability vector $\mathbf{w}$ that is a fixed vector for $\mathbf{P}$, i.e., $\mathbf{wP} = \mathbf{w}$. The component w_j of $\mathbf{w}$ represents the proportion of times, in the long run, that the walker will be in state j. For random walks determined by electric networks, the fixed vector is given by $w_j = C_j/C$, where $C = \sum_x C_x$. (You are asked to prove this

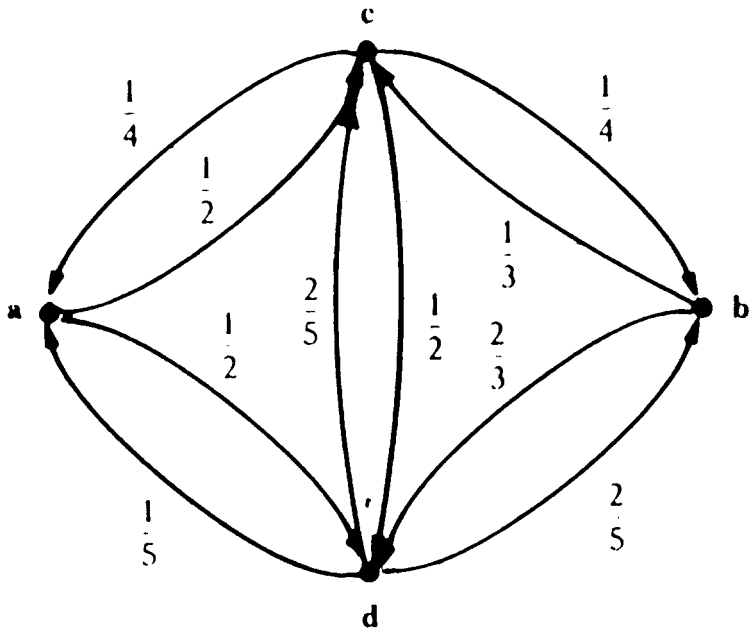


Figure 15: ♣

in Exercise 1.3.2.) For our example $C_a = 2$, $C_b = 3$, $C_c = 4$, $C_d = 5$, and $C = 14$. Thus $\mathbf{w} = (2/14, 3/14, 4/14, 5/14)$. We can check that $\mathbf{w}$ is a fixed vector by noting that

$$\left(\begin{array}{cccc} \frac{2}{14} & \frac{3}{14} & \frac{4}{14} & \frac{5}{14} \end{array}\right) \begin{pmatrix} 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} & 0 \end{pmatrix} = \left(\begin{array}{cccc} \frac{2}{14} & \frac{3}{14} & \frac{4}{14} & \frac{5}{14} \end{array}\right).$$

In addition to being ergodic, Markov chains associated with networks have another property called *reversibility*. An ergodic chain is said to be *reversible* if $w_x P_{xy} = w_y P_{yx}$ for all x, y. That this is true for our network chains follows from the fact that

$$C_x P_{xy} = C_x \frac{C_{xy}}{C_x} = C_{xy} = C_{yx} = C_y \frac{C_{yx}}{C_y} = C_y P_{yx}.$$

Thus, dividing the first and last term by C, we have $w_x P_{xy} = w_y P_{yx}$.

To see the meaning of reversibility, we start our Markov chain with initial probabilities $\mathbf{w}$ (in equilibrium) and observe a few states, for example

$$a \;\; c \;\; b \;\; d.$$

The probability that this sequence occurs is

$$w_a P_{ac} P_{cb} P_{bd} = \frac{2}{14} \cdot \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{2}{3} = \frac{1}{84}.$$

The probability that the reversed sequence

$$d \;\; b \;\; c \;\; a$$

occurs is

$$w_d P_{db} P_{bc} P_{ca} = \frac{5}{14} \cdot \frac{2}{5} \cdot \frac{1}{3} \cdot \frac{1}{4} = \frac{1}{84}.$$

Thus the two sequences have the same probability of occurring.

In general, when a reversible Markov chain is started in equilibrium, probabilities for sequences in the correct order of time are the same as those with time reversed. Thus, from data, we would never be able to tell the direction of time.

If $\mathbf{P}$ is any reversible ergodic chain, then $\mathbf{P}$ is the transition matrix for a random walk on an electric network; we have only to define $C_{xy} = w_x P_{xy}$. Note, however, if $P_{xx} \neq 0$ the resulting network will need a conductance from x to x (see Exercise 1.3.4). Thus reversibility characterizes those ergodic chains that arise from electrical networks. This has to do with the fact that the physical laws that govern the behavior of steady electric currents are invariant under time-reversal (see Onsager [25]).

When all the conductances of a network are equal, the associated random walk on the graph G of the network has the property that, from each point, there is an equal probability of moving to each of the points connected to this point by an edge. We shall refer to this random walk as *simple random walk* on G. Most of the examples we have considered so far are simple random walks. Our first example of a random walk on Madison Avenue corresponds to simple random walk on the graph with points $0, 1, 2, \ldots, N$ and edges the streets connecting these points. Our walks on two dimensional graphs were also simple random walks.

Exercise 1.3.1 Give an example of an ergodic Markov chain that is not regular. (Hint: a chain with two states will do.)

Exercise 1.3.2 Show that, if $\mathbf{P}$ is the transition matrix for a random walk determined by an electric network, then the fixed vector $\mathbf{w}$ is given by $w_x = \frac{C_x}{C}$ where $C_x = \sum_y C_{xy}$ and $C = \sum_x C_x$.

Exercise 1.3.3 Show that, if $\mathbf{P}$ is a reversible Markov chain and a, b, c are any three states, then the probability, starting at a, of the cycle $abca$ is the same as the probability of the reversed cycle $acba$. That is $P_{ab}P_{bc}P_{ca} = P_{ac}P_{cb}P_{ba}$. Show, more generally, that the probability of going around any cycle in the two different directions is the same. (Conversely, if this cyclic condition is satisfied, the process is reversible. For a proof, see Kelly [13].)

Exercise 1.3.4 Assume that $\mathbf{P}$ is a reversible Markov chain with $P_{xx} = 0$ for all x. Define an electric network by $C_{xy} = w_x P_{xy}$. Show that the Markov chain associated with this circuit is $\mathbf{P}$. Show that we can allow $P_{xx} > 0$ by allowing a conductance from x to x.

Exercise 1.3.5 For the *Ehrenfest urn model*, there are two urns that together contain N balls. Each second, one of the N balls is chosen at random and moved to the other urn. We form a Markov chain with states the number of balls in one of the urns. For $N = 4$, the resulting transition matrix is

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}.$$

Show that the fixed vector $\mathbf{w}$ is the binomial distribution $\mathbf{w} = (\frac{1}{16}, \frac{4}{16}, \frac{6}{16}, \frac{4}{16}, \frac{1}{16})$. Determine the electric network associated with this chain.

1.3.2 Voltages for general networks; probabilistic interpretation

We assume that we have a network of resistors assigned to the edges of a connected graph. We choose two points a and b and put a one-volt battery across these points establishing a voltage $v_a = 1$ and $v_b = 0$, as illustrated in Figure 16. We are interested in finding the voltages v_x and the currents i_{xy} in the circuit and in giving a probabilistic interpretation to these quantities.

We begin with the probabilistic interpretation of voltage. It will come as no surprise that we will interpret the voltage as a hitting probability, observing that both functions are harmonic and that they have the same boundary values.

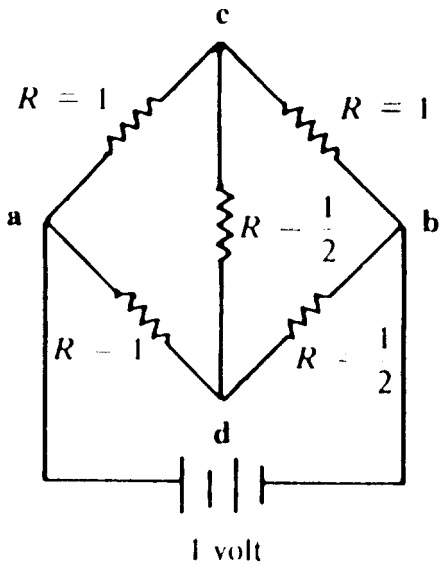


Figure 16: ♣

By Ohm's Law, the currents through the resistors are determined by the voltages by

$$i_{xy} = \frac{v_x - v_y}{R_{xy}} = (V_x - v_y)C_{xy}.$$

Note that $i_{xy} = -i_{yx}$. Kirchhoff's Current Law requires that the total current flowing into any point other than a or b is 0. That is, for $x \neq a, b$

$$\sum_y i_{xy} = 0.$$

This will be true if

$$\sum_y (v_x - v_y)C_{xy} = 0$$

or

$$v_x \sum_y C_{xy} = \sum_y C_{xy} v_y.$$

Thus Kirchhoff's Current Law requires that our voltages have the property that

$$v_x = \sum_y \frac{C_{xy}}{C_x} v_y = \sum_y P_{xy} v_y$$

for $x \neq a, b$. This means that the voltage v_x is harmonic at all points $x \neq a, b$.

Let h_x be the probability, starting at x, that state a is reached before b. Then h_x is also harmonic at all points $x \neq a, b$. Furthermore

$$v_a = h_a = 1$$

and

$$v_b = h_b = 0.$$

Thus if we modify $\mathbf{P}$ by making a and b absorbing states, we obtain an absorbing Markov chain $\bar{\mathbf{P}}$ and v and h are both solutions to the Dirichlet problem for the Markov chain with the same boundary values. Hence $v = h$.

For our example, the transition probabilities $\bar{P}_{xy}$ are shown in Figure 17. The function v_x is harmonic for $\bar{P}$ with boundary values $v_a =$

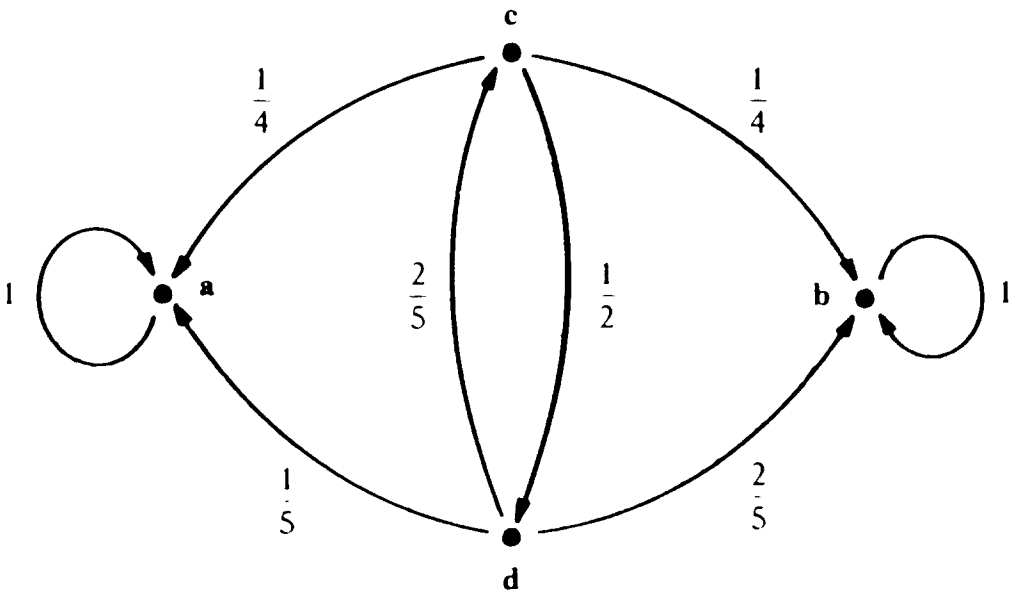


Figure 17: ♣

$1, v_b = 0$.

To sum up, we have the following:

Intrepretation of Voltage. When a unit voltage is applied between a and b, making $v_a = 1$ and $v_b = 0$, the voltage v_x at any point x

represents the probability that a walker starting from x will return to a before reaching b.

In this probabilistic interpretation of voltage, we have assumed a unit voltage, but we could have assumed an arbitrary voltage v_a between a and b. Then the hitting probability h_x would be replaced by an expected value in a game where the player starts at x and is paid v_a if a is reached before b and 0 otherwise.

Let's use this interpretation of voltage to find the voltages for our example. Referring back to Figure 17, we see that

$$v_a = 1$$

$$v_b = 0$$

$$v_c = \frac{1}{4} + \frac{1}{2}v_d$$

$$v_d = \frac{1}{5} + \frac{2}{5}v_c.$$

Solving these equations yields $v_c = \frac{7}{16}$ and $v_d = \frac{3}{8}$. From these voltages we obtain the current i_{xy}. For example $i_{cd} = (\frac{7}{16} - \frac{3}{8}) \cdot 2 = \frac{1}{8}$. The resulting voltages and currents are shown in Figure 18. The voltage at c is $\frac{7}{16}$ and so this is also the probability, starting at c, of reaching a before b.

1.3.3 Probabilistic interpretation of current

We turn now to the probabilistic interpretation of current. This interpretation is found by taking a naive view of the process of electrical conduction: We imagine that positively charged particles enter the network at point a and wander around from point to point until they finally arrive at point b, where they leave the network. (It would be more realistic to imagine negatively charged particles entering at b and leaving at a, but realism is not what we're after.) To determine the current i_{xy} along the branch from x to y, we consider that in the course of its peregrinations the point may pass once or several times along the branch from x to y, and in the opposite direction from y to x. We may now hypothesize that the current i_{xy} is proportional to the expected net number of movements along the edge from x to y, where movements from y back to x are counted as negative. This hypothesis is correct, as we will now show.

The walker begins at a and walks until he reaches b; note that if he returns to a before reaching b, he keeps on going. Let u_x be the

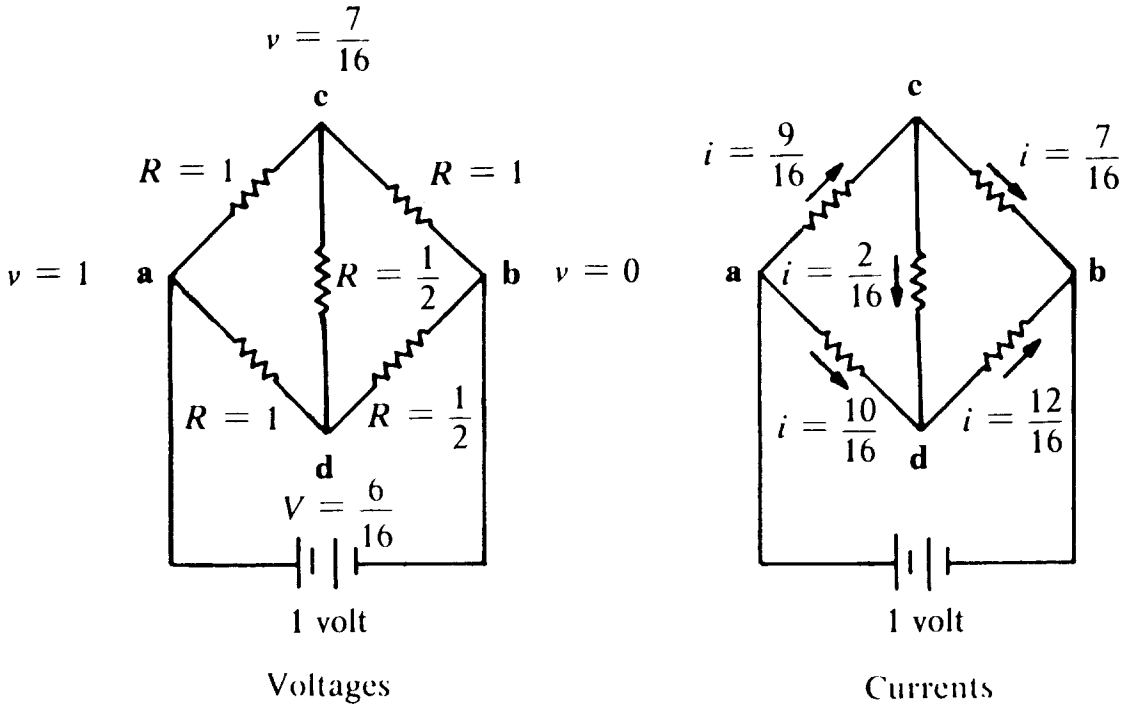

$v = \frac{7}{16}$
c
$R = 1$
$R = 1$
$v = 1$
a
$R = \frac{1}{2}$
b
$v = 0$
$R = 1$
$R = \frac{1}{2}$
d
$V = \frac{6}{16}$
1 volt
Voltages
c
$i = \frac{9}{16}$
$i = \frac{7}{16}$
a
$i = \frac{2}{16}$
b
$i = \frac{10}{16}$
$i = \frac{12}{16}$
d
1 volt
Currents

Figure 18: ♣

expected number of visits to state x before reaching b. Then $u_b = 0$ and, for $x \neq a, b$,

$$u_x = \sum_y u_y P_{yx}.$$

This last equation is true because, for $x \neq a, b$, every entrance to x must come from some y.

We have seen that $C_x P_{xy} = C_y P_{yx}$; thus

$$u_x = \sum_y u_y \frac{P_{xy} C_x}{C_y}$$

or

$$\frac{u_x}{C_x} = \sum_y P_{xy} \frac{u_y}{C_y}.$$

This means that $v_x = u_x/C_x$ is harmonic for $x \neq a, b$. We have also $v_b = 0$ and $v_a = u_a/C_a$. This implies that v_x is the voltage at x when we put a battery from a to b that establishes a voltage u_a/C_a at a and voltage 0 at b. (We remark that the expression $v_x = u_x/C_x$ may be understood physically by viewing u_x as charge and C_x as capacitance; see Kelly [13] for more about this.)

We are interested in the current that flows from x to y. This is

$$i_{xy} = (v_x - v_y) C_{xy} = \left(\frac{u_x}{C_x} - \frac{u_y}{C_y} \right) C_{xy} = \frac{u_x C_{xy}}{C_x} - \frac{u_y C_{yx}}{C_y} = u_x P_{xy} - u_y P_{yx}.$$

Now $u_x P_{xy}$ is the expected number of times our walker will go from x to y and $u_y P_{yx}$ is the expected number of times he will go from y to x. Thus the current i_{xy} is the expected value for the net number of times the walker passes along the edge from x to y. Note that for any particular walk this net value will be an integer, but the expected value will not.

As we have already noted, the currents i_{xy} here are not those of our original electrical problem, where we apply a 1-volt battery, but they are proportional to those original currents. To determine the constant of proportionality, we note the following characteristic property of the new currents i_{xy}: The total current flowing into the network at a (and out at b) is 1. In symbols,

$$\sum_y i_{ay} = 1.$$

Indeed, from our probabilistic interpretation of i_{xy} this sum represents the expected value of the difference between the number of times our

walker leaves a and enters a. This number is necessarily one and so the current flowing into a is 1.

This unit current flow from a to b can be obtained from the currents in the original circuit, corresponding to a 1-volt battery, by dividing through by the total amount of current $\sum_y i_{ay}$ flowing into a; doing this to the currents in our example yields the unit current flow shown in Figure 19.

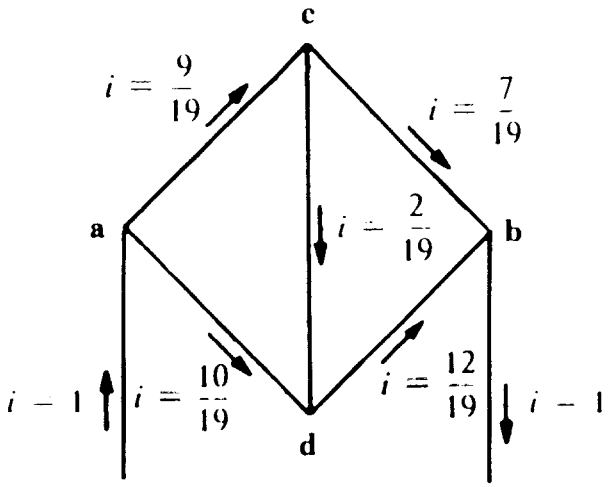


Figure 19: ♣

This shows that the constant of proportionality we were seeking to determine is the reciprocal of the amount of current that flows through the circuit when a I-volt battery is applied between a and b. This quantity, called the effective resistance between a and b, is discussed in detail in Section 1.3.4.

To sum up, we have the following:

Interpretation of Current. When a unit current flows into a and out of b, the current i_{xy} flowing through the branch connecting x to y is equal to the expected net number of times that a walker, starting at a and walking until he reaches b, will move along the branch from x to y. These currents are proportional to the currents that arise when a unit voltage is applied between a and b, the constant of proportionality being the effective resistance of the network.

We have seen that we can estimate the voltages by simulation. We can now do the same for the currents. We have to estimate the expected value for the net number of crossings of xy. To do this, we start a large number of walks at a and, for each one, record the net number of crossings of each edge and then average these as an estimate for the expected value. Carrying out 10,000 such walks yielded the results shown in Figure 20.

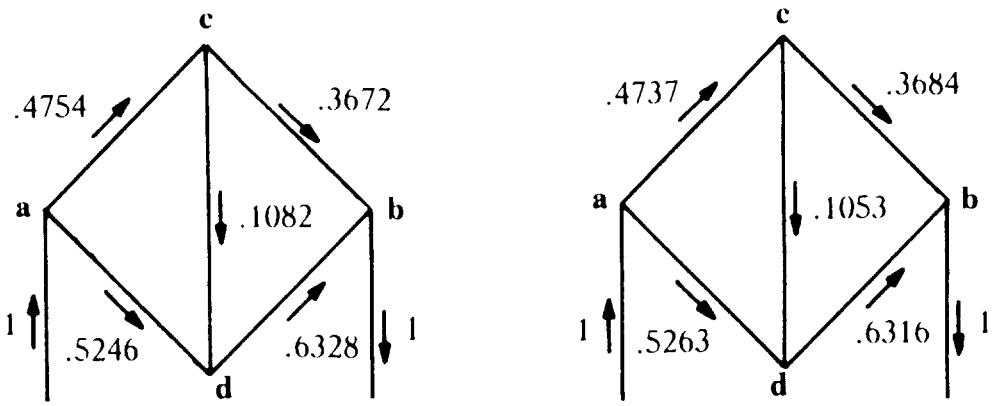


Figure 20: ♣

The results of simulation are in good agreement with the theoretical values of current. As was the case for estimates of the voltages by simulation, we have statistical errors. Our estimates have the property that the total current flowing into a is 1, out of b is 1, and into any other point it is 0. This is no accident; the explanation is that the history of each walk would have these properties, and these properties are not destroyed by averaging.

Exercise 1.3.6 Kingman [17] introduced a different model for current flow. Kelly [13] gave a new interpretation of this model. Both authors use continuous time. A discrete time version of Kelly's interpretation would be the following: At each point of the graph there is a black or a white button. Each second an edge is chosen; edge xy is chosen with probability C_{xy}/C where C is the sum of the conductances. The

buttons on the edge chosen are then interchanged. When a button reaches a it is painted black, and when a button reaches b it is painted white. Show that there is a limiting probability p_x that site x has a black button and that p_x is the voltage v_x at x when a unit voltage is imposed between a and b. Show that the current i_{xy} is proportional to the net flow of black buttons along the edge xy. Does this suggest a hypothesis about the behavior of conduction electrons in metals?

1.3.4 Effective resistance and the escape probability

When we impose a voltage v between points a and b, a voltage $v_a = v$ is established at a and $v_b = 0$, and a current $i_a = \sum_x i_{ax}$ will flow into the circuit from the outside source. The amount of current that flows depends upon the overall resistance in the circuit. We define the *effective resistance* R_{eff} between a and b by $R_{\text{eff}} = v_a/i_a$. The reciprocal quantity $C_{\text{eff}} = 1/R_{\text{eff}} = i_a/v_a$ is the *effective conductance*. If the voltage between a and b is multiplied by a constant, then the currents are multiplied by the same constant, so R_{eff} depends only on the ratio of v_a to i_a.

Let us calculate R_{eff} for our example. When a unit voltage was imposed, we obtained the currents shown in Figure 18. The total current flowing into the circuit is $i_a = 9/16+10/16 = 19/16$. Thus the effective resistance is

$$R_{\text{eff}} = \frac{v_a}{i_a} = \frac{1}{\frac{19}{16}} = \frac{16}{19}.$$

We can interpret the effective conductance probabilistically as an escape probability. When $v_a = 1$, the effective conductance equals the total current i_a flowing into a. This current is

$$i_a = \sum_y (v_a - v_y)C_{ay} = \sum_y (v_a - v_y)\frac{C_{ay}}{C_a}C_a = C_a(1 - \sum_y P_{ay}v_y) = C_a p_{\text{esc}}$$

where p_{esc} is the probability, starting at a, that the walk reaches b before returning to a. Thus

$$C_{\text{eff}} = C_a p_{\text{esc}}$$

and

$$p_{\text{esc}} = \frac{C_{\text{eff}}}{C_a}.$$

In our example $C_a = 2$ and we found that $i_a = 19/16$. Thus

$$p_{\text{esc}} = \frac{19}{32}.$$

In calculating effective resistances, we shall use two important facts about electric networks. First, if two resistors are connected in series, they may be replaced by a single resistor whose resistance is the sum of the two resistances. (See Figure 21.) Secondly, two resistors in parallel

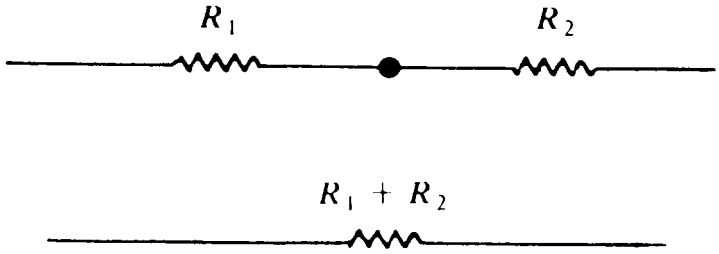


Figure 21: ♣

may be replaced by a single resistor with resistance R such that

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} = \frac{R_1 R_2}{R_1 + R_2}.$$

(See Figure 22.)

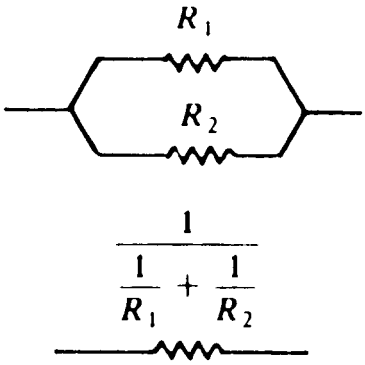


Figure 22: ♣

The second rule can be stated more simply in terms of conductances: If two resistors are connected in parallel, they may be replaced by a single resistor whose conductance is the sum of the two conductances.

We illustrate the use of these ideas to compute the effective resistance between two adjacent points of a unit cube of unit resistors, as shown in Figure 23. We put a unit battery between a and b. Then,

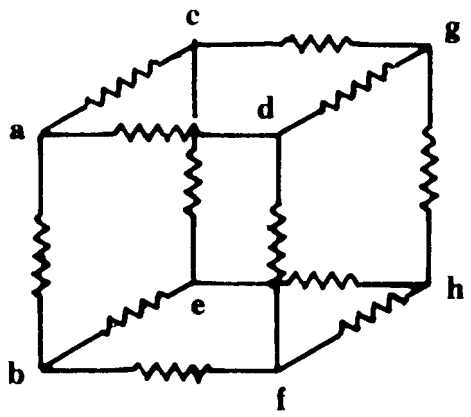


Figure 23: ♣

by symmetry, the voltages at c and d will be the same as will those at e and f. Thus our circuit is equivalent to the circuit shown in Figure 24.

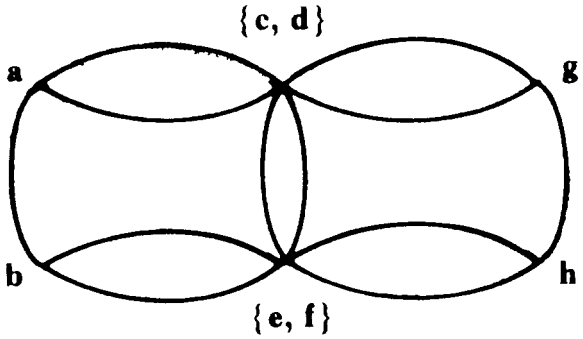


Figure 24: ♣

Using the laws for the effective resistance of resistors in series and parallel, this network can be successively reduced to a single resistor of resistance 7/12 ohms, as shown in Figure 25. Thus the effective resistance is 7/12. The current flowing into a from the battery will be

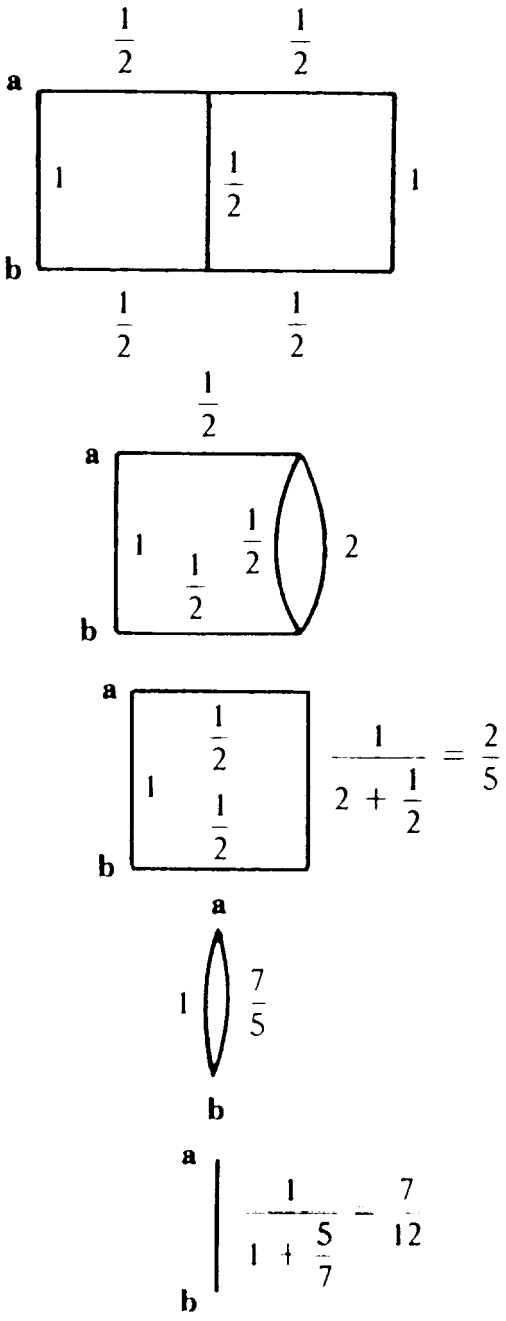

a
b
1
2
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
1
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2}$
1
$\frac{1}{2}$
2
$\frac{1}{2}$
$\frac{1}{2}$
1
$\frac{1}{2}$
$\frac{1}{2}$
$\frac{1}{2 + \frac{1}{2}} = \frac{2}{5}$
1
$\frac{7}{5}$
$\frac{1}{1 + \frac{5}{7}} - \frac{7}{12}$

Figure 25: ♣

$i_a = \frac{1}{R_{\text{eff}}} = 12/7$. The probability that a walk starting at a will reach b before returning to a is

$$p\text{esc} = \frac{i_a}{C_a} = \frac{\frac{12}{7}}{3} = \frac{4}{7}.$$

This example and many other interesting connections between electric networks and graph theory may be found in Bollobas [2].

Exercise 1.3.7 A bug walks randomly on the unit cube (see Figure 26). If the bug starts at a, what is the probability that it reaches food

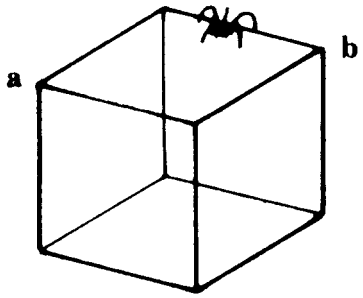


Figure 26: ♣

at b before returning to a?

Exercise 1.3.8 Consider the Ehrenfest urn model with $N = 4$ (see Exercise 1.3.5). Find the probability, starting at 0, that state 4 is reached before returning to 0.

Exercise 1.3.9 Consider the ladder network shown in Figure 27. Show that if R_n is the effective resistance of a ladder with n rungs then $R_1 = 2$ and

$$R_{n+1} = \frac{2 + 2R_n}{2 + R_n}.$$

Use this to show that $\lim_{n\to\infty} R_n = \sqrt{2}$.

Exercise 1.3.10 A drunken tourist starts at her hotel and walks at random through the streets of the idealized Paris shown in Figure 28. Find the probability that she reaches the Arc de Triomphe before she reaches the outskirts of town.

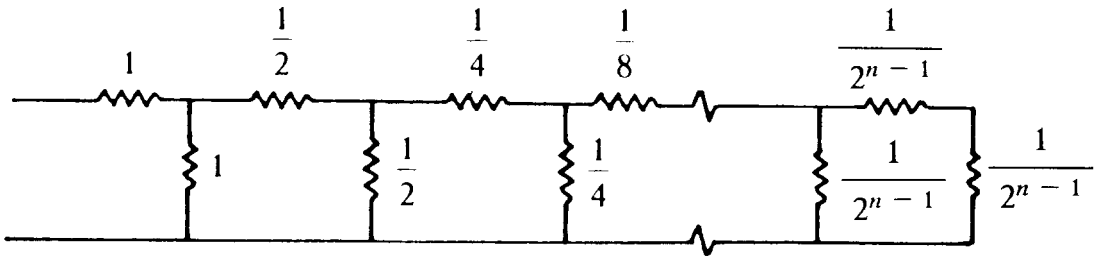


Figure 27: ♣

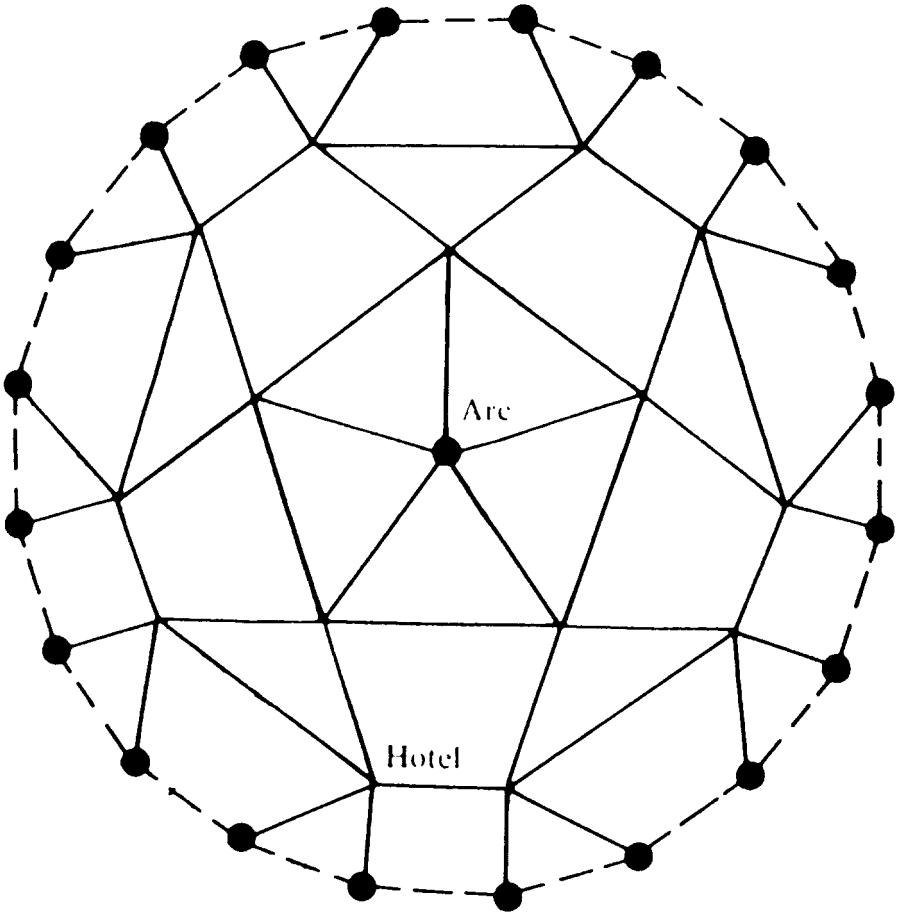


Figure 28: ♣

1.3.5 Currents minimize energy dissipation

We have seen that when we impose a voltage between a and b voltages v_x are established at the points and currents i_{xy} flow through the resistors. In this section we shall give a characterization of the currents in terms of a quantity called *energy dissipation*. When a current i_{xy} flows through a resistor, the energy dissipated is

$$i_{xy}^2 R_{xy};$$

this is the product of the current i_{xy} and the voltage $v_{xy} = i_{xy}R_{xy}$. The *total energy dissipation* in the circuit is

$$E = \frac{1}{2}\sum_{x,y} i_{xy}^2 R_{xy}.$$

Since $i_{xy}R_{xy} = v_x - v_y$, we can also write the energy dissipation as

$$E = \frac{1}{2}\sum_{x,y} i_{xy}(v_x - v_y).$$

The factor $1/2$ is necessary in this formulation since each edge is counted twice in this sum. For our example, we see from Figure 18 that

$$E = \left(\frac{9}{16}\right)^2 \cdot 1 + \left(\frac{10}{16}\right)^2 \cdot 1 + \left(\frac{7}{16}\right)^2 \cdot 1 + \left(\frac{2}{16}\right)^2 \cdot \frac{1}{2} + \left(\frac{12}{16}\right)^2 \cdot \frac{1}{2} = \frac{19}{16}.$$

If a source (battery) establishes voltages v_a and v_b at a and b, then the energy supplied is $(v_a - v_b)i_a$ where $i_a = \sum_x i_{ax}$. By conservation of energy, we would expect this to be equal to the energy dissipated. In our example $v_a - v_b = 1$ and $i_a = \frac{19}{16}$, so this is the case. We shall show that this is true in a somewhat more general setting.

Define a *flow* $\mathbf{j}$ from a to b to be an assignment of numbers j_{xy} to pairs xy such that

(a) $j_{xy} = -j_{yx}$

(b) $\sum_y j_{xy} = 0$ if $x \neq a, b$

(c) $j_{xy} = 0$ if x and y are not adjacent.

We denote by $j_x = \sum_y j_{xy}$ the flow into x from the outside. By (b) $j_x = 0$ for $x \neq a, b$. Of course $j_b = -j_a$. To verify this, note that

$$j_a + j_b = \sum_x j_x = \sum_x \sum_y j_{xy} = \frac{1}{2}\sum_{x,y}(j_{xy} + j_{yx}) = 0,$$

since $j_{xy} = -j_{yx}$.

With this terminology, we can now formulate the following version of the principle of conservation of energy:

Conservation of Energy. Let w be any function defined on the points of the graph and $\mathbf{j}$ a flow from a to b. Then

$$(w_a - w_b)j_a = \frac{1}{2}\sum_{x,y}(w_x - w_y)j_{xy}.$$

Proof.

$$\begin{aligned}\sum_{x,y}(w_x - w_y)j_{xy} &= \sum_x(w_x\sum_y j_{xy}) - \sum_y(w_y\sum_x j_{xy}) \\ &= w_a\sum_y j_{ay} + w_b\sum_y j_{by} - w_a\sum_x j_{xa} - w_b\sum_x j_{xb} \\ &= w_a j_a + w_b j_b - w_a(-j_a) - w_b(-j_b) \\ &= 2(w_a - w_b)j_a.\end{aligned}$$

Thus

$$(w_a - w_b)j_a = \frac{1}{2}\sum_{x,y}(w_x - w_y)j_{xy}$$

as was to be proven. $\diamondsuit$

If we now impose a voltage v_a between a and b with $v_b = 0$, we obtain voltages v_x and currents i_{xy}. The currents $\mathbf{i}$ give a flow from a to b and so by the previous result, we conclude that

$$v_a i_a = \frac{1}{2}\sum_{x,y}(v_x - v_y)i_{xy} = \frac{1}{2}\sum_{x,y} i_{xy}^2 R_{xy}.$$

Recall that $R_{\text{eff}} = v_a/i_a$. Thus in terms of resistances we can write this as

$$i_{xy}^2 R_{\text{eff}} = \frac{1}{2}\sum_{x,y} i_{xy}^2 R_{xy}.$$

If we adjust v_a so that $i_a = 1$, we call the resulting flow *the unit current flow* from a to b. The unit current flow from a to b is a particular example of a *unit flow* from a to b, which we define to be any flow i_{xy} from a to b for which $i_a = -i_b = 1$. The formula above shows that the energy dissipated by the unit current flow is just R_{eff}. According to a basic result called Thomson's Principle, this value is smaller than the energy dissipated by any other unit flow from a to b. Before proving this principle, let us watch it in action in the example worked out above.

Recall that, for this example, we found the true values and some approximate values for the unit current flow; these were shown in Figure 20. The energy dissipation for the true currents is

$$E = R_{\text{eff}} = \frac{16}{19} = .8421053.$$

Our approximate currents also form a unit flow and, for these, the energy dissipation is

$$\bar{E} = (.4754)^2 \cdot 1 + (.5246)^2 \cdot 1 + (.3672)^2 \cdot 1 + (.1082)^2 \cdot \frac{1}{2} + (.6328)^2 \cdot \frac{1}{2} = .8421177.$$

We note that $\bar{E}$ is greater than E, though just barely.

Thomson's Principle. (Thomson [33]). If **i** is the unit flow from a to b determined by Kirchhoff's Laws, then the energy dissipation $\frac{1}{2}\sum_{x,y} i_{xy}^2 R_{xy}$ minimizes the energy dissipation $\frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy}$ among all unit flows **j** from a to b.

Proof. Let **j** be any unit flow from a to b and let $d_{xy} = j_{xy} - i_{xy}$. Then **d** is a flow from a to b with $d_a = \sum_x d_{ax} = 1 - 1 = 0$.

$$\begin{aligned}
\sum_{x,y} j_{xy}^2 R_{xy} &= \sum_{x,y} (i_{xy} + d_{xy})^2 R_{xy} \\
&= \sum_{x,y} i_{xy}^2 R_{xy} + 2\sum_{x,y} i_{xy} R_{xy} d_{xy} + \sum_{x,y} d_{xy}^2 R_{xy} \\
&= \sum_{x,y} i_{xy}^2 R_{xy} + 2\sum_{x,y} (v_x - v_y) d_{xy} + \sum_{x,y} d_{xy}^2 R_{xy}.
\end{aligned}$$

Setting $\mathbf{w} = \mathbf{v}$ and $\mathbf{j} = \mathbf{d}$ in the conservation of energy result above shows that the middle term is $4(v_a - v_b)d_a = 0$. Thus

$$\sum_{x,y} j_{xy}^2 R_{xy} = \sum_{x,y} i_{xy}^2 R_{xy} + \sum_{x,y} d_{xy}^2 R_{xy} \geq \sum_{x,y} i_{xy}^2 R_{xy}.$$

This completes the proof. $\diamondsuit$

Exercise 1.3.11 The following is the so-called "dual form" of Thomson's Principle. Let u be any function on the points of the graph G of a circuit such that $u_a = 1$ and $u_b = 0$. Then the energy dissipation

$$\frac{1}{2}\sum_{x,y} (u_x - u_y)^2 C_{xy}$$

is minimized by the voltages v_x that result when a unit voltage is established between a and b, i.e., $v_a = 1$, $v_b = 0$, and the other voltages are determined by Kirchhoff's Laws. Prove this dual principle. This second principle is known nowadays as the *Dirichlet Principle*, though it too was discovered by Thomson.

Exercise 1.3.12 In Section 1.2.4 we stated that, to solve the Dirichlet problem by the method of relaxations, we could start with an arbitrary initial guess. Show that when we replace the value at a point by the average of the neighboring points the energy dissipation, as expressed in Exercise 1.3.11, can only decrease. Use this to prove that the relaxation method converges to a solution of the Dirichlet problem for an arbitrary initial guess.

1.4 Rayleigh's Monotonicity Law

1.4.1 Rayleigh's Monotonicity Law

Next we will study Rayleigh's Monotonicity Law. This law from electric network theory will be an important tool in our future study of random walks. In this section we will give an example of the use of this law.

Consider a random walk on streets of a city as in Figure 29. Let

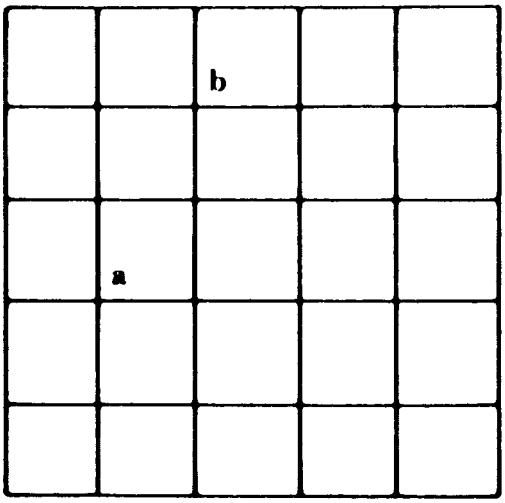


Figure 29: ♣

p_{esc} be the probability that a walker starting from a reaches b before returning to a. Assign to each edge a unit resistance and maintain a voltage of one volt between a and b; then a current i_a will flow into the circuit and we showed in Section 1.3.4 that

$$p_{\text{esc}} = \frac{i_a}{C_a} = \frac{i_a}{4}.$$

Now suppose that one of the streets (not connected to a) becomes blocked. Our walker must choose from the remaining streets if he reaches a corner of this street. The modified graph will be as in Figure 30. We want to show that the probability of escaping to b from a is

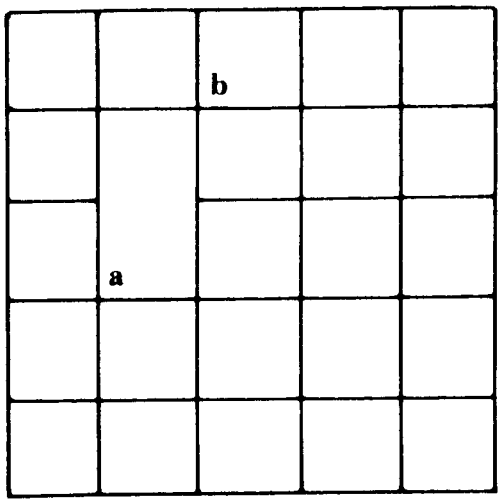


Figure 30: ♣

decreased.

Consider this problem in terms of our network. Blocking a street corresponds to replacing a unit resistor by an infinite resistor. This should have the effect of increasing the effective resistance R_{eff} of the circuit between a and b. If so, when we put a unit voltage between a and b less current will flow into the circuit and

$$p_{\text{esc}} = \frac{i_a}{4} = \frac{1}{4R_{\text{eff}}}$$

will decrease.

Thus we need only show that when we increase the resistance in one part of a circuit, the effective resistance increases. This fact, known as Rayleigh's Monotonicity Law, is almost self-evident. Indeed, the father of electromagnetic theory, James Clerk Maxwell, regarded this to be the case. In his *Treatise on Electricity and Magnetism* ([21], p. 427), he wrote

> If the specific resistance of any portion of the conductor be changed, that of the remainder being unchanged, the

> resistance of the whole conductor will be increased if that of the portion is increased, and diminished if that of the portion is diminished. This principle may be regarded as self-evident ....

Rayleigh's Monotonicity Law. If the resistances of a circuit are increased, the effective resistance R_{eff} between any two points can only increase. If they are decreased, it can only decrease.

Proof. Let **i** be the unit current flow from a to b with the resistors R_{xy}. Let **j** be the unit current flow from a to b with the resistors $\bar{R}_{xy}$ with $\bar{R}_{xy} \geq R_{xy}$. Then

$$\bar{R}_{\text{eff}} = \frac{1}{2}\sum_{x,y} j_{xy}^2 \bar{R}_{xy} \geq \frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy}.$$

But since **j** is a unit flow from a to b, Thomson's Principle tells us that the energy dissipation, calculated with resistors R_{xy}, is bigger than that for the true currents determined by these resistors: that is

$$\frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy} \geq \frac{1}{2}\sum_{x,y} i_{xy}^2 R_{xy} = R_{\text{eff}}.$$

Thus $\bar{R}_{\text{eff}} \geq R_{\text{eff}}$. The proof for the case of decreasing resistances is the same.

Exercise 1.4.1 Consider a graph G and let R_{xy} and $\bar{R}_{xy}$ be two different assignments of resistances to the edges of G. Let $\hat{R}_{xy} = \bar{R}_{xy} + R_{xy}$. Let R_{eff}, $\bar{R}_{\text{eff}}$, and $\hat{R}_{\text{eff}}$ be the effective resistances when R, $\bar{R}$, and $\hat{R}$, respectively, are used. Prove that

$$\hat{R}_{\text{eff}} \geq \bar{R}_{\text{eff}} + R_{\text{eff}}.$$

Conclude that the effective resistance of a network is a concave function of any of its component resistances (Shannon and Hagelbarger [31].)

Exercise 1.4.2 Show that the effective resistance of the circuit in Figure 31 is greater than or equal to the effective resistance of the circuit in Figure 32. Use this to show the following inequality for $R_{ij} \geq 0$:

$$\frac{1}{\frac{1}{R_{11}+R_{12}} + \frac{1}{R_{21}+R_{22}}} \geq \frac{1}{\frac{1}{R_{11}} + \frac{1}{R_{21}}} + \frac{1}{\frac{1}{R_{12}} + \frac{1}{R_{22}}}.$$

See the note by Lehman [18] for a proof of the general Minkowski inequality by this method.

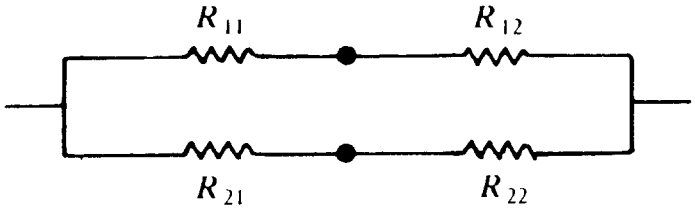


Figure 31: ♣

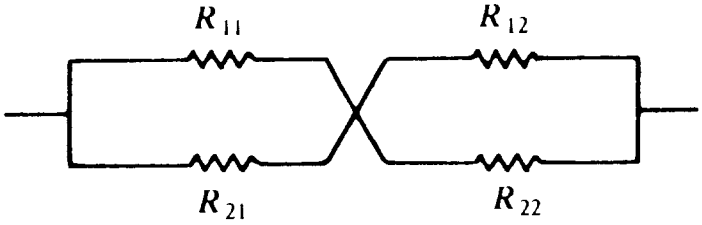


Figure 32: ♣

Exercise 1.4.3 Let $\mathbf{P}$ be the transition matrix associated with an electric network and let a, b, r, s be four points on the network. Let $\bar{\mathbf{P}}$ be a transition matrix defined on the state-space $S = \{a, b, r, s\}$. Let $\bar{P}_{ii} = 0$ and for $i \neq j$ let $\bar{P}_{ij}$ be the probability that, if the chain $\mathbf{P}$ is started in state i, then the next time it is in the set $S - \{i\}$ it is in the state j. Show that $\bar{\mathbf{P}}$ is a reversible Markov chain and hence corresponds to an electric network of the form of a Wheatstone Bridge, shown in Figure 33. Explain how this proves that, in order to prove the Monotonicity Law, it is sufficient to prove that R_{eff} is a monotone function of the component resistances for a Wheatstone Bridge. Give a direct proof of the Monotonicity Law for this special case.

1.4.2 A probabilistic explanation of the Monotonicity Law

We have quoted Maxwell's assertion that Rayleigh's Monotonicity Law may be regarded as self-evident, but one might feel that any argument in terms of electricity is only self-evident if we know what electricity is. In Cambridge, they tell the following story about Maxwell: Maxwell was lecturing and, seeing a student dozing off, awakened him, asking, "Young man, what is electricity?" "I'm terribly sorry, sir," the student

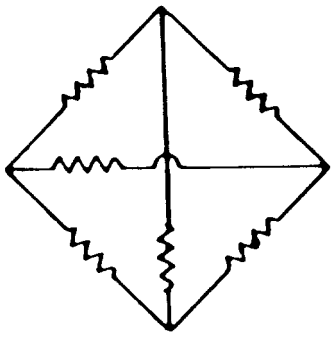


Figure 33: ♣

replied, 'I knew the answer but I have forgotten it." Maxwell's response to the class was, "Gentlemen, you have just witnessed the greatest tragedy in the history of science. The one person who knew what electricity is has forgotten it."

To say that our intuition about the Monotonicity Law is only as solid as our understanding of electricity is not really a valid argument, of course, because in saying that this law is self-evident we are secretly depending on the analogy between electricity and the flow of water (see Feynman [6], Vol. 2, Chapter 12). We just can't believe that if a water main gets clogged the total rate of flow out of the local reservoir is going to increase. But as soon as we admit this, some pedant will ask if we're talking about flows with low Reynolds number, or what, and we'll have to admit that we don't understand water any better than we understand electricity.

Whatever our feelings about electricity or the flow of water, it seems desirable to have an explanation of the Monotonicity Law in terms of our random walker. We now give such an explanation.

We begin by collecting some notation and results from previous sections. As usual, we have a network of conductances (streets) and a walker who moves from point x to point y with probability

$$P_{xy} = \frac{C_{xy}}{C_x}$$

where C_{xy} is the conductance from x to y and $C_x = \sum_y C_{xy}$. We choose two preferred points a and b. The walker starts at a and walks until he reaches b or returns to a. We denote by v_x the probability that the walker, starting at a, reaches a before b. Then $v_a = 1$, $v_b = 0$, and the

function v_x is harmonic at all points $x \neq a, b$. We denote by p_{esc} the probability that the walker, starting at a, reaches b before returning to a. Then

$$p_{\text{esc}} = 1 - \sum_x p_{ax} v_x.$$

Now we have seen that the effective conductance between a and b is

$$C_a p_{\text{esc}}.$$

We wish to show that this increases whenever one of the conductances C_{rs} is increased. If a is different from r or s, we need only show that p_{esc} increases. The case where r or s coincides with a is easily disposed of (see Exercise 1.4.4). The case where r or s coincides with b is also easy (see Exercise 1.4.5). Hence from now on we will assume that $r, s \neq a$ and $r, s \neq b$.

Instead of increasing C_{rs}, we can think of adding a new edge of conductance ϵ between r and s. (See Figure 34.) We will call this

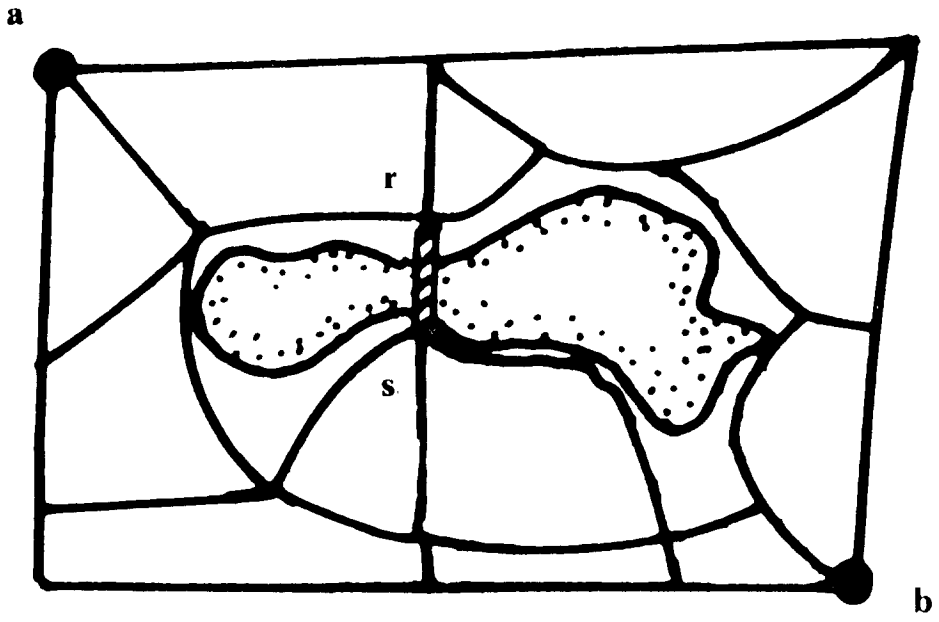


Figure 34: ♣

new edge a "bridge" to distinguish it from the other edges. Note that the graph with the bridge added will have more than one edge between

r and s (unless there was no edge between r and s in the original graph), and this will complicate any expression that involves summing over edges. Everything we have said or will say holds for graphs with "multiple edges" as well as for graphs without them. So far, we have chosen to keep our expressions simple by assuming that an edge is determined by its endpoints. The trade-off is that in the manipulations below, whenever we write a sum over edges we will have to add an extra term to account for the bridge.

Why should adding the bridge increase the escape probability? The first thing you think is, "Of course, it opens up new possibilities of escaping!" The next instant you think, "Wait a minute, it also opens up new possibilities of returning to the starting point. What ensures that the first effect will outweigh the second?" As we shall see, the proper reply is, "Because the walker will cross the bridge more often in the good direction than in the bad direction." To turn this blithe reply into a real explanation will require a little work, however.

