To forge the connection between Bernoulli numbers and sums of powers, we first observe that

$$\begin{aligned}\frac{e^{nx}-1}{e^x-1} &= \sum_{k=0}^{n-1} e^{kx} \\ &= \sum_{k=0}^{n-1}\sum_{p=0}^{\infty}\frac{k^p x^p}{p!} \\ &= \sum_{p=0}^{\infty}\frac{x^p}{p!}\sum_{k=0}^{n-1}k^p \\ &= \sum_{p=0}^{\infty}\frac{x^p}{p!}s_{n-1}^p.\end{aligned}$$

On the other hand, the same quantity satisfies

$$\begin{aligned}\frac{e^{nx}-1}{e^x-1} &= \frac{e^{nx}-1}{x}\frac{x}{e^x-1} \\ &= \sum_{i=1}^{\infty}\frac{n^i x^{i-1}}{i!}\sum_{k=0}^{\infty}B_k\frac{x^k}{k!} \\ &= \sum_{p=0}^{\infty}\frac{x^p}{p!}\sum_{k=0}^{p}\frac{B_k}{k!}\frac{p!}{(p-k+1)!}n^{p-k+1}\end{aligned}$$

after replacing $i+k-1$ by p. Equating coefficients of x^p implies that

$$s_{n-1}^p = \frac{1}{p+1}\sum_{k=0}^{p}\binom{p+1}{k}B_k n^{p-k+1}.$$

One can recover the earlier exact formulas for s_n^p with $1 \le p \le 3$ based on the values $B_0 = 1$, $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, and $B_3 = 0$. A fuller exposition of Bernoulli numbers appears in Wikipedia and the references [**77, 95, 151**].

1.9. Problems

(1) The Goldschmidt method of division reduces the evaluation of a fraction $\frac{a}{b}$ to addition and multiplication. By bit shifting we may assume that $b \in (\frac{1}{2}, 1]$. Replace b by $1-x$ and write

$$\begin{aligned}\frac{a}{1-x} &= \frac{a(1+x)}{1-x^2} \\ &= \frac{a(1+x)(1+x^2)}{1-x^4} \\ &= \frac{a(1+x)(1+x^2)\cdots(1+x^{2^{n-1}})}{1-x^{2^n}}.\end{aligned}$$

Program this algorithm in Julia. How large should n be so that the denominator is effectively 1? Note that the powers of x should be computed by repeated squaring.

(2) Use the significand and exponent functions of Julia and devise a better initial value than $x_0 = 1.0$ for the Babylonian method. Explain your choice, and test it on a few examples.

(3)* To find the square root of $c > 0$, consider the iteration scheme $x_{n+1} = f(x_n)$ with

$$f(x) = \frac{c + ax}{a + x},$$

a positive, and $a^2 > c$. Show that

$$f'(x) = \frac{a^2 - c}{(a + x)^2}$$

and that $|f'(x)| < 1$ for all $x \geq 0$. Now explain in detail why $\sqrt{c}$ is fixed point and why x_n converges to $\sqrt{c}$ regardless of the choice of $x_0 \geq 0$. What is the local rate of convergence at the fixed point?

(4) Dedekind's algorithm for extracting $\sqrt{c}$ iterates according to

$$x_{n+1} = \frac{x_n(x_n^2 + 3c)}{3x_n^2 + c}.$$

Program Dedekind's algorithm in Julia. Demonstrate cubic convergence by deriving the identity

$$x_{n+1}^2 - c = \frac{(x_n^2 - c)^3}{(3x_n^2 + c)^2}.$$

Finally, argue that Dedekind's algorithm converges to $\sqrt{c}$ regardless of the initial value $x_0 > 0$.

(5) Find coefficients (a, b, c) where the standard quadratic formula is grossly inaccurate when implemented in single precision. You will have to look up how to represent single precision numbers in Julia.

(6) Why does the product of the two roots of a quadratic equal $\frac{c}{a}$?