To begin to make sense of the notion that the bridge gets used more often in the good direction than the bad, we will make a preliminary argument that applies to any edge of any graph. Let G be any graph, and let rs be any edge with endpoints not a or b. $v_r > v_s$. Since the walker has a better chance to escape from s than from r, this means that to cross this edge in the good direction is to go from r to s We shall show that the walker will cross the edge from r to s more often on the average than from s to r.

Let u_x be the expected number of times the walker is at x and u_{xy} the expected number of times he crosses the edge xy from x to y before he reaches b or returns to a. The calculation carried out in Section 1.3.3 shows that u_x/C_x is harmonic for $x \neq a, b$ with $u_a/C_a = 1/C_a$ and $b_b/C_b = 0$. But the function v_x/C_a also has these properties, so by the Uniqueness Principle

$$\frac{u_x}{C_x} = \frac{v_x}{C_a}.$$

Now

$$u_{rs} = u_r P_{rs} = u_r \frac{C_{rs}}{C_r} = v_r \frac{C_{rs}}{C_a}$$

and

$$u_{sr} = u_s P_{sr} = u_s \frac{C_{sr}}{C_s} = v_s \frac{C_{sr}}{C_a}.$$

Since $C_{rs} = C_{sr}$, and since by assumption $v_r \geq v_s$, this means that $u_{rs} \geq u_{sr}$. Therefore, we see that any edge leads the walker more often to the more favorable of the points of the edge.

Now let's go back and think about the graph with the bridge. The above discussion shows that the bridge helps in the sense that, on the average, the bridge is crossed more often in the direction that improves the chance of escaping. While this fact is suggestive, it doesn't tell us that we are more likely to escape than if the bridge weren't there; it only tells us what goes on once the bridge is in place. What we need is to make a "before and after" comparison.

Recall that we are denoting the conductance of the bridge by ϵ. To distinguish the quantities pertaining to the walks with and without the bridge, we will put (ϵ) superscripts on quantities that refer to the walk with the bridge, so that, e.g., $p_{\text{esc}}^{(\epsilon)}$ denotes the escape probability with the bridge.

Now let $d^{(\epsilon)}$ denote the expected net number of times the walker crosses from r to s. As above, we have

$$d^{(\epsilon)} = u_r^{(\epsilon)} \frac{\epsilon}{C_r + \epsilon} - u_s^{(\epsilon)} \frac{\epsilon}{C_s + \epsilon} = \left(\frac{u_r^{(\epsilon)}}{C_r + \epsilon} - \frac{u_s^{(\epsilon)}}{C_s + \epsilon} \right) \epsilon.$$

Claim.

$$p_{\text{esc}}^{(\epsilon)} = p_{\text{esc}} + (v_r - v_s) d^{(\epsilon)}.$$

Why. Every time you use the bridge to go from r to s, you improve your chances of escaping by

$$(1 - v_s) - (1 - v_r) = v_r - v_s$$

assuming that you would continue your walk without using the bridge. To get the probability of escaping with the bridge, you take the probability of escaping without the bridge, and correct it by adding in the change attributable to the bridge, which is the difference in the original escape probabilities at the ends of the bridge, multiplied by the net number of times you expect to cross the bridge. ♡

Proof. Suppose you're playing a game where you walk around the graph with the bridge, and your fortune when you're at x is v_x, which is the probability that you would return to a before reaching b if the bridge weren't there. You start at a, and walk until you reach b or return to a.

This is not a fair game. Your initial fortune is 1 since you start at a and $v_a = 1$. Your expected final fortune is

$$1 \cdot (1 - p_{\text{esc}}^{(\epsilon)}) + 0 \cdot p_{\text{esc}}^{(\epsilon)} = 1 - p_{\text{esc}}^{(\epsilon)}.$$

The amount you expect to lose by participating in the game is

$$p_{\text{esc}}^{(\epsilon)}.$$

(Note that escaping has suddenly become a bad thing!)

Let's analyze where it is that you expect to lose money. First of all, you lose money when you take the first step away from a. The amount you expect to lose is

$$1 - \sum_x P_{ax}^{(\epsilon)} v_x = p_{\text{esc}}.$$

Now if your fortune were given by $v_x^{(\epsilon)}$ instead of v_x, the game would be fair after this first step. However, the function v_x is not harmonic for the walk with the bridge; it fails to be harmonic at r and s. Every time you step away from r, you expect to lose an amount

$$v_r - \left(\sum_x \frac{C_{rs}}{C_r + \epsilon} v_x + \frac{\epsilon}{C_r + \epsilon} v_s \right) = (v_r - v_s) \frac{\epsilon}{C_r + \epsilon}.$$

Similarly, every time you step away from s you expect to lose an amount

$$(v_s - v_r) \frac{\epsilon}{C_s + \epsilon}.$$

The total amount you expect to lose by participating in the game is:

$$\begin{aligned}
& \text{expected loss at first step} + \\
& (\text{expected loss at } r) \cdot (\text{expected number of times at } r) + \\
& (\text{expected loss at } s) \cdot (\text{expected number of times at } s) \\
= \; & p_{\text{esc}} + \\
& (v_r - v_s) \frac{\epsilon}{C_r + \epsilon} u_r^{(\epsilon)} + \\
& (v_s - v_r) \frac{\epsilon}{C_s + \epsilon} u_s^{(\epsilon)} \\
= \; & p_{\text{esc}} + (v_r - v_s) d^{(\epsilon)}.
\end{aligned}$$

Equating this with our first expression for the expected cost of playing the game yields the formula we were trying to prove.

According to the formula just established,

$$p_{\text{esc}}^{(\epsilon)} - p_{\text{esc}} = (v_r - v_s) \left(\frac{u_r^{(\epsilon)}}{C_r + \epsilon} - \frac{u_s^{(\epsilon)}}{C_s + \epsilon} \right) \epsilon.$$

For small ϵ, we will have

$$\frac{u_r^{(\epsilon)}}{C_r+\epsilon} - \frac{u_s^{(\epsilon)}}{C_s+\epsilon} \approx \frac{u_r}{C_r} - \frac{u_s}{C_s} = \frac{v_r}{C_a} - \frac{v_s}{C_a},$$

so for small ϵ

$$p_{\rm esc}^{(\epsilon)} - p_{\rm esc} \approx (v_r - v_s)^2 \frac{\epsilon}{C_a}.$$

This approximation allows us to conclude that

$$p_{\rm esc}^{(\epsilon)} \geq p_{\rm esc} \geq 0$$

for small ϵ. But this is enough to establish the monotonicity law, since any finite change in ϵ can be realized by making an infinite chain of graphs each of which is obtained from the one before by adding a bridge of infinitesimal conductance.

To recapitulate, the difference in the escape probabilities with and without the bridge is obtained by taking the difference between the original escape probabilities at the ends of the bridge, and multiplying by the expected net number of crossings of the bridge. This quantity is positive because the walker tends to cross the bridge more often in the good direction than in the bad direction.

Exercise 1.4.4 Give a probabilistic argument to show that $C_a p_{\rm esc}$ increases with C_{ar} for any r. Give an example to show that $p_{\rm esc}$ by itself may actually decrease.

Exercise 1.4.5 Give a probabilistic argument to show that $C_a p_{\rm esc}$ increases with C_{rb} for any r.

Exercise 1.4.6 Show that when $v_r = v_s$, changing the value of C_{rs} does not change $p_{\rm esc}$.

Exercise 1.4.7 Show that

$$\frac{\partial}{\partial R_{rs}} R_{\rm eff} = i_{rs}^2.$$

Exercise 1.4.8 In this exercise we ask you to derive an exact formula for the change in escape probability

$$p_{\rm esc}^{(\epsilon)} - p_{\rm esc},$$

in terms of quantities that refer only to the walk without the bridge.

(a) Let N_{xy} denote the expected number of times in state y for a walker who starts at x and walks around the graph without the bridge until he reaches a or b. It is a fact that

$$u_r^{(\epsilon)} = u_r + u_r^{(\epsilon)} \frac{\epsilon}{C_r + \epsilon}(N_{sr} + 1 - N_{rr}) + u_s^{(\epsilon)} \frac{\epsilon}{C_s + \epsilon}(N_{rr} - N_{sr}).$$

Explain in words why this formula is true.

(b) This equation for $u_r^{(\epsilon)}$ can be rewritten as follows:

$$\frac{C_r}{C_r + \epsilon} u_r^{(\epsilon)} = u_r + d^{(\epsilon)}(N_{sr} - N_{rr}).$$

Prove this formula. (Hint: Consider a game where your fortune at x is N_{xr}, and where you start from a and walk on the graph with the bridge until you reach b or return to a.)

(c) Write down the corresponding formula for $u_s^{(\epsilon)}$, and use this formula to get an expression for $d^{(\epsilon)}$ in terms of quantities that refer to the walk without the bridge.

(d) Use the expression for $d^{(\epsilon)}$ to express $p_{\rm esc}^{(\epsilon)} - p_{\rm esc}$ in terms of quantities that refer to the walk without the bridge, and verify that the value of your expression is ≥ 0 for $\epsilon \geq 0$.

Exercise 1.4.9 Give a probabilistic interpretation of the energy dissipation rate.

1.4.3 A Markov chain proof of the Monotonicity Law

Let $\mathbf{P}$ be the ergodic Markov chain associated with an electric network. When we add an ϵ bridge from r to s, we obtain a new transition matrix $\mathbf{P}^{(\epsilon)}$ that differs from $\mathbf{P}$ only for transitions from r and s. We can minimize the differences between $\mathbf{P}$ and $\mathbf{P}^{(\epsilon)}$ by replacing $\mathbf{P}$ by the matrix $\hat{\mathbf{P}}$ corresponding to the circuit without the bridge but with an ϵ conductance added from r to r and from s to s. This allows the chain to stay in states r and s but does not change the escape probability

from a to b. Thus, we can compare the escape probabilities for the two matrices $\hat{\mathbf{P}}$ and $\mathbf{P}^{(\epsilon)}$, which differ only by

$$\begin{array}{ll} \hat{P}_{rr} = \frac{\epsilon}{C_r+\epsilon} & P^{(\epsilon)}_{rr} = 0 \\ \hat{P}_{rs} = \frac{C_{rs}}{C_r+\epsilon} & P^{(\epsilon)}_{rs} = \frac{C_{rs}+\epsilon}{C_r+\epsilon} \\ \hat{P}_{ss} = \frac{\epsilon}{C_s+\epsilon} & P^{(\epsilon)}_{ss} = 0 \\ \hat{P}_{sr} = \frac{C_{sr}}{C_S+\epsilon} & P^{(\epsilon)}_{sr} = \frac{C_{sr}+\epsilon}{C_s+\epsilon} \end{array} .$$

We make states a and b into absorbing states. Let $\hat{\mathbf{N}}$ and $\mathbf{N}^{(\epsilon)}$ be the fundamental matrices for the absorbing chains obtained from $\hat{\mathbf{P}}$ and $\mathbf{P}^{(\epsilon)}$ respectively. Then $\hat{\mathbf{N}} = (\mathbf{I} - \hat{\mathbf{Q}})^{-1}$ and $\mathbf{N}^{(\epsilon)} = (\mathbf{I} - \mathbf{Q}^{(\epsilon)})^{-1}$ where $\hat{\mathbf{Q}}$ and $\mathbf{Q}^{(\epsilon)}$ differ only for the four components involving only r and s. That is,

$$\mathbf{Q}^{(\epsilon)} = \hat{\mathbf{Q}} + \begin{array}{c} \\ r \\ s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} -\frac{\epsilon}{C_r+\epsilon} & \frac{\epsilon}{C_r+\epsilon} \\ \frac{\epsilon}{C_s+\epsilon} & -\frac{\epsilon}{C_s+\epsilon} \end{pmatrix} \end{array} = \hat{\mathbf{Q}} + \mathbf{hk}$$

where $\mathbf{h}$ is the column vector with only components r and s non-zero

$$\mathbf{h} = \begin{array}{c} \\ r \\ s \end{array} \begin{pmatrix} \frac{\epsilon}{C_r+\epsilon} \\ -\frac{\epsilon}{C_s+\epsilon} \end{pmatrix}$$

and $\mathbf{k}$ is a row vector with only components r and s non-zero

$$\mathbf{k} = \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} -1 & 1 \end{pmatrix} \end{array}.$$

J. G. Kemeny has pointed out to us that if $\mathbf{A}$ is any matrix with inverse $\mathbf{N}$ and we add to $\mathbf{A}$ a matrix of the form $-\mathbf{hk}$, then $\mathbf{A} - \mathbf{hk}$ has an inverse if and only if $\mathbf{kNh} \neq 1$ and, if so, $\bar{\mathbf{N}} = (\mathbf{A} - \mathbf{hk})^{-1}$ is given by

$$\bar{\mathbf{N}} = \mathbf{N} + c(\mathbf{Nh})(\mathbf{kN})$$

where $c = 1/(1 - \mathbf{kNh})$. You are asked to prove this in Exercise 1.4.10. Adding $-\mathbf{hk}$ to $\mathbf{A} = \mathbf{I} - \hat{\mathbf{Q}}$ and using this result, we obtain

$$\mathbf{N}^{(\epsilon)} = \hat{\mathbf{N}} + c(\hat{\mathbf{N}}\mathbf{h})(\mathbf{k}\hat{\mathbf{N}}).$$

Using the simple nature of $\mathbf{h}$ and $\mathbf{k}$ we obtain

$$N^{(\epsilon)}_{ij} = \hat{N}_{ij} + \left(\frac{\hat{N}_{ir}\epsilon}{C_r + \epsilon} - \frac{\hat{N}_{is}\epsilon}{C_s + \epsilon} \right) (\hat{N}_{sj} - \hat{N}_{rj})$$

and

$$c = \frac{1}{1 + \frac{\hat{N}_{rr}\epsilon}{C_r+\epsilon} - \frac{\hat{N}_{sr}\epsilon}{C_r+\epsilon} + \frac{\hat{N}_{ss}\epsilon}{C_s+\epsilon} - \frac{\hat{N}_{rs}\epsilon}{C_s+\epsilon}}.$$

Since $\hat{N}_{rr}$ is the expected number of times in r starting in r and $\hat{N}_{sr}$ is the expected number of times in r starting in s, $\hat{N}_{rr} \geq \hat{N}_{sr}$. Similarly $\hat{N}_{ss} \geq \hat{N}_{rs}$ and so the denominator of c is ≥ 1. In particular, it is positive.

Recall that the absorption probabilities for state b are given by

$$B_{xb} = \sum_y N_{xy} P_{yb}.$$

Since $P^{(\epsilon)}_{xb} = \hat{P}_{xb}$,

$$B^{(\epsilon)}_{xb} = \hat{B}_{xb} + c\left(\frac{\hat{N}_{xr}\epsilon}{C_r+\epsilon} - \frac{\hat{N}_{xs}\epsilon}{C_s+\epsilon}\right)(\hat{B}_{sb} - \hat{B}_{rb}).$$

Since $P^{(\epsilon)}_{ax} = \hat{P}_{ax}$,

$$p^{(\epsilon)}_{\text{esc}} = \hat{p}_{\text{esc}} + c\left(\frac{\hat{u}_r\epsilon}{C_r+\epsilon} - \frac{\hat{u}_s\epsilon}{C_s+\epsilon}\right)(\hat{B}_{sb} - \hat{B}_{rb})$$

where $\hat{u}_x$ is the expected number of times that the ergodic chain $\hat{P}$, started at state a, is in state x before returning to a reaching b for the first time. The absorption probability B_{xa} is the quantity v_x introduced in the previous section. As shown there, reversibility allows us to conclude that

$$\frac{\hat{u}_x}{\hat{C}_x} = \frac{\hat{B}_{xa}}{\hat{C}_a} = \frac{\hat{B}_{xa}}{C_a}$$

so that

$$p^{(\epsilon)}_{\text{esc}} = p_{\text{esc}} + \frac{\epsilon c}{C_a}(\hat{B}_{sb} - \hat{B}_{rb})^2$$

and this shows that the Monotonicity Law is true.

The change from $\mathbf{P}$ to $\hat{\mathbf{P}}$ was merely to make the calculations easier. As we have remarked, the escape probabilities are the same for the two chains as are the absorption probabilities B_{ib}. Thus we can remove the hats and write the same formula.

$$p^{(\epsilon)}_{\text{esc}} = p_{\text{esc}} + \frac{\epsilon c}{C_a}(B_{sb} - B_{rb})^2.$$

The only quantity in this final expression that seems to depend upon quantities from $\hat{\mathbf{P}}$ is c. In Exercise 1.4.11 you are asked to show that c can also be expressed in terms of the fundamental matrix $\mathbf{N}$ obtained from the original $\mathbf{P}$.

Exercise 1.4.10 Let $\mathbf{A}$ be a matrix with inverse $\mathbf{N} = \mathbf{A}^{-1}$ Let $\mathbf{h}$ be a column vector and $\mathbf{k}$ a row vector. Show that

$$\bar{\mathbf{N}} = (\mathbf{A} - \mathbf{hk})^{-1}$$

exists if and only if $\mathbf{kNh} \neq 1$ and, if so,

$$\bar{\mathbf{N}} = \mathbf{N} + \frac{(\mathbf{Nh})(\mathbf{kN})}{1 - \mathbf{kNh}}.$$

Exercise 1.4.11 Show that c can be expressed in terms of the fundamental matrix $\mathbf{N}$ of the original Markov chain $\mathbf{P}$ by

$$c = \frac{1}{1 + \frac{N_{rr}\epsilon}{C_r} - \frac{N_{sr}\epsilon}{C_r} + \frac{N_{ss}\epsilon}{C_s} - \frac{N_{rs}\epsilon}{C_s}}.$$

2 Random walks on infinite networks

2.1 Polya's recurrence problem

2.1.1 Random walks on lattices

In 1921 George Polya [26] investigated random walks on certain infinite graphs, or as he called them, "street networks". The graphs he considered, which we will refer to as lattices, are illustrated in Figure 35.

To construct a d-dimensional lattice, we take as vertices those points $(x_1, \ldots, x_d)$ of $\mathbf{R}^d$ all of whose coordinates are integers, and we join each vertex by an undirected line segment to each of its $2d$ nearest neighbors. These connecting segments, which represent the edges of our graph, each have unit length and run parallel to one of the coordinate axes of $\mathbf{R}^d$. We will denote this d-dimensional lattice by $\mathbf{Z}^d$. We will denote the origin $(0, 0, \ldots, 0)$ by $\mathbf{0}$.

Now let a point walk around at random on this lattice. As usual, by walking at random we mean that, upon reaching any vertex of the graph, the probability of choosing any one of the $2d$ edges leading out of that vertex is $\frac{1}{2d}$. We will call this random walk *simple random walk* in d dimensions.

When $d = 1$, our lattice is just an infinite line divided into segments of length one. We may think of the vertices of this graph as representing the fortune of a gambler betting on heads or tails in a fair coin

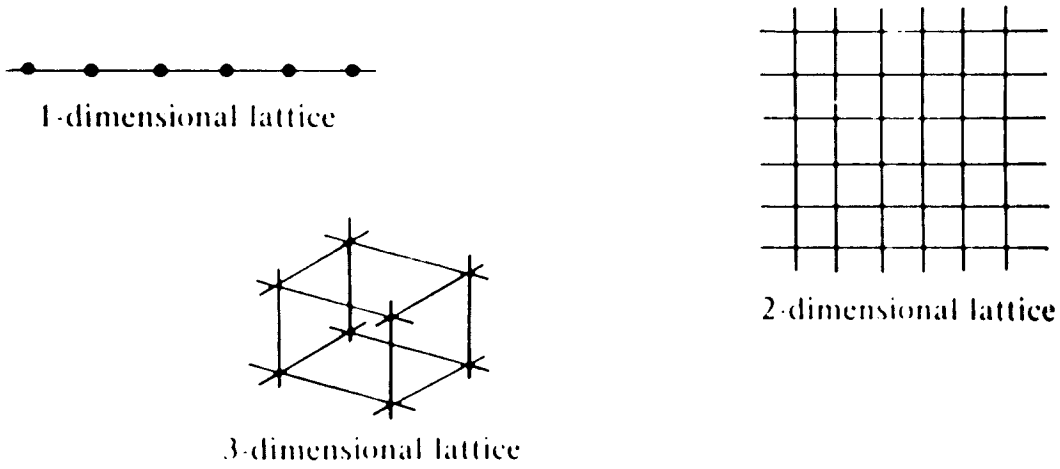


Figure 35: ♣

tossing game. Simple random walk in one dimension then represents the vicissitudes of his or her fortune, either increasing or decreasing by one unit after each round of the game.

When $d = 2$, our lattice looks like an infinite network of streets and avenues, which is why we describe the random motion of the wandering point as a "walk".

When $d = 3$, the lattice looks like an infinite "jungle gym", so perhaps in this case we ought to talk about a "random climb", but we will not do so. It is worth noting that when $d = 3$, the wanderings of our point could be regarded as an approximate representation of the random path of a molecule diffusing in an infinite cubical crystal. Figure 36 shows a simulation of a simple random walk in three dimensions.

2.1.2 The question of recurrence

The question that Polya posed amounts to this: "Is the wandering point certain to return to its starting point during the course of its wanderings?" If so, we say that the walk is *recurrent*. If not, that is, if there is a positive probability that the point will never return to its starting point, then we say that the walk is *transient*.

If we denote the probability that the point never returns to its starting point by p_{esc}, then the chain is recurrent if $p_{\text{esc}} = 0$, and transient if $p_{\text{esc}} > 0$.

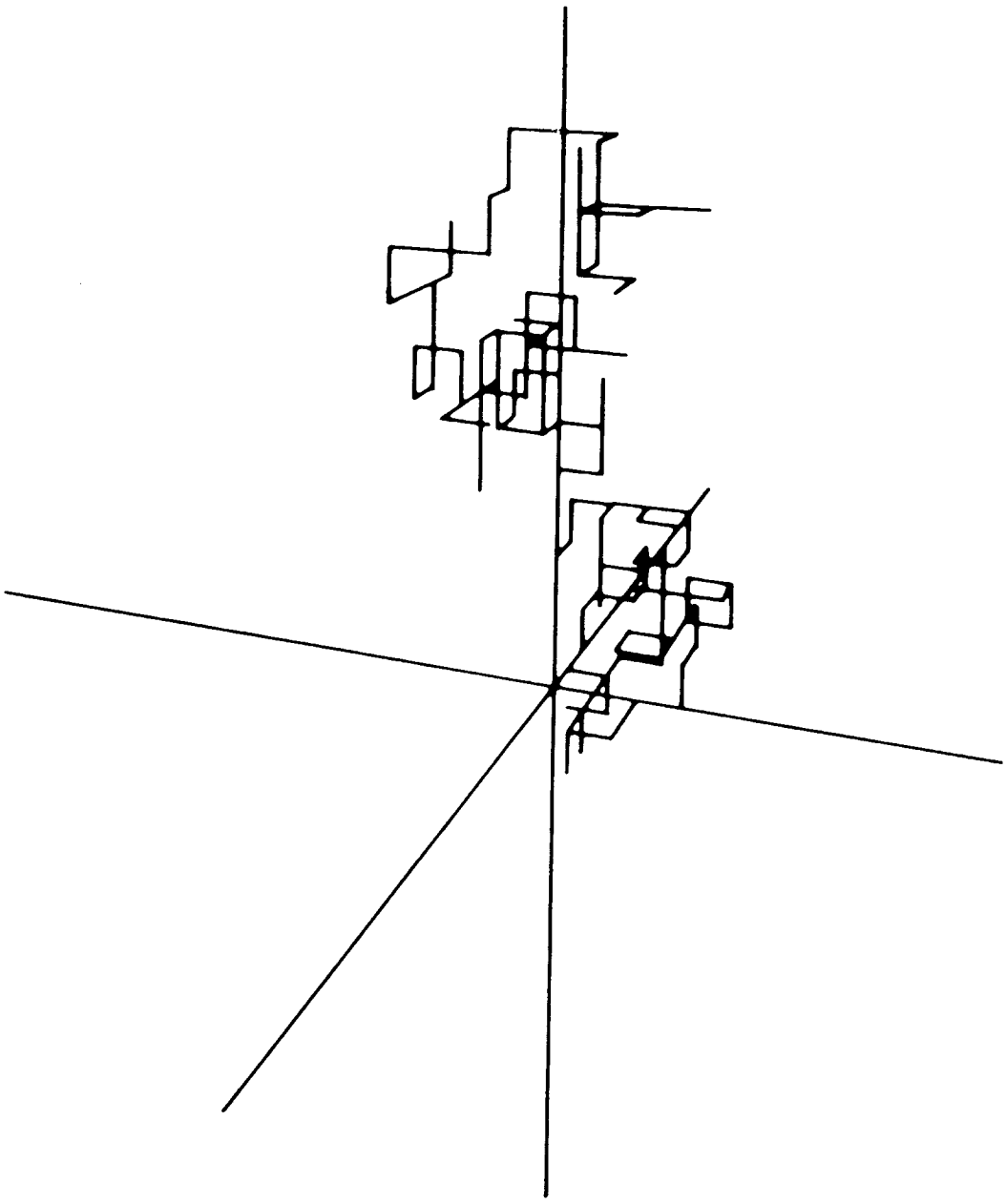


Figure 36: ♣

We will call the problem of determining recurrence or transience of a random walk the *type problem*.

2.1.3 Polya's original question

The definition of recurrence that we have given differs from Polya's original definition. Polya defined a walk to be recurrent if, with probability one, it will pass through every single point of the lattice in the course of its wanderings. In our definition, we require only that the point return to its starting point. So we have to ask ourselves, "Can the random walk be recurrent in our sense and fail to be recurrent in Polya's sense?"

The answer to this question is, "No, the two definitions of recurrence are equivalent." Why? Because if the point must return once to its starting point, then it must return there again and again, and each time it starts away from the origin, it has a certain non-zero probability of hitting a specified target vertex before returning to the origin. And anyone can get a bull's-eye if he or she is allowed an infinite number of darts, so eventually the point will hit the target vertex.

Exercise 2.1.1 Write out a rigorous version of the argument just given.

2.1.4 Polya's Theorem: recurrence in the plane, transience in space

In [26], Polya proved the following theorem:

Polya's Theorem. Simple random walk on a d-dimensional lattice is recurrent for $d = 1, 2$ and transient for $d > 2$.

The rest of this work will be devoted to trying to understand this theorem. Our approach will be to exploit once more the connections between questions about a random walk on a graph and questions about electric currents in a corresponding network of resistors. We hope that this approach, by calling on methods that appeal to our physical intuition, will leave us feeling that we know "why" the theorem is true.

Exercise 2.1.2 Show that Polya's theorem implies that if two random walkers start at $\mathbf{0}$ and wander independently, then in one and two dimensions they will eventually meet again, but in three dimensions there is positive probability that they won't.

Exercise 2.1.3 Show that Polya's theorem implies that a random walker in three dimensions will eventually hit the line defined by $x = 2, z = 0$.

2.1.5 The escape probability as a limit of escape probabilities for finite graphs

We can determine the type of an infinite lattice from properties of bigger and bigger finite graphs that sit inside it. The simplest way to go about this is to look at the lattice analog of balls (solid spheres) in space. These are defined as follows: Let r be an integer—this will be the radius of the ball. Let $G^{(r)}$ be the graph gotten from $\mathbf{Z}^d$ by throwing out vertices whose distance from the origin is $> r$. By "distance from the origin" we mean here not the usual Euclidean distance, but the distance "in the lattice"; that is, the length of the shortest path along the edges of the lattice between the two points. Let $\partial G^{(r)}$ be the "sphere" of radius r about the origin, i.e., those points that are exactly r units from the origin. In two dimensions, $\partial G^{(r)}$ looks like a square. (See Figure 37.) In three dimensions, it looks like an octahedron.

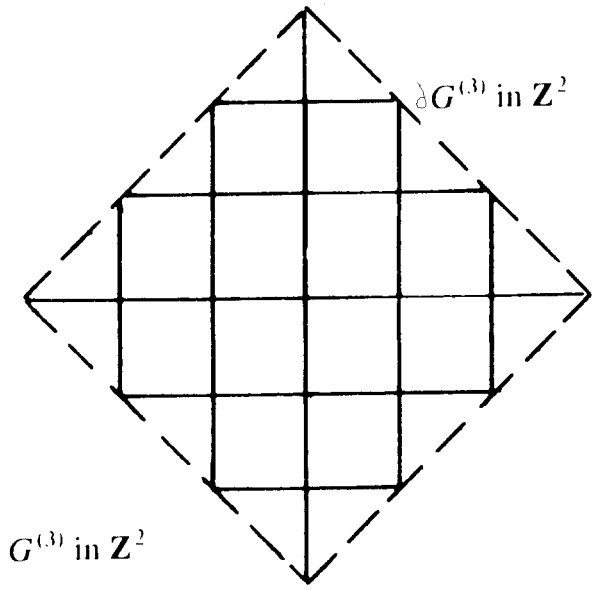


Figure 37: ♣

We define a random walk on $G^{(r)}$ as follows: The walk starts at $\mathbf{0}$ and continues as it would on $\mathbf{Z}^d$ until it reaches a point on $\partial G^{(r)}$

and then it stays at this point. Thus the walk on $G^{(r)}$ is an absorbing Markov chain with every point of $\partial G^{(r)}$ an absorbing state.

Let $p_{\rm esc}^{(r)}$ be the probability that a random walk on $G^{(r)}$ starting at $\mathbf{0}$, reaches $\partial G^{(r)}$ before returning to $\mathbf{0}$. Then $p_{\rm esc}^{(r)}$ decreases as r increases and $p_{\rm esc} = \lim_{r\to\infty} p_{\rm esc}^{(r)}$ is the *escape probability* for the infinite graph. If this limit is 0, the infinite walk is recurrent. If it is greater than 0, the walk is transient.

2.1.6 Electrical formulation of the type problem

Now that we have expressed things in terms of finite graphs, we can make use of electrical methods. To determine $p_{\rm esc}$ electrically, we simply ground all the points of $\partial G^{(r)}$, maintain $\mathbf{0}$ at one volt, and measure the current $i^{(r)}$ flowing into the circuit. (See Figure 38.)

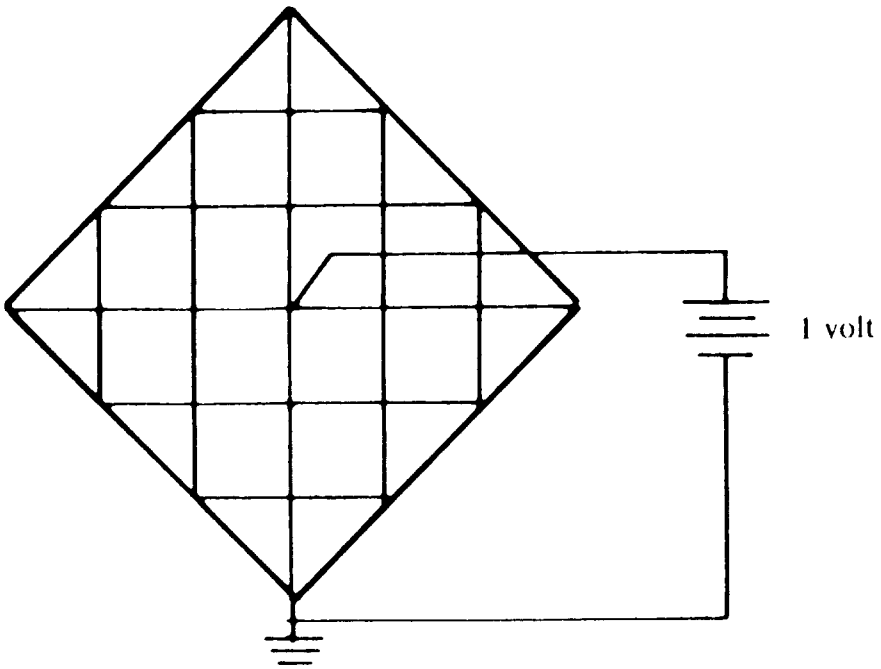


Figure 38: ♣

From Section 1.3.4, we have

$$p_{\rm esc}^{(r)} = \frac{i^{(r)}}{2d},$$

where d is the dimension of the lattice. (Remember that we have to divide by the number of branches coming out of the starting point.) Since the voltage being applied is 1, $i^{(r)}$ is just the effective conductance between $\mathbf{0}$ and $\partial G^{(r)}$, i.e.,

$$i^{(r)} = \frac{1}{R_{\text{eff}}^{(r)}}.$$

where $R_{\text{eff}}^{(r)}$ is the effective resistance from $\mathbf{0}$ to $\partial G^{(r)}$. Thus

$$p_{\text{esc}}^{(r)} = \frac{1}{2dR_{\text{eff}}^{(r)}}.$$

Define R_{eff}, the *effective resistance from the origin to infinity*, to be

$$R_{\text{eff}} = \lim_{r\to\infty} R_{\text{eff}}^{(r)}.$$

This limit exists since $R_{\text{eff}}^{(r)}$ is an increasing function of r. Then

$$p_{\text{esc}} = \frac{1}{2dR_{\text{eff}}}.$$

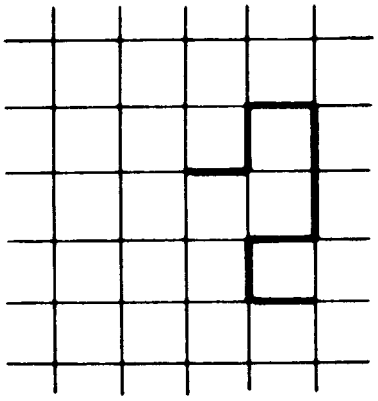
Of course R_{eff} may be infinite; in fact, this will be the case if and only if $p_{\text{esc}} = 0$. Thus the walk is recurrent if and only if the resistance to infinity is infinite, which makes sense.

The success of this electrical formulation of the type problem will come from the fact that the resistance to infinity can be estimated using classical methods of electrical theory.

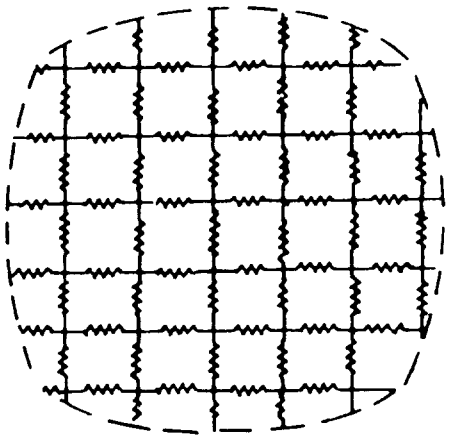
2.1.7 One Dimension is easy, but what about higher dimensions?

We now know that simple random walk on a graph is recurrent if and only if a corresponding network of 1-ohm resistors has infinite resistance "out to infinity". Since an infinite line of resistors obviously has infinite resistance, it follows that simple random walk on the 1-dimensional lattice is recurrent, as stated by Polya's theorem.

What happens in higher dimensions? We are asked to decide whether a d-dimensional lattice has infinite resistance to infinity. The difficulty is that the d-dimensional lattice $\mathbf{Z}^d$ lacks the rotational symmetry of the Euclidean space $\mathbf{R}^d$ in which it sits.



Random walk recurrence?



Resistance to infinity infinite?

Figure 39: ♣

Resistance to infinity infinite!

Figure 40: ♣

To see how this lack of symmetry complicates electrical problems, we determine, by solving the appropriate discrete Dirichlet problem, the voltages for a one-volt battery attached between $\mathbf{0}$ and the points of $\partial G^{(3)}$ in $\mathbf{Z}^2$. The resulting voltages are:

$$\begin{array}{ccccccc} & & & 0 & & & \\ & & 0 & .091 & 0 & & \\ & 0 & .182 & .364 & .182 & 0 & \\ 0 & .091 & .364 & 1 & .364 & .091 & 0\,. \\ & 0 & .182 & .364 & .182 & 0 & \\ & & 0 & .091 & 0 & & \\ & & & 0 & & & \end{array}$$

The voltages at points of $\partial G^{(1)}$ are equal, but the voltages at points of $\partial G^{(2)}$ are not. This means that the resistance from $\mathbf{0}$ to $\partial G^{(3)}$ cannot be written simply as the sum of the resistances from $\mathbf{0}$ to $\partial G^{(1)}$, $\partial G^{(1)}$ to $\partial G^{(2)}$, and $\partial G^{(2)}$ to $\partial G^{(3)}$. This is in marked contrast to the case of a continuous resistive medium to be discussed in Section 2.1.8.

Exercise 2.1.4 Using the voltages given for $G^{(3)}$, find $R^{(3)}_{\text{eff}}$ and $p^{(3)}_{\text{esc}}$.

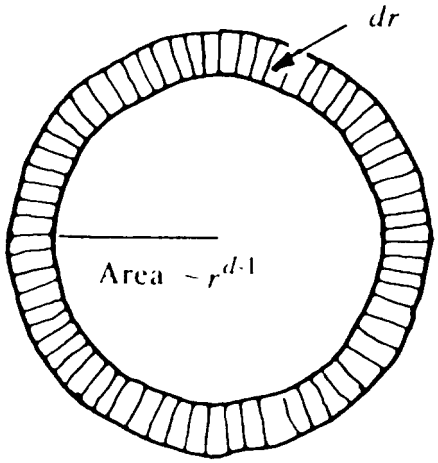
Exercise 2.1.5 Consider a one-dimensional infinite network with resistors $R_n = 1/2^n$ from n to $n+1$ for $n = \ldots, -2, -1, 0, 1, 2, \ldots$. Describe the associated random walk and determine whether the walk is recurrent or transient.

Exercise 2.1.6 A random walk moves on the non-negative integers; when it is in state n, $n > 0$, it moves with probability p_n to $n+1$ and with probability $1 - p_n$, to $n-1$. When at $\mathbf{0}$, it moves to 1. Determine a network that gives this random walk and give a criterion in terms of the p_n for recurrence of the random walk.

2.1.8 Getting around the lack of rotational symmetry of the lattice

Suppose we replace our d-dimensional resistor lattice by a (homogeneous, isotropic) resistive medium filling all of $\mathbf{R}^d$ and ask for the effective resistance to infinity. Naturally we expect that the rotational symmetry will make this continuous problem easier to solve than the original discrete problem. If we took this problem to a physicist, he or she would probably produce something like the scribblings illustrated

in Figure 41, and conclude that the effective resistance is infinite for $d = 1, 2$ and finite for $d > 2$. The analogy to Polya's theorem is obvious, but is it possible to translate these calculations for continuous media into information about what happens in the lattice?



$$R_\infty \sim \int_a^\infty \frac{dr}{r^{d-1}} \quad \begin{cases} = \infty & \text{for } d = 1, 2 \\ < \infty & \text{for } d < 3 \end{cases}$$

Figure 41: ♣

This can indeed be done, and this would certainly be the most natural approach to take. We will come back to this approach at the end of the work. For now, we will take a different approach to getting around the asymmetry of the lattice. Our method will be to modify the lattice in such a way as to obtain a graph that is symmetrical enough so that we can calculate its resistance out to infinity. Of course, we will have to think carefully about what happens to that resistance when we make these modifications.

2.1.9 Rayleigh: shorting shows recurrence in the plane, cutting shows transience in space

Here is a sketch of the method we will use to prove Polya's theorem.

To take care of the case $d = 2$, we will modify the 2-dimensional resistor network by shorting certain sets of nodes together so as to get a new network whose resistance is readily seen to be infinite. As shorting

can only decrease the effective resistance of the network, the resistance of the original network must also be infinite. Thus the walk is recurrent when $d = 2$.

To take care of the case $d = 3$, we will modify the 3-dimensional network by cutting out certain of the resistors so as to get a new network whose resistance is readily seen to be finite. As cutting can only increase the resistance of the network, the resistance of the original network must also be finite. Thus the walk is transient when $d = 3$.

The method of applying shorting and cutting to get lower and upper bounds for the resistance of a resistive medium was introduced by Lord Rayleigh. (Rayleigh [29]; see also Maxwell [21], Jeans [11], PoIya and Szego [28]). We will refer to Rayleigh's techniques collectively as *Rayleigh's short-cut method*. This does not do Rayleigh justice, for Rayleigh's method is a whole bag of tricks that goes beyond mere shorting and cutting—but who can resist a pun?

Rayleigh's method was apparently first applied to random walks by C. St. J. A. Nash-Williams [24], who used the shorting method to establish recurrence for random walks on the 2-dimensional lattice.

2.2 Rayleigh's short-cut method

2.2.1 Shorting and cutting

In its simplest form, Rayleigh's method involves modifying the network whose resistance we are interested in so as to get a simpler network. We consider two kinds of modifications, shorting and cutting. Cutting involves nothing more than clipping some of the branches of the network, or what is the same, simply deleting them from the network. Shorting involves connecting a given set of nodes together with perfectly conducting wires, so that current can pass freely between them. In the resulting network, the nodes that were shorted together behave as if they were a single node.

2.2.2 The Shorting Law and the Cutting Law; Rayleigh's idea

The usefulness of these two procedures (shorting and cutting) stems from the following observations:

Shorting Law. Shorting certain sets of nodes together can only decrease the effective resistance of the network between two given nodes.

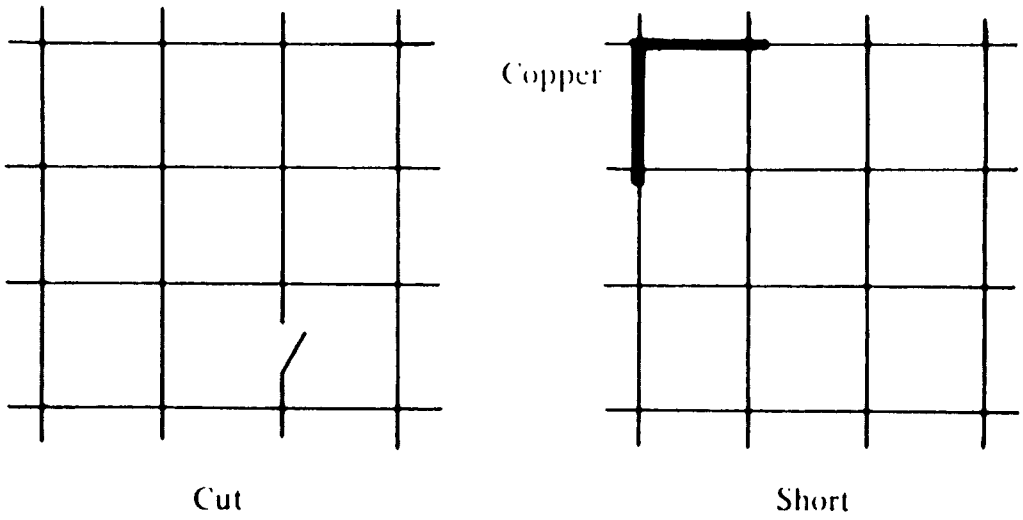


Figure 42: ♣

Cutting Law. Cutting certain branches can only increase the effective resistance between two given nodes.

These laws are both equivalent to Rayleigh's Monotonicity Law, which was introduced in Section 1.4.1 (see Exercise 2.2.1):

Monotonicity Law. The effective resistance between two given nodes is monotonic in the branch resistances.

Rayleigh's idea was to use the Shorting Law and the Cutting Law above to get lower and upper bounds for the resistance of a network. In Section 2.2.3 we apply this method to solve the recurrence problem for simple random walk in dimensions 2 and 3.

Exercise 2.2.1 Show that the Shorting Law and the Cutting Law are both equivalent to the Monotonicity Law.

2.2.3 The plane is easy

When $d = 2$, we apply the Shorting Law as follows: Short together nodes on squares about the origin, as shown in Figure 43. The network we obtain is equivalent to the network shown in Figure 44.

Now as n 1-ohm resistors in parallel are equivalent to a single resistor of resistance $\frac{1}{n}$ ohms, the modified network is equivalent to the network shown in Figure 45. The resistance of this network out to infinity is

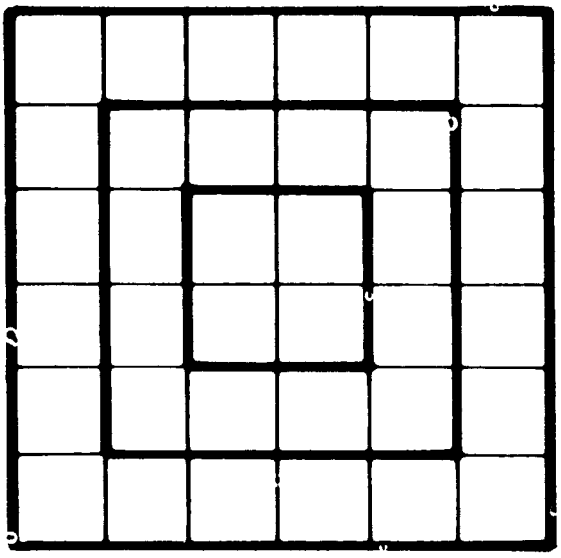


Figure 43: ♣

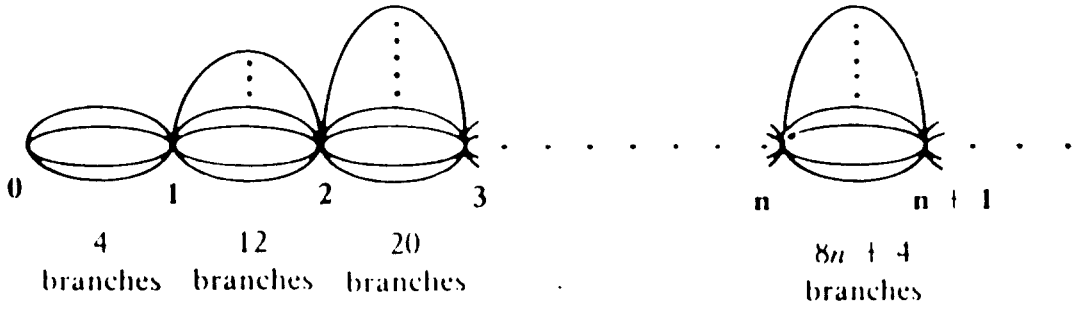


Figure 44: ♣

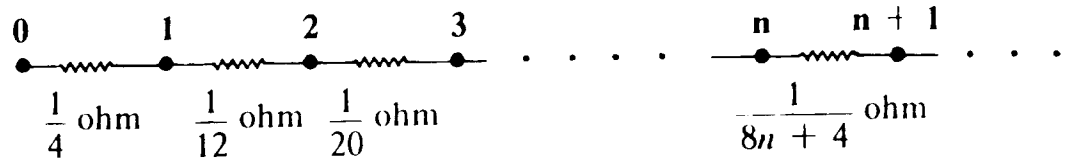


Figure 45: ♣

$$\sum_{n=0}^{\infty} \frac{1}{8n+4} = \infty.$$

As the resistance of the old network can only be bigger, we conclude that it too must be infinite, so that the walk is recurrent when $d = 2$.

Exercise 2.2.2 Using the shorting technique, give an upper bound for $p_{\rm esc}^{(3)}$, and compare this with the exact value obtained in Exercise 2.1.4.

2.2.4 Space: searching for a residual network

When $d = 3$, what we want to do is delete certain of the branches of the network so as to leave behind a residual network having manifestly finite resistance. The problem is to reconcile the "manifestly" with the "finite". We want to cut out enough edges so that the effective resistance of what is left is easy to calculate, while leaving behind enough edges so that the result of the calculation is finite.

2.2.5 Trees are easy to analyze

Trees—that is, graphs without circuits—are undoubtedly the easiest to work with. For instance, consider the full binary tree, shown in Figure 46. Notice that sitting inside this tree just above the root are two copies of the tree itself. This self-similarity property can be used to compute the effective resistance R_∞ from the root out to infinity. (See Exercise 2.2.3.) It turns out that $R_\infty = 1$. We will demonstrate this below by a more direct method.

To begin with, let us determine the effective resistance R_n between the root and the set of nth generation branch points. To do this, we should ground the set of branch points, hook the root up to a 1-volt

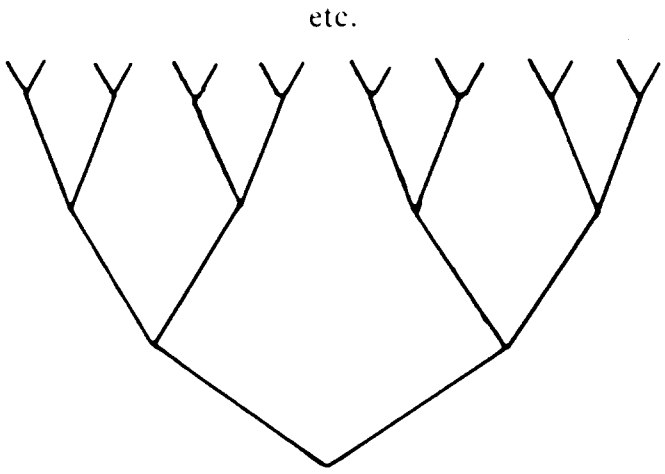


Figure 46: ♣

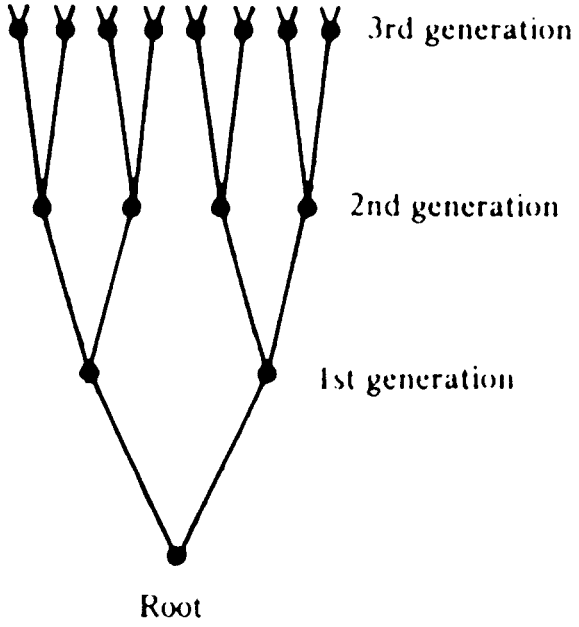


Figure 47: ♣

battery, and compute

$$R_n = \frac{1}{\text{current through battery}}.$$

For $n = 3$, the circuit that we would obtain is shown in Figure 48.

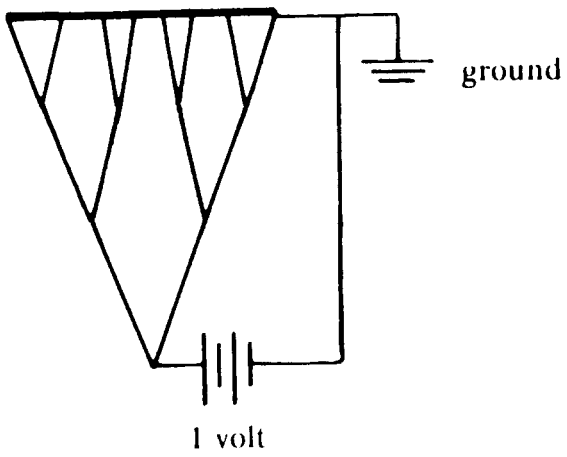


Figure 48: ♣

In the resulting circuit, all branch points of the same generation are at the same voltage (by symmetry). Nothing happens when you short together nodes that are at the same potential. Thus shorting together branch points of the same generation will not affect the distribution of currents in the branches. In particular, this modification will not affect the current through the battery, and we conclude that

$$R_n = \frac{1}{\text{current in original circuit}} = \frac{1}{\text{current in modified circuit}}.$$

For $n = 3$, the modified circuit is shown in Figure 49. This picture shows that

$$R_3 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = 1 - \frac{1}{2^3}.$$

More generally,

$$R_n = \frac{1}{2} + \frac{1}{4} + \ldots + \frac{1}{2^n} = 1 - \frac{1}{2^n}.$$

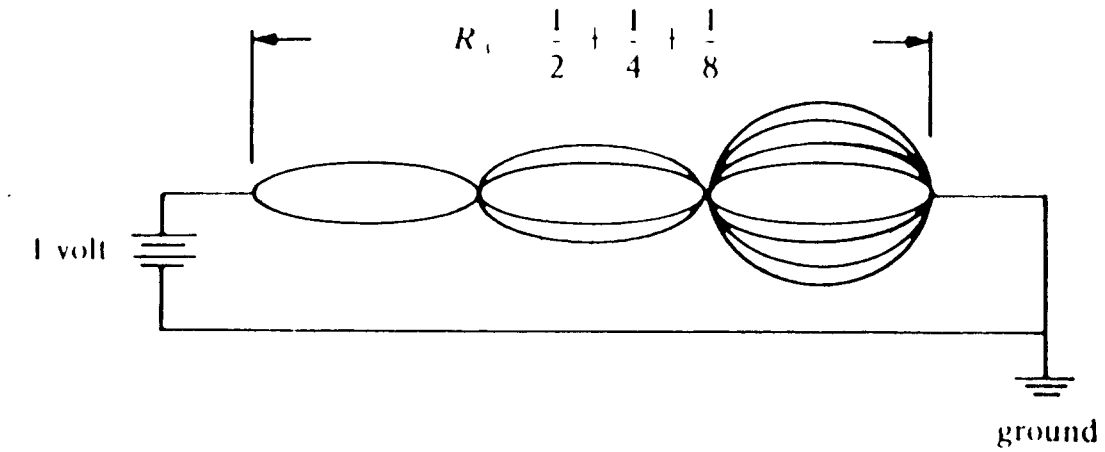


Figure 49: ♣

Letting $n \to \infty$, we get

$$R_\infty = \lim_{n\to\infty} R_n = \lim_{n\to\infty} 1 - \frac{1}{2^n} = 1.$$

Figure 50 shows another closely related tree, the *tree homogeneous of degree three*: Note that all nodes of this tree are similar—there is no intrinsic way to distinguish one from another. This tree is obviously a close relative of the full binary tree. Its resistance to infinity is 2/3.

Exercise 2.2.3 (a) Show, using the self-similarity of the full binary tree, that the resistance R_∞ to infinity satisfies the equation

$$R_\infty = \frac{R_\infty + 1}{2}$$

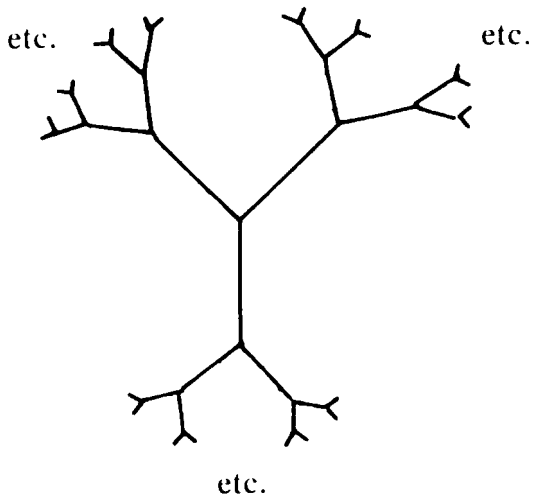
and conclude that either $R_\infty = 1$ or $R_\infty = \infty$.

(b) Again using the self-similarity of the tree, show that

$$R_{n+1} = \frac{R_n + 1}{2}$$

where R_n denotes the resistance out to the set of the nth-generation branch points. Conclude that

$$R_\infty = \lim_{n\to\infty} R_n = 1.$$



The tree homogeneous of degree 3

Figure 50: ♣

2.2.6 The full binary tree is too big

Nothing could be nicer than the two trees we have just described. They are the prototypes of networks having manifestly finite resistance to infinity. Unfortunately, we can't even come close to finding either of these trees as a subgraph of the three-dimensional lattice. For in these trees, the number of nodes in a "ball" of radius r grows exponentially with r, whereas in a d-dimensional lattice, it grows like r, i.e., much slower. (See Figure 51.) There is simply no room for these trees in any finite-dimensional lattice.

2.2.7 NT_3: a "three-dimensional" tree

These observations suggest that we would do well to look for a nice tree NT_3 where the number of nodes within a radius r of the root is on the order of r^3. For we might hope to find something resembling NT_3 in the 3-dimensional lattice, and if there is any justice in the world, this tree would have finite resistance to infinity, and we would be done.

Before introducing NT_3, let's have a look at NT_2, our choice for the tree most likely to succeed in the 2-dimensional lattice (see Figure 52). The idea behind NT_2 is that, since a ball of radius r in the graph ought to contain something like r^2 points, a sphere of radius r ought

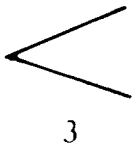
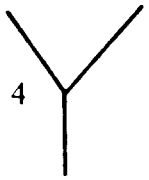
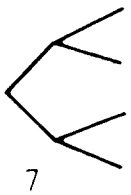
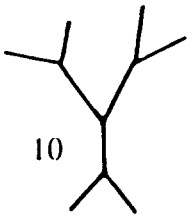
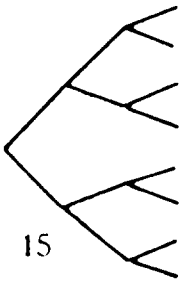
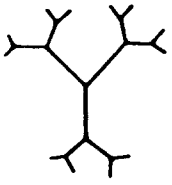
Radius r	*Nodes within r of origin*	
	Full binary tree	Tree homogeneous of degree 3
1	 3	 4
2	 7	 10
3	 15	 22
r	$2^{r+1} - 1$	$3 \cdot 2^r - 2$

Figure 51: ♣

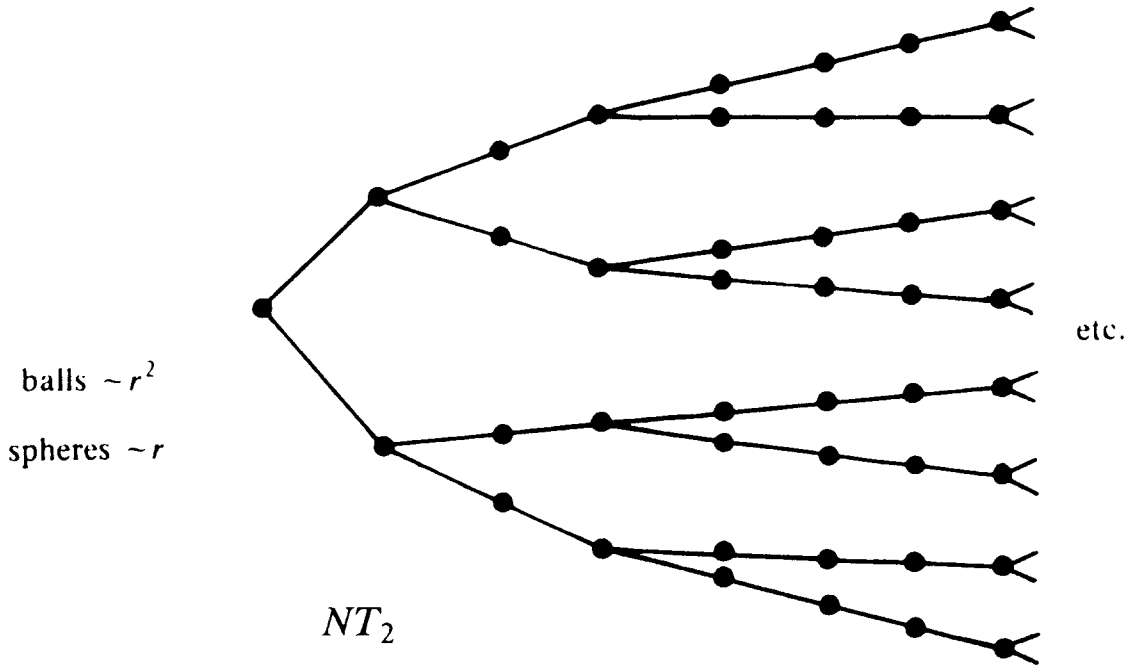


Figure 52: ♣

to contain something like r points, so the number of points in a sphere should roughly double when the radius of the sphere is doubled. For this reason, we make the branches of our tree split in two every time the distance from the origin is (roughly) doubled.

Similarly, in a 3-dimensional tree, when we double the radius, the size of a sphere should roughly quadruple. Thus in NT_3, we make the branches of our tree split in four where the branches of NT_2 would have split in two. NT_3 is shown in Figure 53. Obviously, NT_3 is none too

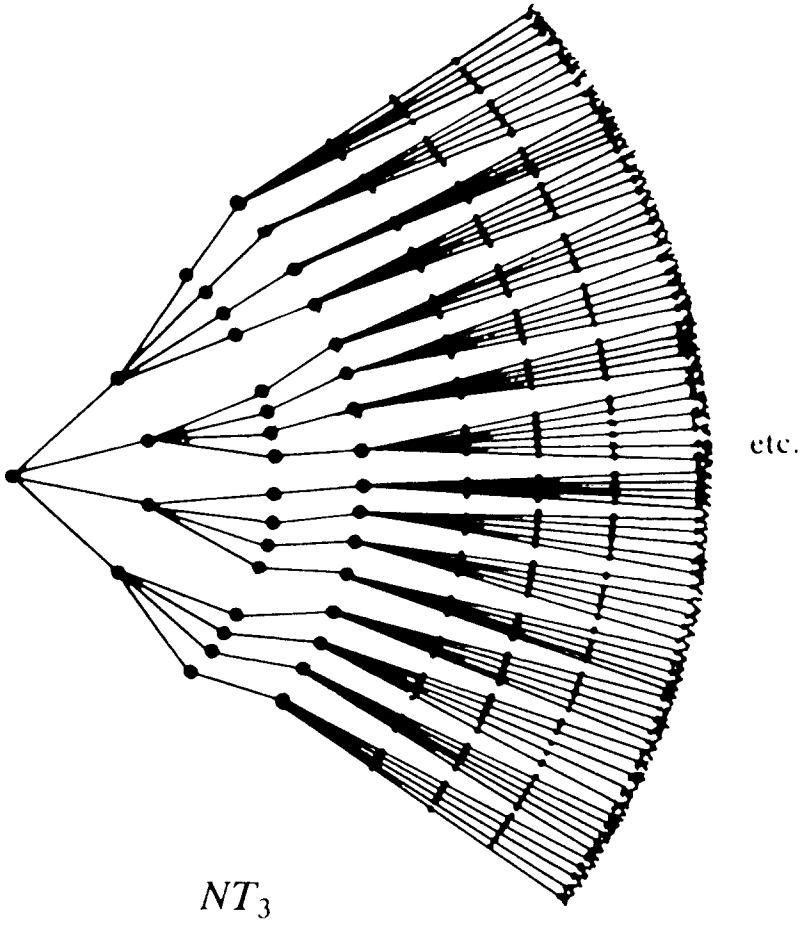


Figure 53: ♣

happy about being drawn in the plane.

2.2.8 NT_3 has finite resistance

To see if we're on the right track, let's work out the resistance of our new trees. These calculations are shown in Figures 54 and 55.

As we would hope, the resistance of NT_2 is infinite, but the resistance of NT_3 is not.

Exercise 2.2.4 Use self-similarity arguments to compute the resistance of NT_2 and NT_3.

2.2.9 But does NT_3 fit in the three-dimensional lattice?

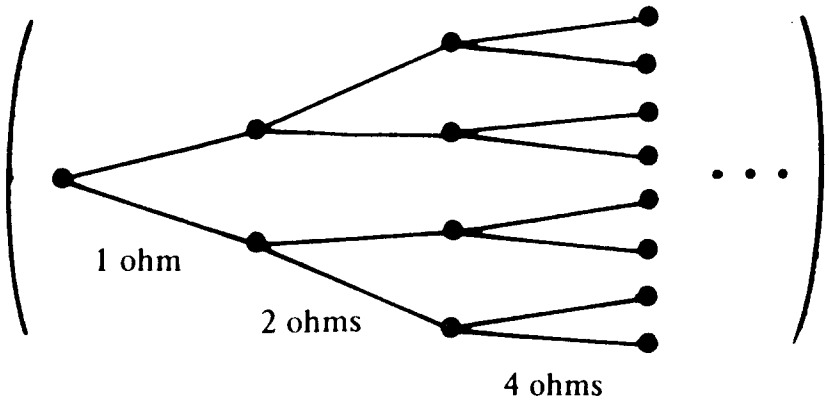
We would like to embed NT_3 in $\mathbf{Z}^3$. We start by trying to embed NT_2 in $\mathbf{Z}^2$. The result is shown in Figure 56.

To construct this picture, we start from the origin and draw 2 rays, one going north, one going east. Whenever a ray intersects the line $x+y=2^n-1$ for some n, it splits into 2 rays, one going north, and one going east. The sequence of pictures in Figure 57 shows successively larger portions of the graph, along with the corresponding portions of NT_2.

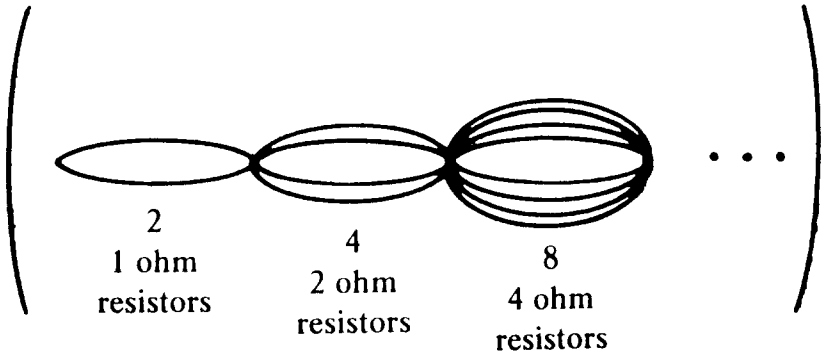
Of course this isn't really an embedding, since certain pairs of points that were distinct in NT_2 get identified, that is, they are made to correspond to a single point in the picture. In terms of our description, sometimes a ray going north and a ray going east pass through each other. This could have been avoided by allowing the rays to "bounce" instead of passing through each other, at the expense of embedding not NT_2 but a close relative—see Exercise 2.2.7. However, because the points of each identified pair are at the same distance from the root of NT_2, when we put a battery between the root and the nth level they will be at the same potential. Hence, the current flow is not affected by these identifications, so the identifications have no effect on R_{eff}. For our purposes, then, we have done just as well as if we had actually embedded NT_2.

To construct the analogous picture in three dimensions, we start three rays off from the origin going north, east, and up. Whenever a ray intersects the plane $x+y+z=2^n-1$ for some n, it splits into three rays, going north, east, and up. This process is illustrated in Figure 58. Surprisingly, the subgraph of the 3-dimensional lattice obtained in this way is not NT_3! Rather, it represents an attempt to embed the tree shown in Figure 59. We call this tree $NT_{2.5849\ldots}$ because it is $2.5849\ldots$-dimensional in the sense that when you double the radius

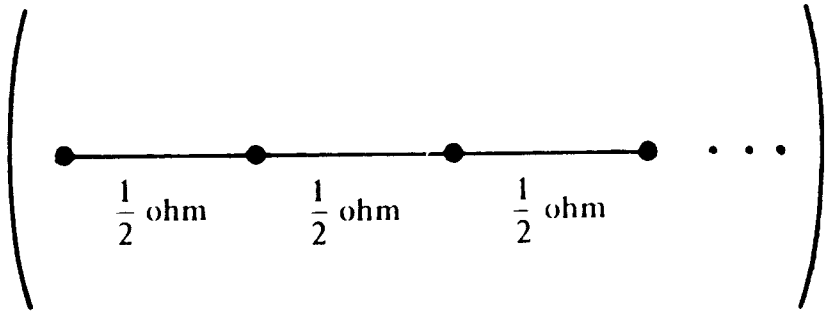
$$R_{\text{root},\infty}(NT_2) = R_{\text{root},\infty}$$



$$= R_{\text{root},\infty}$$



$$= R_{\text{root},\infty}$$

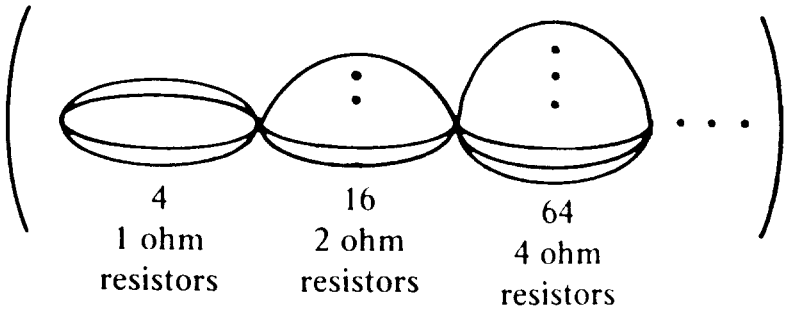


$$= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots$$

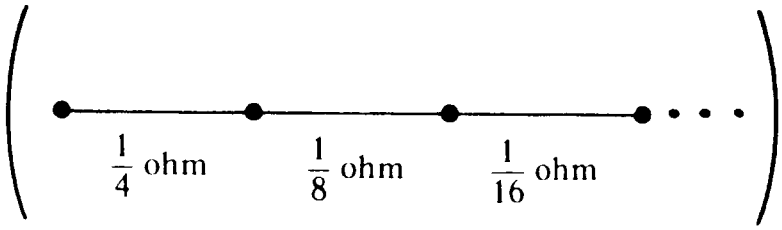
$$= \infty \cdot$$

Figure 54: ♣

$$R_{\text{root},\infty}(NT_3) = R_{\text{root},\infty}$$



$$= R_{\text{root},\infty}$$



$$= \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots$$

$$= \frac{1}{2} .$$

Figure 55: ♣

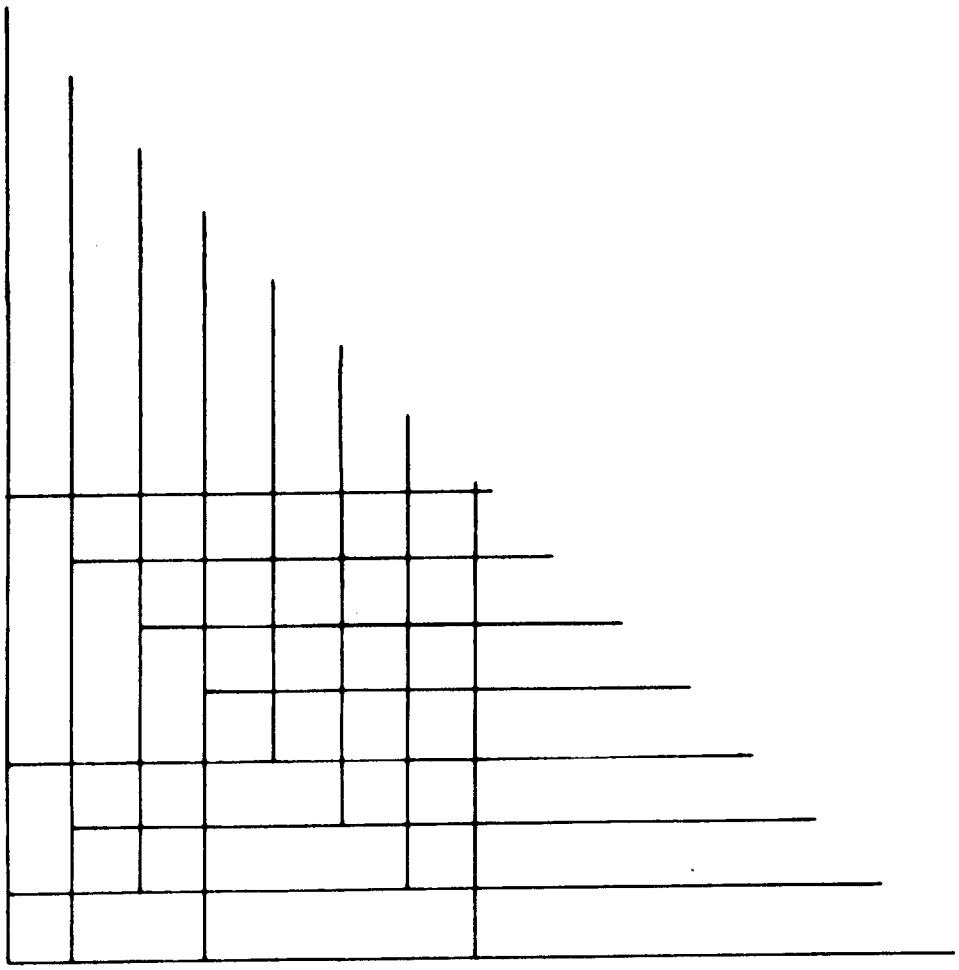


Figure 56: ♣

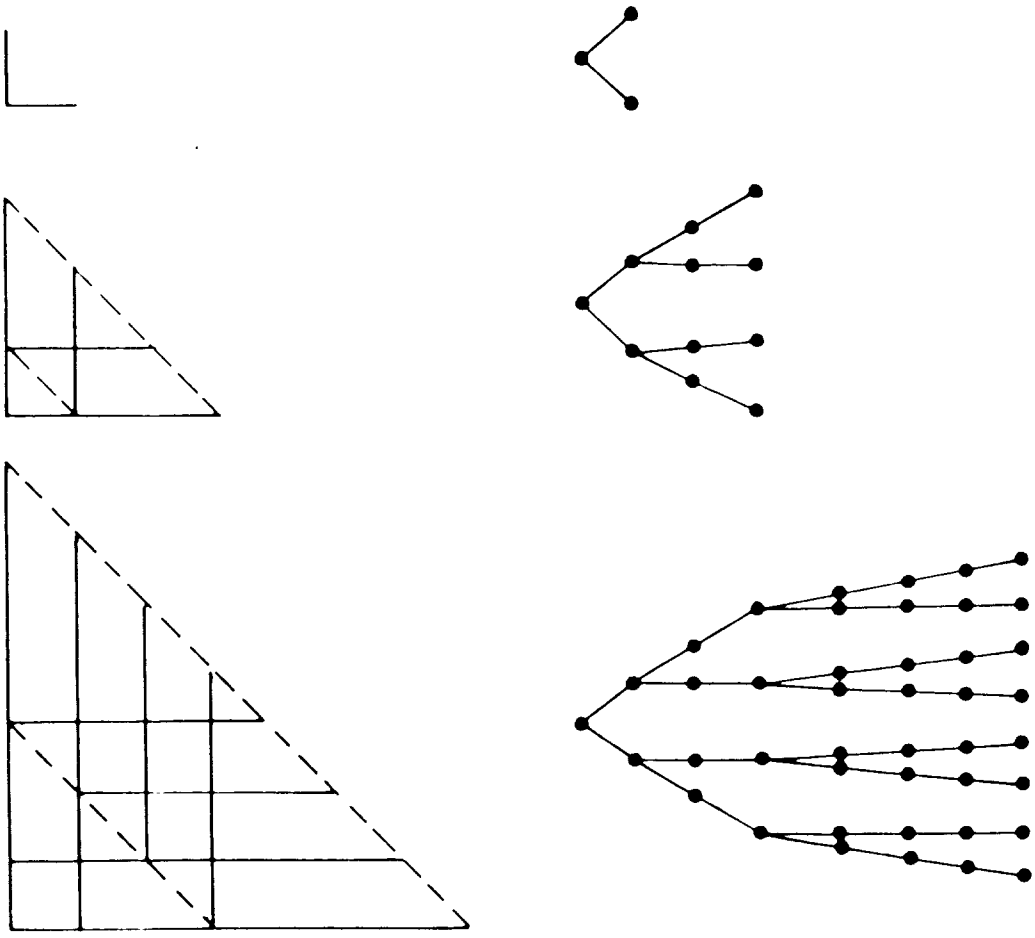


Figure 57: ♣

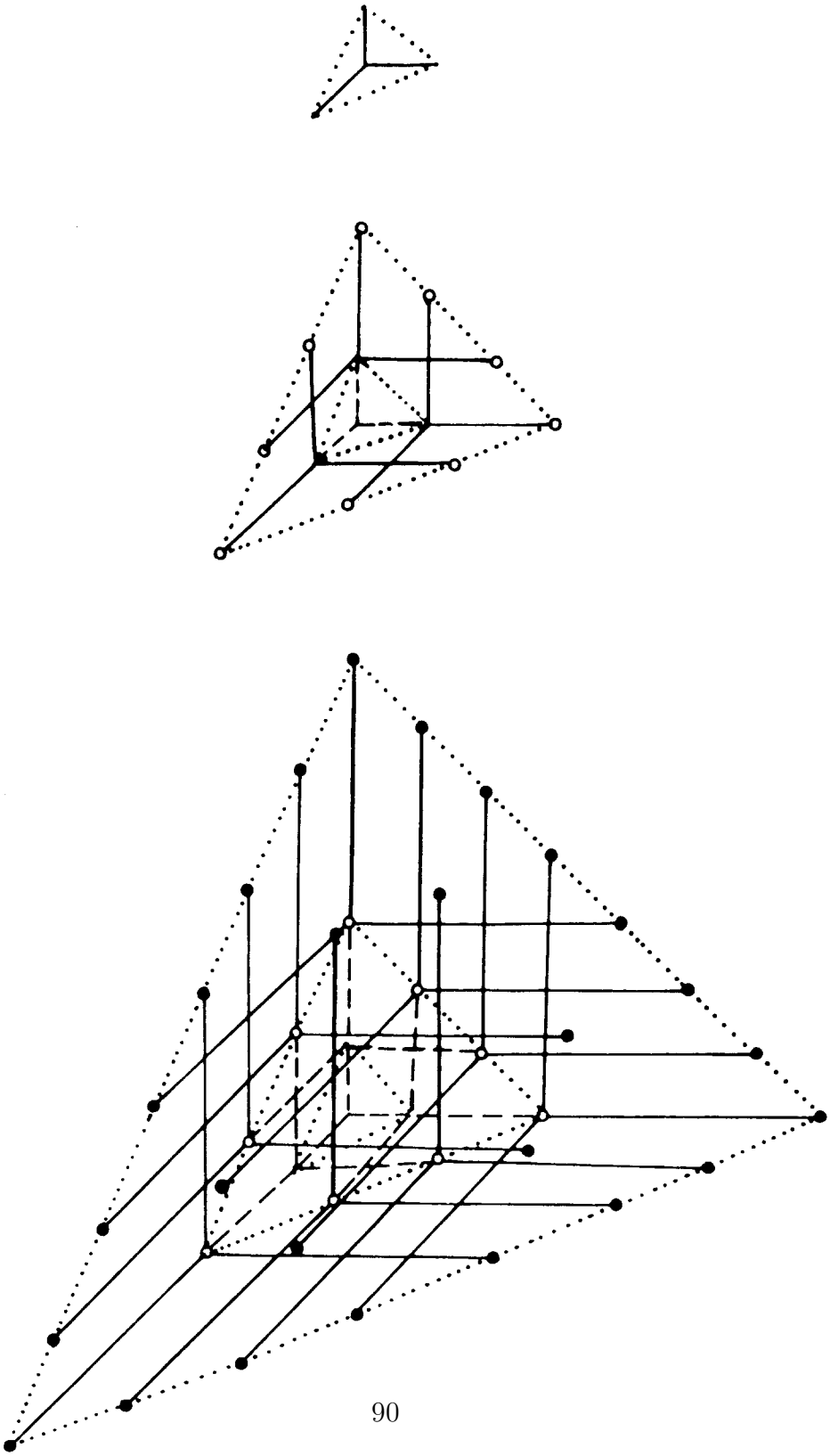


Figure 58: ♣

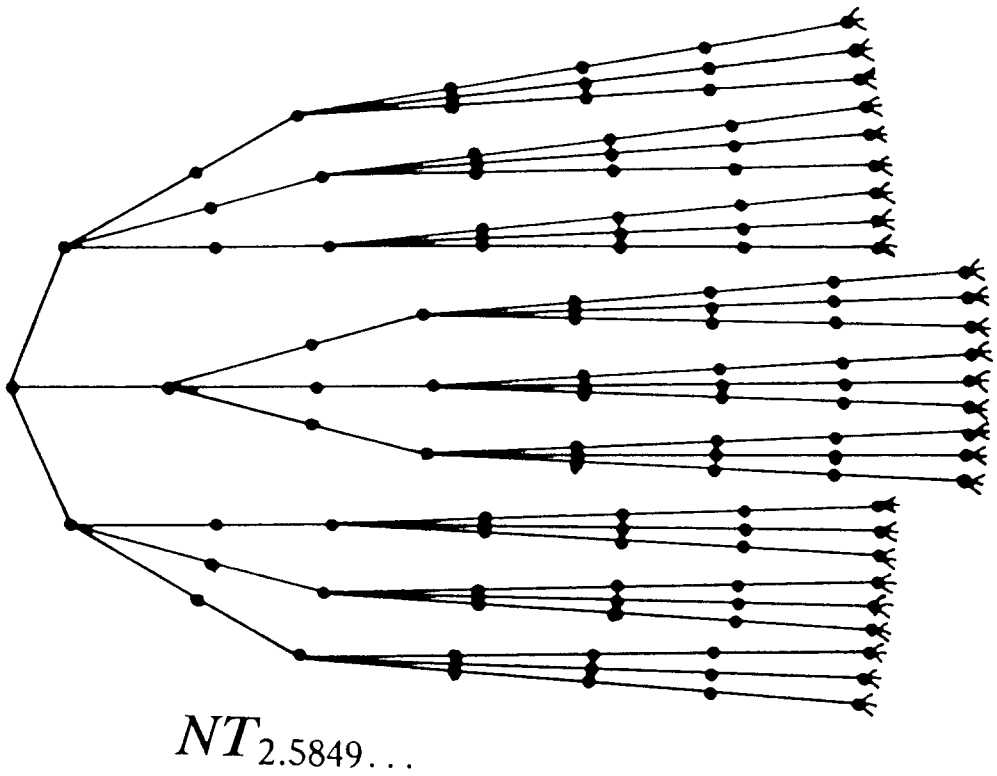


Figure 59: ♣

of a ball, the number of points in the ball gets multiplied roughly by 6 and

$$6 = 2^{\log_2 6} = 2^{2.5849\ldots}.$$

Again, certain pairs of points of $\mathrm{NT}_{2.5849\ldots}$ have been allowed to correspond to the same point in the lattice, but once again the intersections have no effect on R_{eff}.

So we haven't come up with our embedded NT_3 yet. But why bother? The resistance of $\mathrm{NT}_{2.5849\ldots}$ out to infinity is

$$\frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \ldots = \frac{1}{3}\left(1 + \frac{2}{3} + (\frac{2}{3})^2 + \ldots\right) = \frac{1}{3}\frac{1}{1-\frac{2}{3}} = 1.$$

Thus we have found an infinite subgraph of the 3-dimensional lattice having finite resistance out to infinity, and we are done.

Exercise 2.2.5 This exercise deals with the escape probability p_{esc} for simple random walk in 3 dimensions. The idea is to turn upper and lower bounds for the resistance of the lattice into bounds for p_{esc}. Bounds are the best we can ask for using our method. The determination of the exact value will be discussed in Section 2.3.5. It is roughly .66.

(a) Use a shorting argument to get an upper bound for p_{esc}.

(b) We have seen that the resistance of the 3-dimensional lattice is at most one ohm. Show that the corresponding lower bound for p_{esc} is 1/6 Show that this lower bound can be increased to 1/3 with no extra effort.

Exercise 2.2.6 Prove that simple random walk in any dimension $d > 3$ is transient.

Exercise 2.2.7 Show how the not-quite embeddings of NT_2 and $\mathrm{NT}_{2.5849\ldots}$ can be altered to yield honest-to-goodness embeddings of "stretched-out" versions of these trees, obtained by replacing each edge of the tree by three edges in series. (Hint: "bounce".)

2.2.10 What we have done; what we will do

We have finally finished our electrical proof of Polya's theorem. The proof of recurrence for $d = 1, 2$ was straight-forward, but this could hardly be said of the proof for $d = 3$. After all, we started out trying

to embed NT_3 and ended up by not quite embedding something that was not quite NT_3!

This is not bad in itself, for one frequently sets out to do something and in the process of trying to do it gets a better idea. The real problem is that this explicit construction is just too clever, too artificial. We seem to be saying that a simple random walk in 3 dimensions is transient because it happens to contain a beautifully symmetrical subgraph that is in some sense 2.5849...-dimensional! Fine, but what if we hadn't stumbled upon this subgraph? Isn't there some other, more natural way?

We will see that indeed there are more natural approaches to showing transience for $d = 3$. One such approach uses the same idea of embedding trees, but depends on the observation that one doesn't need to be too careful about sending edges to edges. Another approach, based on relating the lattice not to a tree but to Euclidean space, was already hinted at in Section 2.1.8. The main goal for the rest of this work will be to explore these more natural electrical approaches to Polya's theorem.

Before jumping into this, however, we are going to go back and take a look at a classical—i.e., probabilistic—approach to Polya's theorem. This will give us something to compare our electrical proofs with.

2.3 The classical proofs of Polya's Theorem

2.3.1 Recurrence is equivalent to an infinite expected number of returns

For the time being, all of our random walks will be simple. Let u be the probability that a random walker, starting at the origin, will return to the origin. The probability that the walker will be there exactly k times (counting the initial time) is $u^k(1-u)$. Thus, if m is the expected number of times at the origin,

$$m = \sum_{k=1} \infty k u^{k-1}(1-u) = \frac{1}{1-u}.$$

If $m = \infty$ then $u = 1$, and hence the walk is recurrent. If $m < \infty$ then $u < 1$, so the walk is transient. Thus m determines the type of the walk.

We shall use an alternate expression for m. Let u_n be the probability that the walk, starting at $\mathbf{0}$, is at $\mathbf{0}$ on the nth step. Since the walker

starts at $\mathbf{0}$, $u_0 = 1$. Let e_n be a random variable that takes on the value 1 if, at time n the walker is at $\mathbf{0}$ and 0 otherwise. Then

$$T = \sum_{n=0}^{\infty} e_n$$

is the total number of times at $\mathbf{0}$ and

$$m = \mathbf{E}(T) = \sum_{n=0}^{\infty} \mathbf{E}(e_n).$$

But $\mathbf{E}(e_n) = 1 \cdot u_n + 0 \cdot (1 - u_n) = u_n$. Thus

$$m = \sum_{n=0}^{\infty} u_n.$$

Therefore, the walk will be recurrent if the series $\sum_{n=0}^{\infty}$ diverges and transient if it converges.

Exercise 2.3.1 Let $N_{\mathbf{xy}}$ be the expected number of visits to $\mathbf{y}$ for a random walker starting in $\mathbf{x}$. Show that $N_{\mathbf{xy}}$is finite if and only if the walk is transient.

2.3.2 Simple random walk in one dimension

Consider a random walker in one dimension, started at $\mathbf{0}$. To return to $\mathbf{0}$, the walker must take the same number of steps to the right as to the left; hence, only even times are possible. Let us compute u_{2n}. Any path that returns in $2n$ steps has probability $1/2^n$. The number of possible paths equals the number of ways that we can choose the n times to go right from the $2n$ possible times. Thus

$$u_{2n} = \binom{2n}{n} \frac{1}{2^{2n}}.$$

We shall show that $\sum_n u_{2n} = \infty$ by using Stirling's approximation:

$$n! \sim \sqrt{2\pi n} e^{-n} n^n.$$

Thus

$$u_{2n} = \frac{(2n!)}{n!n!} \frac{1}{2^{2n}} \sim \frac{\sqrt{2\pi \cdot 2n} e^{-2n} (2n)^{2n}}{(\sqrt{2\pi n} e^{-n} n^n)^2 2^{2n}} = \frac{1}{\sqrt{\pi n}}.$$

Therefore,

$$\sum_n u_{2n} \sim \sum_n \frac{1}{\sqrt{\pi n}} = \infty$$

and a simple random walk in one dimension is recurrent.

Recall that this case was trivial by the resistor argument.

Exercise 2.3.2 We showed in Section 1.1.5 that a random walker starting at x with $0 < x < N$ has probability x/N of reaching N before 0. Use this to show that a simple random walk in one dimension is recurrent.

Exercise 2.3.3 Consider a random walk in one dimension that moves from n to $n+1$ with probability p and to $n-1$ with probability $q = 1-p$. Assume that $p > 1/2$. Let h_x be the probability, starting at x, that the walker ever reaches 0. Use Exercise 1.1.9 to show that $h_x = (q/p)^x$ for $x \geq 0$ and $h_x = 1$ for $x < 0$. Show that this walk is transient.

Exercise 2.3.4 For a simple random walk in one dimension, it follows from Exercise 1.1.7 that the expected time, for a walker starting at x with $0 < x < N$, to reach 0 or n is $x(N-x)$. Prove that for the infinite walk, the expected time to return to 0 is infinite.

Exercise 2.3.5 Let us regard a simple random walk in one dimension as the fortune of a player in a penny matching game where the players have unlimited credit. Show that the result is a martingale (see Section 1.1.6). Show that you can describe a stopping system that guarantees that you make money.

2.3.3 Simple random walk in two dimensions

For a random walker in two dimensions to return to the origin, the walker must have gone the same number of times north and south and the same number of times east and west. Hence, again, only even times for return are possible. Every path that returns in $2n$ steps has probability $1/4^{2n}$. The number of paths that do this by taking k steps to the north, k south, $n-k$ east and $n-k$ west is

$$\binom{2n}{k, k, n-k, n-k} = \frac{2n!}{k!k!(n-k)!(n-k)!}.$$

Thus

$$\begin{aligned} u_{2n} &= \frac{1}{4^{2n}} \sum_{k=0}^{n} \frac{(2n)!}{k!k!(n-k)!(n-k)!} \\ &= \frac{1}{4^{2n}} \sum_{k=0}^{n} \frac{(2n)!}{n!n!} \frac{n!n!}{k!k!(n-k)!(n-k)!} \\ &= \frac{1}{4^{2n}} \binom{2n}{n} \sum_{k=0}^{n} \binom{n}{k}^2. \end{aligned}$$

But $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$ (see Exercise 2.3.6). Hence

$$u_{2n} = \left(\frac{1}{2^{2n}} \binom{2n}{n} \right)^2.$$

This is just the square of the one dimension result (not by accident, either—see Section 2.3.6). Thus we have

$$m = \sum_n u_{2n} \approx \sum_n \frac{1}{\pi n} = \infty.$$

Recall that the resistor argument was very simple in this case also.

Exercise 2.3.6 Show that $\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$ (Hint: Think of choosing n balls from a box that has $2n$ balls, n black and n white.)

2.3.4 Simple random walk in three dimensions

For a walker in three dimensions to return, the walker must take an equal number of steps back and forth in each of the three different directions. Thus we have

$$u_{2n} = \frac{1}{6^{2n}} \sum_{j,k} \frac{(2n)!}{j!j!k!k!(n-j-k)!(n-j-k)!}$$

where the sum is taken over all j, k with $j + k \leq n$. Following Feller [5], we rewrite this sum as

$$u_{2n} = \frac{1}{2^{2n}} \binom{2n}{n} \sum_{j,k} \left(\frac{1}{3^n} \frac{n!}{j!k!(n-j-k)!} \right)^2.$$

Now consider placing n balls randomly into three boxes, A, B, C. The probability that j balls fall in A, k in B, and $n-j-k$ in C is

$$\frac{1}{3^n}\binom{n}{j,k,n-j-k} = \frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!}.$$

Intuitively, this probability is largest when j, k, and $n-j-k$ are as near as possible to $n/3$, and this can be proved (see Exercise 2.3.7). Hence, replacing one of the factors in the square by this larger value, we have:

$$u_{2n} \leq \frac{1}{2^{2n}}\binom{2n}{n}\left(\frac{1}{3^n}\frac{n!}{\left\lfloor\frac{n}{3}\right\rfloor!\left\lfloor\frac{n}{3}\right\rfloor!\left\lfloor\frac{n}{3}\right\rfloor!}\right)\left(\sum_{j,k}\frac{1}{3^n}\frac{n!}{j!k!(n-j-k)!}\right),$$

where $\lfloor n/3 \rfloor$ denotes the greatest integer $\leq n/3$. The last sum is 1 since it is the sum of all probabilities for the outcomes of putting n balls into three boxes. Thus

$$u_{2n} \leq \frac{1}{2^{2n}}\binom{2n}{n}\left(\frac{1}{3^n}\frac{n!}{\left\lfloor\frac{n}{3}\right\rfloor!^3}\right).$$

Applying Stirling's approximation yields

$$u_{2n} \leq \frac{K}{n^{3/2}}$$

for suitable constant K. Therefore

$$m = \sum_n u_{2n} \leq K\sum_n \frac{1}{n^{3/2}} < \infty,$$

and a simple random walk in three dimensions is recurrent.

While this is a complex calculation, the resistor argument was also complicated in this case. We will try to make amends for this presently.

Exercise 2.3.7 Prove that $\binom{n}{j,k,n-j-k}$ is largest when j, k, and $n-j-k$ are as close as possible to $n/3$.

Exercise 2.3.8 Find an appropriate value for the "suitable constant" K that was mentioned above, and derive an upper bound for m. Use this to get a lower bound for the probability of escape for simple random walk in three dimensions.

2.3.5 The probability of return in three dimensions: exact calculations

Since the probability of return in three dimensions is less than one, it is natural to ask, "What is this probability?" For this we need an exact calculation. The first such calculation was carried out in a pioneering paper of McCrea and Whipple [22]. The solution outlined here follows Feller [5], Exercise 28, Chapter 9, and Montroll and West [23].

Let $p(a, b, c; n)$ be the probability that a random walker, starting at $\mathbf{0}$, is at (a, b, c) after n steps. Then $p(a, b, c; n)$ is completely determined by the fact that

$$p(0, 0, 0; 0) = 1$$

and

$$\begin{aligned} p(a, b, c; n) &= \frac{1}{6}p(a-1, b, c; n-1) + \frac{1}{6}p(a+1, b, c; n-1) + \\ &\quad \frac{1}{6}p(a, b-1, c; n-1) + \frac{1}{6}p(a, b+1, c; n-1) + \\ &\quad \frac{1}{6}p(a, b, c-1; n-1) + \frac{1}{6}p(a, b, c+1; n-1). \end{aligned}$$

Using the technique of generating functions, it is possible to derive a solution of these equations as

$$\begin{aligned} & p(a, b, c; n) \\ = \; & \frac{1}{(2\pi)^3} \cdot \\ & \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \left(\frac{\cos x + \cos y + \cos z}{3}\right)^n \cos(xa)\cos(yb)\cos(zc)dxdydz. \end{aligned}$$

Of course, we can just verify that this formula satisfies our equations once someone has suggested it. Having this formula, we put $a = b = c = 0$ and sum over n to obtain the expected number of returns as

$$m = \frac{3}{(2\pi)^3}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{1}{3 - (\cos x + \cos y + \cos z)}dxdydz.$$

This integral was first evaluated by Watson [35] in terms of elliptic integrals, which are tabulated. A simpler result was obtained by Glasser and Zucker [8] who evaluated this integral in terms of gamma functions. Using this result, we get

$$m = \frac{\sqrt{6}}{32\pi^3}\Gamma\left(\frac{1}{24}\right)\Gamma\left(\frac{5}{24}\right)\Gamma\left(\frac{7}{24}\right)\Gamma\left(\frac{11}{24}\right) = 1.516386059137\ldots,$$

where

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

is Euler's gamma function. (Incidentally, the value given by Glasser and Zucker [8] for the integral above needs to be corrected by multiplying by a factor of $1/(384\pi)$.)

Recall that $m = 1/(1-u)$ so that $u = (m-1)/m$. This gives

$$u = .340537329544\ldots.$$

2.3.6 Simple random walk in two dimensions is the same as two independent one-dimensional random walks

We observed that the probability of return at time $2n$ in two dimensions is the square of the corresponding probability in one dimension. Thus it is the same as the probability that two independent walkers, one walking in the x direction and the other in the y direction, will, at time $2n$, both be at 0. Can we see that this should be the case? The answer is yes. Just change our axes by 45 degrees to new axes $\bar{x}$ and $\bar{y}$ as in Figure 60.

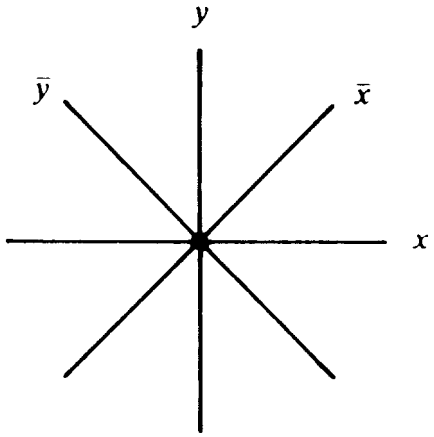


Figure 60: ♣

Look at the possible outcomes for the first step using the x, y coor-

dinates and the $\bar{x}, \bar{y}$ coordinates. We have

x, y coordinates	$\bar{x}, \bar{y}$ coordinates
$(0, 1)$	$(1/\sqrt{2}, 1/\sqrt{2})$
$(0, -1)$	$(-1/\sqrt{2}, -1/\sqrt{2})$.
$(1, 0)$	$(1/\sqrt{2}, -1/\sqrt{2})$
$(-1, 0)$	$(-1/\sqrt{2}, 1/\sqrt{2})$

Assume that we have two independent walkers, one moving with step size $\frac{1}{\sqrt{2}}$ randomly along the $\bar{x}$ axis and the other moving with the same step size along the $\bar{y}$ axis. Then, if we plot their positions using the x, y axes, the four possible outcomes for the first step would agree with those given in the second column of the table above. The probabilities for each of the four pairs of outcomes would also be $(1/2) \cdot (1/2) = 1/4$. Thus, we cannot distinguish a simple random walk in two dimensions from two independent walkers along the $\bar{x}$ and $\bar{y}$ axes making steps of magnitude $1/\sqrt{2}$. Since the probability of return does not depend upon the magnitude of the steps, the probability that our two independent walkers are at $(0, 0)$ at time $2n$ is equal to the product of the probabilities that each separately is at 0 at time $2n$, namely $(1/2^{2n})\binom{2n}{n}$. Therefore, the probability that the standard walk will be at $(0, 0)$ at time $2n$ is $((1/2^{2n})\binom{2n}{n})^2$ as observed earlier.

2.3.7 Simple random walk in three dimensions is not the same as three independent random walks

In three dimensions, the probability that three independent walkers are each back to 0 after time $2n$ is

$$u_{2n} = \left(\binom{2n}{n} \frac{1}{2^{2n}} \right)^3 .$$

This does not agree with our result for a simple random walk in three dimensions. Hence, the same trick cannot work. However, it is interesting to consider a random walk which is the result of three independent walkers. Let (i, j, k) be the position of three independent random walkers. The next position is one of the eight possibilities $(i \pm 1, j \pm 1, k \pm 1)$ Thus we may regard their progress as a random walk on the lattice points in three dimensions. If we center a cube of side 2 at (i, j, k), then the walk moves with equal probabilities to each of the eight corners of the cube. It is easier to show that this random walk is transient

(using classical methods) than it is for simple random walk. This is because we can again use the one-dimension calculation. The probability u_{2n} for return at time $2n$ is

$$u_{2n} = \left(\binom{2n}{n} \frac{1}{2^{2n}} \right)^3 \sim \left(\frac{1}{\sqrt{\pi n}} \right)^3 .$$

Thus

$$m = \sum_n u_{2n} \sim \sum_n \left(\frac{1}{\pi n} \right)^{3/2} < \infty$$

and the walk is transient.

The fact that this three independent walkers model and simple random walk are of the same type suggests that when two random walks are "really about the same", they should either both be transient or both be recurrent. As we will soon see, this is indeed the case. Thus we may infer the transience of simple random walk in 3 dimensions from the transience of the three independent walkers model without going through the involved calculation of Section 2.3.4.

2.4 Random walks on more general infinite networks

2.4.1 Random walks on infinite networks

From now on we assume that G is an infinite connected graph. We assume that it is of *bounded degree*, by which we mean that there is some integer E such that the number of edges from any point is at most E. We assign to each edge xy of G a conductance $C_{xy} > 0$ with $R_{xy} = \frac{1}{C_{xy}}$. The graph G together with the conductances $\mathbf{C} = (C_{xy})$ is called a *network* and denoted by $(G, \mathbf{C})$. Given a network $(G, \mathbf{C})$, we define a random walk by

$$P_{xy} = \frac{C_{xy}}{C_x}$$

where $C_x = \sum_y C_{xy}$. When all the conductances are equal, we obtain a random walk that moves along each edge with the same probability: In agreement with our previous terminology, we call this walk *simple random walk* on G.

We have now a quite general class of infinite-state Markov chains. As in the case of finite networks, the chains are reversible Markov chains: That is, there is a positive vector $\mathbf{w}$ such that $w_x P_{xy} = w_y P_{yx}$. As in the finite case, we can take $w_x = C_x$, since $C_x P_{xy} = C_{xy} = C_{yx} = C_y P_{yx}$.

2.4.2 The type problem

Let $(G, \mathbf{C})$ be an infinite network with random walk $\mathbf{P}$. Let $\mathbf{0}$ be a reference point. Let p_{esc} be the probability that a walk starting at $\mathbf{0}$ will never return to $\mathbf{0}$. If $p_{\text{esc}} = 0$ we say that $\mathbf{P}$ is *recurrent*, and if $p_{\text{esc}} > 0$ we say that it is *transient*. You are asked to show in Exercise 2.4.1 that the question of recurrence or transience of $\mathbf{P}$ does not depend upon the choice of the reference point. The *type problem* is the problem of determining if a random walk (network) is recurrent or transient.

In Section 2.1.5 we showed how to rephrase the type problem for a lattice in terms of finite graphs sitting inside it. In Section 2.1.6 we showed that the type problem is equivalent to an electrical network problem by showing that simple random walk on a lattice is recurrent if and only if the lattice has infinite resistance to infinity. The same arguments apply with only minor modifications to the more general infinite networks as well. This means that we can use Rayleigh's short-cut method to determine the type of these more general networks.

Exercise 2.4.1 Show that the question of recurrence or transience of $\mathbf{P}$ does not depend upon the choice of the reference point.

2.4.3 Comparing two networks

Given two sets of conductances $\mathbf{C}$ and $\bar{\mathbf{C}}$ on G, we say that $(G, \bar{\mathbf{C}}) < (G, \mathbf{C})$ if $\bar{C}_{xy} < C_{xy}$ for all xy, or equivalently, if $\bar{R}_{xy} > R_{xy}$ for all xy. Assume that $(G, \bar{\mathbf{C}}) < (G, \mathbf{C})$. Then by the Monotonicity Law, $\bar{R}_{\text{eff}} \geq R_{\text{eff}}$. Thus if random walk on $(G, \bar{\mathbf{C}})$ is transient, i.e., if $\bar{R}_{\text{eff}} < \infty$, then random walk on $(G, \mathbf{C})$ is also transient. If random walk on $(G, \mathbf{C})$ is recurrent, i.e., if $R_{\text{eff}} = \infty$, then random walk on $(G, \bar{\mathbf{C}})$ is also recurrent.

Theorem. If $(G, \mathbf{C})$ and $(G, \bar{\mathbf{C}})$ are networks, and if there exist constants u, v with $0 < u \leq v < \infty$ such that

$$uC_{xy} \leq \bar{C}_{xy} \leq vC_{xy}$$

for all x and y, then random walk on $(G, \bar{\mathbf{C}})$ is of the same type as random walk on $(G, \mathbf{C})$.

Proof. Let $U_{xy} = uC_{xy}$ and $V_{xy} = vC_{xy}$. Then $(G, \mathbf{U}) \leq (G, \bar{\mathbf{C}}) \leq (G, \mathbf{V})$. But the random walks for $(G, \mathbf{U})$ and $(G, \mathbf{V})$ are the same as random walk on $(G, \mathbf{C})$. Thus random walk for $(G, \bar{\mathbf{C}})$ is of the same type as random walk on $(G, \mathbf{C})$. $\diamondsuit$

Corollary. Let $(G, \mathbf{C})$ be a network. If for every edge xy of G we have $0 < u < C_{xy} < v < \infty$ for some constants u and v, then the random walk on $(G, \mathbf{C})$ has the same type as simple random walk on G.

Exercise 2.4.2 Consider the two-dimensional lattice. For each edge, we toss a coin to decide what kind of resistor to put across this edge. If heads turns up, we put a two-ohm resistor across this edge; if tails turns up, we put a one-ohm resistor across the edge. Show that the random walk on the resulting network is recurrent.

Exercise 2.4.3 Consider the analogous problem to Exercise 2.4.2 in 3 dimensions.

2.4.4 The k-fuzz of a graph

For any integer k, the *k-fuzz* of a graph G is the graph G_k obtained from G by adding an edge xy if it is possible to go from x to y in at most k steps. For example, the 2-fuzz of the two-dimensional lattice is shown in Figure 61; please note that horizontal and vertical edges of length 2, such as those joining $(0, 0)$ to $(0, 2)$, have not been indicated.

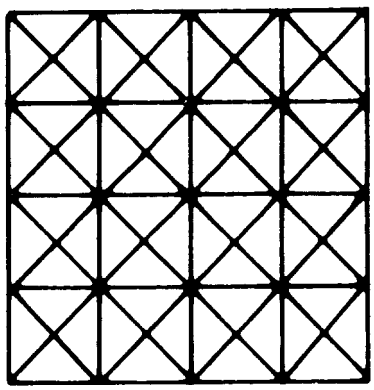


Figure 61: ♣

Theorem. Simple random walk on G and on the k-fuzz G_k of G have the same type.

Proof. Let $\mathbf{P}$ be simple random walk on G. Define $\bar{\mathbf{P}} = (\mathbf{P} + \mathbf{P}^2 + \ldots + \mathbf{P}^k)/k$. Then $\bar{\mathbf{P}}$ may be considered to be $\mathbf{P}$, watched at one of the first k steps chosen at random, then at a time chosen at random from the next k steps after this time, etc. Thinking of $\bar{\mathbf{P}}$ in this way, we see that $\mathbf{P}$ is in state $\mathbf{0}$ at least once for every time $\bar{\mathbf{P}}$ in state $\mathbf{0}$. Hence, if $\bar{\mathbf{P}}$ is recurrent so is $\mathbf{P}$. Assume now that $\bar{\mathbf{P}}$ is transient. Choose a finite set S so that $\mathbf{0}$ cannot be reached in k steps from a point outside of S. Then, since the walk $\bar{\mathbf{P}}$ will be outside S from some time on, the walk $\mathbf{P}$ cannot be at $\mathbf{0}$ after this time, and $\mathbf{P}$ is also transient. Therefore, $\mathbf{P}$ and $\bar{\mathbf{P}}$ are of the same type.

Finally, we show that $\bar{\mathbf{P}}$ has the same type as simple random walk on G_k. Here it is important to remember our restriction that G is of bounded degree, so that for some E no vertex has degree $> E$. We know that $\mathbf{P}$ is reversible with $\mathbf{wP} = \mathbf{w}$, where w_x is the number of edges coming out of x. From its construction, $\bar{\mathbf{P}}$ is also reversible and $\mathbf{w}\bar{\mathbf{P}} = \mathbf{w}$. $\bar{\mathbf{P}}$ is the random walk on a network $(G_k, \bar{\mathbf{C}})$ with $\bar{C}_{xy} = w_x \bar{P}_{xy}$. If $\bar{P}_{xy} > 0$, there is a path $x, x_1, x_2, \ldots, x_{m-1}, y$ in G from x to y of length $m \leq k$. Then

$$\bar{P}_{xy} \geq \frac{1}{k}(\frac{1}{E})^m \geq \frac{1}{k}(\frac{1}{E})^k.$$

Thus

$$0 < \frac{1}{k}(\frac{1}{E})^k \leq \bar{P}_{xy} \leq 1$$

and

$$0 < \frac{1}{k}(\frac{1}{E})^k \leq \bar{C}_{xy} \leq E.$$

Therefore, by the theorem on the irrelevance of bounded twiddling proven in Section 2.4.3, $\bar{\mathbf{P}}$ and simple random walk on G_k are of the same type. So G and G_k are of the same type.

NOTE: This is the only place where we use probabilistic methods of proof. For the purist who wishes to avoid probabilistic methods, Exercise 2.4.9 indicates an alternative electrical proof.

We show how this theorem can be used. We say that a graph G can be *embedded* in a graph $\overline{G}$ if the points x of G can be made to correspond in a one-to-one fashion to points $\bar{x}$ of $\overline{G}$ in such a way that if xy is an edge in G, then $\bar{x}\bar{y}$ is an edge in $\overline{G}$.

Theorem. If simple random walk on G is transient, and if $\overline{G}$ can be embedded in a k-fuzz $\overline{G}_k$ of $\overline{G}$ then simple random walk on $\overline{G}$ is also transient. Simple random walk on G and $\overline{G}$ are of the same type if each graph can be embedded in a k-fuzz of the other graph.

Proof. Assume that simple random walk on G is transient and that G can be embedded in a k-fuzz $\overline{G}_k$ of $\overline{G}$. Since R_{eff} for G is finite and G can be embedded in $\overline{G}_k$, R_{eff} for $\overline{G}_k$ is finite. By our previous theorem, the same is true for $\overline{G}$ and simple random walk on $\overline{G}$ is transient.

If we can embed G in $\overline{G}_k$ and $\overline{G}$ in G_k, then the random walk on G is transient if and only if the random walk on $\overline{G}$ is. $\diamondsuit$

Exercise 2.4.4 We have assumed that there is a bound E for the number of edges coming out of any point. Show that if we do not assume this, it is not necessarily true that G and G_k are of the same type. (Hint: Consider a network something like that shown in Figure 62.)

Figure 62: $\clubsuit$

2.4.5 Comparing general graphs with lattice graphs

We know the type of simple random walk on a lattice $\mathbf{Z}^d$. Thus to determine the type of simple random walk on an arbitrary graph G, it is natural to try to compare G with $\mathbf{Z}^d$. This is feasible for graphs that can be drawn in some Euclidean space $\mathbf{R}^d$ in a civilized manner.

Definition. A graph G can be drawn in a Euclidean space $\mathbf{R}^d$ in a *civilized manner* if its vertices can be embedded in $\mathbf{R}^d$ so that for some $r < \infty$, $s > 0$

(a) The length of each edge is $\leq r$.

(b) The distance between any two points is $> s$.

Note that we make no requirement about being able to draw the edges of G so they don't intersect.

Theorem. If a graph can be drawn in $\mathbf{R}^d$ in a civilized manner, then it can be embedded in a k-fuzz of the lattice $\mathbf{Z}^d$.

Proof. We carry out the proof for the case $d = 2$. Assume that G can be drawn in a civilized manner in $\mathbf{R}^2$. We want to show that G can be embedded in a k-fuzz of $\mathbf{Z}^2$. We have been thinking of $\mathbf{Z}^2$ as

being drawn in $\mathbf{R}^2$ with perpendicular lines and adjacent points a unit distance apart on these lines, but this embedding is only one particular way of representing $\mathbf{Z}^2$. To emphasize this, let's talk about L^2 instead of $\mathbf{Z}^2$. Figure 63 shows another way of drawing L^2 in $\mathbf{R}^2$. From a

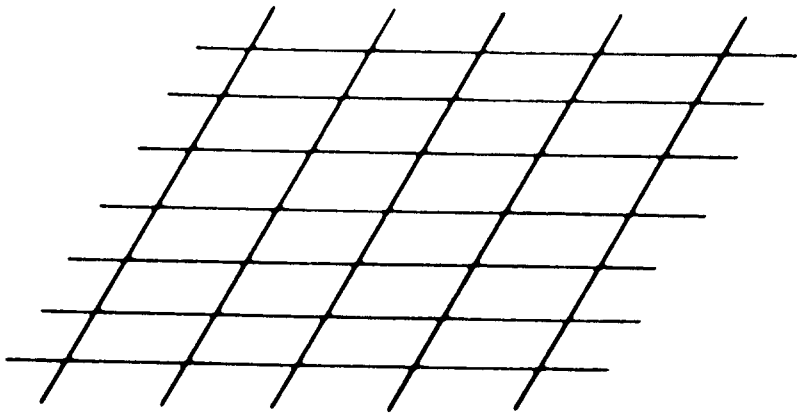


Figure 63: ♣

graph-theoretical point of view, this is the same as $\mathbf{Z}^2$. In trying to compare G to L, we take advantage of this flexibility by drawing L^2 so small that points of G can be moved onto points of L^2 without bumping into each other.

Specifically, let L^2 be a two-dimensional rectangular lattice with lines a distance $s/2$ apart. In any square of L^2, there is at most one point of G. Move each point x of G to the southwest corner $\bar{x}$ of the square that it is in, as illustrated in Figure 64.

Now since any two adjacent points x, y in G were within r of each other in $\mathbf{R}^2$, the corresponding points $\bar{x}$, $\bar{y}$ in L^2 will have Euclidean distance $< r + 2s$. Choose k so that any two points of L^2 whose Euclidean distance is $< r + 2s$ can be connected by a path in L^2 of at most k steps. Then $\bar{x}$ and $\bar{y}$ will be adjacent in L^2_k and—since the prescription for k does not depend on x and y—we have embedded G in the k-fuzz L^2_k.

Corollary. If G can be drawn in a civilized manner in $\mathbf{R}^1$ or $\mathbf{R}^2$, then simple random walk on G is recurrent.

Proof. Assume, for example, that G can be drawn in a civilized manner in $\mathbf{R}^2$. Then G can be embedded in a k-fuzz $\mathbf{Z}^2_k$ of $\mathbf{Z}^2$. If simple random walk on G were transient, then the same would be true for $\mathbf{Z}^2_k$

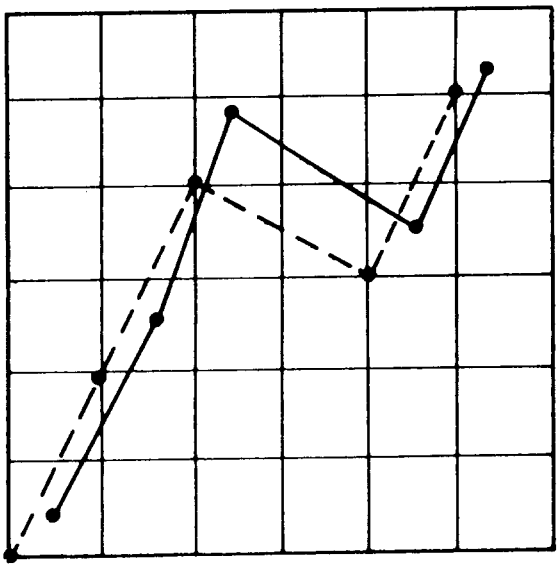


Figure 64: ♣

and $\mathbf{Z}^2$. But we know that simple random walk on $\mathbf{Z}^2$ is recurrent. Thus simple random walk on G is recurrent. $\diamondsuit$

Our first proof that random walk in three dimensions is transient consisted in showing that we could embed a transient tree in $\mathbf{Z}^3$. We now know that it would have been sufficient to show how to draw a transient tree in $\mathbf{R}^3$ in a civilized manner: This is easier (see Exercise 2.4.5).