(7) Solving a cubic equation $ax^3 + bx^2 + cx + d = 0$ is much more complicated than solving a quadratic. Demonstrate that (a) the substitution $x = y - \frac{b}{3a}$ reduces the cubic to $y^3 + ey + f = 0$ for certain coefficients e and f, (b) the further substitution $y = z - \frac{e}{3z}$ reduces this equation to $z^6 + fz^3 - \frac{e^3}{27}$, and (c) the final substitution $w = z^3$ reduces the equation in z to a quadratic in w, which can be explicitly solved. One can now reverse these substitutions and capture six roots, which collapse in pairs to at most three unique roots. Program your algorithm in Julia, and make sure that it captures complex as well as real roots.

(8) Write a Julia program to find the integers c and d in Bézout's identity

$$\gcd(a, b) = ca + db.$$

(9) The prime number theorem says that the number of primes $\pi(n)$ between 1 and n is asymptotic to $\frac{n}{\ln n}$. Use the Sieve of Eratosthenes to check how quickly the ratio $\frac{\pi(n)\ln(n)}{n}$ tends to 1.

(10) A Pythagorean triple (a, b, c) satisfies $a^2 + b^2 = c^2$. Given an array $\boldsymbol{x}$ of positive integers, write a Julia program to find all Pythagorean triples in $\boldsymbol{x}$. (Hint: Replace the entries of $\boldsymbol{x}$ by their squares and sort the result.)

(11) Show that the perimeter lengths a_n and b_n in Archimedes' algorithm satisfy

$$a_n = m \tan \frac{\pi}{m} \quad \text{and} \quad b_n = m \sin \frac{\pi}{m},$$

where $m = 2 \cdot 2^n$ is the number of sides of the two regular polygons. Use this representation and appropriate trigonometric identities to prove the recurrence relations (1.3) and (1.4).

(12) Based on the trigonometric representations of the previous problem, show that $\frac{1}{3}a_n + \frac{2}{3}b_n$ is a much better approximation to π than either a_n or b_n [**193**]. Check your theoretical conclusions by writing a Julia program that tracks all three approximations to π.

(13) Consider evaluation of the polynomial

$$p(x) \quad = \quad a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

for a given value of x. If one proceeds naively, then it takes $n-1$ multiplications to form the powers $x^k = x \cdot x^{k-1}$ for $2 \leq k \leq n$, n multiplications to multiply each power x^k by its coefficient a_{n-k}, and n additions to sum the resulting terms. This amounts to $3n-1$ operations in all. A more efficient method exploits the fact that $p(x)$ can be expressed as

$$\begin{aligned} p(x) &= x(a_0x^{n-1} + a_1x^{n-2} + \cdots + a_{n-1}) + a_n \\ &= xb_{n-1}(x) + a_n. \end{aligned}$$

Since the polynomial $b_{n-1}(x)$ of degree $n-1$ can be similarly reduced, a complete recursive scheme for evaluating $p(x)$ is given by

$$b_0(x) \quad = \quad a_0, \quad b_k(x) \quad = \quad xb_{k-1}(x) + a_k, \quad k = 1, \ldots, n.$$

This scheme requires only n multiplications and n additions in order to compute $p(x) = b_n(x)$. Program the scheme and extend it to the simultaneous evaluation of the derivative $p'(x)$ of $p(x)$.

(14)* Consider a sequence $x_1, \ldots, x_n$ of n real numbers. After you have computed the sample mean and variance

$$\mu_n \quad = \quad \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \sigma_n^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_n)^2,$$

suppose you are presented with a new observation x_{n+1}. It is possible to adjust the sample mean and variance without revisiting all of the previous observations. Verify theoretically and then code the updates

$$\begin{aligned} \mu_{n+1} &= \frac{1}{n+1}(n\mu_n + x_{n+1}) \\ \sigma_{n+1}^2 &= \frac{n}{n+1}\sigma_n^2 + \frac{1}{n}(x_{n+1} - \mu_{n+1})^2. \end{aligned}$$

(15)* Prove the identity

$$\binom{n}{k} \quad = \quad \frac{n!}{k!(n-k)!}.$$

(16)* Verify the identity

$$\log(\sqrt{1+x^2} - 1) \quad = \quad 2\log(x) - \log(\sqrt{1+x^2} + 1).$$

Evaluate the two sides in single precision for x close to 0. Compared to double precision, which expansion is more accurate? Explain why.

(17)* Show that the Bernoulli number $B_{2n+1} = 0$ for all $n \geq 1$. (Hint: Show that the function $f(x) = \frac{x}{e^x - 1} + \frac{x}{2}$ is even.)