The corollary implies that simple random walk on any sufficiently symmetrical graph in $\mathbf{R}^2$ is recurrent. For example, simple random walk on the regular graph made up of hexagons shown in Figure 65 is recurrent.

We can even consider very irregular graphs. For example, on the cover of the January 1977 Scientific American, there is an example due to Conway of an infinite non-periodic tiling using Penrose tiles of the form shown in Figure 66. It is called the cartwheel pattern; part of it is shown in Figure 67. A walker walking randomly on the edges of this very irregular infinite tiling will still return to his or her starting point.

Assume now that G can be drawn in a civilized manner in $\mathbf{R}^3$. Then to show that simple random walk on G is of the same type as $\mathbf{Z}^3$, namely transient, it is sufficient to show that we can embed $\mathbf{Z}^3$ in a k-fuzz of

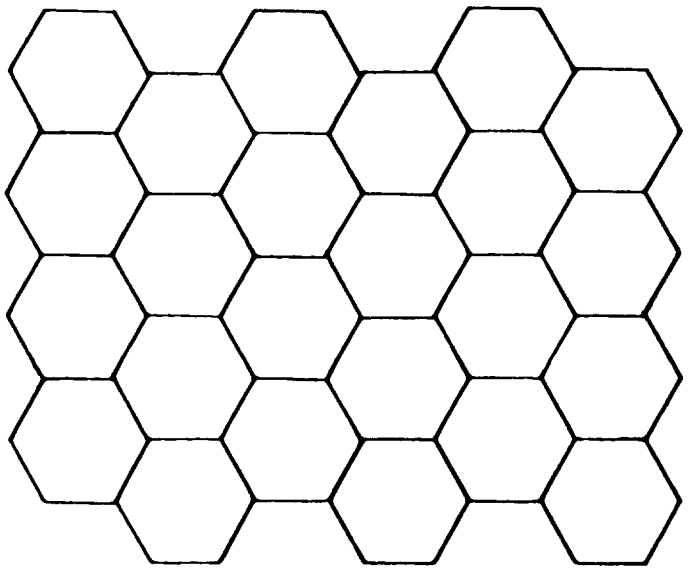


Figure 65: ♣

Figure 66: ♣

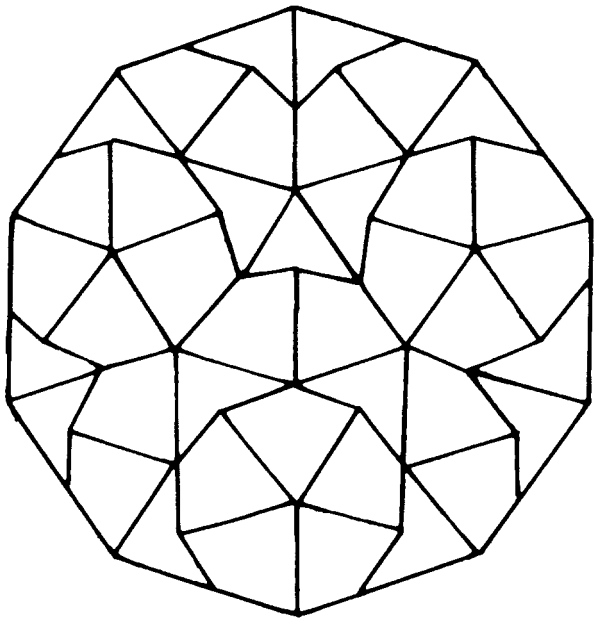


Figure 67: ♣

G. This is clearly possible for any regular lattice in $\mathbf{R}^3$. The three lattices that have been most studied and for which exact probabilities for return have been found are called the SC, BCC, and FCC lattices. The SC (simple cubic) lattice is just $\mathbf{Z}^3$. The walker moves each time to a new point by adding a random choice from the six vectors

$$(\pm 1, 0, 0), (0, \pm 1, 0), (0, 0, \pm 1).$$

For the BCC (body-centered cubic) lattice, the choice is one of the eight vectors

$$(\pm 1, \pm 1, \pm 1).$$

This was the walk that resulted from three independent one-dimensional walkers. For the FCC (face-centered cubic) lattice, the random choice is made from the twelve vectors

$$(\pm 1, \pm 1, 0), (\pm 1, 0, \pm 1), (0, \pm 1, \pm 1).$$

For a discussion of exact calculations for these three lattices, see Montroll and West [23]

As we have seen, once the transience of any one of these three walks is established, no calculations are necessary to determine that the other

walks are transient also. Thus we have yet another way of establishing Polya's theorem in three dimensions: Simply verify transience of the walk on the BCC lattice via the simple three-independent-walkers computation, and infer that walk on the SC lattice is also transient since the BCC lattice can be embedded in a k-fuzz of it.

Exercise 2.4.5 When we first set out to prove Polya's theorem for $d = 3$, our idea was to embed NT_3 in $\mathbf{Z}^3$. As it turned out, what we ended up embedding was not NT_3 but $\mathrm{NT}_{2.5849\ldots}$, and we didn't quite embed it at that. We tried to improve the situation by finding (in Exercise 2.2.7) an honest-to-goodness embedding of a relative of $\mathrm{NT}_{2.5849\ldots}$, but NT_3 was still left completely out in the cold. Now, however, we are in a position to embed NT_3, if not in $\mathbf{Z}^3$ then at least in a k-fuzz of it. All we need to do is to draw NT_3 in $\mathbf{R}^3$ in a civilized manner. Describe how to do this, and thereby give one more proof of Polya's theorem for $d = 3$.

Exercise 2.4.6 Find a graph that can be embedded in a civilized manner in $\mathbf{R}^3$ but not in $\mathbf{R}^2$, but is nonetheless recurrent.

Exercise 2.4.7 Assume that G is drawn in a civilized manner in $\mathbf{R}^3$. To show that simple random walk on G is transient, it is enough to know that $\mathbf{Z}^3$ can be embedded in a k-fuzz of G. Try to come up with a nice condition that will guarantee that this is possible. Can you make this condition simple, yet general enough so that it will settle all reasonably interesting cases? In other words, can you make the condition nice enough to allow us to remember only the condition, and forget about the general method lying behind it?

2.4.6 Solving the type problem by flows: a variant of the cutting method

In this section we will introduce a variant of the cutting method whereby we use Thomson's Principle directly to estimate the effective resistance of a conductor.

Thomson's Principle says that, given any unit flow through a resistive medium, the dissipation rate of that flow gives an upper bound for the effective resistance of the medium. This suggests that to show that a given infinite network is transient, it should be enough to produce a unit flow out to infinity having finite energy dissipation.

In analogy with the finite case, we say that $\mathbf{j}$ is a *flow from* $\mathbf{0}$ *to infinity* if

(a) $j_{xy} = j_{yx}$.

(b) $\sum_y j_{xy} = 0$ if $x \neq \mathbf{0}$.

We define $j_{\mathbf{0}} = \sum_y j_{\mathbf{0}y}$. If $j_{\mathbf{0}} = 1$, we say that $\mathbf{j}$ is a *unit flow to infinity*. Again in analogy with the finite case, we call $\frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy}$ the *energy dissipation* of the flow $\mathbf{j}$.

Theorem. The effective resistance R_{eff} from $\mathbf{0}$ to ∞ is less than or equal to the energy dissipation of any unit flow from $\mathbf{0}$ to infinity.

Proof. Assume that we have a unit flow $\mathbf{j}$ from $\mathbf{0}$ to infinity with energy dissipation

$$E = \frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy}.$$

We claim that $R_{\text{eff}} \leq E$. Restricting j_{xy} to the edges of the finite graph $G^{(r)}$, we have a unit flow from $\mathbf{0}$ to $\partial G^{(r)}$ in $G^{(r)}$. Let $i^{(r)}$ be the unit current flow in $G^{(r)}$ from 0 to $\partial G^{(r)}$. By the results of Section 1.3.5,

$$\mathbf{R}_{\text{eff}}^{(r)} = \frac{1}{2}\sum_{G^{(r)}} (i_{xy}^{(r)})^2 R_{xy} \leq \frac{1}{2}\sum_{G^{(r)}} j_{xy}^2 R_{xy} \leq \frac{1}{2}\sum_{x,y} j_{xy}^2 R_{xy} = E,$$

where $\sum_{G^{(r)}}$ indicates the sum over all pairs x, y such that xy is an edge of $G^{(r)}$.

Exercise 2.4.8 We have billed the method of using Thomson's Principle directly to estimate the effective resistances of a network as a variant of the cutting method. Since the cutting method was derived from Thomson's Principle, and not vice versa, it would seem that we have got the cart before the horse. Set this straight by giving an informal ("heuristic") derivation of Thomson's Principle from the cutting method. (Hint: see Maxwell [21], Chapter VIII, Paragraph 307.) For more on this question, see Onsager [25].

Exercise 2.4.9 Let G be an infinite graph of bounded degree and G_k the k-fuzz of G. Using electric network arguments, show that $R_{\text{eff}} < \infty$ for G if and only if $R_{\text{eff}} < \infty$ for G_k.

2.4.7 A proof, using flows, that simple random walk in three dimensions is transient

We now apply this form of the cutting method to give another proof that simple random walk on the threedimensional lattice is transient. All we need is a flow to infinity with finite dissipation. The flow we are going to describe is not the first flow one would think of. In case you are curious, the flow described here was constructed as a side effect of an unsuccessful attempt to derive the isoperimetric inequality (see Polya [27]) from the "max-flow min-cut" theorem (Ford and Fulkerson [7]). The idea is to find a flow in the positive orthant having the property that the same amount flows through all points at the same distance from $\mathbf{0}$.

Again, it is easiest to show the construction for the two-dimensional case. Let G denote the part of $\mathbf{Z}^2$ lying in the first quadrant. The graph $G^{(4)}$ is shown in Figure 68.

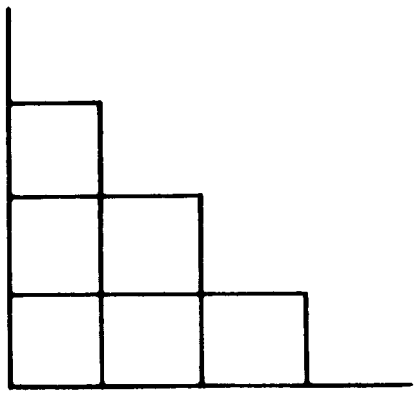


Figure 68: ♣

We choose our flow so that it always goes away from $\mathbf{0}$. Into each point that is not on either axis there are two flows, one vertical and one horizontal. We want the sum of the corresponding values of j_{xy} to be the same for all points the same distance from $\mathbf{0}$. These conditions completely determine the flow. The flow out of the point (x, y) with $x + y = n$ is as shown in Figure 69. The values for the currents out to the fourth level are shown in Figure 70. In general, the flow out of a

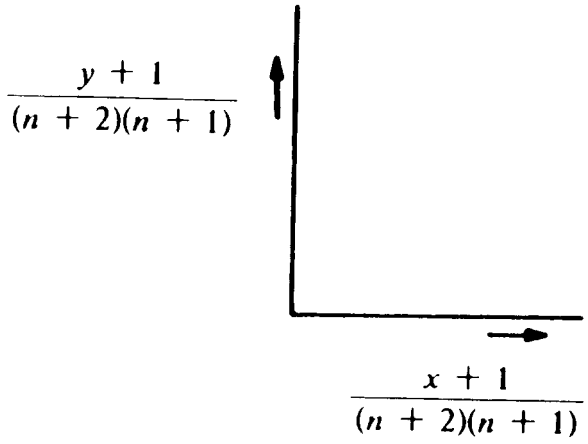


Figure 69: ♣

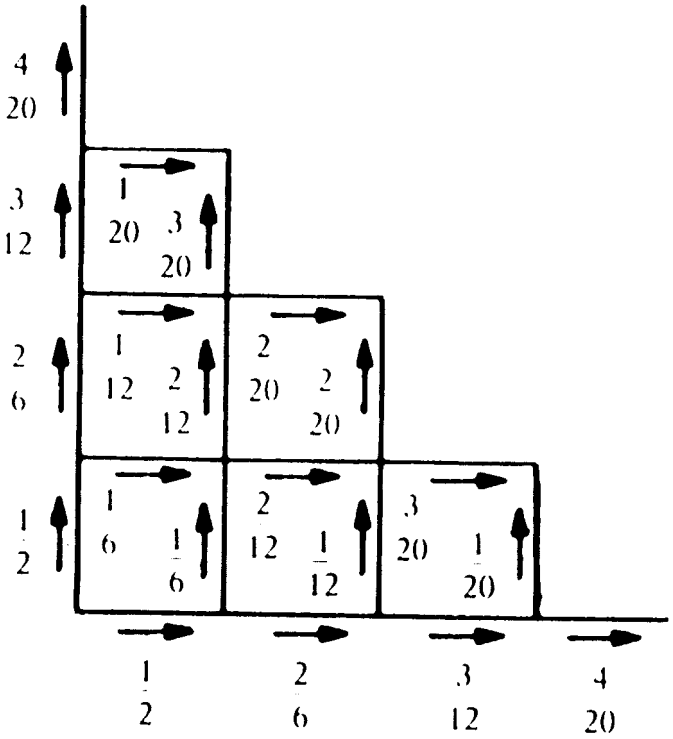


Figure 70: ♣

point (x, y) with $x + y = n$ is

$$\frac{x+1}{(n+2)(n+1)} + \frac{y+1}{(n+2)(n+1)} = \frac{1}{n+1}$$

and the flow into this point is

$$\frac{x}{n(n+1)} + \frac{y}{n(n+1)} = \frac{1}{n+1}$$

Thus the net flow at (x, y) is 0. The flow out of $\mathbf{0}$ is $(1/2) + (1/2) = 1$. For this two-dimensional flow, the energy dissipation is infinite, as it would have to be. For three dimensions, the uniform flow is defined as follows: Out of (x, y, z) with $x + y + z = n$ we have the flow indicated in Figure 71. The total flow out of (x, y, z) is then

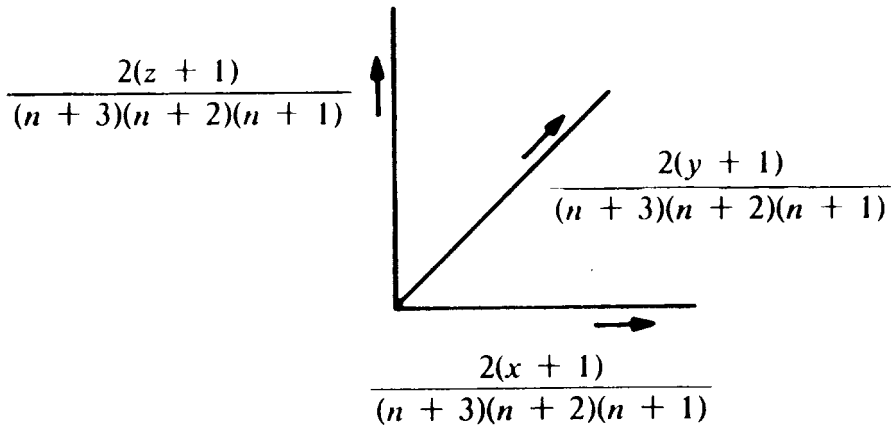


Figure 71: ♣

$$\begin{aligned} & \frac{2(x+1)}{(n+3)(n+2)(n+1)} + \frac{2(y+1)}{(n+3)(n+2)(n+1)} + \frac{2(z+1)}{(n+3)(n+2)(n+1)} \\ = \quad & \frac{2}{(n+2)(n+1)}. \end{aligned}$$

The flow into (x, y, z) comes from the points $(x-1, y, z)$, $(x, y-1, z)$, $(x, y, z-1)$ and, hence, the total flow into (x, y, z) is

$$\frac{2x}{(n+2)(n+1)n} + \frac{2y}{(n+2)(n+1)n} + \frac{2z}{(n+2)(n+1)n} = \frac{2}{(n+2)(n+1)}.$$

Thus the net flow for (x, y, z) is 0. The flow out of $\mathbf{0}$ is $(1/3) + (1/3) + (1/3) = 1$ We have now to check finiteness of energy dissipation. The flows coming out of the edges at the nth level are all $\leq 2/(n+1)^2$. There are $(n+1)(n+2)/2$ points a distance n from $\mathbf{0}$, and thus there are $(3/2)(n+1)(n+2) \leq 3(n+1)^2$ edges coming out of the nth level. Thus the energy dissipation E has

$$E \leq \sum_n 3(n+1)^2 \left(\frac{2}{(n+1)^2}\right)^2 = 12 \sum_n \frac{1}{(n+1)^2} < \infty,$$

and the random walk is transient.

2.4.8 The end

We have come to the end of our labors, and it seems fitting to look back and try to say what it is we have learned.

To begin with, we have seen how phrasing certain mathematical questions in physical terms allows us to draw on a large body of physical lore, in the form of established methods and ways of thought, and thereby often leads us to the answers to those questions.

In particular, we have seen the utility of considerations involving energy. In took hundreds of years for the concept of energy to emerge and take its rightful place in physical theory, but it is now recognized as perhaps the most fundamental concept in all of physics. By phrasing our probabilistic problems in physical terms, we were naturally led to considerations of energy, and these considerations showed us the way through the difficulties of our problems.

As for Polya's theorem and the type problem in general, we have picked up a bag of tricks, known collectively as "Rayleigh's short-cut method", which we may expect will allow us to determine the type of almost any random walk we are likely to embark on. In the process, we have gotten some feeling for the connection between the dimensionality of a random walk and its type. Furthermore, we have settled one of the main questions likely to occur to someone encountering Polya's theorem, namely: "If two walks look essentially the same, and if one has been shown to be transient, must not the other also be transient?"

Another question likely to occur to someone contemplating Polya's theorem is the question raised in Section 2.1.8: "Since the lattice $\mathbf{Z}^d$ is in some sense a discrete analog of a resistive medium filling all of $\mathbf{R}^d$, should it not be possible to go quickly and naturally from the trivial computation of the resistance to infinity of the continuous medium to

a proof of Polya's theorem?" Our shorting argument allowed us to do this in the two-dimensional case; that leaves the case of three (or more) dimensions. Again, it is considerations of energy that allow us to make this connection. The trick is to start with the flow field that one gets by solving the continuous problem, and adapt it to the lattice, so as to get a lattice flow to infinity having finite dissipation rate. We leave the working out of this as an exercise, so as not to rob readers of the fun of doing it for themselves.

Exercise 2.4.10 Give one final proof of Polya's theorem in 3 dimensions by showing how to adapt the $1/r^2$ radial flow field to the lattice. (Hint: "cubes".)

Acknowledgements

This work is derived from the book *Random Walks and Electric Networks*, originally published in 1984 by the Mathematical Association of American in their Carus Monographs series. We are grateful to the MAA for permitting this work to be freely redistributed under the terms of the GNU General Public License. (See Figure 72.)

Subject: Carus Monograph
From: J. Laurie Snell <jlsnell@Dartmouth.EDU>
Date: 15 October 1999

Dear Don,

Peter and I are delighted that we have the MAA's permission to distribute our Carus Monograph 'Random Walks and Electric Networks' under the terms of the GNU General Public License, as published by the Free Software Foundation. We believe that in allowing the public to distribute and modify this work freely, the MAA will be contributing to the dissemination of mathematical knowledge in a way that is entirely consistent with the goals of the Carus Monograph Fund.

For our records, we would appreciate it if you would sign an return to us the enclosed copy of this letter.

Thanks so much for your help!

Sincerely yours,

J. Laurie Snell

Donald J. Albers
Director of Publications
The Mathematical Association of America
11/23/99

Figure 72: ♣

References

[1] Edwin Abbott. *Flatland*. 1899.

[2] B. Bollobás. *Graph Theory*. 1979.

[3] R. Courant, K. Friedrichs, and H. Lewy. Über die pertiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928.

[4] J. L. Doob. *Stochastic Processes*. 1953.

[5] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume I. 1968.

[6] R. P. Feynman. *The Feynmann Lectures on Physics*. 1964.

[7] L. R. Ford and D. R. Fulkerson. Maximal flow through a network. *Can. J. Math.*, 8:399–404, 1956.

[8] M. L. Glasser and I. J. Zucker. Extended Watson integrals for the cubic lattice. *Proc. Natl. Acad. Sci., USA*, 74:1800–1801, 1977.

[9] D. Griffeath and T. M. Liggett. Critical phenomena for Spitzer's reversible nearest particle systems. *Ann. Probab.*, 10:881–895, 1982.

[10] R. Hersh and R. J. Griego. Brownian motion and potential theory. *Scientific American*, pages 66–74, March 1969.

[11] J. Jeans. *The Mathematical Theory of Electricity and Magnetism, 5th Edition*. 1966.

[12] S. Kakutani. Markov processes and the Dirichlet problem. *Proc. Jap. Acad.*, 21:227–233, 1945.

[13] F. Kelly. *Reversibility and Stochastic Networks*. 1979.

[14] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chains*. 1966.

[15] J. G. Kemeny, J. L. Snell, and G. L. Thompson. *Finite Markov Chains, 3rd Edition*. 1974.

[16] Harry Kesten. *Percolation Theory for Mathematicians*. 1982.

[17] J. F. C. Kingman. Markov population processes. *J. Appl. Prob.*, 6:1–18, 1969.

[18] A. Lehman. A resistor network inequality, Problem 60-5. *SIAM Review*, 4:150–154, 1965.

[19] Paul Lévy. *Théorie de l'addition des variable aléatoires*. 1937.

[20] T. J. Lyons. A simple criterion for transience of a reversible Markov chain. *Ann. Probab.*, 11:393–402, 1983.

[21] J. C. Maxwell. *Treatise on Electricity and Magnetism, 3rd Edition*. 1891.

[22] W. H. McCrea and F. J. W. Whipple. Random paths in two and three dimensions. *Proc. of the Royal Soc. of Edinburgh*, 60:281–298, 1940.

[23] E. W. Montroll and B. J. West. Fluctuation phenomena. In *Studies in Statistical Mathematics*, volume 7, pages 61–175. 1979.

[24] C. St. J. A. Nash-Williams. Random walk and electric currents in networks. *Proc. Camb. Phil. Soc.*, 55:181–194, 1959.

[25] L. Onsager. Reciprocal relations in irreversible processes I. *Phys. Rev.*, 37:405–426, 1931.

[26] G. Polya. Über eine Aufgabe betreffend die Irrfahrt im Strassennetz. *Math. Ann.*, 84:149–160, 1921.

[27] G. Polya. *How to Solve It, 2nd Edition*. 1957.

[28] G. Polya and G. Szego. *Isoperimetric Inequalities of Mathematical Physics*. 1951.

[29] J. W. S. Rayleigh. On the theory of resonance. In *Collected scientific papers*, volume 1, pages 33–75. 1899.

[30] H. L. Royden. Harmonic functions on open Riemann surfaces. *Trans. Amer. Math. Soc.*, 73:40–94, 1952.

[31] C. E. Shannon and D. W. Hagelbarger. Concavity of resistance function. *J. of Appl. Phys.*, 27:42–43, 1956.

[32] J. L. Snell. Probability and martingales. *The Mathematical Intelligencer*, 4, 1982.

[33] W. Thomson and P. G. Tait. *Treatise on Natural Philosophy*. 1879.

[34] J. Ville. *Étude critique de la notion de collectif*. 1939.

[35] G. N. Watson. Three triple integrals. *Quarterly J. Math.*, 10:266–276, 1939.

Неравенство Шапиро

А. Храбров

Задачу представляли И.Богданов, В.Бугаенко, К.Куюмжиян, К.Кохась, А.Скопенков, Г.Челноков

1 Неравенство Шапиро

В октябре 1954 г. в журнале "American Mathematical Monthly" появилась задача американского математика Гарольда Шапиро:

Для положительных чисел x_1, x_2, ..., x_n докажите неравенство

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\geqslant\frac{n}{2}\,, \tag{1}$$

причем равенство может достигаться, только если все знаменатели равны между собой.

В "Monthly" в отличие, например, от журнала "Квант" допускалась публикация задач, которые никто не умел решать, причем читателей об этом не предупреждали. Так было и на этот раз. У автора было решение только для $n=3$ и 4.

В предлагаемых ниже задачах вместо положительности всех чисел x_k можно требовать, чтобы все числа x_k были неотрицательными, а все знаменатели — ненулевыми. Если неравенство для положительных чисел уже доказано, то из него несложно вывести неравенство для неотрицательных чисел, для которых знаменатели не обращаются в нуль. Обозначим

$$f(x_1,x_2,\ldots,x_n)=\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\,.$$

1.1. Докажите неравенство (1) при $n=3$, 4, 5, 6.

1.2. Докажите, что неравенство (1) неверно:
a) при $n=20$; b) при $n=14$; c) при $n=25$.

1.3. Докажите неравенство (1) для монотонных последовательностей.

1.4. Докажите, что если неравенство (1) неверно при $n=m$, то оно неверно и при $n=m+2$.

1.5. Докажите, что если неравенство (1) неверно при $n=m$, где m нечетно, то оно неверно и при всех n, бóльших m.

1.6. Докажите, что неравенство (1) верно при $n=8$, 10, 12 и $n=7$, 9, 11, 13, 15, 17, 19, 21, 23. Как следует из утверждения задачи 1.4, достаточно доказать неравенство лишь при $n=12$ и $n=23$.

1.7. Докажите, что $f(x_1,x_2,\ldots,x_n)+f(x_n,x_{n-1},\ldots,x_1)\geqslant n$.

1.8. Предположим, что в точке a_1, a_2, ..., $a_n>0$ функция $f(x_1,x_2,\ldots,x_n)$ имеет локальный минимум.
a) Если n четно, докажите, что $f(a_1,a_2,\ldots,a_n)=n/2$.
b*) Докажите аналогичное утверждение для нечетных n.
c) Докажите с помощью пунктов а) и б) неравенство для $n=7$ и $n=8$.

1.9. Докажите неравенство $f(x_1,x_2,\ldots,x_n)\geqslant cn$ для следующих значений c
a) $c=1/4$; b) $c=\left(\sqrt{2}-1\right)$; c) $c=5/12$.

2 Полезные и родственные неравенства

Докажите следующие неравенства в предположении, что все числа x_k положительны. Проверьте, что выделенные жирным шрифтом константы нельзя заменить на бóльшие (при каждом n).

2.1. Неравенство Морделла.
a) Для любых неотрицательных чисел x_1, x_2, ..., x_n имеет место неравенство

$$\left(\sum_{k=1}^{n}x_k\right)^2\geqslant\mathbf{min}\left\{\frac{\mathbf{n}}{\mathbf{2}},\mathbf{3}\right\}\cdot\sum_{k=1}^{n}x_k(x_{k+1}+x_{k+2}).$$

b) Установите, для каких неотрицательных чисел x_1, x_2, ..., x_n неравенство Морделла обращается в равенство.

2.2. Для всех неотрицательных чисел x_1, x_2, …, x_n докажите неравенство

$$\left(\sum_{k=1}^{n} x_k\right)^2 \geqslant \mathbf{min}\left\{\frac{\mathbf{n}}{\mathbf{3}}, \frac{\mathbf{8}}{\mathbf{3}}\right\} \cdot \sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2} + x_{k+3}).$$

2.3. а) При $n \leqslant 8$ докажите неравенство

$$\frac{x_1}{x_2 + x_3 + x_4} + \frac{x_2}{x_3 + x_4 + x_5} + \ldots + \frac{x_{n-1}}{x_n + x_1 + x_2} + \frac{x_n}{x_1 + x_2 + x_3} \geqslant \frac{n}{3}\,.$$

b*) Верно ли это неравенство еще при каких-нибудь натуральных n?

2.4. $(x_1 + x_2 + \ldots + x_n)^2 \geqslant 4(x_1x_2 + x_2x_3 + \ldots + x_{n-1}x_n + x_nx_1)\,; \quad n \geqslant 4.$

2.5. $\displaystyle\sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} \geqslant \sum_{k=1}^{n} \frac{x_{k+1}}{x_k + x_{k+1}}.$

2.6. $\displaystyle\frac{x_1}{x_n + x_2} + \frac{x_2}{x_1 + x_3} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_n} + \frac{x_n}{x_{n-1} + x_1} \geqslant \mathbf{2}\,; \quad n \geqslant 4.$

2.7. $\displaystyle\frac{x_1 + x_2}{x_1 + x_3} + \frac{x_2 + x_3}{x_2 + x_4} + \ldots + \frac{x_{n-1} + x_n}{x_{n-1} + x_1} + \frac{x_n + x_1}{x_n + x_2} \geqslant \mathbf{4}\,; \quad n \geqslant 4.$

2.8. $\displaystyle\frac{x_1}{x_n + x_3} + \frac{x_2}{x_1 + x_4} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_1} + \frac{x_n}{x_{n-1} + x_2} \geqslant \mathbf{3}\,; \quad n \geqslant 6.$

2.9. $\displaystyle\frac{x_2 + x_3}{x_1 + x_4} + \frac{x_3 + x_4}{x_2 + x_5} + \ldots + \frac{x_n + x_1}{x_{n-1} + x_2} + \frac{x_1 + x_2}{x_n + x_3} \geqslant \mathbf{6}\,; \quad n \geqslant 6.$

2.10. $\displaystyle\frac{x_1 + x_2}{x_1 + x_4} + \frac{x_2 + x_3}{x_2 + x_5} + \ldots + \frac{x_{2004} + x_1}{x_{2004} + x_3} \geqslant \mathbf{6}\,.$

2.11. $\displaystyle\frac{x_1}{x_n + x_4} + \frac{x_2}{x_1 + x_5} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_2} + \frac{x_n}{x_{n-1} + x_3} \geqslant \mathbf{4}$, где n — четное число, большее 7.

2.12. $\displaystyle\sum_{k=1}^{n} \frac{x_k^2}{x_{k+1}^2 - x_{k+1}x_{k+2} + x_{k+2}^2} \geqslant \left[\frac{\mathbf{n+1}}{\mathbf{2}}\right].$

Неравенство Шапиро

3 Добавление после промежуточного финиша

1.10. а) Для любого натурального n существует такое число $q_n > 1$, что при всех вещественных числах x_1, x_2, …, $x_n \in [\frac{1}{q_n}; q_n]$ имеет место неравенство (1).

b*) Существует ли такое $q > 1$, что при всех натуральных n и при всех $x_i \in [\frac{1}{q}; q]$ выполнено неравенство (1)?

1.11. Пусть $S = f(x_1, x_2, \ldots, x_n)$ — левая часть неравенства Шапиро. Обозначим через a_1, a_2, …, a_n числа x_2/x_1, x_3/x_2, …, x_n/x_{n-1}, x_1/x_n, расположенные в порядке возрастания.

а) Докажите, что $S \geqslant \frac{1}{a_1(1+a_n)} + \frac{1}{a_2(1+a_{n-1})} + \ldots + \frac{1}{a_n(1+a_1)}$;

б) Пусть $b_k = \begin{cases} \dfrac{1}{a_k a_{n+1-k}}, & a_k a_{n+1-k} \geqslant 1 \\ \dfrac{2}{a_k a_{n+1-k} + \sqrt{a_k a_{n+1-k}}}, & a_k a_{n+1-k} < 1. \end{cases}$ Докажите, что $2S \geqslant b_1 + b_2 + \ldots + b_n$;

в) Пусть g — наибольшая выпуклая функция, не превосходящая функций e^{-x} и $2(e^x + e^{x/2})^{-1}$. Докажите, что $2S \geqslant g(\ln(a_1 a_n)) + g(\ln(a_2 a_{n-1})) + \ldots + g(\ln(a_n a_1)) \geqslant ng(0)$.

г) Докажите, что для любого $\lambda > g(0)/2$ существует такое натуральное число n и такие положительные числа x_1, x_2, …, x_n, что $S \leqslant \lambda n$.

Решения

1.1. $\underline{n = 3}$. Пусть $S = x_1 + x_2 + x_3$. Как нетрудно видеть, функция $f(t) = \frac{t}{S-t}$ выпукла на промежутке $[0; S)$. Запишем для нее неравенство Йенсена

$$\frac{f(x_1) + f(x_2) + f(x_3)}{3} \geqslant f\left(\frac{x_1 + x_2 + x_3}{3}\right) = f\left(\frac{S}{3}\right) = \frac{1}{2}.$$

Это и есть требуемое неравенство.

$\underline{n = 4}$. Неравенство циклическое. Напишем данные числа последовательно в вершинах квадрата. По диагонали проведем стрелочки от меньшего числа к большему. Тогда у одной из сторон квадрата обе вершины — концы стрелочек. Перенумеруем числа так, чтобы это была сторона x_4x_1. Таким образом, можно считать, что $x_1 \geqslant x_3$, $x_4 \geqslant x_2$. Для переменных, упорядоченных этим способом, имеет место очевидное неравенство

$$\frac{x_1}{x_2 + x_3} + \frac{x_3}{x_4 + x_1} \geqslant \frac{x_1}{x_4 + x_3} + \frac{x_3}{x_2 + x_1}.$$

Воспользуемся им для доказательства неравенства Шапиро

$$\frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_4} + \frac{x_3}{x_4 + x_1} + \frac{x_4}{x_1 + x_2} \geqslant \frac{x_1}{x_4 + x_3} + \frac{x_2}{x_3 + x_4} + \frac{x_3}{x_2 + x_1} + \frac{x_4}{x_1 + x_2} = \frac{x_1 + x_2}{x_3 + x_4} + \frac{x_3 + x_4}{x_1 + x_2} = a + a^{-1} \geqslant 2.$$

$\underline{n = 5}$. Заметим, что функция $f(t) = 1/(S - t)$ выпукла на интервале $[0; S)$. Воспользуемся тогда неравенством Йенсена для пяти чисел

$$a_1 f(t_1) + a_2 f(t_2) + a_3 f(t_3) + a_4 f(t_4) + a_5 f(t_5) \geqslant f(a_1 t_1 + a_2 t_2 + a_3 t_3 + a_4 t_4 + a_5 t_5), \tag{2}$$

где $a_i \geqslant 0$, $\sum a_i = 1$. В качестве a_i возьмем $a_i = \frac{x_i}{S}$, и пусть $t_i = x_i + x_{i-1} + x_{i-2}$, $i = 1, \ldots, 5$ (считаем, что нумерация переменных циклическая: $x_0 = x_5$, $x_{-1} = x_4$). Тогда $f(t_i) = \frac{1}{S - t_i} = \frac{1}{x_{i+1} + x_{i+2}}$, и значит, левая часть неравенства (2) есть в точности левая часть неравенства Шапиро. Посмотрим, что представляет собой правая часть.

$$\frac{1}{S - \sum_{i=1}^5 a_i t_i} = \frac{1}{S - \sum_{i=1}^5 \frac{x_i}{S}(x_i + x_{i-1} + x_{i-2})} = \frac{S}{S^2 - \sum_{i=1}^5 x_i(x_i + x_{i-1} + x_{i-2})}$$

Как нетрудно убедиться, раскрыв скобки, знаменатель представляет собой в точности сумму попарных произведений набора переменных x_i. Заметим еще, что силу однородности доказываемого неравенства, можно считать, что $S = x_1 + x_2 + x_3 + x_4 + x_5 = 1$. Итак, правая часть неравенства (2), есть величина, обратная сумме попарных произведений переменных x_i, подчиненных условию $x_1 + x_2 + x_3 + x_4 + x_5 = 1$. Очевидно, минимум правой части достигается в том случае, когда сумма попарных произведений максимальна. Хорошо известно, что это будет в случае, когда все переменные равны. Нетрудно видеть, что правая часть в этом случае равна $5/2$.

Кстати, при $n = 4$ аналогичное доказательство тоже работает.

$\underline{n = 6}$. Аналогично предыдущему решению. Функция $f(t) = 1/(S - t)$ выпукла на интервале $[0; S)$ Рассмотрим неравенство Йенсена для шести чисел

$$\sum_{i=1}^{6} a_i f(t_i) \geqslant f\Big(\sum_{i=1}^{6} a_i t_i\Big).$$

Пусть $a_i = \frac{x_i}{S}$, $t_i = x_i + x_{i-1} + x_{i-2} + x_{i-3}$, $i = 1, \ldots, 6$ (считаем, что нумерация переменных циклическая: $x_0 = x_6$, $x_{-1} = x_5$, $x_{-2} = x_4$). Тогда $f(t_i) = \frac{1}{S - t_i} = \frac{1}{x_{i+1} + x_{i+2}}$, и значит, левая часть неравенства (1.1) есть в точности левая часть неравенства Шапиро. Посмотрим, что представляет собой правая часть.

$$\frac{1}{S - \sum_{i=1}^{6} a_i t_i} = \frac{1}{S - \sum_{i=1}^{6} \frac{x_i}{S}(x_i + x_{i-1} + x_{i-2} + x_{i-3})} = \frac{S}{S^2 - \sum_{i=1}^{6} x_i(x_i + x_{i-1} + x_{i-2} + x_{i-3})}$$

Как нетрудно убедиться, раскрыв скобки, знаменатель представляет собой в точности сумму попарных произведений набора переменных x_i, кроме произведений x_1x_4, x_2x_5, x_3x_6. Эту сумму можно записать в виде $(x_1 + x_4)(x_2 + x_5) + (x_1 + x_4)(x_3 + x_6) + (x_2 + x_5)(x_3 + x_6)$. Обозначив $A = x_1 + x_4$, $B = x_2 + x_5$, $C = x_3 + x_6$, мы можем записать правую часть нашего неравенства Йенсена в виде

$$\frac{A + B + C}{AB + BC + AC}\,. \tag{3}$$

В силу однородности доказываемого неравенства, можно считать, что $S = x_1 + x_2 + x_3 + x_4 + x_5 = A + B + C = 1$. Тогда очевидно, что выражение (3) не меньше 3, в силу неравенства $(A + B + C)^2 \geqslant 3(AB + BC + AC)$. чТД

Замечание. К сожалению, дальше этот способ не работает. Если мы применяем метод множителей Лагранжа, чтобы найти максимум знаменателей в правой части (при фиксированной сумме переменных), то система уравнений получается линейной, следовательно, подозрительная точка единственна. При $n = 6$ квадратичная форма второго дифференциала в этой точке является положительно определенной и вырожденной. При $n = 7$ она уже не является знакоопределенной (Я проверил это в Maple. КК). Таким образом, при $n = 7$ неравенство Йенсена дает слишком грубую оценку снизу.

Другое решение этой задачи получится, если применить неравенство Коши-Буняковского для наборов чисел

$$\begin{matrix} \sqrt{\dfrac{x_1}{x_2 + x_3}}, & \sqrt{\dfrac{x_2}{x_3 + x_4}}, & \ldots, & \sqrt{\dfrac{x_n}{x_1 + x_2}} \\ \sqrt{x_1(x_2 + x_3)}, & \sqrt{x_2(x_3 + x_4)}, & \ldots, & \sqrt{x_n(x_1 + x_2)} \end{matrix} \qquad \text{и}$$

Получится

$$\frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_4} + \ldots + \frac{x_n}{x_1 + x_2} \geqslant \frac{(x_1 + x_2 + \ldots + x_n)^2}{x_1(x_2 + x_3) + x_2(x_3 + x_4) + \ldots + x_n(x_1 + x_2)}\,.$$

По неравенству Морделла (задача 2.1) при $n \leqslant 6$ правая часть этого неравенства не меньше, чем $n/2$.

1.2. a) Если в качестве $x_1, x_2, \ldots, x_{20}$ взять

$$\begin{matrix} 1 + 5\varepsilon, & 6\varepsilon, & 1 + 4\varepsilon, & 5\varepsilon, & 1 + 3\varepsilon, & 4\varepsilon, & 1 + 2\varepsilon, & 3\varepsilon, & 1 + \varepsilon, & 2\varepsilon, \\ 1 + 2\varepsilon, & \varepsilon, & 1 + 3\varepsilon, & 2\varepsilon, & 1 + 4\varepsilon, & 3\varepsilon, & 1 + 5\varepsilon, & 4\varepsilon, & 1 + 6\varepsilon, & 5\varepsilon, \end{matrix}$$

то левая часть неравенства будет меньше чем $10 - \varepsilon^2 + c\varepsilon^3$ для некоторого c. Следовательно, при достаточно малом ε она будет меньше 10. Этот пример принадлежит Лайтхиллу; опубликован в "Monthly" [22].

b) Если в качестве $x_1, x_2, \ldots, x_{14}$ взять

$$1 + 7\varepsilon,\ 7\varepsilon,\ 1 + 4\varepsilon,\ 6\varepsilon,\ 1 + \varepsilon,\ 5\varepsilon,\ 1,\ 2\varepsilon,\ 1 + \varepsilon,\ 0,\ 1 + 4\varepsilon,\ \varepsilon,\ 1 + 6\varepsilon,\ 4\varepsilon,$$

то правая часть неравенства будет меньше чем $7 - 2\varepsilon^2 + c\varepsilon^3$ и, следовательно, при достаточно малом ε она будет меньше 7. Это пример Цулауфа [27].

А вот еще пример [24], целочисленный, в нем некоторые переменные равны нулю.

$$0,\ 42,\ 2,\ 42,\ 4,\ 41,\ 5,\ 39,\ 4,\ 38,\ 2,\ 38,\ 0,\ 40\,.$$

c) Примеры, подтверждающие что при $n = 25$ неравенство неверно, были построены на компьютере в 1970 г. Дейкиным [10] и Малкольмом [18]. Ниже приведен пример Дейкина (он в отличие от примера Малкольма целочисленный):

$$0,\ 85,\ 0,\ 101,\ 0,\ 120,\ 14,\ 129,\ 41,\ 116,\ 59,\ 93,\ 64,71,\ 63,\ 52,\ 60,\ 36,\ 58,\ 23,\ 58,\ 12,\ 62,\ 3,\ 71.$$

А вот еще пример Р. Алексеева и Е. Фошкина (приведен в [3]).

32, 0, 37, 0, 43, 0, 50, 0, 59, 8, 62, 21, 55,29, 44, 32, 33, 31, 24, 30, 16, 29, 10, 29, 4.

1.3. Утверждение задачи опубликовано в [13]. Мы приводим короткое изящное решение.

Пусть $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n > 0$. Заметим, что произведение n дробей $\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}$ равно 1. Тогда из неравенства о средних заключаем, что

$$\sum_{k=1}^{n} \frac{x_k + x_{k+1}}{x_{k+1} + x_{k+2}} \geqslant n = \sum_{k=1}^{n} \frac{x_{k+1} + x_{k+2}}{x_{k+1} + x_{k+2}}.$$

Следовательно,

$$\sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} \geqslant \sum_{k=1}^{n} \frac{x_{k+2}}{x_{k+1} + x_{k+2}} = \sum_{k=1}^{n} \frac{x_{k+1}}{x_k + x_{k+1}}. \tag{4}$$

Теперь мы воспользуемся следующим известным неравенством, которое будем называть транснеравенством: пусть имеются два набора чисел $a_1 \geqslant \ldots \geqslant a_n$ и $b_1 \geqslant \ldots \geqslant b_n$ Тогда для любой перестановки $k_1, \ldots, k_n$ чисел $1, \ldots, n$ имеет место неравенство

$$a_1b_1 + a_2b_2 + \ldots a_nb_n \geqslant a_1b_{k_1} + a_2b_{k_2} + \ldots a_nb_{k_n} \geqslant a_1b_n + a_2b_{n-1} + \ldots a_nb_1$$

Два раза воспользуемся транснеравеством

$$\begin{aligned}\sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} &= \sum_{k=1}^{n-2} \frac{x_k}{x_{k+1} + x_{k+2}} + \frac{x_{n-1}}{x_n + x_1} + \frac{x_n}{x_1 + x_2} \underset{(*)}{\geqslant} \\ &\geqslant \sum_{k=1}^{n-2} \frac{x_k}{x_{k+1} + x_{k+2}} + \frac{x_{n-1}}{x_1 + x_2} + \frac{x_n}{x_n + x_1} \underset{(**)}{\geqslant} \\ &\geqslant \sum_{k=1}^{n} \frac{x_k}{x_k + x_{k+1}}.\end{aligned}$$

Здесь неравенство (*) — это транснеравенство для наборов из двух чисел: $x_{n-1} \geqslant x_n$ и $\frac{1}{x_n+x_1} \geqslant \frac{1}{x_1+x_2}$; а неравенство (**) — это транснеравенство для обратно упорядоченных наборов чисел x_1, x_2, …, x_{n-1} и $\frac{1}{x_1+x_2}$, $\frac{1}{x_2+x_3}$, …, $\frac{1}{x_{n-1}+x_n}$.

Таким образом,

$$2\sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} \geqslant \sum_{k=1}^{n} \frac{x_k}{x_k + x_{k+1}} + \sum_{k=1}^{n} \frac{x_{k+1}}{x_k + x_{k+1}} = n.$$

Для убывающего набора чисел x_i решение аналогично, поскольку применяя неравенство о средних мы вообще не пользовались упорядочением, а применяя транснеравенство, использовали то, что наборы x_i и $\frac{1}{x_i+x_{i+1}}$ упорядочены по-разному.

А вот более прямолинейное решение из [3].

Разберем сначала случай $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n > 0$. Для краткости обозначим левую часть неравенства через $f_n(x_1, x_2, \ldots, x_n)$. Заметим, что $f_2(x_1, x_2) = 1$. Поэтому если мы покажем, что

$$f_{n+1}(x_1, x_2, \ldots, x_n, x_{n+1}) - f_n(x_1, x_2, \ldots, x_n) \geqslant \frac{1}{2},$$

то, воспользовавшись методом математической индукции, получим

$$f_n(x_1, x_2, \ldots, x_n) \geqslant f_{n-1}(x_1, x_2, \ldots, x_{n-1}) + \frac{1}{2} \geqslant f_{n-2}(x_1, x_2, \ldots, x_{n-2}) + 1 \geqslant \cdots \geqslant f_2(x_1, x_2) + \frac{n-2}{2} = \frac{n}{2}.$$

Рассмотрим разность

$$f_{n+1}(x_1, x_2, \ldots, x_n, x_{n+1}) - f_n(x_1, x_2, \ldots, x_n) = \frac{x_{n-1}}{x_n + x_{n+1}} + \frac{x_n}{x_{n+1} + x_1} + \frac{x_{n+1}}{x_1 + x_2} - \frac{x_{n-1}}{x_n + x_1} - \frac{x_n}{x_1 + x_2}.$$

Она не зависит от чисел x_3, x_4, …, x_{n-2}, поэтому мы можем считать их любыми числами из интервала $[x_2; x_{n-1}]$. Слагаемые, содержащие x_{n-1}, имеют вид

$$\frac{x_{n-1}}{x_n + x_{n+1}} - \frac{x_{n-1}}{x_n + x_1} = x_{n-1} \cdot \frac{x_1 - x_{n+1}}{(x_n + x_{n+1})(x_n + x_1)},$$

т. е. равны x_{n-1}, умноженному на неотрицательное число. Следовательно, при уменьшении x_{n-1} разность может только уменьшиться. Поэтому можно считать, что $x_{n-1} = x_n$. Слагаемые, содержащие x_2, дают

$$\frac{x_{n+1} - x_n}{x_1 + x_2},$$

что уменьшается при уменьшении x_2, поэтому можно считать, что $x_2 = x_{n-1} = x_n$. Следовательно, достаточно доказать неравенство

$$0 \leqslant f_{n+1}(x_1, x_n, \ldots, x_n, x_n, x_{n+1}) - f_n(x_1, x_n, \ldots, x_n, x_n) - \frac{1}{2} =$$
$$= \frac{x_{n+1} - 2x_n}{x_1 + x_n} + \frac{x_n}{x_n + x_{n+1}} + \frac{x_n}{x_1 + x_{n+1}} - \frac{1}{2} = \frac{(x_n - x_{n+1})(x_1^2 + 2x_n^2 + x_n x_{n+1} - x_1 x_n - x_1 x_{n+1} - 2x_{n+1}^2)}{(x_1 + x_n)(x_1 + x_{n+1})(x_n + x_{n+1})}.$$

Таким образом, достаточно доказать, что

$$x_1^2 + 2x_n^2 + x_n x_{n+1} - x_1 x_n - x_1 x_{n+1} - 2x_{n+1}^2 \geqslant 0.$$

Последнее выражение уменьшается при увеличении x_{n+1}, поэтому можно считать, что $x_{n+1} = x_n$. Но в этом случае неравенство превращается в очевидное: $x_1^2 + x_n^2 - 2x_1 x_n \geqslant 0$.

Теперь разберем случай $0 < x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$. Заметим, что

$$f_n(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2}{x_2 + x_3} + \frac{x_2 + x_3}{x_3 + x_4} + \ldots + \frac{x_n + x_1}{x_1 + x_2} - n + \frac{x_2}{x_1 + x_2} + \frac{x_3}{x_2 + x_3} + \ldots + \frac{x_1}{x_n + x_1}.$$

Первая сумма не меньше n, так как произведение слагаемых равно 1. Поэтому достаточно показать, что

$$g_n(x_1, x_2, \ldots, x_n) = \frac{x_2}{x_1 + x_2} + \frac{x_3}{x_2 + x_3} + \ldots + \frac{x_1}{x_n + x_1} \geqslant \frac{n}{2}.$$

Так как $g_2(x_1, x_2) = 1$, то достаточно проверить, что

$$g_n(x_1, x_2, \ldots, x_n) - g_{n-1}(x_1, x_2, \ldots, x_{n-1}) \geqslant \frac{1}{2}.$$

Это действительно так:

$$g_n(x_1, x_2, \ldots, x_n) - g_{n-1}(x_1, x_2, \ldots, x_{n-1}) - \frac{1}{2} = \frac{x_n}{x_{n-1} + x_n} + \frac{x_1}{x_n + x_1} - \frac{x_1}{x_{n-1} + x_1} - \frac{1}{2} =$$
$$= \frac{(x_{n-1} - x_1)(x_n - x_{n-1})(x_n - x_1)}{2(x_{n-1} + x_n)(x_n + x_1)(x_1 + x_{n-1})} \geqslant 0.$$

Замечание. Для монотонно возрастающей последовательности чисел неравенство не может быть доказано по индукции без дополнительных трюков, поскольку

$$f_n(0, 1, \ldots, 1) = \frac{n+1}{2} \quad \text{и} \quad f_{n+1}(0, 1, \ldots, 1, 1\tfrac{1}{10}) = \frac{n-2}{2} + 1\tfrac{1}{10} + \frac{1}{1\frac{1}{10}} + \frac{1}{2\frac{1}{10}} \approx \frac{n+1}{2} + 0{,}4852 < \frac{n+1}{2} + \frac{1}{2}.$$

1.4. [3] Как нетрудно убедиться, $f_{n+2}(x_1, x_2, \ldots, x_n, x_1, x_2) = f_n(x_1, x_2, \ldots, x_n) + 1$. Отсюда сразу следует, что если $f_n(x_1, x_2, \ldots, x_n) < n/2$, то $f_{n+2}(x_1, x_2, \ldots, x_n, x_1, x_2) < (n+2)/2$.

1.5. [3] Предположим, что $f_m(x_1, x_2, \ldots, x_m) < \frac{m}{2}$. Вычислим разность

$$f_{m+1}(x_1, \ldots, x_k, x_k, x_{k+1}, \ldots, x_m) - f_m(x_1, x_2, \ldots, x_m) - \frac{1}{2} =$$
$$= \frac{x_{k-1}}{2x_k} + \frac{x_k}{x_k + x_{k+1}} - \frac{x_{k-1}}{x_k + x_{k+1}} - \frac{1}{2} = \frac{(x_k - x_{k-1})(x_k - x_{k+1})}{2x_k(x_k + x_{k+1})}.$$

Если $(x_k - x_{k-1})(x_k - x_{k+1}) \leqslant 0$, то

$$f_{n+1}(x_1, x_2, \ldots, x_k, x_k, x_{k+1}, \ldots, x_m) < \frac{m+1}{2}$$

и утверждение доказано. При нечетном m такой индекс k обязательно найдется, поскольку если при всех k $(x_k - x_{k-1})(x_{k+1} - x_k) < 0$, то перемножив эти неравенства (их нечетное число!), получим, что

$$(x_2 - x_1)^2 (x_3 - x_2)^2 \ldots (x_m - x_{m-1})^2 (x_1 - x_m)^2 < 0.$$

Итак, если для нечетного m неравенство неверно, то и для $m+1$ оно тоже неверно. Осталось воспользоваться утверждением предыдущей задачи.

1.6. Наиболее короткие из известных доказательств для $7 \leqslant n \leqslant 12$ опубликованы в [7, 8]. Для бóльших n доказательства опираются на компьютерный перебор.

1.7. [28] Пусть $y_k = x_k + x_{k+1}$. Тогда

$$\frac{x_1+x_4}{x_2+x_3} + \frac{x_2+x_5}{x_3+x_4} + \ldots + \frac{x_n+x_3}{x_1+x_2} = \sum_{k=1}^{n} \frac{y_k - y_{k+1} + y_{k+2}}{y_{k+1}} = \sum_{k=1}^{n} \frac{y_k}{y_{k+1}} + \sum_{k=1}^{n} \frac{y_{k+2}}{y_{k+1}} - n \geqslant n,$$

поскольку каждая из сумм по неравенству о средних не меньше n.

1.8. Утверждения a), b) взяты из [21]. Положим для краткости $a = (a_1, a_2, \ldots, a_n)$, $x = (x_1, x_2, \ldots, x_n)$ и $u = (-1, 1, -1, 1, \ldots, -1, 1)$.

a) Заметим, что

$$f(a+tu) = f(a) + t\left(\frac{-1}{x_2+x_3} + \frac{1}{x_3+x_4} + \ldots\right)$$

Таким образом, $f(a+tu)$ — линейная функция. В точке a у нее наблюдается локальный минимум, следовательно, она постоянна и нестрогий локальный минимум наблюдается во всех точках $a+tu$, имеющих положительные координаты. Так как

$$\frac{\partial f}{\partial x_k}(x) = \frac{1}{x_{k+1}+x_{k+2}} - \frac{x_{k-2}}{(x_{k-1}+x_k)^2} - \frac{x_{k-1}}{(x_k+x_{k+1})^2}.$$

и в наших точках минимума

$$\frac{\partial f}{\partial x_k}(a) = 0, \qquad \cdot\frac{\partial f}{\partial x_k}(a+tu) = 0,$$

выполняются соотношения

$$\frac{1}{a_{k+1}+a_{k+2}} - \frac{a_{k-2}}{(a_{k-1}+a_k)^2} - \frac{a_{k-1}}{(a_k+a_{k+1})^2} = 0$$

и

$$\frac{1}{a_{k+1}+a_{k+2}} - \frac{a_{k-2}+t(-1)^{k-2}}{(a_{k-1}+a_k)^2} - \frac{a_{k-1}+t(-1)^{k-1}}{(a_k+a_{k+1})^2} = 0.$$

Вычитая из второго равенства первое получим соотношение

$$\frac{t}{(a_{k-1}+a_k)^2} - \frac{t}{(a_k+a_{k+1})^2} = 0.$$

Таким образом,

$$a_{k-1} + a_k = a_k + a_{k+1}.$$

Стало быть,

$$a_1 = a_3 = a_5 = \cdots = a_{n-1} \quad \text{и} \quad a_2 = a_4 = a_6 = \cdots = a_n.$$

А значит, $f(a) = n/2$.

b) Это короткое доказательство опубликовано в [7]. Положим для краткости $a = (a_1, a_2, \ldots, a_n)$, $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, $z = (z_1, z_2, \ldots, z_n)$, где $y_k = x_k + x_{k+1}$ и $z_k = 1/y_{n+1-k}$.

Положим

$$S(x) = \frac{x_1}{x_2+x_3} + \frac{x_2}{x_3+x_4} + \ldots + \frac{x_{n-1}}{x_n+x_1} + \frac{x_n}{x_1+x_2} = \sum_{k=0}^{n-1} \frac{x_k}{y_{k+1}}.$$

Заметим, что

$$\frac{\partial f}{\partial x_k}(x) = \frac{1}{x_{k+1}+x_{k+2}} - \frac{x_{k-2}}{(x_{k-1}+x_k)^2} - \frac{x_{k-1}}{(x_k+x_{k+1})^2}.$$

Легко проверить тождество

$$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d} + \frac{\dfrac{a}{b^2} + \dfrac{c}{d^2}}{\dfrac{1}{b} + \dfrac{1}{d}}.$$

Таким образом,

$$\frac{x_{k-2}}{x_{k-1}+x_k}+\frac{x_{k-1}}{x_k+x_{k+1}}=\frac{x_{k-2}+x_{k-1}}{(x_{k-1}+x_k)+(x_k+x_{k+1})}+\frac{\dfrac{x_{k-2}}{(x_{k-1}+x_k)^2}+\dfrac{x_{k-1}}{(x_k+x_{k+1})^2}}{\dfrac{1}{x_{k-1}+x_k}+\dfrac{1}{x_k+x_{k+1}}}=$$

$$=\frac{y_{k-2}}{y_{k-1}+y_k}+\frac{z_{n-k}-\dfrac{\partial f}{\partial x_k}(x)}{z_{n-k+1}+z_{n-k+2}}.$$

Следовательно,

$$2S(x)=S(y)+S(z)-\sum_{k=1}^{n}\frac{\dfrac{\partial f}{\partial x_k}(x)}{z_{n-k+1}+z_{n-k+2}}.$$

Если в точке x достигается локальный минимум, то $2S(x)=S(y)+S(z)$. Таким образом, $S(x)=S(y)=S(z)$.

Обозначим среднее арифметическое чисел x_1, x_2, …, x_n через u. Рассмотрим преобразование

$$M(x)=\Big(\frac{x_1+x_2}{2},\frac{x_2+x_3}{2},\dots,\frac{x_n+x_1}{2}\Big).$$

Через $M_k(x)$ обозначим его k-ю итерацию. Заметим, что $S(x)=S(y)=S(M(x))=\dots=S(M_k(x))$. Ясно, что $\lim\limits_{k\to\infty}M_k(x)=(u,u,\dots,u)$. Тогда

$$S(x)=\lim_{k\to\infty}S(M_k(x))=S((u,u,\dots,u))=\frac{n}{2}.$$

c) [16], [7, 8]

1.9. Решения всех пунктов мы взяли из [3].

a) Задача предлагалась на Третьей Всесоюзной олимпиаде по математике, 1969 г. Туда она, видимо, попала из [14]; решение из этой статьи опубликовано по-русски в [3].

Пусть x_{i_1} — наибольшее из чисел x_1, x_2, …, x_n; x_{i_2} — наибольшее из двух следующих за x_{i_1} чисел; x_{i_3} — наибольшее из двух следующих за x_{i_2} чисел и т. д. Будем строить эту последовательность чисел до тех пор пока не дойдем до такого k, что наибольшее из двух следующих за x_{i_k} чисел — это x_{i_1}.

Ясно, что $k\geqslant n/2$, а также

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_n}{x_1+x_2}\geqslant\frac{x_{i_1}}{2x_{i_2}}+\frac{x_{i_2}}{2x_{i_3}}+\ldots+\frac{x_{i_k}}{2x_{i_1}}.$$

По неравенству о среднем арифметическом и среднем геометрическом последнее выражение не меньше $k/2$, а значит, не меньше $n/4$.

b) Каждую из дробей $\frac{x_k}{x_{k+1}+x_{k+2}}$, $k=1$, 2, …, n, запишем в виде

$$\frac{x_k}{x_{k+1}+x_{k+2}}=\frac{x_k+\frac12x_{k+1}}{x_{k+1}+x_{k+2}}+\frac{\frac12x_{k+1}+x_{k+2}}{x_{k+1}+x_{k+2}}-1\,.$$

Мы получим $2n$ дробей. Сгруппируем их по парам — первую с $2n$-й, вторую с третьей, четвертую с пятой и т. д. Оценим снизу сумму дробей из одной пары:

$$\frac{\frac12x_k+x_{k+1}}{x_k+x_{k+1}}+\frac{x_k+\frac12x_{k+1}}{x_{k+1}+x_{k+2}}\geqslant2\sqrt{\frac{(\frac12x_k+x_{k+1})(x_k+\frac12x_{k+1})}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}}=$$

$$=2\sqrt{\left(\frac12+\frac{x_kx_{k+1}}{4(x_k+x_{k+1})^2}\right)\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}>\sqrt2\cdot\sqrt{\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}$$

Так как произведение n чисел $\sqrt{\frac{x_1+x_2}{x_2+x_3}}$, $\sqrt{\frac{x_2+x_3}{x_3+x_4}}$, …, $\sqrt{\frac{x_n+x_1}{x_1+x_2}}$ равно 1, то из неравенства Коши следует, что их сумма не меньше n. Поэтому исходная сумма в левой части неравенства Шапиро не меньше, чем $\sqrt2n-n=(\sqrt2-1)n$.

c) Аналогично предыдущему пункту каждую из дробей $\frac{x_k}{x_{k+1}+x_{k+2}}$, $k=1$, 2, …, n, запишем в виде

$$\frac{x_k}{x_{k+1}+x_{k+2}}=\frac{x_k+\beta x_{k+1}}{x_{k+1}+x_{k+2}}+\alpha\cdot\frac{\beta x_{k+1}+x_{k+2}}{x_{k+1}+x_{k+2}}-\alpha\,,$$

а параметры α и β подберем так, чтобы это равенство оказалось верным. Как нетрудно видеть, для этого нужно, чтобы $\beta+\alpha\beta=\alpha$, т. е. $\beta=\alpha/(\alpha+1)$. Тогда

$$\frac{x_k+\beta x_{k+1}}{x_{k+1}+x_{k+2}}+\alpha\cdot\frac{\beta x_k+x_{k+1}}{x_k+x_{k+1}}\geqslant 2\sqrt{\alpha\frac{(x_k+\beta x_{k+1})(\beta x_k+x_{k+1})}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}}=$$
$$=2\sqrt{\alpha\frac{\beta(x_k+x_{k+1})^2+(\beta-1)^2x_kx_{k+1}}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}}>2\sqrt{\alpha\beta\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}=\frac{2\alpha}{\sqrt{\alpha+1}}\cdot\sqrt{\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}\,.$$

Значит,

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\geqslant\frac{2\alpha}{\sqrt{\alpha+1}}\left(\sqrt{\frac{x_1+x_2}{x_2+x_3}}+\sqrt{\frac{x_2+x_3}{x_3+x_4}}+\ldots+\sqrt{\frac{x_n+x_1}{x_1+x_2}}\right)-\alpha n>$$
$$>\frac{2\alpha}{\sqrt{\alpha+1}}n-\alpha n=\left(\frac{2\alpha}{\sqrt{\alpha+1}}-\alpha\right)n\,.$$

Максимум выражения $g(\alpha)=\frac{2\alpha}{\sqrt{\alpha+1}}-\alpha$ достигается при $\alpha=\alpha_0\approx 1.1479$ (корень кубического уравнения $g'(\alpha)=0$), при этом $g(\alpha_0)\approx 0.4186$. При $\alpha=\frac{5}{4}$ $g(\alpha)=\frac{5}{12}\approx 0.416$ — неплохое приближение.

1.10. а) Утверждение принадлежит В. Кыртоаже [9].

Положим для краткости $y_k=x_k+x_{k+1}$. В новых обозначениях неравенство (1) примет вид

$$\frac{x_1}{y_2}+\frac{x_2}{y_3}+\ldots+\frac{x_n}{y_1}\geqslant\frac{n}{2}\,.$$

Еще немного его преобразуем:

$$\sum_{k=1}^{n}\frac{2q_n^2x_k-y_{k+1}}{y_{k+1}}\geqslant n(q_n^2-1)\,,$$

здесь q_n — это параметр, значение которого мы подберем позже так, чтобы все рассматриваемые неравенства оказались верными. Поскольку

$$2q_n^2x_k-y_{k+1}=(q_n^2x_k-x_{k+1})+(q_n^2x_k-x_{k+2})\geqslant 0\,,$$

по неравенству Коши–Буняковского для наборов чисел

$$\left\{\sqrt{\frac{2q_n^2x_k-y_{k+1}}{y_{k+1}}}\right\}\ \text{ и }\ \left\{\sqrt{(2q_n^2x_k-y_{k+1})y_{k+1}}\right\}$$

имеем

$$\sum_{k=1}^{n}\frac{2q_n^2x_k-y_{k+1}}{y_{k+1}}\geqslant\frac{\left(\sum\limits_{k=1}^{n}(2q_n^2x_k-y_{k+1})\right)^2}{\sum\limits_{k=1}^{n}(2q_n^2x_k-y_{k+1})y_{k+1}}.$$

Таким образом, достаточно показать, что

$$A^2=\left(\sum_{k=1}^{n}(2q_n^2x_k-y_{k+1})\right)^2\geqslant n(q_n^2-1)\sum_{k=1}^{n}(2q_n^2x_k-y_{k+1})y_{k+1}=n(q_n^2-1)B.$$

Поскольку $\sum\limits_{k=1}^{n}y_k=2\sum\limits_{k=1}^{n}x_k$, имеем равенства

$$A=(q_n^2-1)\sum_{k=1}^{n}y_k\,,$$
$$B=2q_n^2\sum_{k=1}^{n}x_ky_{k+1}-\sum_{k=1}^{n}y_k^2=2q_n^2\sum_{k=1}^{n}y_ky_{k+1}-(q_n^2+1)\sum_{k=1}^{n}y_k^2\,.$$

Следовательно, осталось доказать, что

$$(q_n^2-1)\left(\sum_{k=1}^{n}y_k\right)^2\geqslant n\left(2q_n^2\sum_{k=1}^{n}y_ky_{k+1}-(q_n^2+1)\sum_{k=1}^{n}y_k^2\right).\qquad(5)$$

Преобразуем левую часть с помощью соотношения

$$\left(\sum_{k=1}^{n} y_k\right)^2 = n\sum_{k=1}^{n} y_k^2 - \sum_{i<k}(y_i - y_k)^2,$$

неравенство (5) примет вид

$$n\sum_{k=1}^{n}(y_k - y_{k+1})^2 \geqslant \left(1 - \frac{1}{q_n^2}\right)\sum_{i<k}(y_i - y_k)^2.$$

По неравенству Коши–Буняковского

$$\sum_{k=1}^{n}(y_k - y_{k+1})^2 \geqslant \sum_{j=i}^{k-1}(y_j - y_{j+1})^2 \geqslant \frac{1}{k-j}\left(\sum_{j=i}^{k-1}(y_j - y_{j+1})\right)^2 = \frac{1}{k-j}(y_i - y_k)^2 \geqslant \frac{1}{n-1}(y_i - y_k)^2.$$

Следовательно,

$$\frac{n(n-1)}{2}\sum_{k=1}^{n}(y_k - y_{k+1})^2 \geqslant \frac{1}{n-1}\sum_{i<k}(y_i - y_k)^2$$

и можно взять $1 - \frac{1}{q_n^2} = \frac{2}{(n-1)^2}$, т. е. $q_n = \frac{n-1}{\sqrt{n^2-2n-1}} > 1$.

Замечание. С ростом n найденные q_n стремятся к единице.

b)

1.11.

1.11. (a) Обозначим $k_i := x_{i+1}/x_i$. Тогда

$$S = \frac{1}{k_1(k_2+1)} + \frac{1}{k_2(k_3+1)} + \cdots + \frac{1}{k_n(k_1+1)} \geqslant \frac{1}{a_1(a_n+1)} + \frac{1}{a_2(a_{n-1}+1)} + \cdots + \frac{1}{a_n(a_1+1)}.$$

(b) Неравенство справедливо, поскольку

$$\frac{1}{a_i(a_{n+1-i}+1)} + \frac{1}{a_{n+1-i}(a_i+1)} = \frac{1 + \frac{a_i a_{n+1-i}-1}{(1+a_i)(1+a_{n+1-i})}}{a_i a_{n+1-i}} \geqslant b_i$$

где последнее неравенство справедливо, поскольку $(1+a_i)(1+a_{n+1-i}) \geqslant (1+\sqrt{a_i a_{n+1-i}})^2$.

(c) Первое неравенство $2S \geqslant g(\ln(a_1 a_n)) + g(\ln(a_2 a_{n-1})) + \cdots + g(\ln(a_n a_1))$ справедливо, поскольку $g(x)$ меньше и e^{-x}, и $2(e^x+e^{x/2})^{-1}$. Второе неравенство справедливо, по неравенству Йенсена, поскольку g выпукла.

(d) [2]

2.1. Это неравенство доказано в статье [20].

a) При $n = 3$ и $n = 5$ после раскрытия скобок получим неравенство

$$(n-1)(a_1 + a_2 + \ldots + a_n)^2 \geqslant 2n\sum_{i<k} a_i a_k. \tag{6}$$

Оно легко выводится из неравенства Коши–Буняковского. Действительно, напишем неравенство Коши–Буняковского для наборов $a_1, a_2, \ldots, a_n$ и $1, 1, \ldots, 1$:

$$n(a_1^2 + a_2^2 + \ldots + a_n^2) \geqslant (a_1 + a_2 + \ldots + a_n)^2.$$

Далее заметим, что

$$n(a_1 + a_2 + \ldots + a_n)^2 = n(a_1^2 + a_2^2 + \ldots + a_n^2) + 2n\sum_{i<k} a_i a_k \geqslant (a_1 + a_2 + \ldots + a_n)^2 + 2n\sum_{i<k} a_i a_k\,,$$

откуда и следует неравенство (6).

При $n = 4$ нужно проверить неравенство

$$(x_1 + x_2 + x_3 + x_4)^2 \geqslant 2x_1x_2 + 2x_2x_3 + 2x_3x_4 + 2x_4x_1 + 4x_1x_3 + 4x_2x_4\,.$$

Раскроем скобки и приведем подобные слагаемые, получим очевидное неравенство

$$x_1^2 + x_2^2 + x_3^2 + x_4^2 \geqslant 2x_1x_3 + 2x_2x_4\,.$$

Перейдем теперь к случаю $n \geqslant 6$. Передвинув, если нужно, числа по циклу, можно добиться того, что $x_3 \geqslant x_1$ и $x_3 \geqslant x_2$, например, сделав x_3 наибольшим. При $r = 1$, 2 или 3 обозначим через a_r сумму всех чисел x_k, для которых $k \equiv r \pmod 3$, $k \leqslant n$. Тогда $x_1+x_2+\ldots+x_n = a_1+a_2+a_3$ и, следовательно, по неравенству (6)

$$(x_1 + x_2 + \ldots + x_n)^2 = (a_1 + a_2 + a_3)^2 \geqslant 3(a_1a_2 + a_2a_3 + a_3a_1) = 3 \cdot \sum_{(i-k)\not\vdots 3} x_i x_k\,.$$

Положим для краткости

$$A = \sum_{(i-k)\not\vdots 3} x_i x_k \quad \text{и} \quad B = \sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2})$$

и проверим, что $A \geqslant B$. Действительно, при $n \equiv 0 \pmod 3$ все слагаемые из суммы B содержатся в сумме A; при $n \equiv 1 \pmod 3$ в сумме A по сравнению с суммой B недостает лишь слагаемого $x_n x_1$, которое не превосходит содержащегося в ней слагаемого $x_n x_3$; и наконец, при $n \equiv 2 \pmod 3$ в сумме A по сравнению с суммой B недостает слагаемых $x_{n-1}x_1$ и $x_n x_2$, которые не превосходят соответственно слагаемых $x_{n-1}x_3$ и $x_n x_3$.

Итак, во всех случаях $A \geqslant B$. Таким образом,

$$(x_1 + x_2 + \ldots + x_n)^2 \geqslant 3A \geqslant 3B = 3\sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2}).$$

Неулучшаемость константы $\min\left\{\frac{n}{2}, 3\right\}$ очевидна. При $n \leqslant 6$ достаточно положить $x_1 = x_2 = \ldots = x_n = 1$, а при $n \geqslant 6$ — $x_1 = x_2 = x_3 = 1$ и $x_4 = x_5 = \ldots = x_n = 0$.

b) Случай $n < 6$ тривиален. Для $n = 6$ равенство достигается при $x_1 + x_4 = x_2 + x_5 = x_3 + x_6$. Для $n \geqslant 6$ равенство достигается на множествах вида $(t, 1, 1, 1 - t, 0, \ldots, 0)$, где $t \in [0, 1]$, и их циклических сдвигах.

2.2. Это неравенство доказано в статье [20]. Начнем с $n \leqslant 8$.

При $n = 4$ и $n = 7$ это частный случай неравенства (6).

При $n = 5$ неравенство совпадает с неравенством $\sum(x_k - 2x_{k+2} + x_{k+4})^2 \geqslant 0$.

При $n = 6$ после раскрытия скобок и приведения подобных слагаемых получится очевидное неравенство $x_1^2 + x_2^2 + \ldots + x_6^2 \geqslant 2x_1x_4 + 2x_2x_5 + 2x_3x_6$.

Для доказательства неравенства в случае $n = 8$ раскроем скобки в очевидном следствии неравенства Коши–Буняковского

$$4(x_1^2 + x_2^2 + x_3^2 + x_4^2) \geqslant (x_1 + x_2 + x_3 + x_4)^2$$

и получим

$$3(x_1^2 + x_2^2 + x_3^2 + x_4^2) \geqslant 2(x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4).$$

Следовательно,

$$3(x_1 + x_2 + x_3 + x_4)^2 \geqslant 8(x_1x_2 + x_1x_3 + x_1x_4 + x_2x_3 + x_2x_4 + x_3x_4), \tag{7}$$

что совпадает с требуемым неравенством для $n = 8$.

Перейдем теперь к случаю $n > 8$. Передвинув, если нужно, числа по циклу, можно добиться того, что $x_4 \geqslant x_1$, $x_4 \geqslant x_2$ и $x_4 \geqslant x_3$, например, сделав x_4 наибольшим. При $r = 1, 2, 3$ или 4 обозначим через a_r сумму всех чисел x_k, для которых $k \equiv r \pmod 4$, $k \leqslant n$. Тогда $x_1 + x_2 + \ldots + x_n = a_1 + a_2 + a_3 + a_4$ и, следовательно, по неравенству (7)

$$3(x_1 + x_2 + \ldots + x_n)^2 = 3(a_1 + a_2 + a_3 + a_4)^2 \geqslant 8(a_1a_2 + a_2a_3 + a_3a_4 + a_4a_1) \geqslant 8 \cdot \sum_{(i-k)\not\vdots 4} x_i x_k\,.$$

Положим для краткости

$$A = \sum_{(i-k)\not\vdots 4} x_i x_k \quad \text{и} \quad B = \sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2} + x_{k+3})$$

и проверим, что $A \geqslant B$. Действительно, при $n \equiv 0 \pmod 4$ все слагаемые из суммы B содержатся в сумме A; при $n \equiv 1 \pmod 4$ в сумме A по сравнению с суммой B недостает лишь слагаемого $x_n x_1$, которое не превосходит содержащегося в ней слагаемого $x_n x_4$; при $n \equiv 2 \pmod 4$ в сумме A по сравнению с суммой B недостает слагаемых $x_{n-1}x_1$ и $x_n x_2$, которые не превосходят соответственно слагаемых $x_{n-1}x_4$ и $x_n x_4$; и наконец, при $n \equiv 3 \pmod 4$ в сумме A по сравнению с суммой B недостает слагаемых $x_{n-2}x_1$, $x_{n-1}x_2$ и $x_n x_3$, которые не превосходят соответственно слагаемых $x_{n-2}x_4$, $x_{n-1}x_4$ и $x_n x_4$.

Итак, во всех случаях $A \geqslant B$. Таким образом,

$$3(x_1 + x_2 + \ldots + x_n)^2 \geqslant 8A \geqslant 8B = 8\sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2} + x_{k+3}).$$

2.3. a) Это неравенство доказано в статье [11] с использованием преобразования Фурье. Мы приводим элементарное рассуждение. По неравенству Коши–Буняковского имеем

$$\frac{x_1}{x_2 + x_3 + x_4} + \frac{x_2}{x_3 + x_4 + x_5} + \ldots + \frac{x_{n-1}}{x_n + x_1 + x_2} + \frac{x_n}{x_1 + x_2 + x_3} \geqslant \frac{(x_1 + x_2 + \ldots + x_n)^2}{\sum x_k(x_{k+1} + x_{k+2} + x_{k+3})} \geqslant \frac{n}{3},$$

последнее по задаче 2.2.

b)

2.4. Задача Б. Гинзбурга [1, задача 187], предлагалась на Всесоюзной олимпиаде по математике 1972 г. С помощью циклической перестановки чисел можно добиться того, что $x_1 \leqslant x_2$. Пусть $S = x_1 + x_2 + \ldots + x_n$, $S_1 = x_1 + x_3 + \ldots$, $S_2 = x_2 + x_4 + \ldots$. Тогда $S_1^2 + S_2^2 \geqslant (S_1 + S_2)^2/2 = S^2/2$, откуда

$$\frac{S^2}{2} \geqslant S^2 - S_1^2 - S_2^2 = 2 \sum_{(i-k)\not\vdots 2} x_i x_k. \tag{8}$$

Если n — четно, то в последней сумме содержатся все слагаемые вида $x_k x_{k+1}$, а если нечетно, то отсутствует слагаемое $x_n x_1$, зато вместо него имеется большее слагаемое $x_n x_2$. Таким образом,

$$\frac{S^2}{2} \geqslant 2(x_1x_2 + x_2x_3 + \ldots + x_nx_1).$$

2.5. См. решение задачи 1.3 до неравенства (4) (к этому моменту в упомянутом решении упорядочение переменных еще не использовалось).

2.6. Это задача А. Прокопьева, Турнир городов, 1981–82, [4], также опубликована в журнале "Квант", 1982, № 6, задача М749.

Обозначим левую часть неравенства через L_n. При $n = 4$

$$L_4 = \frac{x_1 + x_3}{x_2 + x_4} + \frac{x_2 + x_4}{x_1 + x_3} = a + a^{-1} \geqslant 2.$$

При $n > 4$ рассуждаем по индукции. Так как неравенство циклическое, можно считать, что x_{n+1} — наименьшее из всех чисел. Тогда отбросим в сумме L_{n+1} последнее слагаемое, а потом уменьшим два других, получим

$$L_{n+1} \geqslant \frac{x_1}{x_{n+1} + x_2} + \ldots + \frac{x_n}{x_{n-1} + x_{n+1}} \geqslant \frac{x_1}{x_n + x_2} + \ldots + \frac{x_n}{x_{n-1} + x_n} = L_n\,.$$

Чтобы показать, что константа в правой части точная, возьмем

$$x_1 = x_2 = 1, \quad x_3 = t, \quad x_4 = t^2, \quad \ldots, \quad x_n = t^{n-2}\,.$$

При $t \to +0$ первые два слагаемых стремятся к 1, остальные — к 0.

Используя неравенство Коши–Буняковского подобно тому как это делается в решении следующей задачи, читатель без труда придумает другое доказательство, сводящее данное неравенство к неравенству из задачи 2.4.

2.7. Мы почерпнули условие этой задачи в статье [10]. Положим для краткости $S = x_1 + x_2 + \ldots + x_n$. Воспользуемся неравенством Коши–Буняковского для наборов чисел $\left\{\frac{x_k + x_{k+1}}{x_k + x_{k+2}}\right\}$ и $\{(x_k + x_{k+1})(x_k + x_{k+2})\}$, получим

$$\frac{x_1 + x_2}{x_1 + x_3} + \frac{x_2 + x_3}{x_2 + x_4} + \ldots + \frac{x_{n-1} + x_n}{x_{n-1} + x_1} + \frac{x_n + x_1}{x_n + x_2} \geqslant \frac{4(x_1 + x_2 + \ldots + x_n)^2}{\sum\limits_{k=1}^{n} (x_k + x_{k+1})(x_k + x_{k+2})}.$$

Таким образом, достаточно установить неравенство

$$S^2 \geqslant \sum_{k=1}^{n} (x_k + x_{k+1})(x_k + x_{k+2}) = \sum_{k=1}^{n} x_k^2 + 2\sum_{k=1}^{n} x_k x_{k+1} + \sum_{k=1}^{n} x_k x_{k+2},$$

которое получается с помощью раскрытия скобок в левой части, ибо при $n \geqslant 4$ слагаемые $x_k x_{k+1}$ и $x_k x_{k+2}$ при $k = 1, 2, \ldots, n$ различны.

Покажем, что константу 4 нельзя увеличить. Положим $x_k = a^{k-1}$ при $k = 1, 2, \ldots, n-1$ и $x_n = a^{n-2}$. При $a \to \infty$ первые $n - 3$ слагаемых стремятся к нулю, а оставшиеся к 1, 2 и 1.

2.8. Мы почерпнули условие этой задачи в статье [6]. Воспользуемся неравенством Коши–Буняковского для наборов чисел $\left\{\frac{x_k}{x_{k-1} + x_{k+2}}\right\}$ и $\{x_k(x_{k-1} + x_{k+2})\}$, получим

$$\frac{x_1}{x_n + x_3} + \frac{x_2}{x_1 + x_4} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_1} + \frac{x_n}{x_{n-1} + x_2} \geqslant \frac{(x_1 + x_2 + \ldots + x_n)^2}{(x_1x_2 + x_2x_3 + \ldots + x_nx_1) + (x_1x_3 + x_2x_4 + \ldots + x_nx_2)}.$$

Правая часть полученного неравенства не меньше 3 в силу неравенства Морделла (задача 2.1).

Покажем, что константу 3 нельзя увеличить. Возьмем набор чисел $x_k = a^{k-1}$ при $k = 1, 2, \ldots, n-2$ и $x_{n-1} = x_n = 1$. При $a \to 0$ первое и два последних слагаемых стремятся к единице, а остальные к нулю.

2.9. Мы почерпнули условие этой задачи в статье [5]. Сложим два неравенства из 2.8 (для прямого и обратного порядков чисел).

Покажем, что константу 6 нельзя увеличить. Возьмем набор чисел $x_k = a^{k-1}$ при $k = 1, 2, \ldots, n-2$ и $x_{n-1} = x_n = 1$. При $a \to 0$ последние четыре слагаемых стремятся соответственно к 1, 2, 2, 1; остальные слагаемые стремятся к нулю.

2.10. В [19] это утверждение для произвольного n высказано в качестве гипотезы.

Авторы следующего доказательства — участники конференции P.Milošević и M. Bukić.

Доказываемое неравенство есть сумма двух неравенств при $n = 2004$ — неравенства из задачи 2.8 и неравенства

$$\frac{x_1}{x_1+x_4} + \frac{x_2}{x_2+x_5} + \ldots + \frac{x_n}{x_n+x_3} \geqslant 3\,.$$

Докажем последнее неравенство. При $n = 3m$ оно является суммой трех неравенств:

$$\frac{x_1}{x_1+x_4} + \frac{x_4}{x_4+x_7} + \ldots + \frac{x_{n-2}}{x_{n-2}+x_1} \geqslant 1\,.$$
$$\frac{x_2}{x_2+x_5} + \frac{x_5}{x_5+x_8} + \ldots + \frac{x_{n-1}}{x_{n-1}+x_2} \geqslant 1\,.$$
$$\frac{x_3}{x_3+x_6} + \frac{x_6}{x_6+x_9} + \ldots + \frac{x_n}{x_n+x_3} \geqslant 1\,.$$

Каждое из этих неравенств записывается в виде

$$\frac{1}{1+a_1} + \frac{1}{1+a_3} + \ldots + \frac{1}{1+a_m} \geqslant 1 \qquad \text{где} \quad a_1 a_2 \ldots a_m = 1.$$

Это неравенство доказывается по индукции. База $m = 2$

$$\frac{1}{1+a_1} + \frac{1}{1+\frac{1}{a_1}} = 1 \geqslant 1\,.$$

Для обоснования перехода проверяем что

$$\frac{1}{1+b} + \frac{1}{1+c} \geqslant \frac{1}{1+bc}\,.$$

Это делается непосредственно с помощью домножения на знаменатели и раскрытия скобок.

Приводим доказательство А. Храброва. Будем доказывать неравенство

$$Z = \frac{x_1+x_2}{x_1+x_4} + \frac{x_2+x_3}{x_2+x_5} + \ldots + \frac{x_{3n}+x_1}{x_{3n}+x_3} \geqslant 6.$$

Для удобства будем считать, что нумерация переменных циклическая: $x_{3n+k} = x_k$. Положим для краткости ($r = 0$, 1 и 2)

$$S_r = \sum_{k=1}^{n} x_{3k+r}, \quad X_r = \sum_{k=1}^{n} \frac{x_{3k+r}}{x_{3k+r}+x_{3k+3+r}} \quad \text{и} \quad Y_r = \sum_{k=1}^{n} \frac{x_{3k+r+1}}{x_{3k+r}+x_{3k+3+r}},$$

Докажем сначала, что $X_r \geqslant 1$. Чтобы не усложнять формулы, ограничимся случаем $r = 0$. Заметим, что

$$X_0 S_0^2 \geqslant X_0 \left(\sum_{k=1}^{n} x_{3k}^2 + \sum_{k=1}^{n} x_{3k} x_{3k+3} \right) = X_0 \left(\sum_{k=1}^{n} x_{3k}(x_{3k} + x_{3k+3}) \right) \geqslant S_0^2,$$

последнее — по неравенству Коши–Буняковского. Поэтому, $X_0 \geqslant 1$.

Далее проверим неравенство $Y_r \geqslant S_{r+1}/S_r$ (мы полагаем $S_3 = S_0$). Опять рассмотрим лишь случай $r = 0$.

$$Y_0 S_0 S_1 \geqslant Y_0 \left(\sum_{k=1}^{n} x_{3k} x_{3k+1} + \sum_{k=1}^{n} x_{3k+1} x_{3k+3} \right) = Y_0 \left(\sum_{k=1}^{n} x_{3k+1}(x_{3k} + x_{3k+3}) \right) \geqslant S_1^2,$$

в последнем неравенстве мы снова применили неравенство Коши–Буняковского. Стало быть, $Y_0 \geqslant S_1/S_0$.

Сложим все доказанные неравенства и воспользуемся неравенством о средних, получим

$$Z = X_0 + X_1 + X_2 + Y_0 + Y_1 + Y_2 \geqslant 3 + \frac{S_1}{S_0} + \frac{S_2}{S_1} + \frac{S_0}{S_2} \geqslant 6.$$

Покажем, что константу 6 нельзя увеличить. Возьмем набор чисел $x_1 = x_2 = x_3 = 1$, $x_k = a^{n-k+1}$ при $k = 3, 4, \ldots, n$. При $a \to 0$ первое и второе слагаемое стремятся к 2, третье, а также последнее — к 1, остальные слагаемые стремятся к нулю.

2.11. Доказательство А. Храброва. Положим для краткости $S = x_1 + x_2 + \ldots + x_n$ и $T = \sum\limits_{(i-k)\not\vdots 2} x_i x_k$. По неравенству Коши–Буняковского для наборов чисел $\left\{\frac{x_k}{x_{k-1}+x_{k+3}}\right\}$ и $\{x_k(x_{k-1}+x_{k+3})\}$ имеем

$$\frac{x_1}{x_n+x_4}+\frac{x_2}{x_1+x_5}+\ldots+\frac{x_{n-1}}{x_{n-2}+x_2}+\frac{x_n}{x_{n-1}+x_3}\geqslant\frac{(x_1+x_2+\ldots+x_n)^2}{(x_1x_2+x_2x_3+\ldots+x_nx_1)+(x_1x_4+x_2x_5+\ldots+x_nx_3)}.$$

Таким образом, достаточно показать, что

$$S^2\geqslant 4(x_1x_2+x_2x_3+\ldots+x_nx_1)+4(x_1x_4+x_2x_5+\ldots+x_nx_3).$$

В решении задачи 2.4 мы установили неравенство $S^2 \geqslant 4T$, см. (8). Поэтому достаточно показать, что

$$T\geqslant(x_1x_2+x_2x_3+\ldots+x_nx_1)+(x_1x_4+x_2x_5+\ldots+x_nx_3). \tag{9}$$

Поскольку n четно, то все слагаемые из правой суммы содержатся и в левой сумме.

Покажем, что константу 4 нельзя увеличить. Возьмем набор чисел $x_k = a^{k-1}$ при $k = 1, 2, \ldots, n-3$ и $x_{n-2} = x_{n-1} = x_n = 1$. При $a \to +0$ первое слагаемое и три последних стремятся к единице, а остальные к нулю.

2.12. Мы взяли эту задачу в статье [14].

Заметим, что $a^2 - ab + b^2 \leqslant \max\{a, b\}^2$.

Пусть x_{i_1} — наибольшее из чисел $x_1, x_2, \ldots, x_n$; x_{i_2} — наибольшее из двух следующих за x_{i_1} чисел; x_{i_3} — наибольшее из двух следующих за x_{i_2} чисел и т. д. Будем строить эту последовательность чисел до тех пор пока не дойдем до такого k, что наибольшее из двух следующих за x_{i_k} чисел — это x_{i_1}.

Ясно, что $k \geqslant n/2$ и поэтому $k \geqslant \left[\frac{n+1}{2}\right]$.

$$\sum_{k=1}^{n}\frac{x_k^2}{x_{k+1}^2-x_{k+1}x_{k+2}+x_{k+2}^2}\geqslant\sum_{j=1}^{k}\frac{x_{i_j}^2}{x_{i_{j+1}}^2}\geqslant k,$$

последнее по неравенству о среднем арифметическом и среднем геометрическом.

Выражение $\left[\frac{n+1}{2}\right]$ в правой части нельзя увеличить, поскольку если положить $x_k = 1$ при нечетных k и $x_k = 0$ при четных k, то левая часть будет в точности равна $\left[\frac{n+1}{2}\right]$.

2.13. a) Мы взяли эту задачу в статье [10].

Можно считать, что x_3 — наибольшее. Тогда первое слагаемое не меньше единицы. Кроме того заметим, что сумма двух соседних слагаемых тоже не меньше единицы:

$$\frac{x_k+x_{k+2}}{x_k+x_{k+1}}+\frac{x_{k+1}+x_{k+3}}{x_{k+1}+x_{k+2}}=1+\frac{x_kx_{k+1}+x_{k+2}^2+x_kx_{k+3}+x_{k+1}x_{k+3}}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}\geqslant 1.$$

b) Мы взяли эту задачу в статье [17].

c) Случай $n \leqslant 4$ разобран в [10], случай $n = 5$ — в [25].

d) Вот контрпримеры из [25]. $n = 6$: 381, 0, 334, 29, 340, 49;
$n = 13$: 41, 0, 28, 0, 19, 4, 17, 10, 18, 18, 20, 29, 18.

2.14. a) ([11]) и b) ([12]) следуют из неравенства Коши–Бушяковского и задач 2.2 a) и b).

c) [11]

d, e, f)

2.15. Мы взяли эту задачу в статье [15].

Пусть x_{i_1} — наибольшее из чисел $x_1, x_2, \ldots, x_n$; x_{i_2} — наибольшее из m следующих за x_{i_1} чисел; x_{i_3} — наибольшее из m следующих за x_{i_2} чисел и т. д. Будем строить эту последовательность чисел до тех пор пока не дойдем до такого k, что наибольшее из m следующих за x_{i_k} чисел — это x_{i_1}.

Ясно, что $k \geqslant n/m$ и поэтому $k \geqslant \left[\frac{n+m-1}{m}\right]$. Таким образом,

$$\sum_{k=1}^{n}\frac{x_k}{x_{k+1}+x_{k+2}+\ldots+x_{k+m}}\geqslant\sum_{j=1}^{k}\frac{x_{i_j}}{mx_{i_{j+1}}}\geqslant\frac{k}{m},$$

последнее — по неравенству о среднем арифметическом и среднем геометрическом.

2.16. a) Мы взяли эту задачу в статье [20].

При $n = m$ имеем очевидное равенство. При $n = m + 1$ и $n = 2m + 1$ это неравенство — частный случай неравенства (6). При $n = 2m$ неравенство можно переписать в виде

$$(x_1 - x_{m+1})^2 + (x_2 - x_{m+2})^2 + \ldots + (x_m - x_{2m})^2 \geqslant 0,$$

в котором оно очевидно.

Пусть теперь $n = m + 2$. Положим для краткости $s = x_1 + x_2 + \ldots + x_n$. Нам нужно проверить неравенство

$$\frac{n-2}{n} s^2 \geqslant \sum_{k=1}^{n} x_k(x_{k+1} + x_{k+2} + \ldots + x_{k+m}) = \sum_{k=1}^{n} x_k(s - x_k - x_{k+n-1}) = s^2 - \sum_{k=1}^{n} x_k(x_k + x_{k+n-1}).$$

Или, что тоже самое,

$$\frac{2s^2}{n} \leqslant \sum_{k=1}^{n} x_k(x_k + x_{k+n-1}) = \frac{1}{2} \sum_{k=1}^{n} (x_k + x_{k+n-1})^2.$$

Последнее неравенство является очевидным следствием неравенства Коши–Буняковского:

$$n \sum_{k=1}^{n} (x_k + x_{k+n-1})^2 \geqslant \left(\sum_{k=1}^{n} (x_k + x_{k+n-1}) \right)^2.$$

Перейдем теперь к случаю $n = 2m + 2$. При $1 \leqslant r \leqslant m + 1$ обозначим через a_r сумму всех чисел x_k, для которых $k \equiv r \pmod{m+1}$, $k \leqslant n$. Тогда $x_1 + x_2 + \ldots + x_n = a_1 + a_2 + \ldots + a_m$. Заметим, что

$$\sum_{i<k} a_i a_k = \sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2} + \ldots + x_{k+m}}.$$

Таким образом, нужно доказать неравенство

$$(a_1 + a_2 + \ldots + a_{m+1})^2 \geqslant \frac{2m+2}{m} \sum_{i<k} a_i a_k,$$

но это опять неравенство (6).

b, c) Мы взяли эту задачу в статье [12].

2.17. a) [20]
b) [12]
c) [20]
d), e) [12]

2.18. Случай $s = n$ и $m = 1$ разобран в статье [23]; общий случай — в статье [26].

2.19. a) Аналогично 2.12.

Литература

[1] *Васильев Н. Б., Егоров А. А.* Задачи Всесоюзных математических олимпиад. М.: Наука, 1988.

[2] *Дринфельд В. Г.* Об одном циклическом неравенстве // Мат. заметки. 1971. Т. 9. № 2. С. 113–119.

[3] *Курляндчик Л. Д., Файбусович А.* История одного неравенства // Квант. 1991. № 4. С. 14–18.

[4] *Толпыго А. К.* Тысяча задач Международного математического Турнира городов. М.: МЦНМО, 2009.

[5] *Чимэдцэрэн С.* Нэгэн орчилт нийлбэр // Математикийн олимпиадын цуврал. 1999. Т. 22. (На монгольск. яз.).

[6] *Чимэдцэрэн С., Адъяасурен В., Батболд С.* Оценка в одной циклической сумме // Монгол улсын их сургууль, Эрдэм шинжилгээний бичиг. 2000. Т. 7 (168). С. 79–84.

[7] *Bushell P. J.* Shapiro's Cyclic Sum // Bull. London Math. Soc. 1994. Vol. 26. No 6. P. 564–574

[8] *Bushell P. J., McLeod J. B.* Shapiro's cyclic inequality for even n // J. Inequal. & Appl., 2002. Vol. 7(3). P. 331–348

[9] *Cîrtoaje V.* Crux Mathematicorum. 2006. Vol. 32. No. 8. Problem 3195.

[10] *Daykin D. E.* Inequalities for certain cyclic sums // Proc. Edinburgh Math. Soc. (2) 1970/71. Vol. 17. P. 257–262.

[11] *Diananda P. H.* Extensions of an inequality of H. S. Shapiro // Amer. Math. Monthly 1959. Vol. 66. P. 489–491.

[12] *Diananda P. H.* On a conjecture of L. J. Mordell regarding an inequality involoving quadratic forms // J. London Math. Soc. 1961. Vol. 36. P. 185–192.

[13] *Diananda P. H.* Inequalities for a class of cyclic and other sums // J. London Math. Soc. 1962. Vol. 37. P. 424–431.

[14] *Diananda P. H.* Some cyclic and other inequalities // Proc. Cambridge Philos. Soc. 1962. Vol. 58. P. 425–427.

[15] *Diananda P. H.* Some cyclic and other inequalities, II // Proc. Cambridge Philos. Soc. 1962. Vol. 58. P. 703–705.

[16] *Diananda P. H.* On a cyclic sum // Proc. Glasgow Math. Assoc. 1963. Vol. 6. P. 11–13.

[17] *Elbert A.* On a cyclic inequality // Period. Math. Hungarica. 1973. Vol. 4. № 2–3. P. 163–168.

[18] *Malcolm M. A.* A note on a conjecture of L. J. Mordell // Math. Comp. 1971. Vol. 25. P. 375–377.

[19] *Mitrinović D. S., Pečarić J. E., Fink A. M.* Classical and new inequalities in analysis. Kluwer Academic Publishers Group, Dordrecht, 1993. (Mathematics and its Applications (East European Series), Vol. 61).

[20] *Mordell L. J.* On the inequality $\sum\limits_{r=1}^{n} \frac{x_r}{x_{r+1}+x_{r+2}} \geqslant \frac{n}{2}$ and some others // Abh. Math. Sem. Univ. Hamburg. 1958. Vol. 22. P. 229–240.

[21] *Nowosad P.* Isoperimetric eigenvalue problems in algebras // Comm. Pure Appl. Math. 1968. Vol. 21. P. 401–465.

[22] *Shapiro H. S., Northover F. H.* Amer. Math. Monthly. 1956. Vol. 63. № 3. P. 191–192.

[23] *Tanahashi K., Tomiyama J.* Indecomposable positive maps in matrix algebras // Canad. Math. Bull. 1988. Vol. 31. № 3. P. 308–317.

[24] *Troesch B. A.* Full solution of Shapiro's cyclic inequality // Notices Amer. Math. Soc. 1985. Vol. 39. № 4. P. 318.

[25] *Vukmirović J.* A note on an inequality for the cyclic sums introduced by D. E. Daykin // Math. Balk. 1978. Vol. 8. P. 293–297.

[26] *Yamagami S.* Cyclic inequalities // Proc. Amer. Math. Soc. 1993. Vol. 118. № 2. P. 521–527.

[27] *Zulauf A.* Note on a conjecture of L. J. Mordell // Abh. Math. Sem. Univ. Hamburg. 1958. Vol. 22. P. 240–241.

[28] *Zulauf A.* Note on an Inequality // Math. Gazette. 1962. Vol. 46. №355. P. 41–42.

Shapiro's inequality

A. Khrabrov

1 Shapiro's inequality

In October, 1954 the American Mathematical Monthly published the following problem of Harold Shapiro

Prove the following inequality for positive numbers x_1, x_2, ..., x_n:

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\geqslant\frac{n}{2}\,, \tag{1}$$

the equality holds only if all the denominators are equal.

In contrast to, say, "Kvant" magazine, it was allowed to publish problems in the Monthly, which were not solved by the proposer, and the readers had not been informed about this nuance. This time the situation was exactly like that. The author had a solution for partial cases $n=3$ and 4 only.

In the following problems we can replace the condition that all the x_k's are positive with the condition that all the x_k's are nonnegative and all the denominators are nonzero. Indeed, if the inequality is proven for positive numbers, then it is not difficult to deduce the inequality for nonnegative numbers (and nonzero denominators). Let

$$f(x_1,x_2,\ldots,x_n)=\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\,.$$

1.1. Prove the inequality (1) for $n=3$, 4, 5, 6.

1.2. Prove that the inequality (1) is wrong
a) for $n=20$; b) for $n=14$; c) for $n=25$.

1.3. Prove the inequality (1) for monotonic sequences.

1.4. Prove that if the inequality (1) does not hold for $n=m$, then it does not hold for $n=m+2$.

1.5. Prove that if the inequality (1) does not hold for $n=m$, where m is odd, then it does not hold for all $n>m$.

1.6. Prove the inequality (1) for $n=8$, 10, 12 and for $n=7$, 9, 11, 13, 15, 17, 19, 21, 23. Due to the statement of the previous problem it is sufficient to prove the inequality only for $n=12$ and $n=23$.

1.7. Prove that $f(x_1,x_2,\ldots,x_n)+f(x_n,x_{n-1},\ldots,x_1)\geqslant n$.

1.8. Assume that the function $f(x_1,x_2,\ldots,x_n)$ has a local minimum in the point $(a_1,a_2,\ldots,a_n)$, a_1, a_2, ..., $a_n>0$.
a) Prove that $f(a_1,a_2,\ldots,a_n)=n/2$ if n is even.
b*) Prove the same statement for odd n.
c) Use the statements пїSa) and b) to prove the inequality for $n=7$ and $n=8$.

1.9. Prove the inequality $f(x_1,x_2,\ldots,x_n)\geqslant cn$ for the following values of the constant c:
a) $c=1/4$; b) $c=\left(\sqrt{2}-1\right)$; c) $c=5/12$.

2 Useful and related inequalities

Prove the following inequalities assuming that all the x_k's are positive. Prove that the constants printed in bold can not be decreased (for each n).

2.1. Mordell's inequality.
a) $\left(\sum\limits_{k=1}^{n}x_k\right)^2\geqslant\mathbf{min}\left\{\dfrac{\mathbf{n}}{\mathbf{2}},\mathbf{3}\right\}\cdot\sum\limits_{k=1}^{n}x_k(x_{k+1}+x_{k+2})$.
b) Find all n-tuples x_1, x_2, ..., x_n such that the equality is achieved.

2.2. $\left(\sum\limits_{k=1}^{n}x_k\right)^2\geqslant\mathbf{min}\left\{\dfrac{\mathbf{n}}{\mathbf{3}},\dfrac{\mathbf{8}}{\mathbf{3}}\right\}\cdot\sum\limits_{k=1}^{n}x_k(x_{k+1}+x_{k+2}+x_{k+3})$.

2.3. пїSa) Prove that for $n\leqslant 8$

$$\frac{x_1}{x_2+x_3+x_4}+\frac{x_2}{x_3+x_4+x_5}+\ldots+\frac{x_{n-1}}{x_n+x_1+x_2}+\frac{x_n}{x_1+x_2+x_3}\geqslant\frac{n}{3}\,.$$

b) For which $n>8$ this inequality is also true?

2.4. $(x_1 + x_2 + \ldots + x_n)^2 \geqslant 4(x_1x_2 + x_2x_3 + \ldots + x_{n-1}x_n + x_nx_1)$; $\quad n \geqslant 4$.

2.5. $\displaystyle\sum_{k=1}^{n} \frac{x_k}{x_{k+1} + x_{k+2}} \geqslant \sum_{k=1}^{n} \frac{x_{k+1}}{x_k + x_{k+1}}$.

2.6. $\displaystyle\frac{x_1}{x_n + x_2} + \frac{x_2}{x_1 + x_3} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_n} + \frac{x_n}{x_{n-1} + x_1} \geqslant \mathbf{2}$; $\quad n \geqslant 4$.

2.7. $\displaystyle\frac{x_1 + x_2}{x_1 + x_3} + \frac{x_2 + x_3}{x_2 + x_4} + \ldots + \frac{x_{n-1} + x_n}{x_{n-1} + x_1} + \frac{x_n + x_1}{x_n + x_2} \geqslant \mathbf{4}$; $\quad n \geqslant 4$.

2.8. $\displaystyle\frac{x_1}{x_n + x_3} + \frac{x_2}{x_1 + x_4} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_1} + \frac{x_n}{x_{n-1} + x_2} \geqslant \mathbf{3}$; $\quad n \geqslant 4$.

2.9. $\displaystyle\frac{x_2 + x_3}{x_1 + x_4} + \frac{x_3 + x_4}{x_2 + x_5} + \ldots + \frac{x_n + x_1}{x_{n-1} + x_2} + \frac{x_1 + x_2}{x_n + x_3} \geqslant \mathbf{6}$; $\quad n \geqslant 6$.

2.10. $\displaystyle\frac{x_1 + x_2}{x_1 + x_4} + \frac{x_2 + x_3}{x_2 + x_5} + \ldots + \frac{x_{2004} + x_1}{x_{2004} + x_3} \geqslant \mathbf{6}$.

2.11. $\displaystyle\frac{x_1}{x_n + x_4} + \frac{x_2}{x_1 + x_5} + \ldots + \frac{x_{n-1}}{x_{n-2} + x_2} + \frac{x_n}{x_{n-1} + x_3} \geqslant \mathbf{4}$, where $n > 5$ is even.

2.12. $\displaystyle\sum_{k=1}^{n} \frac{x_k^2}{x_{k+1}^2 - x_{k+1}x_{k+2} + x_{k+2}^2} \geqslant \left[\frac{\mathbf{n+1}}{\mathbf{2}}\right]$.

3 After the intermediate finish

1.10. a) Prove that for each n there exists $q_n > 1$, such that for all real x_1, x_2, …, $x_n \in [\frac{1}{q_n}; q_n]$ the inequality (1) holds.

b*) Is it possible to choose $q > 1$, such that for all integers $n > 0$ and for all $x_i \in [\frac{1}{q}; q]$ the inequality (1) holds?

1.11. Let $S = f(x_1, x_2, \ldots, x_n)$ be the left hand side of Shapiro's inequality. Denote by a_1, a_2, …, a_n the numbers x_2/x_1, x_3/x_2, …, x_n/x_{n-1}, x_1/x_n, arranged in increasing order.

a) Prove that $S \geqslant \frac{1}{a_1(1+a_n)} + \frac{1}{a_2(1+a_{n-1})} + \ldots + \frac{1}{a_n(1+a_1)}$;

b) Let $b_k = \begin{cases} \dfrac{1}{a_k a_{n+1-k}}, & a_k a_{n+1-k} \geqslant 1 \\ \dfrac{2}{a_k a_{n+1-k} + \sqrt{a_k a_{n+1-k}}}, & a_k a_{n+1-k} < 1. \end{cases}$ Prove that $2S \geqslant b_1 + b_2 + \ldots + b_n$;

c) Let g be the maximal convex function that does not exceed both functions e^{-x} и $2(e^x + e^{x/2})^{-1}$. Prove that $2S \geqslant g(\ln(a_1a_n)) + g(\ln(a_2a_{n-1})) + \ldots + g(\ln(a_na_1)) \geqslant ng(0)$.

d) Prove that for each $\lambda > g(0)$ there exist a nonnegative integer n and positive numbers x_1, x_2, …, x_n, such that $S \leqslant \lambda n$.

Solutions

1.1. $\underline{n=3}$. Let $S = x_1 + x_2 + x_3$. It is easy to see that the function $f(t) = \frac{t}{S-t}$ is convex on the interval $[0; S)$. Apply the Jensen inequality to it:

$$\frac{f(x_1) + f(x_2) + f(x_3)}{3} \geqslant f\left(\frac{x_1 + x_2 + x_3}{3}\right) = f\left(\frac{S}{3}\right) = \frac{1}{2}.$$

We are done.

$\underline{n=4}$. This inequality is cyclic. Write down the values of x_i's successively at the vertices of a square. Then on each diagonal put an arrow leading from the smaller value to the greater one. Notice that there is a side of the square with two tails on it. Re-number the x_i's in such a manner that this side becomes x_4x_1. Now we may assume that $x_1 \geqslant x_3$, $x_4 \geqslant x_2$. For the variables with these restrictions the following inequality is true:

$$\frac{x_1}{x_2 + x_3} + \frac{x_3}{x_4 + x_1} \geqslant \frac{x_1}{x_4 + x_3} + \frac{x_3}{x_2 + x_1}\,.$$

Indeed, re-write it in the following way:

$$x_1\Big(\frac{1}{x_2 + x_3} - \frac{1}{x_4 + x_3}\Big) \geqslant x_3\Big(\frac{1}{x_2 + x_1} - \frac{1}{x_4 + x_1}\Big)\,.$$

Reduce both hands to a common denominator, cancel $x_4 - x_2$ in both hands (if $x_4 - x_2 = 0$, we already have the equality), and multiply both hands to the product of denominators. We obtain the evident (since $x_1 \geqslant x_3$) inequality

$$x_1(x_2 + x_1)(x_4 + x_1) \geqslant x_3(x_2 + x_3)(x_4 + x_3)\,.$$

Use it to prove Shapiro's inequality:

$$\frac{x_1}{x_2 + x_3} + \frac{x_2}{x_3 + x_4} + \frac{x_3}{x_4 + x_1} + \frac{x_4}{x_1 + x_2} \geqslant \frac{x_1}{x_4 + x_3} + \frac{x_2}{x_3 + x_4} + \frac{x_3}{x_2 + x_1} + \frac{x_4}{x_1 + x_2} = \frac{x_1 + x_2}{x_3 + x_4} + \frac{x_3 + x_4}{x_1 + x_2} = a + a^{-1} \geqslant 2\,.$$

$\underline{n=5}$. Notice that the function $f(t) = 1/(S - t)$ is convex on the interval $[0; S)$. So we can apply the Jensen inequality with $n = 5$:

$$a_1f(t_1) + a_2f(t_2) + a_3f(t_3) + a_4f(t_4) + a_5f(t_5) \geqslant f(a_1t_1 + a_2t_2 + a_3t_3 + a_4t_4 + a_5t_5)\,, \tag{2}$$

where $a_i \geqslant 0$, $\sum a_i = 1$. Take $a_i = \frac{x_i}{S}$, and let $t_i = x_i + x_{i-1} + x_{i-2}$, $i = 1, \ldots, 5$ (we assume that the variables are enumerated cyclically: $x_0 = x_5$, $x_{-1} = x_4$). Then $f(t_i) = \frac{1}{S-t_i} = \frac{1}{x_{i+1}+x_{i+2}}$, and it means that the left-hand side of inequality (2) coincides with the left-hand side of Shapiro's inequality. Now consider the right-hand side of 2:

$$\frac{1}{S - \sum_{i=1}^{5} a_it_i} = \frac{1}{S - \sum_{i=1}^{5} \frac{x_i}{S}(x_i + x_{i-1} + x_{i-2})} = \frac{S}{S^2 - \sum_{i=1}^{5} x_i(x_i + x_{i-1} + x_{i-2})}.$$

Open the brackets. It is easy to see that the denominator is the sum of pairwise products of the set of variables x_i. Since the initial inequality is homogeneous, we may assume that $S = x_1 + x_2 + x_3 + x_4 + x_5 = 1$. Now the right-hand side of inequality (2) is the inverse number to the sum of pairwise products of the variables x_i, satisfying one condition $x_1 + x_2 + x_3 + x_4 + x_5 = 1$. The right-hand side reaches its minimum when the sum of pairwise products reaches its maximum. It is well-known that for it all the variables should be equal. But the right-hand side equals $5/2$ in this point.

The analogous proof also works for $n = 4$.

$\underline{n=6}$. Proceed as above. The function $f(t) = 1/(S - t)$ is convex on the interval $[0; S)$. So we can apply the Jensen inequality with $n = 6$:

$$\sum_{i=1}^{6} a_if(t_i) \geqslant f\Big(\sum_{i=1}^{6} a_it_i\Big).$$

Let $a_i = \frac{x_i}{S}$, $t_i = x_i + x_{i-1} + x_{i-2} + x_{i-3}$, $i = 1, \ldots, 6$ (we assume that the variables are enumerated cyclically: $x_0 = x_6$, $x_{-1} = x_5$, $x_{-2} = x_4$). Then $f(t_i) = \frac{1}{S-t_i} = \frac{1}{x_{i+1}+x_{i+2}}$, and this means that the left-hand side of the inequality (1.1) coincides with the left-hand side of Shapiro's inequality. Now consider the right-hand side of (1.1):

$$\frac{1}{S - \sum_{i=1}^{6} a_it_i} = \frac{1}{S - \sum_{i=1}^{6} \frac{x_i}{S}(x_i + x_{i-1} + x_{i-2} + x_{i-3})} = \frac{S}{S^2 - \sum_{i=1}^{6} x_i(x_i + x_{i-1} + x_{i-2} + x_{i-3})}.$$

Open the brackets. It is easy to see that the denominator is the sum of pairwise products of the variables x_i's but the products x_1x_4, x_2x_5, and x_3x_6. This sum can be re-written as $(x_1+x_4)(x_2+x_5)+(x_1+x_4)(x_3+x_6)+(x_2+x_5)(x_3+x_6)$. Denote $A = x_1 + x_4$, $B = x_2 + x_5$, $C = x_3 + x_6$. The right-hand side of (1.1) can be re-written as

$$\frac{A+B+C}{AB+BC+AC}\,. \tag{3}$$

Since the initial inequality is homogeneous, we may assume that $S = x_1 + x_2 + x_3 + x_4 + x_5 = A + B + C = 1$. Now it is clear that the expression (3) is greater than or equal to 3, since $(A + B + C)^2 \geqslant 3(AB + BC + AC)$. ïiSïiSïiS

Remark. Unfortunately, this method does not work for $n > 6$.

Second solution. Apply the Cauchy-Bunyakovsky inequality to the sets of numbers

$$\sqrt{\frac{x_1}{x_2+x_3}},\quad \sqrt{\frac{x_2}{x_3+x_4}},\quad \ldots,\quad \sqrt{\frac{x_n}{x_1+x_2}} \qquad \text{and}$$
$$\sqrt{x_1(x_2+x_3)},\quad \sqrt{x_2(x_3+x_4)},\quad \ldots,\quad \sqrt{x_n(x_1+x_2)}\quad .$$

We obtain

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_n}{x_1+x_2}\geqslant\frac{(x_1+x_2+\ldots+x_n)^2}{x_1(x_2+x_3)+x_2(x_3+x_4)+\ldots+x_n(x_1+x_2)}\,.$$

Use Mordell's inequality (problem 2.1). When $n \leqslant 6$, it gives us that the right-hand side of this inequality is greater than or equal to $n/2$.

1.2. a) [22] Take as x_1, x_2, …, x_{20} numbers

$$\begin{array}{llllllllll} 1+5\varepsilon, & 6\varepsilon, & 1+4\varepsilon, & 5\varepsilon, & 1+3\varepsilon, & 4\varepsilon, & 1+2\varepsilon, & 3\varepsilon, & 1+\varepsilon, & 2\varepsilon,\\ 1+2\varepsilon, & \varepsilon, & 1+3\varepsilon, & 2\varepsilon, & 1+4\varepsilon, & 3\varepsilon, & 1+5\varepsilon, & 4\varepsilon, & 1+6\varepsilon, & 5\varepsilon. \end{array}$$

Then $f(x_1,\ldots,x_{20}) < 10 - \varepsilon^2 + c\varepsilon^3 < 10$ for some c and small enough ε.

b) [27] Take as x_1, x_2, …, x_{14} numbers

$$1+7\varepsilon,\ 7\varepsilon,\ 1+4\varepsilon,\ 6\varepsilon,\ 1+\varepsilon,\ 5\varepsilon,\ 1,\ 2\varepsilon,\ 1+\varepsilon,\ 0,\ 1+4\varepsilon,\ \varepsilon,\ 1+6\varepsilon,\ 4\varepsilon.$$

Then $f(x_1,\ldots,x_{20}) < 7 - 2\varepsilon^2 + c\varepsilon^3 < 7$ for some c and small enough ε.

An alternative example [24]:

$$0,\ 42,\ 2,\ 42,\ 4,\ 41,\ 5,\ 39,\ 4,\ 38,\ 2,\ 38,\ 0,\ 40\,.$$

c) [10], [18]. Take

$$0,\ 85,\ 0,\ 101,\ 0,\ 120,\ 14,\ 129,\ 41,\ 116,\ 59,\ 93,\ 64,71,\ 63,\ 52,\ 60,\ 36,\ 58,\ 23,\ 58,\ 12,\ 62,\ 3,\ 71.$$

Alternatively, in [3] the following example is given:

$$32,\ 0,\ 37,\ 0,\ 43,\ 0,\ 50,\ 0,\ 59,\ 8,\ 62,\ 21,\ 55,29,\ 44,\ 32,\ 33,\ 31,\ 24,\ 30,\ 16,\ 29,\ 10,\ 29,\ 4.$$

1.3. The statement of the problem is published in [13]. We present here a short nice solution.

Let $x_1 \geqslant x_2 \geqslant \cdots \geqslant x_n > 0$. Observe that the product of n fractions $\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}$ is equal to 1. Then by Cauchy inequality we conclude that

$$\sum_{k=1}^{n}\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}\geqslant n=\sum_{k=1}^{n}\frac{x_{k+1}+x_{k+2}}{x_{k+1}+x_{k+2}}.$$

Hence

$$\sum_{k=1}^{n}\frac{x_k}{x_{k+1}+x_{k+2}}\geqslant\sum_{k=1}^{n}\frac{x_{k+2}}{x_{k+1}+x_{k+2}}=\sum_{k=1}^{n}\frac{x_{k+1}}{x_k+x_{k+1}}. \tag{4}$$

Now we will apply the *rearranging inequality*: Let $a_1 \geqslant \ldots \geqslant a_n$ and $b_1 \geqslant \ldots \geqslant b_n$ be two sets of numbers. Then for each permutation k_1, …, k_n of numbers 1, …, n the following inequality holds

$$a_1b_1 + a_2b_2 + \ldots a_nb_n \geqslant a_1b_{k_1} + a_2b_{k_2} + \ldots a_nb_{k_n} \geqslant a_1b_n + a_2b_{n-1} + \ldots a_nb_1.$$

Use the rearranging inequality twice

$$\sum_{k=1}^{n}\frac{x_k}{x_{k+1}+x_{k+2}}=\sum_{k=1}^{n-2}\frac{x_k}{x_{k+1}+x_{k+2}}+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2}\underset{(*)}{\geqslant}$$
$$\geqslant\sum_{k=1}^{n-2}\frac{x_k}{x_{k+1}+x_{k+2}}+\frac{x_{n-1}}{x_1+x_2}+\frac{x_n}{x_n+x_1}\underset{(**)}{\geqslant}$$
$$\geqslant\sum_{k=1}^{n}\frac{x_k}{x_k+x_{k+1}}.$$

The inequality (*) here is the rearranging inequality for two pairs of numbers: $x_{n-1}\geqslant x_n$ and $\frac{1}{x_n+x_1}\geqslant\frac{1}{x_1+x_2}$; and the inequality (**) is the rearranging inequality for the sets x_1, x_2, …, x_{n-1} and $\frac{1}{x_1+x_2}$, $\frac{1}{x_2+x_3}$, …, $\frac{1}{x_{n-1}+x_n}$ that have opposite ordering.

Thus

$$2\sum_{k=1}^{n}\frac{x_k}{x_{k+1}+x_{k+2}}\geqslant\sum_{k=1}^{n}\frac{x_k}{x_k+x_{k+1}}+\sum_{k=1}^{n}\frac{x_{k+1}}{x_k+x_{k+1}}=n.$$

For the decreasing set x_i the solution is similar because we do not use the order of the variables when we apply the Cauchy inequality, and for the rearranging inequalities we need the fact that the sets x_i and $\frac{1}{x_i+x_{i+1}}$ have different orderings.

1.4. [3] It is easy to see that $f_{n+2}(x_1,x_2,\dots,x_n,x_1,x_2)=f_n(x_1,x_2,\dots,x_n)+1$. Therefore if $f_n(x_1,x_2,\dots,x_n)<n/2$, then $f_{n+2}(x_1,x_2,\dots,x_n,x_1,x_2)<(n+2)/2$.

1.5. [3] Assume that $f_m(x_1,x_2,\dots,x_m)<\frac{m}{2}$. For each k let us calculate the difference

$$f_{m+1}(x_1,\dots,x_k,x_k,x_{k+1},\dots,x_m)-f_m(x_1,x_2,\dots,x_m)-\frac{1}{2}=$$
$$=\frac{x_{k-1}}{2x_k}+\frac{x_k}{x_k+x_{k+1}}-\frac{x_{k-1}}{x_k+x_{k+1}}-\frac{1}{2}=\frac{(x_k-x_{k-1})(x_k-x_{k+1})}{2x_k(x_k+x_{k+1})}.$$

If $(x_k-x_{k-1})(x_k-x_{k+1})\leqslant 0$, then

$$f_{n+1}(x_1,x_2,\dots,x_k,x_k,x_{k+1},\dots,x_m)<\frac{m+1}{2}$$

and we are done. If n is odd, we can always choose k such that $(x_k-x_{k-1})(x_k-x_{k+1})\leqslant 0$ because otherwise the product of the (odd number of) inequalities $(x_k-x_{k-1})(x_{k+1}-x_k)<0$ for all k is

$$(x_2-x_1)^2(x_3-x_2)^2\dots(x_m-x_{m-1})^2(x_1-x_m)^2<0\,.$$

Thus if for odd n the Shapiro inequality is wrong then for $n+1$ it is wrong, too. It remains to apply the statement of the previous problem.

1.6. [7, 8]

1.7. [28] Let $y_k=x_k+x_{k+1}$. Then

$$\frac{x_1+x_4}{x_2+x_3}+\frac{x_2+x_5}{x_3+x_4}+\ldots+\frac{x_n+x_3}{x_1+x_2}=\sum_{k=1}^{n}\frac{y_k-y_{k+1}+y_{k+2}}{y_{k+1}}=\sum_{k=1}^{n}\frac{y_k}{y_{k+1}}+\sum_{k=1}^{n}\frac{y_{k+2}}{y_{k+1}}-n\geqslant n,$$

because by Cauchy inequality each sum is at least n.

1.8. The statements a), b) were published in [21].

a) !!! This short proof is taken from [8].

Denote for brevity $a=(a_1,a_2,\dots,a_n)$, $x=(x_1,x_2,\dots,x_n)$, and $u=(-1,1,-1,1,\dots,-1,1)$.

Observe that

$$\frac{\partial f}{\partial x_k}(x)=\frac{1}{x_{k+1}+x_{k+2}}-\frac{x_{k-2}}{(x_{k-1}+x_k)^2}-\frac{x_{k-1}}{(x_k+x_{k+1})^2}.$$

It is easy to see that we have an identity

$$f(x+tu)=f(x)+t\sum_{k=1}^{n}(-1)^k\frac{\partial f}{\partial x_k}(x).$$

Since a is the minimum point, we have

$$\frac{\partial f}{\partial x_k}(a)=0.$$

Therefore $f(a+tu) = f(a)$ if all the coordinates of the point $a+tu$ are positive. Hence $a+tu$ is the minimum point of the function f as well. Hence,

$$\frac{\partial f}{\partial x_k}(a+tu) = 0.$$

So

$$\frac{1}{a_{k+1}+a_{k+2}} - \frac{a_{k-2}}{(a_{k-1}+a_k)^2} - \frac{a_{k-1}}{(a_k+a_{k+1})^2} = 0$$

and

$$\frac{1}{a_{k+1}+a_{k+2}} - \frac{a_{k-2}+t(-1)^{k-2}}{(a_{k-1}+a_k)^2} - \frac{a_{k-1}+t(-1)^{k-1}}{(a_k+a_{k+1})^2} = 0.$$

Subtract the first equality from the second:

$$\frac{t}{(a_{k-1}+a_k)^2} - \frac{t}{(a_k+a_{k+1})^2} = 0.$$

Therefore,

$$a_{k-1}+a_k = a_k+a_{k+1}.$$

and hence

$$a_1 = a_3 = a_5 = \cdots = a_{n-1} \quad \text{and} \quad a_2 = a_4 = a_6 = \cdots = a_n.$$

Thus, $f(a) = n/2$.

b) This short proof is taken from [7]. Denote for brevity $a = (a_1, a_2, \ldots, a_n)$, $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, $z = (z_1, z_2, \ldots, z_n)$, where $y_k = x_k + x_{k+1}$ and $z_k = 1/y_{n+1-k}$.

Set

$$S(x) = \frac{x_1}{x_2+x_3} + \frac{x_2}{x_3+x_4} + \ldots + \frac{x_{n-1}}{x_n+x_1} + \frac{x_n}{x_1+x_2} = \sum_{k=0}^{n-1} \frac{x_k}{y_{k+1}}.$$

Observe that

$$\frac{\partial f}{\partial x_k}(x) = \frac{1}{x_{k+1}+x_{k+2}} - \frac{x_{k-2}}{(x_{k-1}+x_k)^2} - \frac{x_{k-1}}{(x_k+x_{k+1})^2}.$$

It is easy to check the following identities:

$$\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d} + \frac{\dfrac{a}{b^2} + \dfrac{c}{d^2}}{\dfrac{1}{b} + \dfrac{1}{d}}.$$

Hence,

$$\frac{x_{k-2}}{x_{k-1}+x_k} + \frac{x_{k-1}}{x_k+x_{k+1}} = \frac{x_{k-2}+x_{k-1}}{(x_{k-1}+x_k)+(x_k+x_{k+1})} + \frac{\dfrac{x_{k-2}}{(x_{k-1}+x_k)^2} + \dfrac{x_{k-1}}{(x_k+x_{k+1})^2}}{\dfrac{1}{x_{k-1}+x_k} + \dfrac{1}{x_k+x_{k+1}}} =$$

$$= \frac{y_{k-2}}{y_{k-1}+y_k} + \frac{z_{n-k} - \dfrac{\partial f}{\partial x_k}(x)}{z_{n-k+1}+z_{n-k+2}}.$$

Therefore,

$$2S(x) = S(y) + S(z) - \sum_{k=1}^{n} \frac{\dfrac{\partial f}{\partial x_k}(x)}{z_{n-k+1}+z_{n-k+2}}.$$

If x is a minimum point then we have $2S(x) = S(y) + S(z)$. Hence $S(x) = S(y) = S(z)$.

Let $u := (x_1 + x_2 + \cdots + x_n)/n$. Consider the transformation $M : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$M(x) = \Big(\frac{x_1+x_2}{2}, \frac{x_2+x_3}{2}, \ldots, \frac{x_n+x_1}{2}\Big).$$

Let $M_k(x)$ be its k-th iteration. Observe that $S(x) = S(y) = S(M(x)) = \cdots = S(M_k(x))$. It is clear that $\lim\limits_{k\to\infty} M_k(x) = (u, u, \ldots, u)$. Then

$$S(x) = \lim_{k\to\infty} S(M_k(x)) = S((u, u, \ldots, u)) = \frac{n}{2}.$$

c) [16], [7, 8]

1.9. These solutions are taken from [3].

ПїS) The problem was presented at the Third USSR mathematical olympiad, 1969. Probably it was originally published in [14].

Let x_{i_1} be the maximal number among $x_1, x_2, \ldots, x_n$; x_{i_2} be the maximum of the two next numbers after x_{i_1} (i.e. of x_{i_1+1} and x_{i_1+2}); x_{i_3} be the maximum of the two next numbers after x_{i_2}, and so on. We will continue this sequence till the step number k when the maximum of the two next after x_{i_k} numbers is x_{i_1}.

It is clear that $k \geqslant n/2$. We have

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_n}{x_1+x_2} \geqslant \frac{x_{i_1}}{2x_{i_2}}+\frac{x_{i_2}}{2x_{i_3}}+\ldots+\frac{x_{i_k}}{2x_{i_1}}.$$

The last expression is at least $k/2$ by the Cauchy inequality therefore it is at least $n/4$.

b) Rewrite each of the fractions $\frac{x_k}{x_{k+1}+x_{k+2}}$, $k = 1, 2, \ldots, n$, in the form

$$\frac{x_k}{x_{k+1}+x_{k+2}} = \frac{x_k+\frac{1}{2}x_{k+1}}{x_{k+1}+x_{k+2}}+\frac{\frac{1}{2}x_{k+1}+x_{k+2}}{x_{k+1}+x_{k+2}}-1\,.$$

We obtain $2n$ fractions. Combine them by pairs: the first and the last, the second and the third, the fourth and the fifth and so on. Now estimate the sum of each pair from below

$$\frac{\frac{1}{2}x_k+x_{k+1}}{x_k+x_{k+1}}+\frac{x_k+\frac{1}{2}x_{k+1}}{x_{k+1}+x_{k+2}} \geqslant 2\sqrt{\frac{(\frac{1}{2}x_k+x_{k+1})(x_k+\frac{1}{2}x_{k+1})}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}} =$$

$$= 2\sqrt{\left(\frac{1}{2}+\frac{x_kx_{k+1}}{4(x_k+x_{k+1})^2}\right)\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}} > \sqrt{2}\cdot\sqrt{\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}.$$

Since the product of n numbers $\sqrt{\frac{x_1+x_2}{x_2+x_3}}$, $\sqrt{\frac{x_2+x_3}{x_3+x_4}}$, $\ldots$, $\sqrt{\frac{x_n+x_1}{x_1+x_2}}$ equals 1, then by the Cauchy inequality their sum is at least n. Therefore $f(x_1,\ldots,x_n) \geqslant \sqrt{2}n-n = (\sqrt{2}-1)n$.

c) As in the previous solution rewrite each of the fractions $\frac{x_k}{x_{k+1}+x_{k+2}}$, $k = 1, 2, \ldots, n$, in the form

$$\frac{x_k}{x_{k+1}+x_{k+2}} = \frac{x_k+\beta x_{k+1}}{x_{k+1}+x_{k+2}}+\alpha\cdot\frac{\beta x_{k+1}+x_{k+2}}{x_{k+1}+x_{k+2}}-\alpha\,,$$

where α and β are parameters chosen to make the equality true. For such a choice of α and β we need $\beta+\alpha\beta=\alpha$, i.e. $\beta = \alpha/(\alpha+1)$. Then

$$\frac{x_k+\beta x_{k+1}}{x_{k+1}+x_{k+2}}+\alpha\cdot\frac{\beta x_k+x_{k+1}}{x_k+x_{k+1}} \geqslant 2\sqrt{\alpha\frac{(x_k+\beta x_{k+1})(\beta x_k+x_{k+1})}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}} =$$

$$= 2\sqrt{\alpha\frac{\beta(x_k+x_{k+1})^2+(\beta-1)^2x_kx_{k+1}}{(x_k+x_{k+1})(x_{k+1}+x_{k+2})}} > 2\sqrt{\alpha\beta\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}} = \frac{2\alpha}{\sqrt{\alpha+1}}\cdot\sqrt{\frac{x_k+x_{k+1}}{x_{k+1}+x_{k+2}}}.$$

Therefore

$$\frac{x_1}{x_2+x_3}+\frac{x_2}{x_3+x_4}+\ldots+\frac{x_{n-1}}{x_n+x_1}+\frac{x_n}{x_1+x_2} \geqslant \frac{2\alpha}{\sqrt{\alpha+1}}\left(\sqrt{\frac{x_1+x_2}{x_2+x_3}}+\sqrt{\frac{x_2+x_3}{x_3+x_4}}+\ldots+\sqrt{\frac{x_n+x_1}{x_1+x_2}}\right)-\alpha n >$$

$$> \frac{2\alpha}{\sqrt{\alpha+1}}n-\alpha n = \left(\frac{2\alpha}{\sqrt{\alpha+1}}-\alpha\right)n\,.$$

For $\alpha = \frac{5}{4}$ we have $c = 5/12$.

Remark. This is a good approximation.The expression $g(\alpha) = \frac{2\alpha}{\sqrt{\alpha+1}}-\alpha$ reaches its maximal value at $\alpha = \alpha_0 \approx 1.1479$ (this is a root of the cubic equation $g'(\alpha)=0$), and the minimum value is $g(\alpha_0)\approx 0.4186$. For $\alpha=\frac{5}{4}$ we have $g(\alpha)=\frac{5}{12}\approx 0.416$.

1.10. [9]. Set $y_k = x_k+x_{k+1}$. We need to prove that

$$\frac{x_1}{y_2}+\frac{x_2}{y_3}+\ldots+\frac{x_n}{y_1} \geqslant \frac{n}{2}\,,$$

or

$$\sum_{k=1}^{n} \frac{2q_n^2 x_k - y_{k+1}}{y_{k+1}} \geqslant n(q_n^2 - 1).$$

We suppose that the parameter q_n will be chosen later. Since

$$2q_n^2 x_k - y_{k+1} = (q_n^2 x_k - x_{k+1}) + (q_n^2 x_k - x_{k+2}) \geqslant 0\,,$$

by the Cauchy-Bunyakovsky inequality for sets

$$\left\{ \sqrt{\frac{2q_n^2 x_k - y_{k+1}}{y_{k+1}}} \right\} \text{ and } \left\{ \sqrt{(2q_n^2 x_k - y_{k+1}) y_{k+1}} \right\}$$

we have

$$\sum_{k=1}^{n} \frac{2q_n^2 x_k - y_{k+1}}{y_{k+1}} \geqslant \frac{\Big(\sum\limits_{k=1}^{n} (2q_n^2 x_k - y_{k+1})\Big)^2}{\sum\limits_{k=1}^{n} (2q_n^2 x_k - y_{k+1}) y_{k+1}}.$$

So it suffices to prove that

$$A^2 := \left(\sum_{k=1}^{n} (2q_n^2 x_k - y_{k+1}) \right)^2 \geqslant n(q_n^2 - 1) \sum_{k=1}^{n} (2q_n^2 x_k - y_{k+1}) y_{k+1} =: n(q_n^2 - 1) B.$$

Since $\sum\limits_{k=1}^{n} y_k = 2 \sum\limits_{k=1}^{n} x_k$, we have

$$A = (q_n^2 - 1) \sum_{k=1}^{n} y_k\,,$$

$$B = 2q_n^2 \sum_{k=1}^{n} x_k y_{k+1} - \sum_{k=1}^{n} y_k^2 = 2q_n^2 \sum_{k=1}^{n} y_k y_{k+1} - (q_n^2 + 1) \sum_{k=1}^{n} y_k^2\,.$$

So it remains to prove that

$$(q_n^2 - 1) \left(\sum_{k=1}^{n} y_k \right)^2 \geqslant n \left(2q_n^2 \sum_{k=1}^{n} y_k y_{k+1} - (q_n^2 + 1) \sum_{k=1}^{n} y_k^2 \right). \tag{5}$$

Transform the left-hand side using the relation

$$\left(\sum_{k=1}^{n} y_k \right)^2 = n \sum_{k=1}^{n} y_k^2 - \sum_{i<k} (y_i - y_k)^2.$$

The inequiality (5) will be transformed to

$$n \sum_{k=1}^{n} (y_k - y_{k+1})^2 \geqslant \left(1 - \frac{1}{q_n^2} \right) \sum_{i<k} (y_i - y_k)^2.$$

By the Cauchy-Bunyakovsky inequality

$$\sum_{k=1}^{n} (y_k - y_{k+1})^2 \geqslant \sum_{j=i}^{k-1} (y_j - y_{j+1})^2 \geqslant \frac{1}{k-j} \left(\sum_{j=i}^{k-1} (y_j - y_{j+1}) \right)^2 = \frac{1}{k-j} (y_i - y_k)^2 \geqslant \frac{1}{n-1} (y_i - y_k)^2.$$

Hence

$$\frac{n(n-1)}{2} \sum_{k=1}^{n} (y_k - y_{k+1})^2 \geqslant \frac{1}{n-1} \sum_{i<k} (y_i - y_k)^2.$$

So we can take $1 - \frac{1}{q_n^2} = \frac{2}{(n-1)^2}$, i. e. $q_n = \frac{n-1}{\sqrt{n^2-2n-1}} > 1$.

R e m a r k. When n tends to infinity, the values q_n which are found above tend to 1.

b)

1.11. (a) Denote $k_i := x_{i+1}/x_i$. Then

$$S = \frac{1}{k_1(k_2+1)} + \frac{1}{k_2(k_3+1)} + \cdots + \frac{1}{k_n(k_1+1)} \geqslant \frac{1}{a_1(a_n+1)} + \frac{1}{a_2(a_{n-1}+1)} + \cdots + \frac{1}{a_n(a_1+1)}.$$

(b) The inequality holds because

$$\frac{1}{a_i(a_{n+1-i}+1)}+\frac{1}{a_{n+1-i}(a_i+1)}=\frac{1+\frac{a_ia_{n+1-i}-1}{(1+a_i)(1+a_{n+1-i})}}{a_ia_{n+1-i}}\geqslant b_i$$

where the latter inequality holds because $(1+a_i)(1+a_{n+1-i})\geqslant(1+\sqrt{a_ia_{n+1-i}})^2$.

(c) The first inequality $2S\geqslant g(\ln(a_1a_n))+g(\ln(a_2a_{n-1}))+\dots+g(\ln(a_na_1))$ holds because $g(x)$ is less than both e^{-x} and $2(e^x+e^{x/2})^{-1}$. The second inequality holds by the Jensen inequality because g is convex.

(d) [Dr]

2.1. a) [20]

For $n=4$ we need to prove that

$$(x_1+x_2+x_3+x_4)^2\geqslant 2x_1x_2+2x_2x_3+2x_3x_4+2x_4x_1+4x_1x_3+4x_2x_4\,.$$

This follows from the inequality

$$x_1^2+x_2^2+x_3^2+x_4^2\geqslant 2x_1x_3+2x_2x_4\,.$$

For $n=3$ and $n=5$ re-write the inequality. We need to prove that

$$(n-1)(a_1+a_2+\ldots+a_n)^2\geqslant 2n\sum_{i<k}a_ia_k. \tag{6}$$

Indeed, notice that the Cauchy–Bunyakovsky inequality applied to sets $a_1,a_2,\dots,a_n$ and $1,1,\dots,1$ gives us:

$$n(a_1^2+a_2^2+\ldots+a_n^2)\geqslant(a_1+a_2+\ldots+a_n)^2.$$

Now we have

$$n(a_1+a_2+\ldots+a_n)^2=n(a_1^2+a_2^2+\ldots+a_n^2)+2n\sum_{i<k}a_ia_k\geqslant(a_1+a_2+\ldots+a_n)^2+2n\sum_{i<k}a_ia_k\,,$$

which implies (6).

Now assume that $n\geqslant 6$. We may suppose that $x_3\geqslant x_1$ and $x_3\geqslant x_2$ (e.g. make a cyclic shift of variables such that x_3 becomes the maximum). For $r=1$, 2, and 3 denote by a_r the sum of all x_k such that $k\equiv r\pmod 3$ and $k\leqslant n$. Then $x_1+x_2+\ldots+x_n=a_1+a_2+a_3$. Hence by (6) we have

$$(x_1+x_2+\ldots+x_n)^2=(a_1+a_2+a_3)^2\geqslant 3(a_1a_2+a_2a_3+a_3a_1)=3\cdot\sum_{(i-k)\not\vdots 3}x_ix_k\,.$$

Set

$$A:=\sum_{(i-k)\not\vdots 3}x_ix_k\quad\text{and}\quad B:=\sum_{k=1}^{n}x_k(x_{k+1}+x_{k+2}).$$

We have $A\geqslant B$ because

- for $n\equiv 0\pmod 3$ all the summands of B are contained in A;
- for $n\equiv 1\pmod 3$ the sum A contains all the summands of B except x_nx_1, but x_nx_1 does not exceed x_nx_3;
- for $n\equiv 2\pmod 3$ the sum A contains all the summands of B except $x_{n-1}x_1$ and x_nx_2, but these summands do not exceed $x_{n-1}x_3$ and x_nx_3, respectively.

Hence

$$(x_1+x_2+\ldots+x_n)^2\geqslant 3A\geqslant 3B=3\sum_{k=1}^{n}x_k(x_{k+1}+x_{k+2}).$$

In order to show that $\min\left\{\frac{n}{2},3\right\}$ is the sharp constant for $n\leqslant 6$ we set $x_1=x_2=\ldots=x_n=1$ and for $n\geqslant 6$ we set $x_1=x_2=x_3=1$ and $x_4=x_5=\ldots=x_n=0$.

b) The case $n<6$ is trivial. For $n=6$ the equality is achieved when $x_1+x_4=x_2+x_5=x_3+x_6$. For $n\geqslant 6$ the equality is achieved for the sets of form $(t,1,1,1-t,0,\dots,0)$, where $t\in[0,1]$, and their cyclic shifts.

2.2. [20]

For $n=4$ and $n=7$ this is a particular case of (6).

For $n=5$ the inequality coincides with $\sum(x_k-2x_{k+2}+x_{k+4})^2\geqslant 0$.

For $n=6$ the inequality follows from $x_1^2+x_2^2+\ldots+x_6^2\geqslant 2x_1x_4+2x_2x_5+2x_3x_6$.

For $n=8$ open brackets in the following corollary of the Cauchy–Bunyakovsky inequality

$$4(x_1^2+x_2^2+x_3^2+x_4^2)\geqslant(x_1+x_2+x_3+x_4)^2.$$

We obtain

$$3(x_1^2+x_2^2+x_3^2+x_4^2)\geqslant 2(x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4).$$

Hence

$$3(x_1+x_2+x_3+x_4)^2 \geqslant 8(x_1x_2+x_1x_3+x_1x_4+x_2x_3+x_2x_4+x_3x_4), \tag{7}$$

This is the required inequality for $n=8$.

Now assume that $n>8$. We may suppose that $x_4 \geqslant x_1$, $x_4 \geqslant x_2$, and $x_4 \geqslant x_3$. For $r=1$, 2, 3, and 4 denote by a_r the sum of all x_k such that $k \equiv r \pmod 4$ and $k \leqslant n$. Then $x_1+x_2+\ldots+x_n = a_1+a_2+a_3+a_4$. Hence by (7)

$$3(x_1+x_2+\ldots+x_n)^2 = 3(a_1+a_2+a_3+a_4)^2 \geqslant 8(a_1a_2+a_2a_3+a_3a_4+a_4a_1) \geqslant 8\cdot \sum_{(i-k)\not\vdots 4} x_ix_k\,.$$

Set

$$A := \sum_{(i-k)\not\vdots 4} x_ix_k \quad \text{and} \quad B := \sum_{k=1}^{n} x_k(x_{k+1}+x_{k+2}+x_{k+3}).$$

We have $A \geqslant B$ because

- for $n \equiv 0 \pmod 4$ all the summands of B are contained in A;
- for $n \equiv 1 \pmod 4$ the sum A contains all the summands of B except x_nx_1, but x_nx_1 does not exceed x_nx_4;
- for $n \equiv 2 \pmod 4$ the sum A contains all the summands of B except $x_{n-1}x_1$ and x_nx_2, but these summands do not exceed $x_{n-1}x_4$ and x_nx_4;
- for $n \equiv 3 \pmod 4$ the sum A contains all the summands of B except $x_{n-2}x_1$, $x_{n-1}x_2$ and x_nx_3, but these summands do not exceed $x_{n-2}x_4$, $x_{n-1}x_4$, and x_nx_4.

Hence

$$3(x_1+x_2+\ldots+x_n)^2 \geqslant 8A \geqslant 8B = 8\sum_{k=1}^{n} x_k(x_{k+1}+x_{k+2}+x_{k+3}).$$

2.3. a) Cf. [11]. By the Cauchy–Bunyakovsky inequality and Problem 2.2 we have

$$\frac{x_1}{x_2+x_3+x_4}+\frac{x_2}{x_3+x_4+x_5}+\ldots+\frac{x_{n-1}}{x_n+x_1+x_2}+\frac{x_n}{x_1+x_2+x_3} \geqslant \frac{(x_1+x_2+\ldots+x_n)^2}{\sum x_k(x_{k+1}+x_{k+2}+x_{k+3})} \geqslant \frac{n}{3}.$$

b) ???

2.4. [1, Problem 187]. We may assume that $x_1 \leqslant x_2$. Set

$$S := x_1+x_2+\ldots+x_n, \quad S_1 := x_1+x_3+\ldots, \quad S_2 := x_2+x_4+\ldots.$$

Then $S_1^2+S_2^2 \geqslant (S_1+S_2)^2/2 = S^2/2$. Hence

$$\frac{S^2}{2} \geqslant S^2-S_1^2-S_2^2 = 2\sum_{(i-k)\not\vdots 2} x_ix_k. \tag{8}$$

If n is even, then the last sum contains all the summands of form x_kx_{k+1}. If n is odd, then the summand x_nx_1 is missing, however the sum contains a greater summand x_nx_2. So

$$\frac{S^2}{2} \geqslant 2(x_1x_2+x_2x_3+\ldots+x_nx_1).$$

2.5. See the solution of 1.3 up to the inequality (4).

2.6. Induction on $n \geqslant n=4$. Denote the left-hand side by L_n. We have

$$L_4 = \frac{x_1+x_3}{x_2+x_4}+\frac{x_2+x_4}{x_1+x_3} = a+a^{-1} \geqslant 2.$$

Let us prove the inductive step. We may assume that x_{n+1} is the minimal of all x_i's. Now remove the last summand from L_{n+1}, and then decrease two others. We obtain

$$L_{n+1} \geqslant \frac{x_1}{x_{n+1}+x_2}+\ldots+\frac{x_n}{x_{n-1}+x_{n+1}} \geqslant \frac{x_1}{x_n+x_2}+\ldots+\frac{x_n}{x_{n-1}+x_n} = L_n\,.$$

In order to show that the constant 2 is sharp, take

$$x_1=x_2=1, \quad x_3=t, \quad x_4=t^2, \quad \ldots, \quad x_n=t^{n-2}\,.$$

When $t \to +0$, the first two summands tend to 1 and the remaining tends to 0.

2.7. [10]. Set $S := x_1 + x_2 + \ldots + x_n$. Use the Cauchy–Bunyakovsky inequality for sets $\left\{\frac{x_k+x_{k+1}}{x_k+x_{k+2}}\right\}$ and $\{(x_k + x_{k+1})(x_k + x_{k+2})\}$. We obtain

$$\frac{x_1+x_2}{x_1+x_3}+\frac{x_2+x_3}{x_2+x_4}+\ldots+\frac{x_{n-1}+x_n}{x_{n-1}+x_1}+\frac{x_n+x_1}{x_n+x_2}\geqslant\frac{4(x_1+x_2+\ldots+x_n)^2}{\sum\limits_{k=1}^{n}(x_k+x_{k+1})(x_k+x_{k+2})}.$$

So it suffices to prove that

$$S^2\geqslant\sum_{k=1}^{n}(x_k+x_{k+1})(x_k+x_{k+2})=\sum_{k=1}^{n}x_k^2+2\sum_{k=1}^{n}x_kx_{k+1}+\sum_{k=1}^{n}x_kx_{k+2}.$$

This can be shown by opening brackets in the left-hand side, because for $n \geqslant 4$ all the summands x_kx_{k+1} and x_kx_{k+2}, where $k = 1, 2, \ldots, n$, are different.

In order to show that the constant 4 is sharp, take $x_k = a^{k-1}$ for $k = 1, 2, \ldots, n-1$ and $x_n = a^{n-2}$. When $a \to \infty$, the first $n-3$ summands tend to 0 and the remaining summands tend to 1, 2 and 1.

Using the Cauchy–Bunyakovsky inequality as it is done in the solution of the next problem, the reader will easily find another solution of this problem reducing it to the inequality from Problem 2.4.

2.8. [6]. Use the Cauchy–Bunyakovsky inequality for sets $\left\{\frac{x_k}{x_{k-1}+x_{k+2}}\right\}$ и $\{x_k(x_{k-1} + x_{k+2})\}$. We obtain

$$\frac{x_1}{x_n+x_3}+\frac{x_2}{x_1+x_4}+\ldots+\frac{x_{n-1}}{x_{n-2}+x_1}+\frac{x_n}{x_{n-1}+x_2}\geqslant\frac{(x_1+x_2+\ldots+x_n)^2}{(x_1x_2+x_2x_3+\ldots+x_nx_1)+(x_1x_3+x_2x_4+\ldots+x_nx_2)}.$$

So it suffices to prove that

$$S^2\geqslant 3(x_1x_2+x_2x_3+\ldots+x_nx_1)+3(x_1x_3+x_2x_4+\ldots+x_nx_2)=:3Y,$$

where $S := x_1 + x_2 + \ldots + x_n$. Set

$$S_1 := x_1 + x_4 + \ldots, \quad S_2 = x_2 + x_5 + \ldots \quad \text{and} \quad S_3 = x_3 + x_6 + \ldots.$$

Then $S = S_1 + S_2 + S_3$ and $S_1^2 + S_2^2 + S_3^2 \geqslant S^2/3$. We may assume that $x_3 \geqslant x_1$ and $x_3 \geqslant x_2$. Notice that

$$S^2\geqslant\frac{3}{2}(S^2-S_1^2-S_2^2-S_3^2)=3\sum_{(i-k)\not\vdots 3}x_ix_k=:3Z.$$

- If $n \equiv 0 \pmod 3$, then all the summands of Y are contained in Z.
- If $n \equiv 1 \pmod 3$, then Z contains all the summands of Y except x_nx_1, but this summand does not exceed x_nx_3.
- If $n \equiv 2 \pmod 3$, then Z contains all the summands of Y except $x_{n-1}x_1$ and x_nx_2, but these summands do not exceed $x_{n-1}x_3$ and x_nx_3.

Hence $S^2 \geqslant 3Y \geqslant 3Z$, which proves the initial inequality.

In order to show that the constant 3 is sharp, take $x_k = a^{k-1}$ for $k = 1, 2, \ldots, n-2$ and $x_{n-1} = x_n = 1$. When $a \to 0$, the first and the last two summands tend to 1, while the remaining summands tend to 0.

2.9. [5]. The inequality is obtained by summing two inequalities of 2.8 (for the direct and the opposite order of variables).

In order to show that the constant 6 is sharp, take $x_k = a^{k-1}$ for $k = 1, 2, \ldots, n-2$ and $x_{n-1} = x_n = 1$. When $a \to 0$, the last four summands tend to 1, 2, 2, 1, respectively; the remaining tend to 0.

2.10. This is conjectured in [19].

The following proof is due to P.Milošević и M. Bukić, participants of the Conference.

This inequality can be represented as sum of two inequalities for $n = 2004$ — the inequality from Problem 2.8 and the inequality

$$\frac{x_1}{x_1+x_4}+\frac{x_2}{x_2+x_5}+\ldots+\frac{x_n}{x_n+x_3}\geqslant 3\,.$$

Prove the last inequality. For $n = 3m$ it is the sum of three inequalities:

$$\frac{x_1}{x_1+x_4}+\frac{x_4}{x_4+x_7}+\ldots+\frac{x_{n-2}}{x_{n-2}+x_1}\geqslant 1\,.$$
$$\frac{x_2}{x_2+x_5}+\frac{x_5}{x_5+x_8}+\ldots+\frac{x_{n-1}}{x_{n-1}+x_2}\geqslant 1\,.$$
$$\frac{x_3}{x_3+x_6}+\frac{x_6}{x_6+x_9}+\ldots+\frac{x_n}{x_n+x_3}\geqslant 1\,.$$

Each of these inequalities can be re-written as

$$\frac{1}{1+a_1}+\frac{1}{1+a_3}+\ldots+\frac{1}{1+a_m}\geqslant 1 \qquad \text{where } a_1a_2\ldots a_m=1.$$

This can be shown by induction. The base $m=2$ is the following inequality:

$$\frac{1}{1+a_1}+\frac{1}{1+\frac{1}{a_1}}=1\geqslant 1\,.$$

To prove the induction step, let us check that

$$\frac{1}{1+b}+\frac{1}{1+c}\geqslant\frac{1}{1+bc}\,.$$

This can be done directly by reducing to a common denominator and opening brackets.

Here is the proof of A. Khrabrov. Let us prove that

$$Z:=\frac{x_1+x_2}{x_1+x_4}+\frac{x_2+x_3}{x_2+x_5}+\ldots+\frac{x_{3n}+x_1}{x_{3n}+x_3}\geqslant 6.$$

Set $x_{3n+k}:=x_k$ and, for $r=0,1,2$,

$$S_r:=\sum_{k=1}^{n}x_{3k+r},\quad X_r:=\sum_{k=1}^{n}\frac{x_{3k+r}}{x_{3k+r}+x_{3k+3+r}},\quad \text{and}\quad Y_r:=\sum_{k=1}^{n}\frac{x_{3k+r+1}}{x_{3k+r}+x_{3k+3+r}}.$$

First we prove that $X_r\geqslant 1$. Consider only the case $r=0$. Then

$$X_0S_0^2\geqslant X_0\left(\sum_{k=1}^{n}x_{3k}^2+\sum_{k=1}^{n}x_{3k}x_{3k+3}\right)=X_0\left(\sum_{k=1}^{n}x_{3k}(x_{3k}+x_{3k+3})\right)\geqslant S_0^2,$$

where the last inequality holds by the Cauchy–Bunyakovsky inequality. So $X_0\geqslant 1$.

Now prove that $Y_r\geqslant S_{r+1}/S_r$ (we set $S_3:=S_0$). Consider only the case $r=0$.

$$Y_0S_0S_1\geqslant Y_0\left(\sum_{k=1}^{n}x_{3k}x_{3k+1}+\sum_{k=1}^{n}x_{3k+1}x_{3k+3}\right)=Y_0\left(\sum_{k=1}^{n}x_{3k+1}(x_{3k}+x_{3k+3})\right)\geqslant S_1^2,$$

where the last inequality holds by the Cauchy–Bunyakovsky inequality. So $Y_0\geqslant S_1/S_0$.

Summing up all the proved inequalities we obtain

$$Z=X_0+X_1+X_2+Y_0+Y_1+Y_2\geqslant 3+\frac{S_1}{S_0}+\frac{S_2}{S_1}+\frac{S_0}{S_2}\geqslant 6.$$

In order to show that the constant 6 is sharp, take $x_1=x_2=x_3=1$, $x_k=a^{n-k+1}$ for $k=3,\,4,\,\ldots,\,n$. When $a\to 0$, the first and the second summands tend to 2, the third and the last tend to 1, and the remaining summands tend to 0.

2.11. This proof is due to A. Khrabrov. Set $S=x_1+x_2+\ldots+x_n$ and $T=\sum\limits_{(i-k)\not\vdots 2}x_ix_k$. By the Cauchy–Bunyakovsky inequality for sets $\left\{\dfrac{x_k}{x_{k-1}+x_{k+3}}\right\}$ and $\{x_k(x_{k-1}+x_{k+3})\}$ we have

$$\frac{x_1}{x_n+x_4}+\frac{x_2}{x_1+x_5}+\ldots+\frac{x_{n-1}}{x_{n-2}+x_2}+\frac{x_n}{x_{n-1}+x_3}\geqslant\frac{(x_1+x_2+\ldots+x_n)^2}{(x_1x_2+x_2x_3+\ldots+x_nx_1)+(x_1x_4+x_2x_5+\ldots+x_nx_3)}.$$

So it suffices to prove that

$$S^2\geqslant 4(x_1x_2+x_2x_3+\ldots+x_nx_1)+4(x_1x_4+x_2x_5+\ldots+x_nx_3).$$

In the solution of problem 2.4 we proved that $S^2\geqslant 4T$, see (8). So it suffices to prove that

$$T\geqslant(x_1x_2+x_2x_3+\ldots+x_nx_1)+(x_1x_4+x_2x_5+\ldots+x_nx_3). \tag{9}$$

Since n is even, all the summands of the right-hand sum are contained in the left-hand sum.

In order to show that the constant 6 is sharp, take $x_k=a^{k-1}$ and $k=1,\,2,\,\ldots,\,n-3$ and $x_{n-2}=x_{n-1}=x_n=1$. When $a\to+0$ the first summand and the three last summands tend to 1, and the remaining summands tend to 0.

2.12. [14]. Note that $a^2 - ab + b^2 \leqslant \max\{a, b\}^2$.

Let x_{i_1} be the maximal number of $x_1, x_2, \ldots, x_n$. Let x_{i_2} be the maximal number of x_{i_1+1} and x_{i_1+2}. Let x_{i_3} be the maximal number of x_{i_2+1} and x_{i_2+2}, and so on. There exists a number k such that $x_{i_{k+1}} = x_{i_1}$. Hence

$$\sum_{k=1}^{n} \frac{x_k^2}{x_{k+1}^2 - x_{k+1}x_{k+2} + x_{k+2}^2} \geqslant \sum_{j=1}^{k} \frac{x_{i_j}^2}{x_{i_{j+1}}^2} \geqslant k \geqslant \left[\frac{n+1}{2}\right],$$

where the latter inequality holds because $k \geqslant n/2$.

In order to show that the constant $\left[\frac{n+1}{2}\right]$ is sharp, take $x_k = 1$ for odd k and $x_k = 0$ for even k. Then the left-hand side is $\left[\frac{n+1}{2}\right]$.

References

[1] *Васильев Н. Б., Егоров А. А.* Задачи Всесоюзных математических олимпиад. М.: Наука, 1988.

[2] *Дринфельд В. Г.* Об одном циклическом неравенстве // Мат. заметки. 1971. Т. 9. № 2. С. 113–119.

[3] *Курляндчик Л. Д., Файбусович А.* История одного неравенства // Квант. 1991. № 4. С. 14–18.

[4] *Толпыго А. К.* Тысяча задач Международного математического Турнира городов. М.: МЦНМО, 2009.

[5] *Чимэдцэрэн С.* Нэгэн орчилт нийлбэр // Математикийн олимпиадын цуврал. 1999. Т. 22. (На монгольск. яз.).

[6] *Чимэдцэрэн С., Адъяасурен В., Батболд С.* Оценка в одной циклической сумме // Монгол улсын их сургууль, Эрдэм шинжилгээний бичиг. 2000. Т. 7 (168). С. 79–84.

[7] *Bushell P. J.* Shapiro's Cyclic Sum // Bull. London Math. Soc. 1994. Vol. 26. No 6. P. 564–574

[8] *Bushell P. J., McLeod J. B.* Shapiro's cyclic inequality for even n // J. Inequal. & Appl., 2002. Vol. 7(3). P. 331–348

[9] *Cîrtoaje V.* Crux Mathematicorum. 2006. Vol. 32. No. 8. Problem 3195.

[10] *Daykin D. E.* Inequalities for certain cyclic sums // Proc. Edinburgh Math. Soc. (2) 1970/71. Vol. 17. P. 257–262.

[11] *Diananda P. H.* Extensions of an inequality of H. S. Shapiro // Amer. Math. Monthly 1959. Vol. 66. P. 489–491.

[12] *Diananda P. H.* On a conjecture of L. J. Mordell regarding an inequality involoving quadratic forms // J. London Math. Soc. 1961. Vol. 36. P. 185–192.

[13] *Diananda P. H.* Inequalities for a class of cyclic and other sums // J. London Math. Soc. 1962. Vol. 37. P. 424–431.

[14] *Diananda P. H.* Some cyclic and other inequalities // Proc. Cambridge Philos. Soc. 1962. Vol. 58. P. 425–427.

[15] *Diananda P. H.* Some cyclic and other inequalities, II // Proc. Cambridge Philos. Soc. 1962. Vol. 58. P. 703–705.

[16] *Diananda P. H.* On a cyclic sum // Proc. Glasgow Math. Assoc. 1963. Vol. 6. P. 11–13.

[17] *Elbert A.* On a cyclic inequality // Period. Math. Hungarica. 1973. Vol. 4. № 2–3. P. 163–168.

[18] *Malcolm M. A.* A note on a conjecture of L. J. Mordell // Math. Comp. 1971. Vol. 25. P. 375–377.

[19] *Mitrinović D. S., Pečarić J. E., Fink A. M.* Classical and new inequalities in analysis. Kluwer Academic Publishers Group, Dordrecht, 1993. (Mathematics and its Applications (East European Series), Vol. 61).

[20] *Mordell L. J.* On the inequality $\sum\limits_{r=1}^{n} \frac{x_r}{x_{r+1}+x_{r+2}} \geqslant \frac{n}{2}$ and some others // Abh. Math. Sem. Univ. Hamburg. 1958. Vol. 22. P. 229–240.

[21] *Nowosad P.* Isoperimetric eigenvalue problems in algebras // Comm. Pure Appl. Math. 1968. Vol. 21. P. 401–465.

[22] *Shapiro H. S., Northover F. H.* Amer. Math. Monthly. 1956. Vol. 63. № 3. P. 191–192.

[23] *Tanahashi K., Tomiyama J.* Indecomposable positive maps in matrix algebras // Canad. Math. Bull. 1988. Vol. 31. № 3. P. 308–317.

[24] *Troesch B. A.* Full solution of Shapiro's cyclic inequality // Notices Amer. Math. Soc. 1985. Vol. 39. № 4. P. 318.

[25] *Vukmirović J.* A note on an inequality for the cyclic sums introduced by D. E. Daykin // Math. Balk. 1978. Vol. 8. P. 293–297.

[26] *Yamagami S.* Cyclic inequalities // Proc. Amer. Math. Soc. 1993. Vol. 118. № 2. P. 521–527.

[27] *Zulauf A.* Note on a conjecture of L. J. Mordell // Abh. Math. Sem. Univ. Hamburg. 1958. Vol. 22. P. 240–241.

[28] *Zulauf A.* Note on an Inequality // Math. Gazette. 1962. Vol. 46. №355. P. 41–42.

ТРОПИЧЕСКАЯ ГЕОМЕТРИЯ

А. Заславский, Ф. Нилов, А. Скопенков, М. Скопенков

Краткий обзор. [1]

16-я проблема Гильберта спрашивает, *каким может быть количество и взаимное расположение кривых, образующих подмножество плоскости, заданное уравнением* $\sum_{i+j\le d} a_{ij}x^iy^j = 0$. Более аккуратная формулировка и примеры приводятся в части А. [2] Цель данного цикла задач — частичное решение 16-й проблемы Гильберта для $d = 6$, а именно, *построение* требуемых подмножеств в наиболее содержательном случае (см. Основную Теорему ниже).

Для подмножества плоскости, заданного многочленом от двух переменных с некоторыми конкретными коэффициентами, определить число и расположение кривых не так-то просто (даже вооружившись современным компьютером). При решении задач части В Вы нащупаете формулировку основной леммы, которая позволит легко сделать это для многочленов некоторого специального вида. Вы увидите, как при рисовании подмножеств, заданных уравнениями вида $\sum_{i+j\le d} (a_{ij}x^iy^j)^N = 0$, где N — большое нечетное число, естественно возникает *тропическая геометрия*. Используя ее, Вы сможете строить такие подмножества с различным расположением овалов.

Исходные соображения тропической геометрии элементарны. Заменим умножение на сложение, а сложение — на операцию, связанную со сложением таким же *дистрибутивным* законом, каким сложение связано с умножением. В качестве такой операции можно взять *максимум* $\max\{a, b\}$ пары чисел a и b. При такой замене функция $\sum_{i+j\le d} b_{ij}x^iy^j$ перейдет в функцию вида (проверьте!): $f(x, y) = \max_{i+j\le d}(ix + jy + b_{ij})$. Множество точек излома такой функции называется *тропической кривой*. [3]

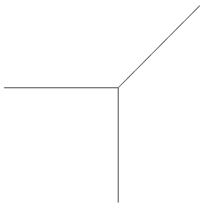


Рис. 1:

Например, прямая на плоскости задается уравнением $Ax + By + C = 0$. Левая часть данного уравнения при нашей замене переходит в функцию вида $f(x, y) = \max\{x + a, y + b, c\}$. Множество точек излома полученной функции $f(x, y)$ имеет вид, показанный на рисунке 1 (проверьте!). Так определяется *тропическая прямая*. Для тропических прямых сохраняются многие свойства обычных прямых. "Экспериментальной" проверке этих свойств посвящена часть С.

А. Примеры алгебраических кривых.

Многочленом (от двух переменных) называется функция $F : \mathbb{R}^2 \to \mathbb{R}$, для которой существуют числа d и a_{ij}, $0 \le i, j \le d$, такие что $F(x, y) = \sum_{i+j\le d} a_{ij}x^iy^j$. Вы можете пользоваться без доказательства следующим фактом: *для данной функции F такие числа единственны с точностью до увеличения d и выбора новых a_{ij} равными нулю.*

Множеством нулей многочлена F называется множество $F^{-1}(0) := \{(x, y) \in \mathbb{R}^2 \mid F(x, y) = 0\}$.

A1. Однозначно ли множество $F^{-1}(0)$ определяет многочлен F?

A2. Какие из следующих множеств являются множествами нулей многочленов?
(a) прямая; (b) окружность; (c) точка; (d) отрезок; (e) объединение двух прямых;
(f) "хрюшка" (объединение 6 окружностей) на рисунке 2.

[1] Не переживайте, если Вам что-нибудь непонятно в этом кратком обзоре. Вы можете пропустить его и начать решать задачи с любой из частей А или С.

[2] При $d \le 5$ ответ был известен еще в 19-м веке. Гильберт формулировал свою проблему для $d = 6$, в этом случае решение было получено Гудковым. Для $d = 7$ проблема была решена О.Я. Виро с использованием методов тропической геометрии. Для $d \ge 8$ данный вопрос до сих пор открыт.

[3] Не следует думать, что многочлену от двух переменных с некоторыми конкретными коэффициентами соответствует какая-то определенная тропическая кривая. Связь между многочленами и тропическими кривыми сложнее. В некотором смысле, тропическая кривая является "пределом" целого *семейства* подмножеств, заданных многочленами от двух переменных — см. часть В.

Степень многочлена — это наименьшее возможное число d, для которого найдутся требуемые a_{ij}. (*Комментарий.* Степень — это наибольшее возможное d, для которого $a_{i,d-i} \neq 0$ для некоторого представления многочлена и некоторого i.)

A3. (a) Сколько точек может быть в пересечении прямой с множеством нулей многочлена степени d?

(b) Множество нулей многочлена нечетной степени неограничено (то есть, не содержится ни в каком в диске).

Многочлен F *приводим*, если $F = G \cdot H$ для некоторых многочленов G и H.

Кривые[4]**.** Функция $\gamma : [a, b] \to \mathbb{R}$ называется *дифференцируемой* в точке t_0, если для некоторого числа A и любого $\varepsilon > 0$ существует δ, такое, что для любого

$$t \in (t_0 - \delta, t_0 + \delta) \quad \text{выполнено} \quad |\gamma(t) - \gamma(t_0) - A(t - t_0)| < \varepsilon |t - t_0|.$$

Отображение $\gamma : [a, b] \to \mathbb{R}^2$ можно рассматривать как упорядоченную пару функций $\gamma_1, \gamma_2 : [a, b] \to \mathbb{R}$. Дифференцируемость отображения $\gamma = (\gamma_1, \gamma_2)$ равносильна дифференцируемости функций γ_1 и γ_2.

(Гладкой) кривой на плоскости называется дифференцируемое отображение $\gamma : [a, b] \to \mathbb{R}^2$ (или $\gamma : \mathbb{R} \to \mathbb{R}^2$). Кривая $\gamma : [a, b] \to \mathbb{R}^2$ называется *замкнутой*, если $\gamma(a) = \gamma(b)$.

В задачах A4.cfg и A7 достаточно привести пример многочлена; доказательство его свойств не требуется.

A4. (a) Существует неприводимый многочлен степени 3, множество нулей которого содержит замкнутую кривую.

(b) То же для степени 4.

(c) Существует неприводимый многочлен степени 4, множество нулей которого содержит две замкнутых кривые, одну внутри другой.

(d) Если множество нулей многочлена степени 4 содержит две замкнутых кривые, одну внутри другой, то это множество не содержит никаких других точек.

(e) Верен ли аналог утверждения (d) для неприводимого многочлена степени 5?

(f) Существует многочлен степени 4, множество нулей которого содержит 4 замкнутые кривые.

(g) Существует многочлен степени 4, множество нулей которого содержит 3 замкнутые кривые.

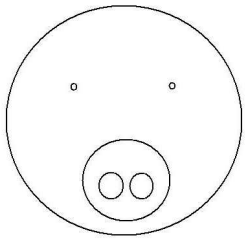


Рис. 2:

Овалы . Компоненты связности, на которые распадается множество нулей многочлена, называются *ветвями*. (Наличие неограниченных ветвей усложняет изучение множеств нулей многочленов.) Для неограниченной ветви B прямые, соединяющие начало координат O с точками на ветви B, стремящимися к "бесконечности", стремятся к некоторой "предельной" прямой. Две неограниченные ветви *элементарно эквивалентны*, если их "предельные" прямые совпадают.

A5. Неограниченные ветви гиперболы $xy = 1$ элементарно эквивалентны.

Две неограниченные ветви *эквивалентны*, если существует соединяющая их цепочка ветвей, в которой каждые две соседние ветви элементарно эквивалентны. Множество нулей многочлена *невырождено*, если все его ветви — гладкие кривые. *Овал* невырожденного множества нулей многочлена — это либо замкнутая кривая (содержащаяся в этом множестве), либо класс эквивалентности неограниченных ветвей. (Заметим, что это определение отлично от "правильного", приводящегося в учебниках.)

A6. Найдите все h, при которых множество нулей невырождено, и посчитайте количество овалов для многочлена

(a) $xy(x + y - 1) - h$. (b) $x^3 - x + h - y^2$. (ответ может зависеть от h).

A7. Существует многочлен степени 5, множество нулей которого невырождено и состоит из 7 овалов.

16-я проблема Гильберта. *Каким может быть количество и взаимное расположение овалов у невырожденного множества нулей многочлена степени d?*

Мы не приводим формального определения "расположения" овалов. Такое определение потребовало бы понятие *проективизации* многочлена.

[4]Следующие определения нужны только для аккуратного обоснования примеров в задаче A4.

Основная теорема. (a) *Существует многочлен степени* 6*, множество нулей которого невырождено и состоит из* 11 *овалов.*

(b) *Существуют три многочлена степени* 6*, множество нулей каждого из которых невырождено и состоит из* 11 *овалов, а расположение овалов для всех трех многочленов различно.*

B. Тропическая кривая как предел алгебраических.

B1. Нарисуйте множество нулей многочлена

(a) $x-y-1$; (a') $x^{1001}-y^{1001}-1$;

(b) $x+y-1$; (b') $x^{1001}+y^{1001}-1$;

(c) $xy=x+y$; (c') $x^{1001}y^{1001}=x^{1001}+y^{1001}$;

(d) $x^2+y^2-4x-4y-2=0$; (d') $x^{2002}+y^{2002}-4^{1001}x^{1001}-4^{1001}y^{1001}-2^{1001}$;

(e') $x^{3003}+2^{1001}x^{1001}y^{2002}-3^{1001}x^{1001}y^{1001}+y^{2002}-x^{1001}-2^{1001}$.

Обозначим через

$$F_N(x,y)=\sum_{i+j\le d}(a_{ij}x^iy^j)^N$$

семейство многочленов, зависящее от *нечетного* числа $N\ge 1$.

При замене переменных $u=x^N, v=y^N$ каждый многочлен F_N переходит в многочлен $\sum\limits_{i+j\le d}a_{ij}^Nu^iv^j$ степени d. Поэтому для решения 16-й проблемы Гильберта полезно научиться определять число и расположения овалов для множества $F_N^{-1}(0)$.

B2. Количество овалов для $F_N^{-1}(0)$ может отличаться от количества овалов у $F_1^{-1}(0)$.

Обозначим через B_R круг радиуса R с центром в начале координат 0.

B3. Для любых $\varepsilon, R>0$ найдется $N_0>0$, такое что для всех нечетных $N>N_0$ пересечение множества нулей многочлена

(a) $x^{2N}-x^N-y^N$ с кругом B_R содержится в ε-окрестности объединения прямых $x=0$, $x=\pm 1$, $x=\pm y$ и парабол $y=\pm x^2$.

(b) F с кругом B_R содержится в ε-окрестности объединения множеств нулей всевозможных многочленов $a_{ij}x^iy^j\pm a_{kl}x^ky^l$, в которых $(i,j)\ne(k,l)$, $i+j\le d$, $k+l\le d$.

Обозначим $\mathbb{R}_+:=(0,+\infty)$ множество положительных чисел и через $\mathbb{R}_+^2:=(0,+\infty)^2$ угол, заданный неравенствами $x>0$, $y>0$. Определим отображение $LOG:\mathbb{R}_+^2\to\mathbb{R}^2$ by $LOG(x,y)=(\log_2 x,\log_2 y)$.

B4. (abcde) Нарисуйте LOG-образ (то есть, образ при отображении LOG) пересечения множеств нулей многочленов (a'b'c'd'e') из задачи B1 с углом $\mathbb{R}_+^2$.

Тропическим многочленом называется функция

$$f(x,y):=\max_{i+j\le d}(ix+jy+b_{ij}).$$

Обозначим множества $f^{pq}=\{(x,y)\in\mathbb{R}^2\mid f(x,y)=px+qy+b_{pq}\}$. Некоторые пары таких множеств пересекаются друг с другом (по границе). Объединение всех таких попарных пересечений называется *тропической кривой (степени* d*).* (Это в точности множество точек, в которых график функции $f(x,y)$ имеет "излом".)

Далее будем считать, что $a_{ij}\ne 0$ при всех i,j, удовлетворяющих неравенству $i+j\le d$. Тропическая кривая *соответствует* семейству многочленов F_N, если $b_{ij}=\log_2|a_{ij}|$. Это определение мотивировано важной задачей B6b ниже.

B5. Нарисуйте тропическую кривую, соответствующую семейству многочленов

(a) $(ax)^N+(by)^N+c^N$? (b) $(ax^2)^N+(2bxy)^N+(cy^2)^N$? (Ответ может зависеть от a,b,c.)

Обозначим через Δ_R треугольник, заданный неравенствами $x\ge -R, y\ge -R, x+y\le R$.

B6. (a) Для любых $\varepsilon, R>0$ найдется N_0, такое что для всех нечетных $N>N_0$ пересечение LOG-образа множества нулей многочлена $x^{2N}-x^N-y^N$ с треугольником Δ_R содержится в ε-окрестности объединения луча $y=2x, x\ge 0$ и луча $x=0, y\ge 0$.

(b) Для любых чисел $\{a_{ij}\}_{i+j\le d}$ и $\varepsilon, R>0$ найдется $N_0>0$, такое что для всех нечетных $N>N_0$ множество $LOG(F_N^{-1}(0))\cap\Delta_R$ содержится в ε-окрестности тропической кривой, соответствующей семейству F_N.

C. Тропические прямые и окружности.

Эта часть цикла задач представляет собой художественный конкурс: предлагается экспериментально проверить некоторые теоремы тропической геометрии путем рисования аккуратных чертежей. Избранные рисунки будут выставлены для всеобщего обозрения. "Задачи" C1-C10 не оцениваются (хотя точные формулировки и доказательства каких-нибудь из этих утверждений будут награждаться "плюсиками"). Часть "задачи" можно пропустить, если упоминаемая теорема евклидовой геометрии Вам неизвестна. Вся часть C не нужна для решения 16-й проблемы Гильберта и может быть опущена.

Рассмотрим плоскость с фиксированной системой координат. *Тропической прямой* (*"лапой"*) называется объединение трех лучей на плоскости, выходящих из одной точки (называемой *вершиной*), один из которых направлен строго влево, другой строго вниз, а третий (называемый *диагональным*) — вправо-вверх под углом ровно 45°.

C1. Существуют две различные тропические прямые, пересекающиеся в двух различных точках.

Будем говорить, что две точки находятся *в общем положении*, если евклидова прямая, проходящая через эти точки, не параллельна координатным осям и прямой $x=y$.

C2. (a) Через две точки общего положения проходит единственная тропическая прямая.

(b) Если вершины двух тропических прямых находятся в общем положении, то эти тропические прямые имеют единственную общую точку.

Будем говорить, что две тропические прямые параллельны, если вершина одной из них лежит на диагональном луче другой.

C3. Пусть точка A и вершина тропической прямой b находятся в общем положении. Тогда через точку A можно провести единственную тропическую прямую, параллельную b.

Будем говорить, что две тропические прямые перпендикулярны, если прямые, которые содержат их диагональные лучи, симметричны относительно прямой $x=y$.

C4. Пусть точка A и вершина тропической прямой b находятся в общем положении. Тогда через точку A можно провести единственную тропическую прямую, перпендикулярную b.

Тропический треугольник (*"паук"*) — это объединение трех тропических прямых, вершины которых находятся (попарно) в общем положении.

C5. Нарисуйте чертежи к тропическим аналогам следующих теорем.

(a) Высоты треугольника пересекаются в одной точке.

(b) Теорема Паппа.

(c) Теорема Дезарга.

Пусть даны две точки A и B. *Тропической окружностью* (*"цаплей"*) назовем ГМТ X таких, что существуют две перпендикулярные тропических прямые, одна из которых проходит через A и X, а другая — через B и X. (Не забывайте, что через две точки A и X, вообще говоря, можно провести несколько тропических прямых!)

C6. (a) Нарисуйте тропическую окружность. Как зависит это множество от расположения точек A и B?

(b) У любого ли треугольника существует описанная окружность?

(c) Теорема Паскаля.

C7*. Придумайте понятие середины отрезка в тропической геометрии, так чтобы выполнялась теорема о медианах треугольника.

ТРОПИЧЕСКАЯ ГЕОМЕТРИЯ

А. Заславский, Ф. Нилов, А. Скопенков, М. Скопенков

Основная серия задач состоит из двух частей - окончания части B и новой части D. Задачи части D не используют (за исключением явно оговоренных случаев) понятий и результатов других частей проекта. Поэтому их можно решать, не принимая участия в других частях проекта.

B. Теорема Виро о склейке.

B3. Для любых $\varepsilon, R > 0$ найдется $N_0 > 0$, такое что для всех нечетных $N > N_0$ пересечение множества нулей многочлена

(c) $x^{2N} - x^N - y^N$ с кругом B_R и с первой четвертью ($x > 0$, $y > 0$) содержится в ε-окрестности объединения множеств

$$\{(1, y) \mid 0 \le y \le 1\}, \quad 0 \le x = y \le 1 \quad \text{и} \quad y = x^2 \ge 1.$$

(d) $x^{2N} - x^N - y^N$ с кругом B_R и со второй четвертью ($x < 0$, $y > 0$) содержится в ε-окрестности множества, симметричного объединению из (c) относительно оси Oy.

B7. Сформулируйте и докажите аналог пункта B3d для третьей и четвертой четвертей.

B8. Пересечение множества нулей многочлена $x^{2N} - x^N - y^N$ с третьей четвертью пусто.

B9. Для любых $\varepsilon, R > 0$ найдется $N_0 > 0$, такое что для всех нечетных $N > N_0$ пересечение множества нулей многочлена $x^{2N} - x^N - y^N$ с кругом B_R и

(a) первой четвертью содержится в ε-окрестности объединения множеств $\{(1, y) \mid 0 \le y \le 1\}$ и $y = x^2 \ge 1$.

(b) второй четвертью содержится в ε-окрестности объединения множеств $0 \le -x = y \le 1$ и $y = x^2 \ge 1$.

B10. Сформулируйте и докажите аналог задачи B9 для четвертой четверти.

Сформулируем теорему Виро о склейке, которая позволяет найти число и взаимное расположение овалов для некоторых алгебраических кривых.

B11. Каждая тропическая кривая является конечным объединением отрезков и лучей.

Определение кривой Виро и ее овалов. Возьмем тропическую кривую, соответствующую набору чисел $a_{ij} \neq 0$. Тропическая кривая является конечным объединением *ребер* (отрезков и лучей), которые пересекаются в *вершинах* (т.е. в общих точках ребер). *Гранью* тропической кривой называется компонента связности ее дополнения в плоскости. Каждой грани соответствует пара (p, q) таких чисел, что $px + qy + \log_2 |a_{pq}| = \max\limits_{i+j \le d,\ a_{ij} \neq 0} (ix + jy + \log_2 |a_{ij}|)$ для точек (x, y) этой грани, а также соответствует знак коэффициента a_{pq}. *В этом определении мы используем не $\{a_{ij}\}$, а тропическую кривую, на гранях которой расставлены пары чисел и знаки*[5].

Перенесем тропическую кривую параллельно, чтобы ее вершины оказались в угле $x > 0, y > 0$. Обозначим через $U_{p,q,00}$ образ при этом переносе грани тропической кривой, на которой стоит пара (p, q). Обозначим $U_{p,q,01}, U_{p,q,10}$ и $U_{p,q,11}$ — образы множества $U_{p,q,00}$ при симметриях относительно оси x, оси y и $(0, 0)$, соответственно. Продолжим расстановку знаков с первой координатной четверти на всю плоскость по следующему правилу: при отражении области $U_{p,q}$ относительно оси Ox знак области умножается на $(-1)^q$, а при отражении относительно оси Oy — на $(-1)^p$. (То есть, $\operatorname{sgn} U_{p,q,st} = (-1)^{sp+tq} \operatorname{sgn} U_{p,q,00}$.) Определим *кривую Виро* как объединение $\cup\{U_\alpha \cap U_\beta \mid \operatorname{sgn} U_\alpha \neq \operatorname{sgn} U_\beta\}$ тех ребер тропической кривой, которые разделяют грани разных знаков. Две неограниченные компоненты связности кривой Виро называются

- *элементарно эквивалентны*, если они содержат лучи, симметричные относительно точки $(0, 0)$.
- *эквивалентны* если существует последовательность компонент от одной к другой, в которой любые две последовательные компоненты элементарно эквивалентны.

Овалом кривой Виро называется либо замкнутая ломаная, содержащаяся в кривой Виро, или класс эквивалентности компонент связности.

Следующей теоремой разрешается пользоваться без доказательства:

Теорема Харнака. *Невырожденное множество нулей многочлена степени d не может содержать больше $\dfrac{(d-1)(d-2)}{2} + 1$ овалов.*

B12.* Теорема Виро о склейке. *Пусть кривая Виро, построенная по набору чисел $a_{ij} \neq 0$, содержит ровно $\dfrac{(d-1)(d-2)}{2} + 1$ овалов. Тогда существует такое N, что множество нулей многочлена*

[5]Пару чисел (p, q), соответствующую грани, легко восстановить по самой тропической кривой (подумайте, как!)

$\sum_{i+j\le d} a_{ij}^N u^i v^j$ *невырождено, а число и взаимное расположение его овалов такое же, как у соответствующей кривой Виро.*

D. Построение примеров в 16-й проблеме Гильберта.

Цель задач серии D — научиться описывать тропические кривые на чисто комбинаторном языке, и тем самым получить комбинаторный метод построения примеров кривых в 16-й проблеме Гильберта.

Напомним, что *тропическая кривая* степени d — это множество точек излома графика функции $\max_{i+j\le d}\{ix+jy+b_{ij}\}$ (подробнее см. выше, абзац после задачи B4).

D1. (a) Проверьте, что тропическая кривая степени 1 выглядит так, как показано на рисунке 1. (Сравните с определением тропической прямой в части C).

(b) Из каждой вершины тропической кривой выходит как минимум 3 ребра.

Припишем каждому ребру тропической кривой *кратность* по следующему правилу. Предположим, что в одной из областей, граничащих с этим ребром, максимальной является величина $ix+jy+b_{ij}$, а другой — величина $i'x+j'y+b_{i'j'}$. Тогда прямая, содержащая данный отрезок, задается уравнением $(i-i')x+(j-j')y+(b_{ij}-b_{i'j'})=0$. Будем считать *кратностью* данного ребра наибольший общий делитель чисел $i-i'$ и $j-j'$.

Будем обозначать на рисунках кратные ребра тропической кривой двойными (тройными и т.д.) линиями.

D2. Тропические кривые степени d обладают следующими свойствами:

(a) Наклон каждого ребра рационален.

(b) В каждой вершине выполняется следующее условие сбалансированности. Обозначим через v_i вектор с началом в данной вершине, имеющий направление i-го ребра, выходящего из вершины, и равный кратчайшему целочисленному вектору с данным направлением, умноженному на кратность ребра. Тогда $\sum v_i = 0$ для каждой вершины.

(c) Имеется $3d$ бесконечных рёбер, взятых с учетом кратностей, d из которых направлены строго влево, d направлены строго вниз, и d направлены вправо-вверх с углом наклона 45°.

D3. (a) Тропический многочлен $\max_{i+j\le d}\{ix+jy+b_{ij}\}$ восстанавливается по своей тропической кривой однозначно с точностью до добавления постоянной.

(b) Всякий граф на плоскости с прямыми ребрами и предписанными кратностями, удовлетворяющий свойствам (a), (b) и (c) задачи D2, является тропической кривой степени d.

Мы говорим, что две тропические кривые имеют *одинаковую конфигурацию*, если у них совпадает комбинаторный тип графа и наклон его рёбер (но не обязательно их длины и положение).

D4. Нарисуйте 5 различных конфигураций тропических кривых второй степени.

Для решения следующих задач достаточно прочитать в предыдущем пункте абзац "Определение кривой Виро и ее овалов".

D5. Из какого максимального числа овалов может состоять кривая Виро при $d=$
(a) 2; (b) 3; (c) 4; (d) 5? (Доказательства максимальности мы не требуем. Сравните ответ с задачами A4f и A7.)

D6*. Напишите программу на компьютере, которая:

(a) рисует все конфигурации тропических кривых данной степени d;

(b) по данной конфигурации тропической кривой и набору знаков, приписанных областям U_{ij} дополнения к ней, определяет количество овалов у кривой Виро.

D7*. (ab) Докажите Основную теорему (разрешается пользоваться теоремой Виро о склейке без доказательства).

РЕШЕНИЯ ЗАДАЧ

A1. Ответ: нет. Например, прямая $x = 0$ является множеством нулей разных многочленов $F(x, y) = x$ и $G(x, y) = x^2$.

A2. Ответ: a, b, c, e, f.

Примеры. (a) Любая прямая на плоскости задается уравнением $Ax+By+C = 0$ с некоторыми числами A, B, C.

(b) Уравнение окружности: $(x-x_0)^2+(y-y_0)^2-R^2 = 0$, где (x_0, y_0) — координаты центра, R — радиус.

(c) Уравнение точки (x_0, y_0): $(x - x_0)^2 + (y - y_0)^2 = 0$.

(e) Уравнение объединения двух прямых: $(Ax + By + C)(ax + by + c) = 0$, где $Ax + By + C = 0$ — уравнение первой, $ax + by + c = 0$ — уравнение второй прямой.

(f) Уравнение объединения 6 окружностей: $\prod_{k=0}^{6} \left((x - x_k)^2 + (y - y_k)^2 - R_k^2\right) = 0$, где $(x - x_k)^2 + (y - y_k)^2 - R_k^2 = 0$ — уравнение k-й окружности.

Невозможность в пункте (d) напрямую следует из задачи A3a.

A3. (a) Ответ: либо от 0 до d, либо прямая содержится в нашем множестве нулей.

Параметризуем прямую l: $x = x_0+\alpha\cdot t$, $y = y_0+\beta\cdot t$. Подставляя эти выражения в многочлен, получим многочлен $P(t)$ степени не более, чем d. Многочлен $P(t)$ либо имеет не более, чем d вещественных корней, либо тождественно равен 0.

Покажем, что для любого $d' < d$, существует кривая степени d и прямая l, которые пересекаются в d' точках. Рассмотрим уравнения d прямых, отличных от l, среди которых ровно $d - d'$ параллельны l. Произведение этих уравнений является нужным многочленом.

Обозначим через d степень данного многочлена $F(x, y) = \sum_{i+j\le d} a_{ij}x^i y^j$. Покажем, что *существует невырожденная замена координат* $x = \alpha_1 x' + \beta_1 y'$, $y = \alpha_2 x' + \beta_2 y'$ *(невырожденность означает, что* $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$*), после которой коэффициент при одночлене* $(x')^d$ *будет ненулевым.*

Коэффициент $A(\alpha_1, \alpha_2)$ при одночлене $(x')^d$ равен $\sum_{i+j\le d} a_{ij}\alpha_1^i\alpha_2^j$. Так как не все a_{ij} равны 0, то существуют такие α_1 и α_2, не равные одновременно нулю, что $A(\alpha_1, \alpha_2) \neq 0$. Подбирая коэффициенты β_1 и β_2, так, чтобы они не были пропорциональны α_1 и α_2 (т.е. чтобы $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$), мы получим нужную замену.

Вернемся к решению задачи. Так как при замене из леммы ограниченные множества переходят в ограниченные, то можно считать, что коэффициент при x^d ненулевой. Так как d нечетно, то для каждого y уравнение $F(x, y) = 0$ имеет решение. Значит, $F^{-1}(0)$ неограничено.

A4. (a) Например, подойдет многочлен $xy(x + y - 1) + \frac{1}{100}$.

Обозначим через ϕ множество его нулей. Покажем, что данный многочлен неприводим. Действительно, иначе он разлагается на произведение многочленов, степень одного из которых равна 1. Поэтому ϕ содержит прямую. Эта прямая обязана пересекать одну из прямых Ox и Oy, которые не пересекают ϕ. Полученное противоречие доказывает неприводимость.

Координаты x точек пересечения прямой $y = c$ с ϕ удовлетворяют уравнению $x^2 + (c - 1)x + \frac{1}{100c} = 0$. Дискриминант $D = D(c)$ этого уравнения равен $(c - 1)^2 - \frac{1}{25c}$. Равенство $D(c) = 0$ равносильно равенству $f(c) := 25c(c-1)^2 - 1 = 0$. Это уравнение третьей степени, которое имеет не более трёх корней. Поскольку $f(\frac{1}{100}) < 0$, $f(\frac{1}{2}) > 0$, $f(1) < 0$, $f(2) > 0$, то два корня c_1 и c_2 уравнения $f(c) = 0$ лежат на интервале $(0, 1)$, а третий корень лежит на интервале $(1, 2)$. Поэтому $D(c) = 0$ ровно в двух точках c_1 и c_2 интервала $(0, 1)$, причём $D(c) > 0$ для любого $c \in (c_1, c_2)$ и $D(c) < 0$ для остальных точек этого интервала. (Для определённости считаем, что $c_1 < c_2$.) Значит, при c, равном c_1 или c_2, прямая $y = c$ пересекает ϕ ровно в одной точке. Поэтому при $c \in (c_1, c_2)$ прямая $y = c$ пересекает ϕ в двух точках $(x_1(c), c)$ и $(x_2(c), c)$, где $x_{1,2}(c) = \dfrac{\pm\sqrt{D} - (c - 1)}{2}$. При остальных значениях $c \in (0, 1)$ прямая $y = c$ не пересекает ϕ.

Определим

$$\gamma : [c_1, 2c_2 - c_1] \to \mathbb{R}^2 \quad \text{формулой} \quad \begin{cases} (x_1(t), t) & t \in [c_1, c_2] \\ (x_2(2c_2 - t), 2c_2 - t) & t \in [c_2, 2c_2 - c_1] \end{cases}$$

Так как функции $x_1(c)$ и $x_2(c)$ дифференцируемые, то отображение $\gamma(t)$ дифференцируемое в точках, отличных от c_2. Так как $2c_2 - t = t$ для $t = c_2$ и $(x_1)'(c_2) = (x_2)'(c_2)$, то отображение $\gamma(t)$ гладкое во всех точках. Теперь ясно, что $\gamma(I)$ есть замкнутая кривая, содержащаяся в кривой ϕ.

(b) Указание. Рассмотрите многочлен $(x + 1)(x - 1)(y + 1)(y - 1) + \frac{1}{100}$.

(c) Указание. Рассмотрите многочлен $(x^2+y^2-1)(x^2+y^2-9)+\frac{1}{100}$.

(d) Предположим противное, пусть есть хотя бы одна другая точка X. Рассмотрим точку Y внутри внутренней замкнутой кривой. Тогда прямая XY пересечёт множество нулей данного многочлена не менее, чем в пяти точках, что противоречит утверждению задачи A3(a).

(e) Ответ. Нет. Указание. Рассмотрите многочлен $x(x^2+y^2-1)(x^2+y^2-9)+\frac{1}{100}$.

(f) Указание. Рассмотрите многочлен $(x^2+2y^2-3)(2x^2+y^2-3)+\frac{1}{100}$.

(g) Указание. Рассмотрите многочлен $(x^2+y^2-1)(x-y-1)(x+y-1)+\frac{1}{100}$.

A5. Направление прямой OM, соединяющей начало координат O с точкой $M(x,y)$ ветви гиперболы $xy=1$, лежащей в положительном квадранте, стремится к направлению прямой Ox при $x\to+\infty$. Поэтому Ox является предельной прямой для ветви гиперболы $xy=1$, лежащей в положительном квадранте. Аналогично она является предельной прямой для другой ветви гиперболы. Поэтому ветви гипербол эквивалентны.

A6. (a) Ответ: При $h<0$ — один овал, при $h\in(0,1/27)$ — два овала, при $h>1/27$ — один овал. При $h=0$ и $h=1/27$ алгебраическая кривая вырождена. Приведем решение.

Обозначим $f(x,y):=xy(x+y-1)+h$. Введём обозначения для точек пересечения прямых Ox, Oy и $x+y-1=0$ и областей, на которые эти прямые разбивают плоскость:

$$A:=(1,0),\quad B:=(0,1),$$

$C:=(0,0)$, $X:=\{(x,y)\mid x>0,y>0,x+y<1\}$, $X_A:=y<0,x+y>1$, $X_B:=x<0,x+y>1$, $X_C:=x<0,y<0$, $Y_A:=x<0,y>0,x+y<1$, $Y_B:=x>0,y<0,x+y<1$, $Y_C:=x>0,y>0,x+y>1$.

Ясно, что $f(x,y)=h$ в точках прямых Ox, Oy и $x+y-1=0$, $f(x,y)<h$ в точках областей X_A, X_B, X_C и X, а также $f(x,y)>h$ в точках областей Y_A, Y_B и Y_C. Поэтому при $h>0$ нули многочлена $f(x,y)$ могут лежать только в X_A, X_B, X_C и X, при $h<0$ они могут лежать только в Y_A, Y_B и Y_C.

Пусть $h<0$. Обозначим $y_A:=Y_A\cap f^{-1}(0)$. Аналогично определим y_B и y_C.

Докажем, что y_A является связной компонентой множества $f^{-1}(0)$ нулей f. Координаты x точек пересечения прямых $y=c$ с $f^{-1}(0)$ удовлетворяют уравнению $x^2+(c-1)x+\frac{h}{c}=0$. Дискриминант $D=D(c)$ этого уравнения равен $(c-1)^2-\frac{4h}{c}$. Поскольку $h<0$, то для любого $c\in R_+$, $D(c)>0$. Значит, каждая из прямых $y=c$, где $c\in R_+$, пересекает F ровно в двух точках $(x_{1,2}(c),c)$, таких, что $x_{1,2}(c)=\dfrac{\pm\sqrt{D}-(c-1)}{2}$. Определим

$$\gamma: R_+\to\mathbb{R}^2\quad\text{формулой}\quad\Big\{(x_2(t),t)\quad t\in R_+\ .$$

Так как функция $x_2(c)$ гладкая, то отображение $\gamma(t)$ гладкое. Поскольку $\gamma(R_+)=y_A$, то y_A является связной компонентой множества $f^{-1}(0)$ нулей f. Аналогично y_B и y_C являются связными компонентами множества $f^{-1}(0)$ нулей f.

Нетрудно проверить, что направление прямой Ox является предельным для ветви y_C.

Аналогично, это направление является предельным для ветви y_A. Поэтому ветви y_A и y_C элементарно эквивалентны. Аналогично, ветви y_A и y_B элементарно эквивалентны, поскольку направление прямой $x+y-1=0$ является для этих ветвей предельным. Значит, ветви y_A, y_B и y_C эквивалентны и поэтому образуют один овал.

При $h=0$ алгебраическая кривая f вырождена.

Пусть $h>0$. Введём обозначения

$$x:=X\cap f^{-1}(0),\quad x_A:=X_A\cap f^{-1}(0),\quad x_B:=X_B\cap f^{-1}(0)\quad\text{и}\quad x_C:=X_C\cap f^{-1}(0).$$

Связность и эквивалентность x_A, x_B и x_C устанавливается аналогично случаю $h<0$. Поэтому они образуют один овал при любом $h>0$. Если множество x не пусто и не представляет собой одну точку, то оно является овалом (доказательство аналогично решению задачи A4a).

Покажем, что множество точек x не пусто только при $h\in(0,\frac{1}{27}]$. Ясно, что x пусто тогда и только тогда, когда $D(c)<0$ при любом $c\in(0,1)$. Производная $D'(c)>0$ при $c\in(0,1/3)$, $D'(c)=0$ при $c=1/3$, $D'(c)<0$ при $c\in(1/3,1)$. Поэтому в точке $c=1/3$ достигается максимум функции $D(c)$ на интервале $(0,1)$. Значит, $D(c)<0$ при любом $c\in(0,1)$ тогда и только тогда, когда $D(1/3)=4/9-\frac{4h}{c}<0$, т.е. $h>1/27$. При $h=1/27$ множество x состоит из одной точки. Таким образом, при $h\in(0,1/27)$ множество нулей $f^{-1}(0)$ состоит из двух овалов, при $h=1/27$ алгебраическая кривая вырождена, при $h>1/27$ множество $f^{-1}(0)$ состоит из одного овала.

(b) Ответ: Один овал при $h\in(-\frac{2}{3\sqrt{3}},\frac{2}{3\sqrt{3}})$, два овала при $h\in(-\infty,-\frac{2}{3\sqrt{3}})$ и $h\in(\frac{2}{3\sqrt{3}},\infty)$, алгебраическая кривая $x^3-x+h-y^2$ вырождена при $h=\pm\frac{2}{3\sqrt{3}}$. Указание. Аналогично (a).

A7. Указание. Рассмотрите многочлен $x((x-1)^2+y^2-2)((x+1)^2+y^2-2)+\frac{1}{100}$.

B1. (a') *Указание.* См. рисунок 3.a'. Обоснуем рисунок. Ясно, что множество нулей многочлена $x^{2001}-y^{2001}-1$ лежит "ниже" прямой $y=x$ и симметрично относительно прямой $x+y=0$. Поэтому мы можем рассматривать только случай $y>-x$. Точки $(0,-1)$ и $(1,0)$ являются точками пересечения нашего множества с осями координат. Если $x=1+\epsilon$, где $\epsilon>0$, то для y выполнено неравенство $(1001\epsilon)^{\frac{1}{1001}}<y<1+\epsilon$. Поэтому, при достаточно малых значениях ϵ, значения y могут изменяться от 0 до 1. При $\epsilon>1/1001$ значения x и y приблизительно равны. Если $1-\epsilon<x<1$, где ϵ достаточно мало, значения y могут изменяться от -1 до 0. Аналогично разбирается случай $y<-x$.

(b') Указание. См. рисунок 3.b'. Аналогично (a').

(c') Указание. См. рисунок 3.c'. Ясно, что множество нулей многочлена $x^{1001}y^{1001}-x^{1001}-y^{1001}$ симметрично относительно прямой $y=x$. Поэтому можно рассматривать только случай $x\geqslant y$. При $x<0$ выполнено $y>x$. При $x\in[0,1000/1001]$ значение y приблизительно равно x. При $x\in[1000/1001,1]$ значение y изменяется от -1 до $-\infty$. При $x\in(1,1002/1001)$ выполнено $y>x$. При $x\geqslant 1002/1001$ значение y приблизительно равно 1.

(d') Указание. См. рисунок 3.d'. Ясно, что множество нулей многочлена $x^{2002}+y^{2002}-4^{1001}x^{1001}-4^{1001}y^{1001}-2^{1001}$ симметрично относительно прямой $y=x$. Поэтому можно рассматривать только случай $x\geqslant y$. Точек с координатой $x\geqslant 4$ среди множества нулей нашего многочлена нет. При $x\in(-1/2,1/2]$ значение y приблизительно равно -1/2. При $x\in(1/2,4003/1001]$ значение y приблизительно равно $-x$. При $x\in(4003/1001,4)$ значение y изменяется от -4 до 4.

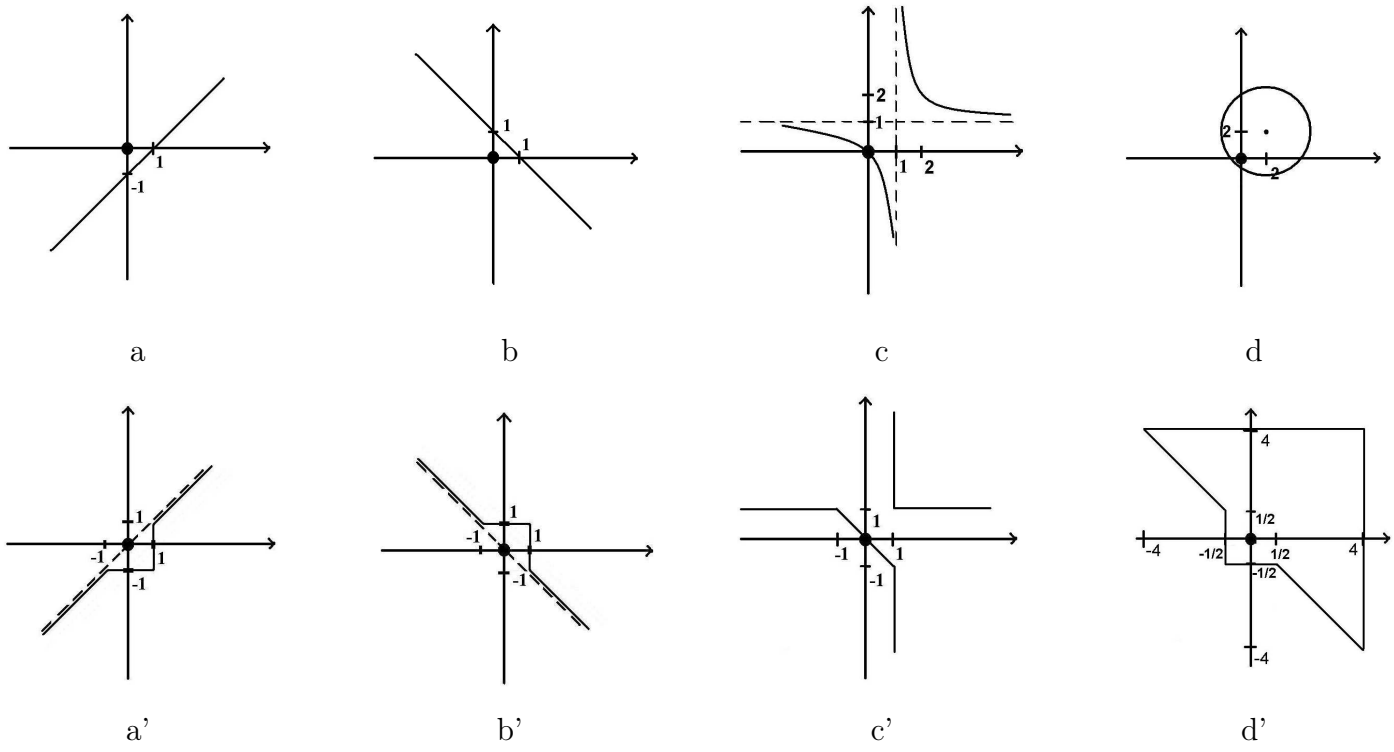


Рис. 3:

B2. Пусть $F(x,y)=x^3-px+q-y^2$, где $p,q>0$. Тогда множество нулей многочлена F состоит из двух овалов, если многочлен $f(x)=x^3-px+q$ имеет три действительных корня, и из одного овала, если $f(x)$ имеет один действительный корень. Приравняв производную многочлена f к нулю, убеждаемся, что он имеет локальный максимум в точке $x_1=-\sqrt{p/3}$ и локальный минимум в точке $x_2=\sqrt{p/3}$. Соответственно, $f(x)$ имеет три корня тогда и только тогда, когда $f(x_2)<q<f(x_1)$, т.е. $4p^3>27q^2$. Аналогично получаем, что множество нулей многочлена $F_N(x,y)$ состоит из двух овалов при $4p^{3N}>27q^{2N}$ и из одного овала в противном случае. Очевидно, что при $1<\frac{p^3}{q^2}<\frac{27}{4}$ первое неравенство не выполняется, а второе выполняется при достаточно больших N.

B3. (a) Пункт (a) является частным случаем пункта (b).

(b) *Указание.* Предположим, что в некоторой точке (x,y) значения всех одночленов $a_{ij}x^iy^j$ различны по абсолютной величине, причем $|a_{kl}x^ky^l|>|a_{ij}x^iy^j|$ для всех пар $(i,j)\neq(k,l)$. Тогда при $N\to\infty$ $|\frac{a_{ij}x^iy^j}{a_{kl}x^ky^l}|^N\to 0$. Следовательно, при всех достаточно больших N $|a_{kl}x^ky^l|^N$ превосходит сумму абсолютных

величин остальных одночленов и равенство $F_N(x,y) = 0$ невозможно. Поэтому множество $F_N^{-1}(0)$ при больших N стремится к некоторому подмножеству объединения множеств, задаваемых равенствами вида $|a_{ij}x^iy^j| = |a_{kl}x^ky^l|$.

Решение. Пусть дано $\epsilon, R > 0$. Обозначим через Γ объединение всех кривых, задаваемых уравнениями $a_{ij}x^iy^j \pm a_{kl}x^ky^l = 0$. Рассмотрим произвольную точку $(x_0, y_0) \in B_R$ на расстоянии больше ϵ от Γ. Будем считать, что эта точка лежит в той части $\mathbb{R}^2 - \Gamma$, в которой максимальное по модулю значение принимает одночлен $a_{kl}x^ky^l$. Пусть $(x_1, y_0) \in \Gamma$ — ближайшая к (x_0, y_0) точка на прямой $y = y_0$. Тогда $|y_1 - y_0| > \epsilon$ и $|a_{ij}x_1^iy_0^j| = |a_{kl}x_1^ky_0^l|$ для некоторых i, j.

Оценим отношение

$$\frac{a_{ij}x_0^iy_0^j}{a_{kl}x_0^ky_0^l} = \frac{a_{ij}}{a_{kl}}x_0^{i-k}y_0^{j-l} = (\frac{x_0}{x_1})^{i-k}$$

Так как $|a_{kl}x_0^ky_0^l| > |a_{ij}x_0^iy_0^j|$, то $|x_0/x_1| < 1$. Поскольку $|x_0|, |x_1| < R$, то $|x_0/x_1| < 1 - \epsilon/R$. Значит, для любых i, j наше отношение

$$\frac{|a_{ij}x_0^iy_0^j|}{|a_{kl}x_0^ky_0^l|} < 1 - \epsilon/R.$$

Следовательно, при всех N, удовлетворяющих неравенству $d^2\left(1 - \frac{\epsilon}{R}\right)^N < 1$ одночлен $|a_{kl}x_0^ky_0^l|^N$ превосходит сумму абсолютных величин остальных одночленов и равенство $F_N(x_0, y_0) = 0$ невозможно. Задача решена.

(c) Непосредственно следует из утверждения предыдущей задачи, примененного к многочлену $F_N = x^{2N} - x^N - y^N$.

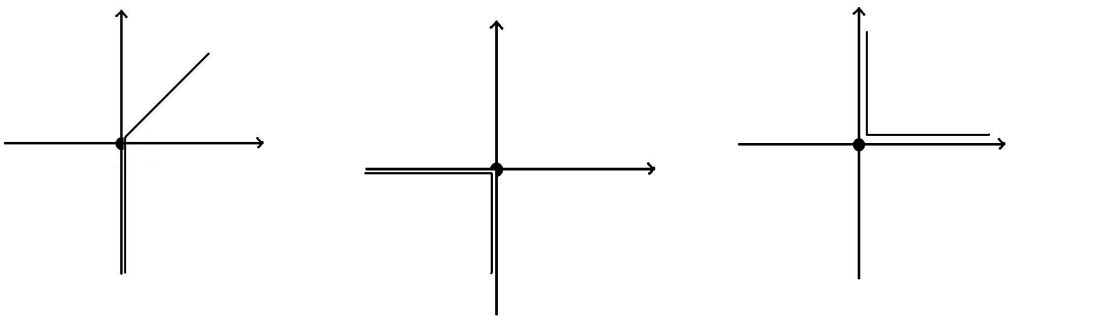
(d) Множество нулей многочлена $x^{2N} - x^N - y^N$, лежащих во второй четверти, симметрично относительно оси ординат множеству нулей многочлена $x^{2N} + x^N - y^N$, лежащих в первой четверти. Поэтому требуемое утверждение непосредственно следует из утверждения пункта b)

B4. (a) *Указание.* См. рис. 4.a. Обоснуем рисунок. Ясно, что множество нулей функции $f(x,y) := 2^{1001x} - 2^{1001y} - 1$ лежит "правее" оси Oy. При $y < 0$ значение x приблизительно равно 0, а при $y > 0$ значение x приблизительно равно y.

(b) Указание. См. рис. 4.b. Обоснование аналогично (a).

(c) При $x \in (0, 1/1001)$ значения y изменяются от 0 до $+\infty$, при $x \geqslant 1/1001$ значение y приблизительно равно 0.

(d) Ясно, что множество нулей функции $f(x,y) := 2^{2002x} + 2^{2002y} - 4^{1001}2^{1001x} - 4^{1001}2^{1001y} - 2^{1001}$ симметрично относительно прямой $y = x$. Поэтому можно рассматривать только случай $x \geqslant y$. При $x \geqslant 2003/1001$ значение y приблизительно равно $x/2 + 1$, при $x \in (2, 2003/1001)$ значение y изменяется от $-\infty$ до 2.



a b c d

Рис. 4:

B6. (a) Рассуждая, как в задаче B3b, получаем, что пересечение множества нулей с первой координатной четвертью лежит вблизи объединения множеств $x^2 = y \geq x$, $x^2 = x \geq y$, $x = y \geq x^2$. Однако, коэффициенты при одночленах x^N и y^N у данного многочлена имеют одинаковые знаки, так что в окрестности последнего множества F_N не может обращаться в нуль. При логарифмическом отображении первое из указанных множеств переходит в луч $y = 2x$, $x \geq 0$, а второе — в луч $x = 0$, $y \leq 0$.

(b) Аналогично предыдущему пункту получаем, что множество нулей многочлена F_N при больших N лежит в окрестности объединения множеств, задаваемых соотношениями $a_{ij}x_iy^j = a_{kl}x^ky^l \geq a_{pq}x^py^q$.

При логарифмическом отображении эти множества переходят в множества вида $ix + jy = kx + ly \geq px+qy$, каждое из которых является отрезком или лучом. Объединение этих множеств задает тропическую кривую.

B7. Рассуждая аналогично задаче B3d, получаем, что множества, содержащие пересечения множества нулей с первой и четвертой четвертью, симметричны относительно оси абсцисс, а с первой и третьей — относительно начала координат.

B8. Так как в третьей четверти $x < 0$, $y < 0$, то все одночлены, входящие в многочлен F_N положительны, т.е. равенство $F_N(x, y) = 0$ невозможно.

B9. (a) Решение аналогично решению задачи B6a.

(b) Из задачи B3d следует, что пересечение множества нулей с второй четвертью лежит в окрестности множеств $x^2 = y \geq -x$, $x^2 = -x \geq y$, $-x = y \geq x^2$. Поскольку знаки одночленов $-x^N$ и x^{2N} во второй четверти совпадают, то пересечение множества нулей со второй четвертью содержится в окрестности только первого и третьего из указанных множеств.

B10. Рассуждая, как в задаче B9b, получаем, что искомое пересечение содержится в окрестности объединения множеств $(1, y)$, $0 \geq y \geq -1$ и $0 \leq x = -y \leq 1$.

B11. Каждое ребро тропической кривой задается системой из одного уравнения и нескольких неравенств вида $ix+jy+b_{ij} = kx+ly+b_{kl} \geq px+qy+b_{pq}$. Если эта система совместна, то уравнение определяет прямую, а неравенства высекают на этой прямой отрезок или луч.

B12. *Указание.* Действительно, рассмотрим отдельно поведение множества нулей многочлена $F_N(x, y)$ в каждом из четырёх квадрантов. Отображение $LOG : (\mathbb{R} - \{0\})^2 \to \mathbb{R}^2$, $(x, y) \mapsto (\log_2 |x|, \log_2 |y|)$ переводит каждый из квадрантов на плоскость взаимно однозначно. Выберем один из квадрантов (например, $x, y > 0$), и отождествим его с плоскостью указанным отображением. Тропическая кривая, задаваемая тропическим многочленом $\max\limits_{i+j\leq d}\{ix + jy + b_{ij}\}$, $b_{ij} = \log_2 |a_{ij}|$ разбивает тропическую плоскость на области. Внутри каждой области поведение многочлена $F_N(x, y)$ определяется поведением одного из мономов $(a_{ij}x_iy_j)^N$, и в зависимости от знака коэффициента a_{ij} (а также выбранного квадранта) многочлен F_N в данной области либо положителен, либо отрицателен. Закрасим каждую из областей дополнения к тропической кривой в один из двух цветов, в соответствии со знаком многочлена F_N в этой области. Если две соседние области, граничащие вдоль некоторого ребра, окрашены в разные цвета, то по тереме о промежуточном значении вдоль этого ребра проходит ветвь множества нулей многочлена F_N. Если же обе соседние области окрашены в одинаковые цвета, то вблизи этого ребра нет вещественных точек кривой. Таким образом, для больших нечетных значений параметра N множество нулей многочлена F_N (в выбранном квадранте) приближённо изображается набором некоторого количества явно перечисляемых рёбер тропической кривой, а множество нулей многочлена F_N во всей плоскости приближенно задается кривой Виро.

В принципе, множество нулей многочлена F_N могло бы иметь больше ветвей, чем у кривой Виро — например, могли бы существовать "маленькие" овалы вблизи вершин тропической кривой. Отсутствие "лишних" ветвей (и овалов) гарантирует нам предположение, что число овалов у кривой Виро равно $(d-1)(d-2)/2+1$ и теорема Харнака.

Замечание. Авторам задачи неизвестно, остается ли верной теорема о склейке Виро без предположения, что число овалов у кривой Виро равно $(d-1)(d-2)/2+1$.

C5. (a) См. рис. 5.a

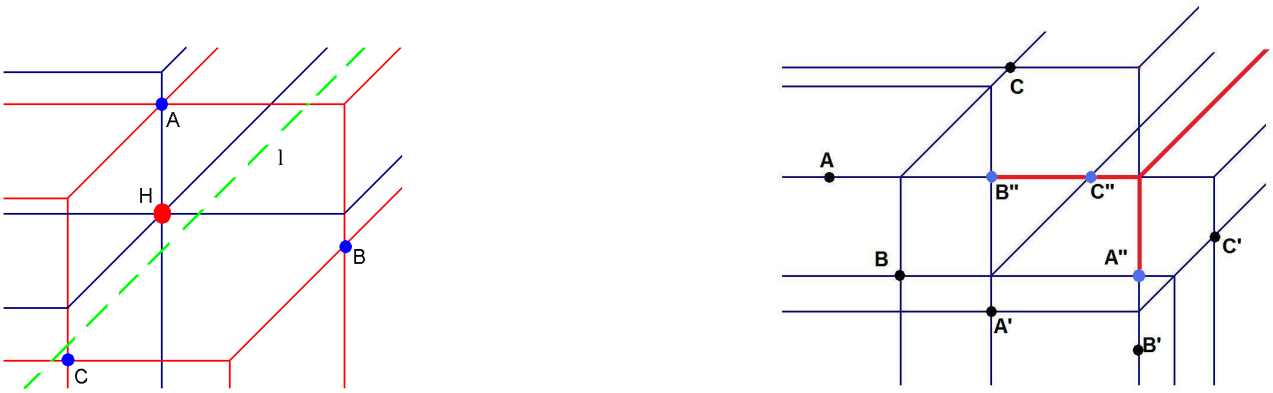
(b) См. рис. 5.b.

(c) См. рис. 6.

D1. (a) *Указание.* Функция $\max\{x+a, y+b, c\}$ имеет следующее поведение. При x и y отрицательных и больших по абсолютной величине максимальным из трёх величин является постоянное значение c. При увеличении x значение функции не меняется до тех пор, пока точка (x, y) не пересечет вертикальную прямую $x+a = c$. Правее этой прямой максимальной является величина $x+a$. Аналогично, при движении точки (x, y) вверх переход к величине $y + b$ осуществляется на горизонтальной прямой $y + b = c$, вдоль которой максимум достигается на двух конкурирующих выражениях $y+q$ и c. Наконец, области, в которых значение функции $f(x, y)$ совпадает с выражениями $x+a$ и $y+b$, разделяются лучом прямой $x+a = y+b$, имеющей наклон 1. Все три полученных луча сходятся в точке $(c-a; c-b)$.

D2. (a) Очевидно.

(b) *Указание.* Рассмотрим вершину кривой и предположим, что к этой вершине подходит r областей (дополнения к тропической кривой), в которых максимум достигается на функциях $i_1x+j_1y+b_{i_1j_1}, \ldots, i_rx+$



a b

Рис. 5:

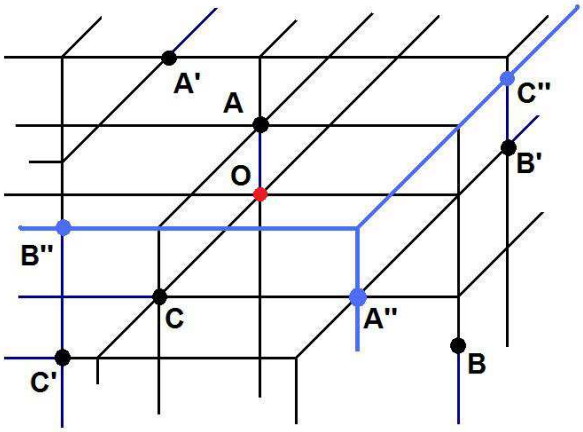


Рис. 6:

$j_r y + b_{i_r j_r}$, соответственно (мы считаем области занумерованными против часовой стрелки). Тогда, очевидно, выполняется векторное равенство

$$\binom{i_2 - i_1}{j_2 - j_1} + \cdots + \binom{i_r - i_{r-1}}{j_r - j_{r-1}} + \binom{i_1 - i_r}{j_1 - j_r} = 0.$$

Осталось заметить только, что вектор $\binom{i_{s+1}-i_s}{j_{s+1}-j_s}$ отличается от вектора v_s, участвующего в условии сбалансированности, только лишь поворотом на $90°$.

(c) *Указание.* Докажем, например, что тропическая кривая степени d имеет (с учетом кратности) ровно d горизонтальных лучей, направленных влево. Действительно, будем рассматривать только ту часть плоскости, в которой координата x отрицательна и очень велика по абсолютной величине. Ясно, что в этой части максимальной может быть только одна из величин $jy + a_{0j}$, $j = 0, 1, \ldots, d$. Ясно также, что в этой части при больших по абсолютной величине отрицательных y максимальна величина a_{00}, а при больших по абсолютной величине положительных y максимальна величина $dy + a_{0d}$. Пусть при увеличении y максимальными последовательно становятся величины a_{00}, $j_1 y + a_{0j_1}$, $j_2 y + a_{0j_2}$, $\ldots$, $j_k y + a_{0j_k}$, $dy + a_{0d}$. Легко видеть, что $0 < j_1 < j_2 < \cdots < j_k < d$. Тогда кратности горизонтальных ребер равны j_1, $j_2 - j_1$, $\ldots$, $d - j_k$. Поэтому количество горизонтальных ребер с учетом кратности равно $(j_1) + (j_2 - j_1) + \cdots + (d - j_k) = d$.

D3. (ab) *Указание.* Действительно, предположим, что в некоторой области тропический многочлен совпадает с линейной функцией $ix + jy + b_{ij}$. Пусть прямая, содержащая отрезок границы этой области, имеет уравнение $px+qy+r = 0$. Тогда в соседней области, граничащей с исходной вдоль отрезка, многочлен совпадает с линейной функцией $(i+p)x+(j+q)y+(b_{ij}+r)$. Иными словами, мы устанавливаем равенство $b_{i+p,j+q} = b_{i,j} + r$. Продолжая таким же образом, мы восстанавливаем весь многочлен область за областью по индукции. Условие сбалансированности гарантирует нам, что в процессе построения мы никогда не придём к противоречию. Условие поведения тропической кривой на бесконечности обеспечит наличие только тех "тропических мономов", полученных в процессе построения, которые только и возможны для тропических многочленов данной степени.

D4. *Указание.* Как и для обычной гиперболы, тропическую кривую второй степени можно получить, пошевелив слегка объединение двух тропических прямых. Объединение двух тропических прямых задается суммой двух тропических многочленов первой степени. У графа, являющегося множеством точек излома такой суммы, имеется вершина валентности 4, в которой максимум достигается одновременно на четырёх конкурирующих линейных функциях. После небольшого шевеления одной из этих линейных функций точка валентности четыре распадается на две точки валентности три. Некоторые из возможных тропических кривых второй степени приведены на рис. .

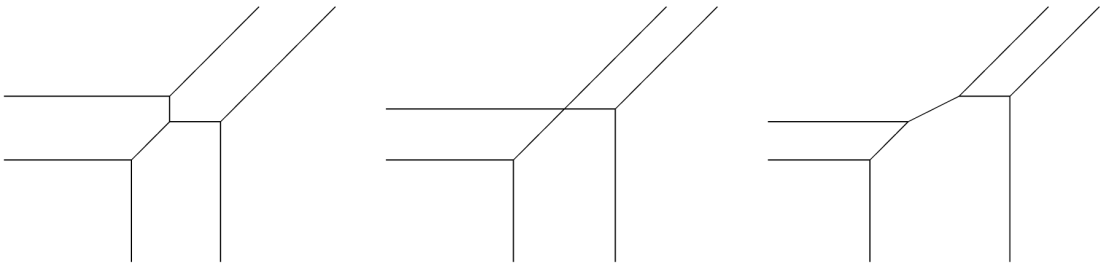


Рис. 7: Вырожденная тропическая кривая второй степени и два её шевеления

D5. *Ответ.* (a) 1; (b) 2; (c) 4; (d) 7.

D7. (b) **Разбиения диаграммы Ньютона.** При решении задачи части D7 может оказаться полезным следующее "двойственное" описание конфигураций тропических кривых. Рассмотрим на плоскости треугольник с вершинами $(0;0)$, $(0,d)$ и $(d,0)$. Этот треугольник называется *треугольником Ньютона* тропического многочлена. Со всякой тропической кривой связывается некоторое разбиение треугольника Ньютона на выпуклые многоугольники с целочисленными вершинами. А именно, рассмотрим область в дополнении к тропической кривой, в которой максимальной является величина $ix+jy+b_{ij}$. Этой области мы сопоставим вершину с координатами (i,j) на треугольнике Ньютона. Ребру тропической кривой, разделяющему две области, сопоставляется отрезок на диаграмме Ньютона, соединяющий вершины, отвечающие этим двум областям. Наконец, всякой вершине тропической кривой, к которой подходит r областей её дополнения, соответствует r-угольник на диаграмме Ньютона, вершины которого отвечают этим r областям. В частности, бесконечным областям соответствуют вершины разбиения, входящие в границу диаграммы, а бесконечным ребрам — отрезки границы диаграммы. Полезно отметить, что направление всякого ребра тропической кривой ортогонально направлению двойственного ребра разбиения диаграммы.

Алгоритм построения кривых Виро. Процедуру построения кривых Виро удобно переформулировать на двойственном языке диаграмм Ньютона. Эта процедура, носящая название "patchworking" (склейка Виро), состоит в последовательном выполнении следующих шагов (см. результат на рис.).

1. Выбираем произвольную триангуляцию диаграммы Ньютона Δ с вершинами в целых точках;

2. Расставляем в вершинах триангуляции знаки, + или -, произвольным образом.

3. Отразив диаграмму Ньютона вместе с её триангуляцией последовательно относительно координатных осей, получаем триангуляцию квадрата $|i| + |j| \leq d$, называемого *расширенной диаграммой Ньютона.*

4. Продолжим расстановку знаков на вершины расширенной диаграммы Ньютона, используя следующее правило: знак вершины $(e_1 i, e_2 j)$ отличается от знака вершины (i,j) множителем $e_1^i e_2^j$, где $e_1, e_2 = \pm 1$.

5. В каждом из треугольников построенной триангуляции расширенной диаграммы Ньютона соединим отрезком середины тех сторон, на концах которых стоят разные знаки (если таковые имеются). Объединение всех построенных отрезков задает ломаную линию на расширенной диаграмме Ньютона. Эта линия и является комбинаторной моделью кривой Виро.

6. Отождествим между собой противоположные точки границы расширенной диаграммы Ньютона. Тогда некоторые ветви комбинаторной модели кривой Виро склеятся в *овалы*.

Литература.

[1] М.Э. Казарян, Тропическая геометрия, Материал курса летней школы "Современная математика". http://www.mccme.ru/dubna/2006/notes/Kazaryan.pdf

[2] O. Ya. Viro, Introduction into Topology of Real Algebraic Varieties. http://www.math.uu.se/ oleg/es/index.html.

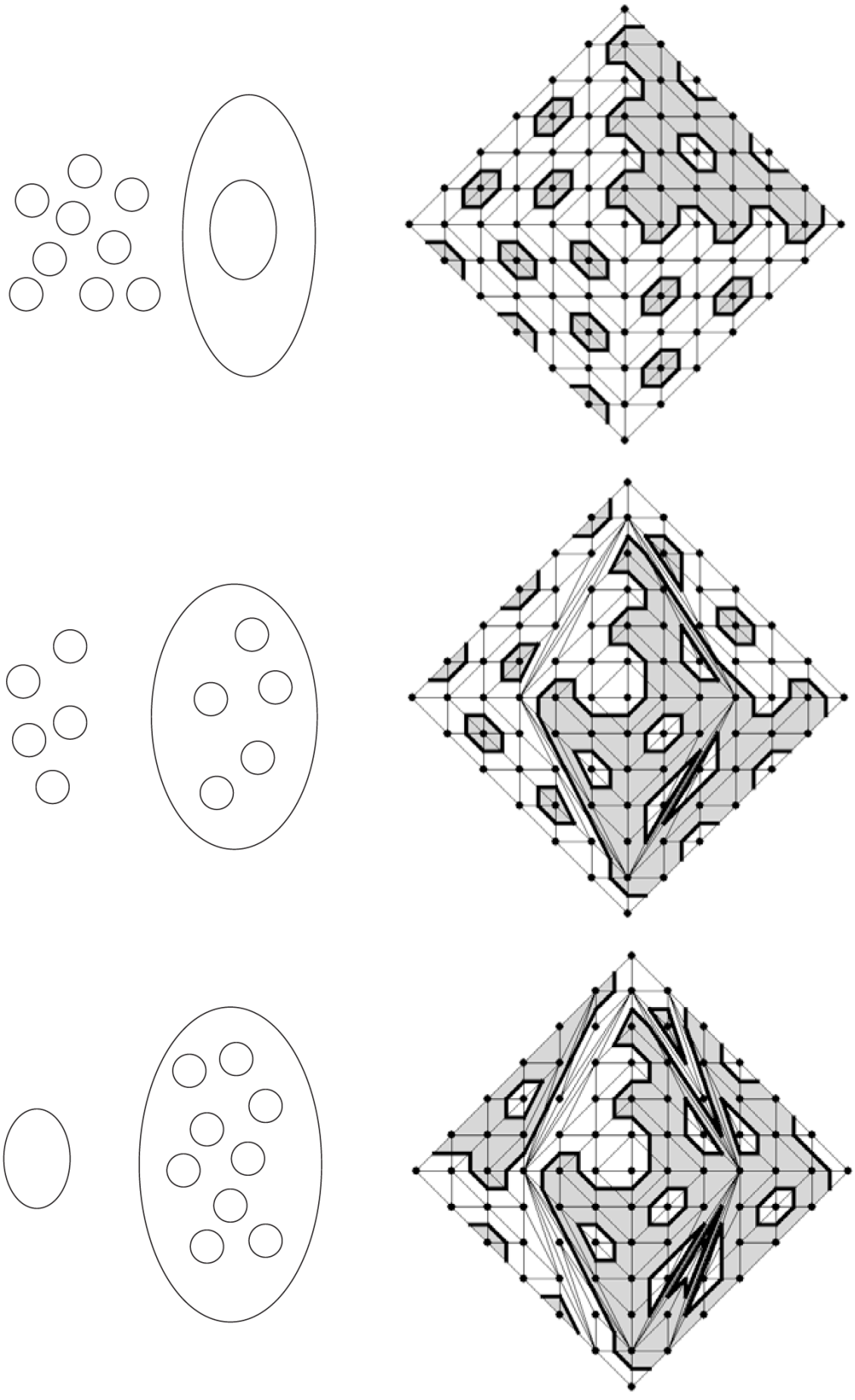


Рис. 8:

TROPICAL GEOMETRY

F. Nilov, A. Skopenkov, M. Skopenkov and A. Zaslavsky

A brief overview. [6]

Hilbert's 16th problem asks *what could be the number and mutual arrangement of curves which form the subset of the plane given by an equation* $\sum_{i+j\le d} a_{ij}x^iy^j = 0$. A more accurate statement and examples are given in part A. [7] The aim of this series of problems is to outline an approach to the "existence" part of Hilbert's 16th problem for $d = 6$ (see Main Theorem below).

It is not easy to determine the number and mutual arrangement of curves for the subset of the plane given by a polynomial in two variables with certain concrete coefficients (even using a modern computer). While solving the problems of part B you will find the statement of main lemma which allows to do it for polynomials of certain specific type. You will see how *tropical geometry* appears naturally while drawing of subsets given by the equations of type $\sum_{i+j\le d} (a_{ij}x^iy^j)^N = 0$, where N is a large odd number. Using tropical geometry you will be able to construct such subsets with distinct mutual arrangement of ovals.

The basic ideas of tropical geometry are elementary. Replace multiplication by addition, and addition by an operation related to addition via the same *distributive* law, like multiplication is related to addition. As such an operation one can take *maximum* $\max\{a, b\}$ of the pair of numbers a and b. Under this transformation the function $\sum_{i+j\le d} b_{ij}x^iy^j = 0$ transforms to the function (check it!): $f(x, y) = \max_{i+j\le d}(ix + jy + b_{ij})$. The set of "break points" of the function is called *the tropical curve.*

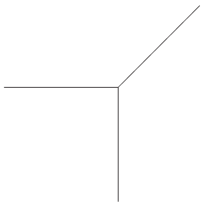


Figure 1.

For instance, a line in the plane is given by the equation $Ax+By+C = 0$. Left part of this equation turns to the function $f(x, y) = \max\{x+a, y+b, c\}$ under our transformation. the set of "break" points of the function $f(x, y)$ looks like shown in figure 1 (check it!). This way *the tropical line* is defined. Tropical lines have many properties of Euclidean lines. Part C of the project deals with "experimental" investigation of these properties.

A. Examples of algebraic curves.

A *polynomial* (in two variables) is a function $F : \mathbb{R}^2 \to \mathbb{R}$ for which there exist numbers d and a_{ij}, $0 \le i, j \le d$, such that $F(x, y) = \sum_{i+j\le d} a_{ij}x^iy^j$. You can use without proof the following non-trivial fact: *for given function F such numbers are unique up to increasing d and taking all the new a_{ij} to be zeroes.*

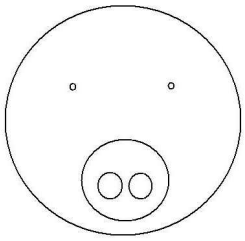


Figure 2.

The *zero set* of the polynomial F is $F^{-1}(0) := \{(x, y) \in \mathbb{R}^2 \mid F(x, y) = 0\}$.

A1. Is F uniquely determined by $F^{-1}(0)$?

A2. Which of the following sets are zero sets of polynomials?

(a) a line; (b) a circle; (c) a point; (d) a segment; (e) the union of 2 lines;

(f) the "pig" (union of 6 circles) in figure 2.

The *degree* of a polynomial is the least possible d for which there exist the required a_{ij}. (The degree is the maximal d such that $a_{i,d-i} \ne 0$ for some representation of the polynomial and for some number i.)

A3. (a) How many points there could be in the intersection of the zero set of a polynomial of degree d and a line?

(b) The zero set of a polynomial of odd degree is unbounded (i.e. is not contained in a disk).

[6] Do not worry if you do not understand something in this brief overview. You can omit it and start solving problems from either part A or part C.

[7] For $d \le 5$ the answer was known as early as in 19th century. Hilbert stated his problem for $d = 6$. The solution for this case was obtained by Gudkov. For $d = 7$ the problem was solved by Viro using methods of tropical geometry. For $d \ge 8$ the problem is open.

A polynomial F is *reducible*, if $F = G \cdot H$ for some polynomials G and H.

Curves[8]**.** A function $\gamma : [a, b] \to \mathbb{R}$ is *differentiable* at the point t_0, if for some number A and any $\varepsilon > 0$ there exist δ such that for any

$$t \in (t_0 - \delta, t_0 + \delta) \quad \text{we have} \quad |\gamma(t) - \gamma(t_0) - A(t - t_0)| < \varepsilon |t - t_0|.$$

A map $\gamma : [a, b] \to \mathbb{R}^2$ can be considered as an ordered pair of functions $\gamma_1, \gamma_2 : [a, b] \to \mathbb{R}$. A map $\gamma : [a, b] \to \mathbb{R}^2$ is *differentiable* if both functions γ_1, γ_2 are differentiable.

A *(smooth) curve* in the plane is a differentiable map $\gamma : \mathbb{R} \to \mathbb{R}^2$ (or $\gamma : [a, b] \to \mathbb{R}^2$).

In problems A4.cfg and A7 you only need to give an example of a polynomial; proving its properties is not required.

A4. (a) There is an irreducible polynomial of degree 3 whose zero set contains a closed curve.

(b) The same for degree 4.

(c) There is an irreducible polynomial of degree 4 whose zero set consists of two closed curves one inside the other.

(d) If the zero set of a polynomial of degree 4 contains two closed curves one inside the other, then the zero set contains no other points.

(e) Is the analogue of (d) correct for an irreducible polynomial of degree 5?

(f) There is a polynomial of degree 4 whose zero set contains 4 closed curves.

(g) There is a polynomial of degree 4 whose zero set contains 3 closed curves.

Ovals . Connected components of the zero set of a polynomial are called *branches*. (Existence of unbounded branches makes the investigation of zero sets harder.) For an unbounded branch B the lines joining the origin O with the points of B have a "limit" line. Two unbounded branches are *elementary equivalent* if their "limit" lines coincide.

A5. The infinite branches of hyperbola $xy = 1$ are elementary equivalent.

Two infinite branches are *equivalent* if there is a sequence of branches joining them, in which sequence each two consecutive branches are elementary equivalent. A zero set is *nondegenerate* if it is a disjoint union of smooth curves. An *oval* of a non-degenerate zero set of a polynomial is either a closed curve (contained in the zero set) or an equivalence class of unbounded branches. (Note that this definition is different from the "correct" one given in textbooks.)

A6. Find all h such that the zero set is non-degenerate and find the number of ovals for the polynomial
(a) $xy(x + y - 1) - h$. (b) $x^3 - x + h - y^2$. (the answer could depend on h).

A7. There is a polynomial of degree 5 whose zero set is non-degenerate and consists of 7 ovals.

Hilbert's 16th problem. *What could be the number and mutual arrangement of ovals of a non-degenerate zero set of a polynomial of degree d?*

We do not assign any formal meaning to the words 'mutual arrangement'. Such a meaning can be assigned, but requires *projectivization* of a polynomial.

Main Theorem. (a) *There is a polynomial of degree* 6 *whose zero set is non-degenerate and consists of* 11 *ovals.*

(b) *There are three polynomials of degree* 6 *each whose zero sets are non-degenerate and consist of* 11 *ovals each, with different mutual arrangement of ovals.*

B. Tropical curve as a limit of algebraic curves.

B1. Draw the zero sets of
(a) $x - y - 1$; (a') $x^{1001} - y^{1001} - 1$;
(b) $x + y - 1$; (b') $x^{1001} + y^{1001} - 1$;
(c) $xy = x + y$; (c') $x^{1001}y^{1001} = x^{1001} + y^{1001}$;
(d) $x^2 + y^2 - 4x - 4y - 2 = 0$; (d') $x^{2002} + y^{2002} - 4^{1001}x^{1001} - 4^{1001}y^{1001} - 2^{1001}$;
(e') $x^{3003} + 2^{1001}x^{1001}y^{2002} - 3^{1001}x^{1001}y^{1001} + y^{2002} - x^{1001} - 2^{1001}$.

Denote by

$$F_N(x, y) = \sum_{i+j \le d} (a_{ij} x^i y^j)^N$$

[8]These definitions are required only for the accurate proofs of problem A4.

a family of polynomials depending on an *odd* number $N \geq 1$. Under the transformation of variables $u = x^N, v = y^N$ each polynomial F_N goes to the polynomial $\sum_{i+j\leq d} a_{ij}^N u^i v^j$ of degree d. So for solution of the Hilbert 16th problem it is worth to determine the number and mutual arrangement of ovals of $F_N^{-1}(0)$.

B2. The number of ovals of $F_N^{-1}(0)$ can be different from that of $F_1^{-1}(0)$.

Denote by B_R the ball of radius R centered at 0.

B3. (a) For each $\varepsilon, R > 0$ there is $N_0 > 0$ such that for each odd $N > N_0$ the intersection of the zero set of $x^{2N} - x^N - y^N$ with B_R is contained in ε-neighborhood of the union of the lines $x = 0$, $x = 1$, $x = y$ and the parabola $y = x^2$.

(b) For each $\varepsilon, R > 0$ there is $N_0 > 0$ such that for each odd $N > N_0$ the set $F_N^{-1}(0) \cap B_R$ is contained in ε-neighborhood of the union of the zero sets of all the polynomials $a_{ij}x^iy^j - a_{kl}x^ky^l$, in which $(i, j) \neq (k, l)$, $i + j \leq d$, $k + l \leq d$.

Denote by $\mathbb{R}_+ := [0, +\infty)$ the set of positive numbers and by $\mathbb{R}^2_+ := [0, +\infty)^2$ the angle defined by the inequalities $x > 0$, $y > 0$. Define a map $LOG : \mathbb{R}^2_+ \to \mathbb{R}^2$ by $LOG(x, y) = (\log_2 x, \log_2 y)$.

B4. (abcde) Draw the *LOG*-image of the intersection with $\mathbb{R}^2_+$ of the zero sets of polynomials (a'b'c'd'e') of B1.

A *tropical polynomial* is a function

$$f(x, y) := \max_{i+j\leq d}(ix + jy + b_{ij}).$$

$$\text{Let} \quad f^{pq} = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) = px + qy + b_{pq}\}.$$

The union of intersections of different f^{pq} is *a tropical curve*. (This is the set of "break points" of f.)

Assume further that all $a_{ij} \neq 0$ for $i + j \leq d$. The tropical curve *corresponds* to the family of polynomials F_N, if $b_{ij} = \log_2 |a_{ij}|$. This definition is motivated by important problem B6b below.

B5. Draw the tropical curve corresponding to the family of polynomials
(a) $(ax)^N + (by)^N + c^N$? (b) $(ax^2)^N + (2bxy)^N + (cy^2)^N$? (the answer could depend on a, b, c.)

Denote by Δ_R the triangle given by the inequalities $x \geq -R, y \geq -R, x + y \leq R$.

B6. (a) For each $\varepsilon, R > 0$ there is N_0 such that for $N > N_0$ the intersection of the *LOG*-image of the zero set of the polynomial $x^{2N} - x^N - y^N$ with the triangle Δ_R is contained in ε-neighborhood of the union of the ray $y = 2x, x \geq 0$ and the ray $x = 0, y \geq 0$.

(b) For each numbers $\{a_{ij}\}_{i+j\leq d}$ and $\varepsilon, R > 0$ there is $N_0 > 0$ such that for $N > N_0$ the set $LOG(F_N^{-1}(0) \cap \mathbb{R}^2_+) \cap \Delta_R$ is contained in ε-neighborhood of the intersection of the tropical curve corresponding to F_N.

C. Tropical lines and circles.

This part of the project is an contest in art: it is suggested to check the theorems of tropical geometry experimentally by drawing accurate figures. Selected figures will be exposed for public viewing. "Problems" C1-C10 are not graded (although for accurate statement and proving some of these assertions additional points would be awarded). Ignore part of a "problem" if you do not know the corresponding theorem of Euclidean geometry. The whole part C of the project is not required for the solution of the Hilbert 16th problem and can be skipped.

Consider the plane with fixed Cartesian coordinate system. A *tropical line* (*"leg"*) is a union of three rays with common origin (called the *vertex*), one of them going "west", the other going "south" and the third going "north-east".

C1. There are different tropical lines intersecting at two different points.

Two points are *in general position* if the Euclidean line passing through these points is not parallel either to coordinate axes or to the line $x = y$.

C2. (a) For each two points in general position there is a unique tropical line passing through these points.
(b) If the vertices of two tropical lines are in general position, then the lines have the only common point.

Two tropical lines are *parallel* if the vertex of one lies on the "north-eastern" ray of the other.

C3. If a point A is in general position with the vertex of a tropical line b, then there is a unique tropical line passing through A and parallel to b.

Two tropical lines are *perpendicular* if the Euclidean lines containing their "north-eastern" rays are symmetric with respect to the line $x = y$.

C4. If a point A is in general position with the vertex of a tropical line b, then there is a unique tropical line passing through A and perpendicular to b.

A *tropical triangle* (*"spider"*) is the union of three tropical lines whose vertices are (pairwise) in general position.

C5. Draw figures to tropical analogues of the following theorems.
(a) The heights of a triangle intersect at a common point.
(b) the Pappus theorem.
(c) The Desargue theorem.
(d) The Sondat theorem.

For given points A and B a *tropical circle* (*"heron"*) is the set of points X for which there are orthogonal tropical lines, one of them passing through A and X and the other passing through B and X. (Recall that there could be different tropical lines passing through A and X.)

C6. (a) Draw a tropical circle. How does this set depend on A, B?
(b) Is it true that each tropical triangle has a circumscribed tropical circle?
(c) The Pascal theorem.

C7.* Define the *tropical middle point of a tropical segment* so that the tropical medians of a tropical triangle would intersect in a common point.

TROPICAL GEOMETRY

F. Nilov, A. Skopenkov, M. Skopenkov, A. Zaslavsky

The main problem's complect consists of two parts: the end of Part B and a new part D. Problems of part D use (except some explicitly indicated cases) neither notions nor results of previous parts of the project. So one may solve these problems without taking part in previous parts.

B. The Viro patchworking theorem.

B3. For each $\varepsilon, R > 0$ there is $N_0 > 0$ such that for each odd $N > N_0$ the intersection of the zero set of the polynomial

(c) $x^{2N} - x^N - y^N$ with the disk B_R and with the first coordinate quarter ($x > 0$, $y > 0$) is contained in the ε-neighborhood of the union of the sets

$$\{(1, y) \mid 0 \le y \le 1\}, \quad 0 \le x = y \le 1 \quad \text{and} \quad y = x^2 \ge 1.$$

(d) $x^{2N} - x^N - y^N$ with the disk B_R and with the second coordinate quarter ($x < 0$, $y > 0$) is contained in the ε-neighborhood of the union of the set that is symmetric to the union from (c) w.r.t. Oy.

B7. State and prove the analogues of B3d for the third and the fourth coordinate quarters.

B8. The intersection of the zero set of the polynomial $x^{2N} - x^N - y^N$ with the third coordinate quarter is empty.

B9. For each $\varepsilon, R > 0$ there is $N_0 > 0$ such that for each odd $N > N_0$ the intersection of the zero set of the polynomial $x^{2N} - x^N - y^N$ with the disk B_R and with

(a) the first coordinate quarter is contained in the ε-neighborhood of the union of the sets $\{(1, y) \mid 0 \le y \le 1\}$ and $y = x^2 \ge 1$.

(b) the second coordinate quarter is contained in the ε-neighborhood of the union of the sets $0 \le -x = y \le 1$ and $y = x^2 \ge 1$.

B10. State and prove the analogue of B9 for the fourth coordinate quarter.

Let us state the Viro patchworking theorem that allows to find the number and mutual arrangement of ovals for certain special algebraic curves.

B11. Each tropical curve is a finite union of segments and rays.

Definition of the Viro curve and its ovals. Take the tropical curve corresponding to $\{a_{ij}\}$. The tropical curve is a finite union of *edges* (segments and rays) that intersect at *vertices* (i.e. at common points of edges). A *face* of the tropical curve is a connected component if its complement in the plane. To each face there corresponds a pair (p, q) of integers such that $px + qy + \log_2 |a_{pq}| = \max\limits_{i+j \le d,\ a_{ij} \ne 0} (ix + jy + \log_2 |a_{ij}|)$ for points (x, y) of this face, and the sign of $a_{p,q}$. *In this definition we use not* $\{a_{ij}\}$ *but the tropical curve whose faces are marked with pairs of integers and signs.*

Make a parallel transfer so that the vertices of the tropical curve would move into the angle $x > 0, y > 0$. Define $U_{p,q,00}$ to be the image of the face of the tropical curve marked by (p, q) under this parallel transfer. Let $U_{p,q,01}, U_{p,q,10}$ and $U_{p,q,11}$ be the symmetric images of $U_{p,q,00}$ under the symmetries with respect to the x-axis, y-axis and $(0, 0)$, respectively. Extend the given disposition of signs from the first coordinate quarter to the whole plane as follows: under the symmetry of U_{pq} w.r.t. the x-axis the sign is multiplied by $(-1)^q$, while under the symmetry of U_{ij} w.r.t. the y-axis the sign is multiplied by $(-1)^p$. (Thus $\operatorname{sgn} U_{pq,st} = (-1)^{ps+qt} \operatorname{sgn} U_{pq,00}$.) Define *the Viro curve* to be the union $\cup\{U_\alpha \cap U_\beta \mid \operatorname{sgn} U_\alpha \ne \operatorname{sgn} U_\beta\}$ of those edges of the tropical curve that split faces of different signs (see Figure). Two unbounded connected components of the Viro curve are

- *elementary equivalent* if they contain rays symmetric w.r.t the origin $(0, 0)$.
- *equivalent* if there is a sequence of components joining them, in which sequence each two consecutive components are elementary equivalent.

An *oval* of the Viro curve is either a closed broken line contained in the Viro curve or an equivalence class of unbounded connected components.

You can use the following theorem without proof:

The Harnak Theorem. *A non-degenerate zero set of a polynomial of degree d cannot have more than* $\frac{(d-1)(d-2)}{2}+1$ *ovals.*

B12.* The Viro patchworking theorem. *Let the Viro curve assigned to the family of polynomials* $F_M = \sum_{i+j\le d} (a_{ij}x^iy^j)^M$ *with all* $a_{ij} \neq 0$ *contain exactly* $\frac{(d-1)(d-2)}{2}+1$ *ovals. Then there exist* N *such that the zero set of the polynomial* $\sum_{i+j\le d} a_{ij}^N u^i v^j$ *is non-degenerate, and the number and mutual arrangement of the ovals are the same as those of the corresponding Viro curve.*

D. Construction of examples in the Hilbert 16th problem.

The aim of part D is to describe tropical curves using purely combinatorial method and to obtain a purely combinatorial construction of examples in 16th Hilbert problem.

Let us recall that *tropical curve of degree d* is the set of "break points" of graph of the function $\max_{i+j\le d}\{ix+jy+b_{ij}\}$ (see the details above, after problem B4).

D1. (a) Check that a tropical curve of degree 1 looks like picture 1. (Compare with the definition of a tropical line in part C).

(b) Each vertex of a tropical curve is contained in at least 3 edges.

To any edge of a tropical curve assign its *multiplicity* as follows. Suppose that value $ix+jy+b_{ij}$ is maximal in one of faces bounded by this edge, and value $i'x+j'y+b_{i'j'}$ is maximal in the other one. So the line, which contains the given segment, has the equation $(i-i')x+(j-j')y+(b_{ij}-b_{i'j'})=0$. We define *multiplicity* of the given edge as the greatest common divisor of numbers $i-i'$ and $j-j'$.

In pictures we shall denote the multiplicate edges of a tropical curve with double (triple, and so on) lines.

D2. Tropical curves of degree d have the following properties:

(a) The slope of any edge is a rational number.

(b) For any vertex the following balance condition holds. Denote by v_i a vector beginning at the given vertex parallel to i-th edge starting from the vertex, and equal to the shortest vector with integer coordinates and given direction, multiplied by edge's multiplicity. Then $\sum v_i = 0$.

(c) There are $3d$ infinite edges (counted with multiplicity), d of them are directed (strictly) to the "west", d — to the "south" and d — to the "north-east" with slope angle 45°.

D3. (a) One may uniquely restorate a tropical polynomial $\max_{i+j\le d}\{ix+jy+b_{ij}\}$ (up to adding a constant) by its tropical curve.

(b) If the edges of a graph in the plane are segments and rays with given multiplicities, and the conditions (a), (b), (c) of problem D2 are satisfied, then the graph is a tropical curve of degree d.

If two tropical curves have the same combinatorial type of their graphs and the same slopes of their edges (but not nesessary their lengths and positions), we shall say that these curves have it the same configuration.

D4. Draw 5 different configurations of tropical curves of degree two.

All information required for solving the following problems you can find in the paragraph "Definition of Viro curve and its ovals" contained in the previous part.

D5. What maximal number of ovals may have Viro curve if $d=$ (a) 2; (b) 3; (c) 4; (d) 5? (We do not require the proof of maximality. Compare your answer with problems A4f and A7).

D6*. Write down a computer program which:

(a) draws all configurations of tropical curves of given degree d;

(b) given a tropical curve configuration and given the set of signs "plus, minus" assigned to all the faces U_{ij} of its complement — the program checks the number of Viro curve ovals.

D7*. (ab) Prove the Main Theorem (you may use Viro patchworking theorem without proof).

SOLUTIONS

A1. Answer:no. For example, the line $x = 0$ is a set of zeros for different polynomials $F(x, y) = x$ and $G(x, y) = x^2$.

A2. Answer: a, b, c, e, f.

Examples. (a) Any line on the plane has an equation $Ax + By + C = 0$ for some numbers A, B, C.

(b) Equation of a circle: $(x - x_0)^2 + (y - y_0)^2 - R^2 = 0$, where (x_0, y_0) are coordinates of centre, R is radius.

(c) Equation of a point (x_0, y_0): $(x - x_0)^2 + (y - y_0)^2 = 0$.

(e) Equation of a unite of two lines: $(Ax + By + C)(ax + by + c) = 0$, where $Ax + By + C = 0$ is an equation of the first line, $ax + by + c = 0$ — of a second one.

(f) Equation of a unite of 6 circles: $\prod_{k=0}^{6} \left((x - x_k)^2 + (y - y_k)^2 - R^k\right) = 0$, where $(x-x_k)^2+(y-y_k)^2-R_k^2 = 0$ is an equation of k-th circle.

Impossibility in point (d) is consequence of Problem A3a.

A3. (a) Let us parametrize the line l: $x = x_0 + \alpha \cdot t$, $y = y_0 + \beta \cdot t$. Substituting these formulas in the polynomial, we'll get a new polynomial $P(t)$, its degree no more than d. So polynomial $P(t)$ has no more than d real roots, or equals to zero everywhere. Now let us prove that for any $d' < d$, there exist a curve of degree d and a line l such one, that they have d' points of intersection. Consider d lines, which are differ from l, and such that exactly $d - d'$ of them are parallel to l. The product of their equations is the polynomial we need.

(b) Let d be the degree of given polynomial $F(x, y) = \sum_{i+j\le d} a_{ij}x^i y^j$. We'll show that there exist some non-degenerate change of coordinates $x = \alpha_1 x' + \beta_1 y'$, $y = \alpha_2 x' + \beta_2 y'$ (the word "non-degenerate"means that $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$), such that after it the monomial $(x')^d$ will have non-zero coefficient.

Coefficient $A(\alpha_1, \alpha_2)$ of monomial $(x')^d$ equals $\sum_{i+j\le d} a_{ij}\alpha_1^i\alpha_2^j$. Numbers a_{ij} aren't equal zero (at least, some of them), so, there exist such α_1 and α_2, that at least one of them isn't equal 0, and $A(\alpha_1, \alpha_2) \neq 0$. Now we take coefficients β_1 and β_2 not proportional to α_1 and α_2 (i.e., $\alpha_1\beta_2 - \alpha_2\beta_1 \neq 0$), and it will be the change we seek for.

Now let us return to solving our problem. The change from the Lemma transforms bounded sets to bounded ones, so we may suppose that monomial x^d has non-zero coefficient. As d is odd number, so for any y the equation $F(x, y) = 0$ has some solution. So $F^{-1}(0)$ is unlimited.

A4. (a) For instance, take the polynomial $f = xy(x + y - 1) + \frac{1}{100}$.

Denote by ϕ the zero set of this polynomial. Let us prove that this polynomial is irreducible. Indeed, otherwise there are polynomials g and h, such that $f = gh$. Then one of them is a polynomial of degree 1 and so ϕ contains a line. This line must have a common point with one of the lines Ox and Oy. But it can't be true because ϕ is disjoint with Ox and Oy. Thus f is irreducible.

Coordinates x of the intersection of line $y = c$ with ϕ satisfy the equation $x^2 + (c - 1)x + \frac{1}{100c} = 0$. The discriminant $D = D(c)$ of this equation is equal to $(c - 1)^2 - \frac{1}{25c}$. The equation $D(c) = 0$ is equivalent to the equation $f(c) := 25c(c - 1)^2 - 1 = 0$. This equation has degree 3 and thus has no more than 3 roots. Since $f(\frac{1}{100}) < 0$, $f(\frac{1}{2}) > 0$, $f(1) < 0$, $f(2) > 0$, it follows that two roots c_1 and c_2 of the equation $f(c) = 0$ belong to the interval $(0, 1)$, and the third root belongs to the interval $(1, 2)$. Therefore $D(c) = 0$ precisely in two points c_1 and c_2 of the interval $(0, 1)$, and $D(c) > 0$ for any $c \in (c_1, c_2)$ and $D(c) < 0$ for remaining points of the interval $(0, 1)$. (We assume w. l. g. that $c_1 < c_2$.) Thus for c equal either c_1 or c_2 the straight line $y = c$ intersects ϕ exactly at one point. Therefore for $c \in (c_1, c_2)$ the straight line $y = c$ intersects ϕ at two points $(x_1(c), c)$ and $(x_2(c), c)$, where $x_{1,2}(c) = \dfrac{\pm\sqrt{D} - (c - 1)}{2}$. For remaining values $c \in (0, 1)$ the straight line $y = c$ does not intersect ϕ.

Define the curve

$$\gamma : [c_1, 2c_2 - c_1] \to \mathbb{R}^2 \quad \text{by the formula} \quad \begin{cases} (x_1(t), t) & t \in [c_1, c_2] \\ (x_2(2c_2 - t), 2c_2 - t) & t \in [c_2, 2c_2 - c_1] \end{cases}$$

Since the functions $x_1(c)$ and $x_2(c)$ are differentiable, it follows that the map $\gamma(t)$ is differentiable at all points except c_2. Since $2c_2 - t = t$ for $t = c_2$ and $(x_1)'(c_2) = (x_2)'(c_2)$, the map $\gamma(t)$ is smooth at all points. Now it is clear that $\gamma(I)$ is a closed curve contained in ϕ.

(b) Hint. Consider the polynomial $(x + 1)(x - 1)(y + 1)(y - 1) + \frac{1}{100}$.

(c) Hint. Consider the polynomial $(x^2 + y^2 - 1)(x^2 + y^2 - 9) + \frac{1}{100}$.

(d) Suppose the contrary: there exist at least one other point X. Consider some point Y inside the inner closed curve. Then the line XY intersects the set of zeros of the given polynomial in 5 or more points. It contradicts the statement of Problem A3(a).

(e) Answer: no. Hint. Consider the polynomial $x(x^2+y^2-1)(x^2+y^2-9)+\frac{1}{100}$.

(f) Hint. Consider the polynomial $(x^2+2y^2-3)(2x^2+y^2-3)+\frac{1}{100}$.

(g) Hint. Consider the polynomial $(x^2+y^2-1)(x-y-1)(x+y-1)+\frac{1}{100}$.

A5. Direction of a line OM, on which lie points 0 (beginning of coordinates) and $M(x,y)$ on hyperbola branch in the first quadrante, tends to direction of the line Ox (axis) when $x \to +\infty$. So Ox is "limit line"for hyperbola $xy=1$ branch in the 1st quadrante. Similarly, this line is a "limit line"for the other branch of hyperbola. So hyperbola's branches are equivalent.

Definition. Two unbounded branches are *elementary equivalent*, if they have a common "limit line".

A6. (a) Answer: one oval, if $h<0$; two ovals, if $h \in (0,1/27)$, one oval, if $h>1/27$. If $h=0$ or $h=1/27$, the algebraic curve is degenerate. Here is the proof.

Denote $f(x,y) := xy(x+y-1)+h$. Further, denote the points of intersection of lines Ox, Oy and $x+y-1=0$ and regions, into which the plane is divided by the lines as follows:

$$A := (1,0), \quad B := (0,1),$$

$C := (0,0)$, $X := \{(x,y) \mid x>0, y>0, x+y<1\}$, $X_A := y<0, x+y>1$, $X_B := x<0, x+y>1$, $X_C := x<0, y<0$, $Y_A := x<0, y>0, x+y<1$, $Y_B := x>0, y<0, x+y<1$, $Y_C := x>0, y>0, x+y>1$.

Obviously $f(x,y)=h$ if (x,y) belongs to one of lines Ox, Oy or $x+y-1=0$, and $f(x,y)<h$ when (x,y) belongs to one of regions X_A, X_B, X_C or X, and $f(x,y)>h$ when (x,y) belongs to one of regions Y_A, Y_B и Y_C. So if $h>0$ then zeros of polynomial $f(x,y)$ may lie only in X_A, X_B, X_C and X, and if $h<0$ they may lie only in Y_A, Y_B and Y_C.

Suppose $h<0$. Denote $y_A := Y_A \cap f^{-1}(0)$. Definitions of y_B and of y_C are similar.

Let's prove that y_A is a connected componenta of the set $f^{-1}(0)$ of zeros of f. Coordinates x of points of intersection lines $y=c$ and $f^{-1}(0)$ are roots of the equation $x^2+(c-1)x+\frac{h}{c}=0$. The discriminant $D=D(c)$ of this equation equals to $(c-1)^2-\frac{4h}{c}$. As $h<0$ so for any $c \in R_+$, $D(c)>0$. It follows, that any line $y=c$, where $c \in R_+$, intersects F in two points (no more no less) namely $(x_{1,2}(c),c)$ such that $x_{1,2}(c) = \dfrac{\pm\sqrt{D}-(c-1)}{2}$. Denote

$$\gamma : R_+ \to \mathbb{R}^2 \quad \text{by formula} \quad \Big\{(x_2(t),t) \quad t \in R_+ \,.$$

The function $x_2(c)$ is smooth, so the curve $\gamma(t)$ is smooth also. $\gamma(R_+)=y_A$ implies that y_A is a connected componenta of the set $f^{-1}(0)$ of zeros of f. Similarly y_B и y_C are connected componentas of the same set. It is easy to prove that the direction of line Ox is a "limit direction" for the branch y_C. Similarly, this direction is a "limit direction" for branch y_A. So branches y_A and y_C are elementary equivalent. Similarly, branches y_A and y_B are elementary equivalent, as for these branches the direction of line $x+y-1=0$ is a "limit one". So, branches y_A, y_B and y_C are equivalent and form one oval.

If $h=0$, then an algebraic curve f is a degenerate one. Suppose $h>0$. Denote

$$x := X \cap f^{-1}(0), \quad x_A := X_A \cap f^{-1}(0), \quad x_B := X_B \cap f^{-1}(0) \quad \text{and} \quad x_C := X_C \cap f^{-1}(0).$$

One proves that x_A, x_B and x_C are connected and equivalent in the same way as in the case $h<0$. So they form one oval for any $h>0$. If the set x is non-empty and has more than one point, it is an oval (the proof is similar to solution of Problem A4(a)).

Let us prove that the set of points x isn't empty only if $h \in (0, \frac{1}{27}]$. It is clear that x is empty if and only if $D(c)<0$ for any $c \in (0,1)$. Derivative $D'(c)>0$ if $c \in (0,1/3)$, $D'(c)=0$ if $c=1/3$, $D'(c)<0$ if $c \in (1/3,1)$. So the function $D(c)$ has its maximum on intervale $(0,1)$ in the point $c=1/3$. So $D(c)<0$ for any $c \in (0,1)$ if and only if $D(1/3)=4/9-\frac{4h}{c}<0$, i.e. $h>1/27$. If $h=1/27$, then the set x consists from only one point. So, if $h \in (0,1/27)$, then the set of zeros $f^{-1}(0)$ consists from two ovals, if $h=1/27$, then the algebraic curve is degenerate, if $h>1/27$ then the set $f^{-1}(0)$ has only one oval.

(b) Answer: one oval if $h \in (-\frac{2}{3\sqrt{3}}, \frac{2}{3\sqrt{3}})$, two ovals if $h \in (-\infty, -\frac{2}{3\sqrt{3}})$ and $h \in (\frac{2}{3\sqrt{3}}, \infty)$, an algebraic curve $x^3-x+h-y^2$ is degenerate if $h=\pm\frac{2}{3\sqrt{3}}$. Hint. The situation is similar to (a).

A7. Hint. Consider the polynomial $x((x-1)^2+y^2-2)((x+1)^2+y^2-2)+\frac{1}{100}$.

B1. (a') *Hint.* See figure 3.a'.Why the picture is right one? It is clear that the line $x+y=0$ is an axis of symmetry for the set of zeros of our polynomial. So we may study only the case $y>-x$. Moreover. it lies under

the line $y = x$. The set of zeros intersects with coordinate axes in points $(0, -1)$ and $(1, 0)$. If $x = 1 + \epsilon$ ($\epsilon > 0$) holds $(1001\epsilon)^{\frac{1}{1001}} < y < 1 + \epsilon$. So, if ϵ is sufficiently small, y may take values from 0 to 1. If $\epsilon > 1/1001$, then x is approximately equal to y. If $1 - \epsilon < x < 1$ (ϵ is sufficiently small) y may take values from -1 to 0. The case $y < x$ is analogous.

(b') *Hint.* See figure 3.b'.Everything is analogous to (a').

(c') *Hint.* See figure 3.c'.

(d') *Hint.* See figure 3.d'.

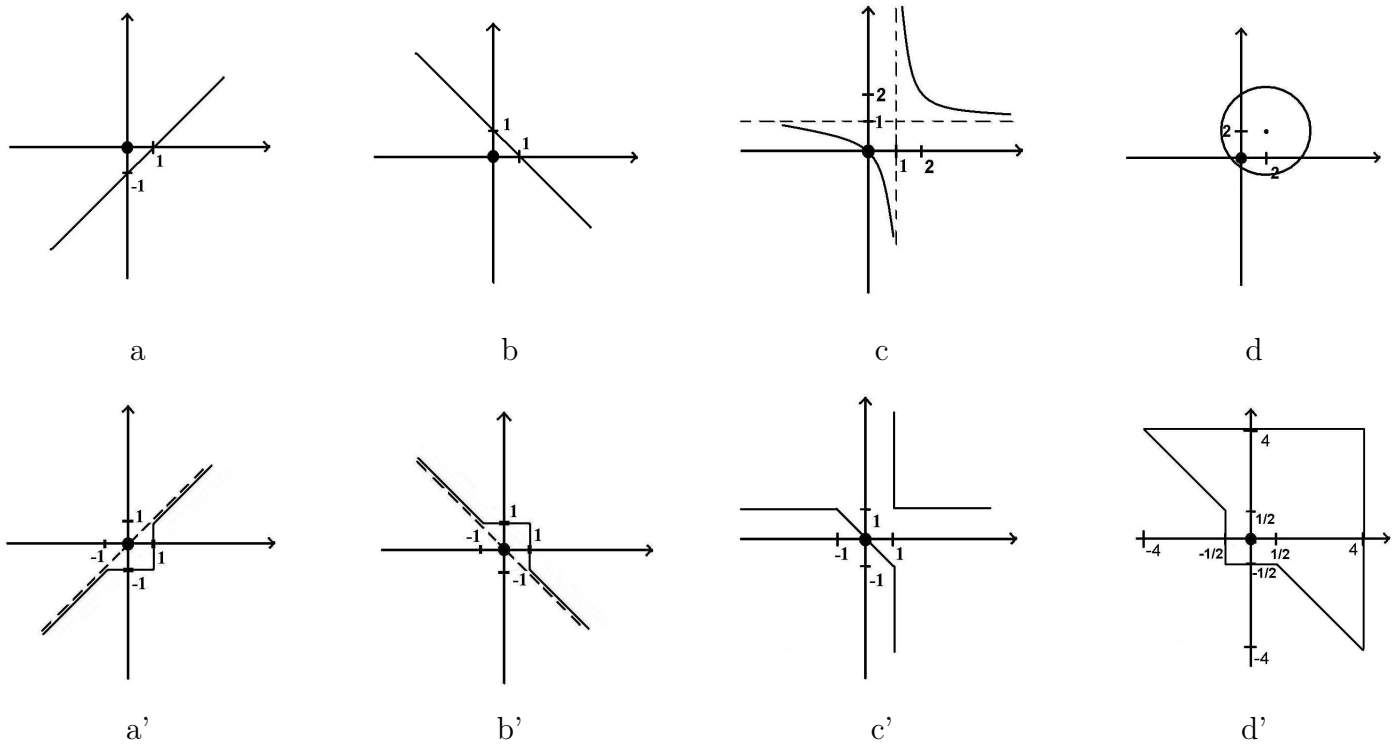


Figure 3.

B2. Let $F(x, y) = x^3 - px + q - y^2$, where $p, q > 0$. Then the set of zeros of the polynomial F consists of two ovals, if a polynomial $f(x) = x^3 - px + q$ has three real roots, and it consists of one oval, if $f(x)$ has one real root. After solving the equation $f'(x) = 0$, we'll see, that $f(x)$ has a local maximum in the point $x_1 = -\sqrt{p/3}$ and a local minimum in point $x_2 = \sqrt{p/3}$. It follows, that $f(x)$ has three roots if and only if $f(x_2) < q < f(x_1)$, i.e., $4p^3 > 27q^2$. Analogously, we can prove that a set of zeros of a polynomial $F_N(x, y)$ consists of two ovals if $4p^{3N} > 27q^{2N}$ and of one oval otherwise. It is obvious that if $1 < \frac{p^3}{q^2} < \frac{27}{4}$ the first inequality doesn't hold, and the second holds for sufficiently big N.

B3. (a) Point (a) is a specific case of (b).

(b) *Hint.* Suppose that there is a point (x, y) such that in it values of all monomials $a_{ij}x^iy^j$ differ by their modules, and $|a_{kl}x^ky^l| > |a_{ij}x^iy^j|$ for all pairs $(i, j) \neq (k, l)$. Then, when $N \to \infty$, we have $\left|\frac{a_{ij}x^iy^j}{a_{kl}x^ky^l}\right|^N \to 0$. So for any sufficiently big N $|a_{ij}x^iy^j|^N$ is more than the sum of modules of all the rest monomials, so the equality $F_N(x, y) = 0$ is impossible. So, for sufficiently big N the set $F_N^{-1}(0)$ tends to some subset of the union of sets, which are defined by equalities of the type $|a_{ij}x^iy^j| = |a_{kl}x^ky^l|$.

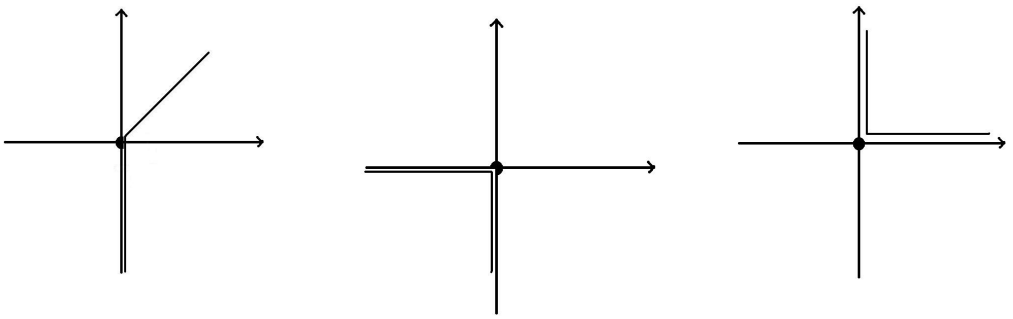
(c) It follows from the statement of previous problem. One must consider it for a polynomial $F_N = x^{2N} - x^N - y^N$.

(d) A set of zeros of a polynomial $x^{2N} - x^N - y^N$, which lie in a second quadrant, is symmetric with respect to ordinate axis to the set of zeros of a polynomial $x^{2N} + x^N - y^N$, which lie in a first quadrant. So our statement is a consequence of the statement of point (b).

B4. (a) *Hint.* See figure 4.a. Let us explain why the picture is correct. It is clear that the set of zeros of the function $f(x, y) := 2^{1001x} - 2^{1001y} - 1$ lies on the right of the axis Oy. If $y < 0$, then x is approximately equals to zero, and If $y > 0$, then x is approximately equals to y.

(b) *Hint.* See figure 4.b. Everything is analogous to (a).

(c) *Hint.* See figure 4.c.



a b c d

Figure 4.

B6. (a) We may follow the method used in problem B3b, and we'll see that the intersection of zero set with the 1st coordinate quadrant lies near the union of sets $x^2 = y \geq x$,$x^2 = x \geq y$, $x = y \geq x^2$. But given polynomial has the same signs of coefficients of monomials x^N and y^N, so F_N cannot equal zero near the last set. Logarithmical map brings the first named set to the ray $y = 2x$, $x \geq 0$, and the second one — to the ray $x = 0$, $y \leq 0$.

B7. Arguing, as in Problem B3d, we see that the intersection of zero set with the first quadrant, is symmetric with respect to abscissa axis to the intersection of zero set with the fourth quadrant, and is symmetric with respect to coordinate beginning to the intersection with third quadrant.

B8. In the third quadrant $x < 0$, $y < 0$. It means that all monomials in F_N are positive, so the equality $F_N(x, y) = 0$ is impossible.

B9. (a) The solution is analogous to the solution of Problem B6a.

(b) It follows from the problem B3d, that the intersection of a set of zeros with the second quadrante lies near sets $x^2 = y \geq -x$, $x^2 = -x \geq y$, $-x = y \geq x^2$. Signs of monomials $-x^N$ and x^{2N} coincide in the second quadrant. So the intersection of a set of zeros with the second quadrant lie only near the first and the third of the named sets.

B10. Arguing, as in Problem B9b, we see that the intersection we study lies near the unite of sets $(1, y)$, $0 \geq y \geq -1$ and $0 \leq x = -y \leq 1$.

B11. Any edge of a tropical curve may be defined by the system which consists of one equation and some inequalities of type$ix + jy + b_{ij} = kx + ly + b_{kl} \geq px + qy + b_{pq}$. If this system is compatible, then the equation defines some line, and inequalities show that one must take a ray or segment instead of all line.

B12. *Hint.* Really, let us study zeros of a polynomial $F_N(x, y)$ in each quadrant separately. The map $LOG : (\mathbb{R} - \{0\})^2 \to \mathbb{R}^2$, $(x, y) \mapsto (\log_2 |x|, \log_2 |y|)$ is a bijection of each quadrant to the plane. Let us take any quadrant (for example $x, y > 0$), and let us identify it with the plane with this map. A tropical curve, corresponding to a tropical polynomial $\max\limits_{i+j \leq d} \{ix + jy + b_{ij}\}$, $b_{ij} = \log_2 |a_{ij}|$, divides a tropical plane to some areas. In each of this areas one of the monomials $(a_{ij} x_i y_j)^N$ defines the behavior of the polynomial $F_N(x, y)$, and it is positive or negative, correspondingly to the sign of a coefficients a_{ij} (of course, it depends also from the quadrant). Let us paint each area of the complement to tropical curve in one color, if F_N is positive in this area, and in other color, if F_N is negative in it. If two areas i-th common edge are painted in different colors, then, by the Theorem on intermediate value near this edge lies some branch of the set of zeros of F_N. Now, if such two areas are painted in the same color, then no real point of the curve lies near this edge. So, for big odd values of N the set of zeros of F_N (in chosen quadrant) may be approximately shown as a set of some edges of a tropical curve (which may be implicitly named), and the set of zeros of F_N in all the plane is, approximately, a Viro curve.

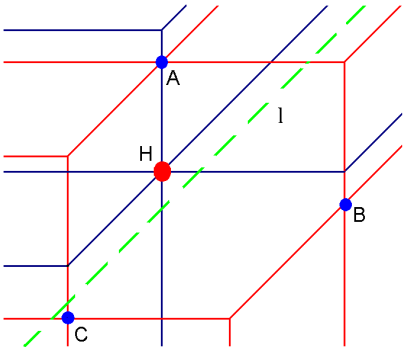
Principally, the set of zeros of F_N could have more branches, than Viro curve — for example, there could be some "small" ovals near vertices of tropical curve. But the supposition about the number of Viro curve ovals (we suppose it has $(d-1)(d-2)/2 + 1$ ovals), combined with Harnak theorem guarantees us from superfluous branches and ovals.

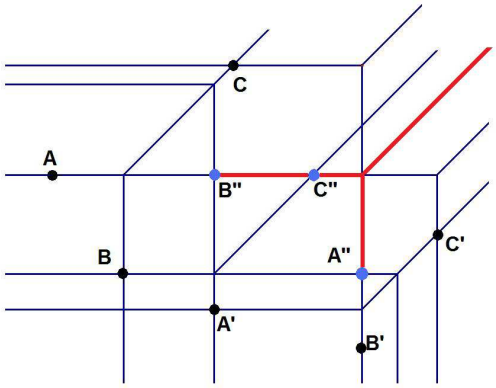
Remark. Authors of the problem don't know, is Viro patchworking theorem is true without the supposition that Viro curve has $(d-1)(d-2)/2 + 1$ ovals.

C5. (a) See figure 5.a.

(b) See figure 5.b.

(c) See figure 6.





a b

Figure 5.

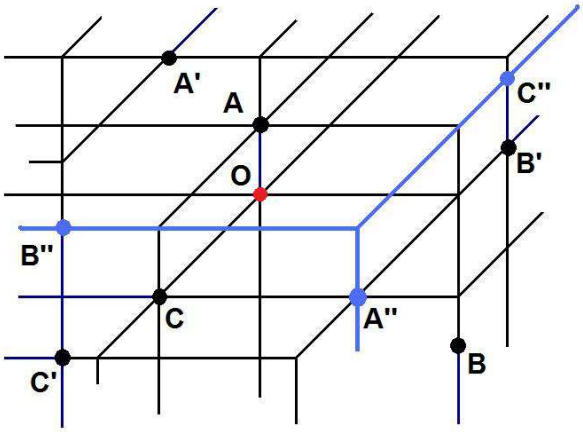


Figure 6.

D1. (a) *Hint.* The behavior of a function $\max\{x+a, y+b, c\}$ is such one. If x and y are negative and big (by module) then the constanta c is the biggest of our three values. When x increases, nothing changes till the point (x, y) will intersect the vertical line $x+a=c$. After such intersection the value $x+a$ is maximal. Similarly, when the point (x, y) moves up, maximal value is still c till it would reach the horizontal line $y+b=c$. On it both values $y+b$ and c are maximal, later - only $y+b$. At last, the areas, in which maximal value is $x+a$ or $y+b$, are divided by the ray of the line $x+a=y+b$. All three rays have the common point $(c-a, c-b)$

D2. (a) It is obvious.

(b) *Hint.* Let us take any vertice of the curve. Suppose that there are r areas (supplements to tropical curve) which are near this vertice, and that in these areas maximal are functions $i_1x+j_1y+b_{i_1j_1}, \dots, i_rx+j_ry+b_{i_rj_r}$, respectively (we suppose that the areas are numerated in positive direction, against the clock needle). Then the equality is obvious:

$$\binom{i_2-i_1}{j_2-j_1}+\dots+\binom{i_r-i_{r-1}}{j_r-j_{r-1}}+\binom{i_1-i_r}{j_1-j_r}=0.$$

Now one has to notice only, that the vector $\binom{i_{s+1}-i_s}{j_{s+1}-j_s}$ differs from the vector v_s in the "balance condition" only by turn on $90°$.

(c) *Hint.* Let us prove, for example, that the tropical curve of degree d has exactly d horizontal rays (counting with multiplicity, of course). Let us study only the part of the plane, where coordinate x is negative and very big by module. Obviously in this part only one of the values $jy+a_{0j}$, $j=0,1,\dots,d$ may be the maximal one. It is

obvious also, that in this part, when y is negative and big by module, then a_00 is maximal, and when y is positive and big by module, then $dy+a_{0d}$ is maximal. Let y grow, and suppose, that maximal value will be (successively) a_{00}, $j_1y + a_{0j_1}$, $j_2y + a_{0j_2}$, $\dots$, $j_ky + a_{0j_k}$, $dy + a_{0d}$. One easily sees, that $0 < j_1 < j_2 < \cdots < j_k < d$. Then multiplicities of horizontal edges are equal to j_1, $j_2 - j_1$, $\dots$, $d - j_k$. So their number (counting with multiplicity) equals $(j_1) + (j_2 - j_1) + \cdots + (d - j_k) = d$.

D3. (ab) *Hint.* Really, suppose, that in some area a tropical polynomial coincides with a linear function $ix + jy + b_{ij}$. Consider the line, which contains a segment of boundary of this area; let $px + qy + r = 0$ be its equation. Then in the neighbor area (which borders with the first one by the segment) our polynomial coincides with linear function $(i + p)x + (j + q)y + (b_{ij} + r)$. In other words, we set the equality $b_{i+p,j+q} = b_{i,j} + r$. Proceeding in the same way, we'll restore all the polynomial, area by area, by induction. The "balance condition" guarantees, that we'll never come to contradiction. The condition on behavior of tropical curve on infinity guarantees existence only such "tropical monoms", which we have got in the process, which are only possible for tropical polynomials of given degree.

D4. *Hint.* A tropical curve of degree two one may get, as usual hyperbola, by little stirring of unite of two tropical lines. Unite of two tropical lines may be defined by the sum of two tropical polynomials of degree one. A graph — set of break points of such sum — has a vertice of 4 valency, in it maximal are four functions at once. When we'll stir one of these function (small stirring), a point of 4 valency will break on two points of 3 valency. Some of such possible tropical curves of degree two are given on Figure 7.

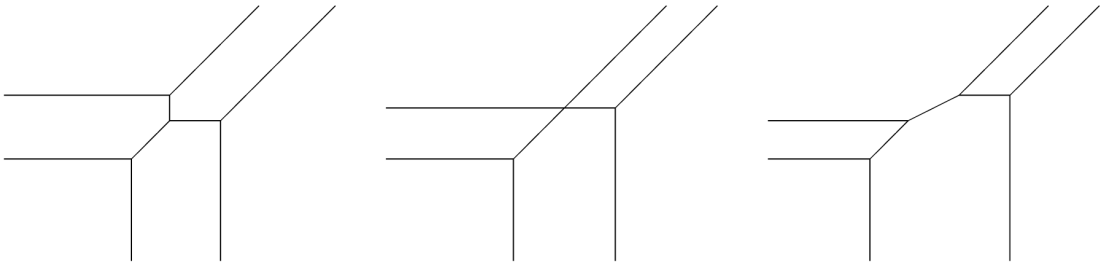


Figure 7.

D5. *Answer.* (a) 1; (b) 2; (c) 4; (d) 7.

D7. (b) **Dividing of Newton diagram.** When solving problems of part D7 it may be useful to remember such a "dual" description of tropical curves configurations. Consider a triangle on the plane whose vertices are $(0; 0)$, $(0, d)$ and $(d, 0)$. This triangle is called it a Newton triangle of a tropical polynomial. If you have any tropical curve, you have also the corresponding division of Newton triangle to a number of convex polygons with integer vertices. Namely, consider an area in complement of a tropical curve, in which the value $ix + jy + b_{ij}$ is maximal. We'll juxtapose to it a vertex with coordinates (i, j) on Newton triangle. If some edge divides two areas, we'll juxtapose to it the segment in Newton diagram from one vertex to other. At last, any vertex of tropical curve, in which r areas meet, corresponds the polygon with r corresponding vertices. In particular, if some area is infinite, the corresponding point will lie on the border of diagram, and if the edge is infinite — the corresponding segment lies on the border. It is useful to remember, that the direction of any edge of tropical curve is orthogonal to the direction of "dual" edge on diagram.

An algorithm of drawing of Viro curves. It is convenient to reformulate the procedure of drawing of Viro curves on the "dual" language of Newton diagrams. This procedure, named "Viro patchworking", consists in such successive steps (look the result in figure 8).

1. Take any triangulation of Newton diagram Δ with integer vertices;
2. In vertices of this triangulation we pose signs + or -, in arbitrary way.
3. Reflecting the Newton diagram with its triangulation respectively from coordinate axes, we get the triangulation of a square $|i| + |j| \le d$, (this square has the name of it expanded Newton diagram).
4. Now we continue posing signs on vertices of the expanded Newton diagram, as follows: sign of vertice $(e_1 i, e_2 j)$ differs from the sign of vertice (i, j) by the factor $e_1^i e_2^j$, where $e_1, e_2 = \pm 1$.
5. In every triangle of our triangulation of expanded Newton diagram we'll join by a segment midpoints of edges, on whose ends signs are different (if one has such edges). Unite of all this segments is a broken line on expanded Newton diagram. This line is a combinatorial model of Viro curve.

6. Let us identify the opposite points of the border of expanded Newton diagram. Then some branches of combinatorial model of Viro curve will patch in it ovals.

References.

[1] M. Kazaryan, Tropical geometry, Lecture notes of a course in school "Contemporary mathematics". http://www.mccme.ru/dubna/2006/notes/Kazaryan.pdf

[2] O. Ya. Viro, Introduction into Topology of Real Algebraic Varieties. http://www.math.uu.se/ oleg/es/index.html.

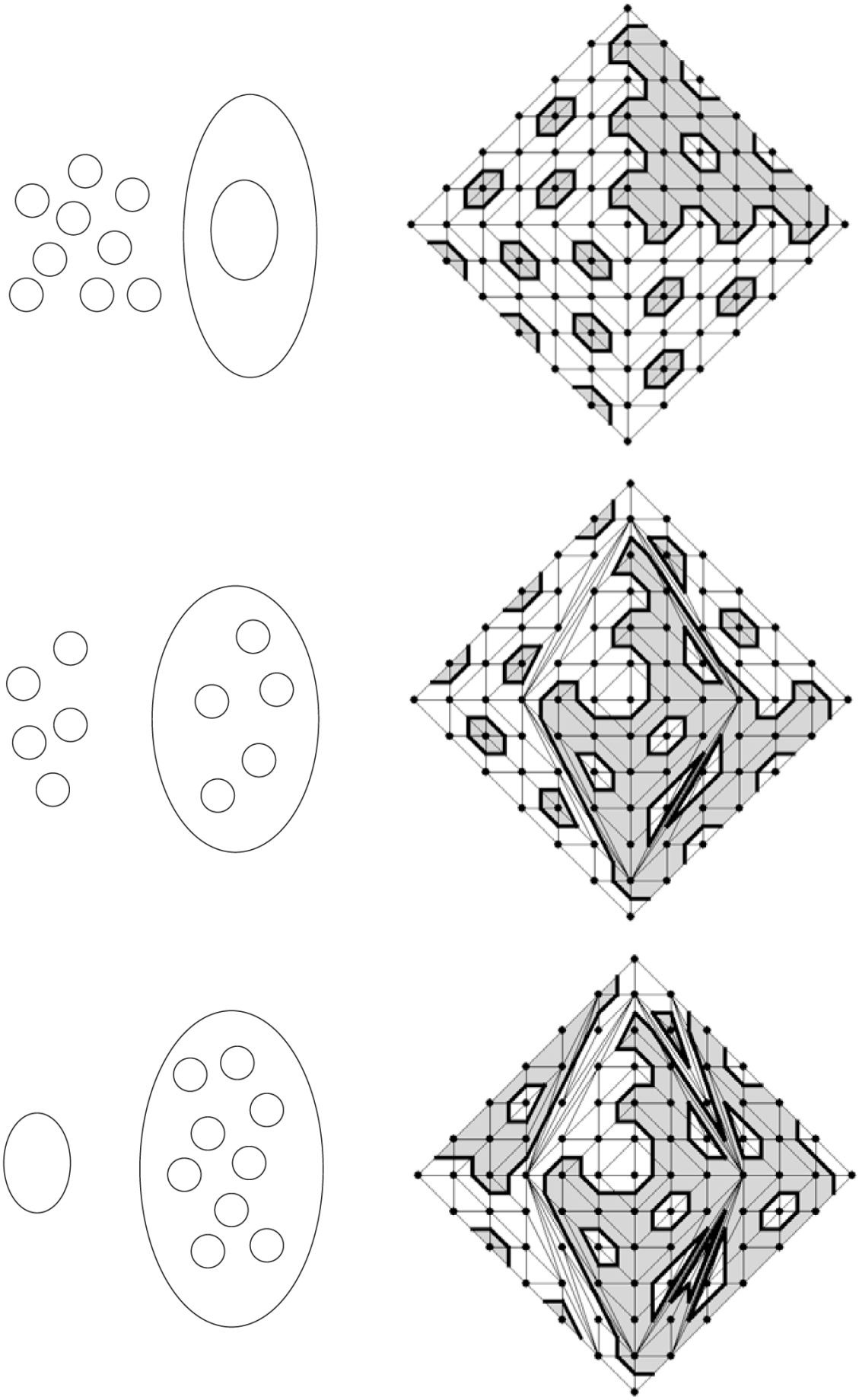


Figure 8.