(18)* Based on the result of the previous problem, verify the identities

$$x \coth x = \frac{2x}{e^{2x}-1} + x = \sum_{n=0}^{\infty} \frac{4^n B_{2n}}{(2n)!} x^{2n}.$$

Show that this in turn implies

$$x \cot x = \sum_{n=0}^{\infty} \frac{(-1)^n 4^n B_{2n}}{(2n)!} x^{2n}.$$

The Bernoulli numbers also figure in the Taylor expansion of $\tan x$ [**95**].

(19)* Prove that $\cos(2x) = 2\cos^2 x - 1$. Use this fact to code an algorithm for computing $\cos x$ from its Taylor expansion around 0.

(20)* Show that $\log(y) = 2\sum_{k=0}^{\infty} \frac{1}{2k+1}\left(\frac{y-1}{y+1}\right)^{2k+1}$. Based on this rapidly converging expansion and the identity $\ln(y) = 2\ln(\sqrt{y})$, design, implement, and test an algorithm for computing $\ln(y)$. (Hint: Verify that the series expansion has value 0 at $y = 1$ and derivative y^{-1}.)

(21)* One can approximate e^x by the truncated series $\sum_{i=0}^{n} x^i/i!$ for n small. If $|x|$ is large, truncation can lead to serious errors. If the truncated expansion is accurate for all $|x| \leq c$, then one can exploit the property $e^{x+y} = e^x e^y$ of the exponential function. Thus, if $|x| > c$, take the smallest integer $m \geq 0$ such that $2^{-m}|x| \leq c$ and approximate $e^{2^{-m}x}$ by the truncated series. Applying the multiplicative property, compute e^x by squaring $e^{2^{-m}x}$, squaring the result $e^{2^{-m+1}x}$, squaring the result of this, and so forth, a total of m times. See the reference [**5**] for variations and safeguards of this algorithm. Implement and test the basic algorithm.

(22) Write a fast accurate function to evaluate $\tan x$ exploiting the facts

$$\begin{aligned} \tan(x+\pi) &= \tan x \\ \tan(2x) &= \frac{2\tan x}{1-\tan^2 x} \\ \tan x &\approx x + x^3\left(\frac{1}{3} + \frac{2}{15}x^2 + \frac{17}{315}x^4\right) \text{ as } x \downarrow 0. \end{aligned}$$

(Hint: First reduce x to a point in $[0, \frac{\pi}{2})$.)

(23) If $t = \tan\frac{1}{2}x$, then demonstrate that

$$\begin{aligned} \tan x &= \frac{2t}{1-t^2} \\ \sin x &= \frac{2t}{1+t^2} \\ \cos x &= \frac{1-t^2}{1+t^2}. \end{aligned}$$

Thus, any algorithm for calculating $\tan x$, can be converted into an algorithm for calculating any of the six trignometric functions.

```
C = randn(s, p);
d = randn(s);
(x, u) = constrained_lsq(A, C, b, d, 1.0e10);
println(norm(x - u[1:length(x)])) # compare two methods
println(x) # solution
```

The two methods produce very similar answers for the choice $\gamma = 10^{10}$ assumed in the code.

5.8. Problems

(1) The various ways of computing the product $\boldsymbol{C} = \boldsymbol{AB}$ of two matrices depend on computer architecture and are not equally fast. Time Julia code for computing via
   (a) $\boldsymbol{C} = \boldsymbol{A} * \boldsymbol{B}$
   (b) $\boldsymbol{C}[:, k] = \boldsymbol{A} * \boldsymbol{B}[:, k]$ for all k
   (c) $\boldsymbol{C}[i, :] = \boldsymbol{A}[i : i, :] * \boldsymbol{B}$ for all i
   (d) $\boldsymbol{C}[i : i, k] = \boldsymbol{A}[i : i, :] * \boldsymbol{B}[:, k]$ for all i and k
   (e) $\boldsymbol{C}[i, k] = \sum_j \boldsymbol{A}[i, j] * \boldsymbol{B}[j, k]$ for all i and k

   for $\boldsymbol{A}$ and $\boldsymbol{B}$ large. Here the unusual index notation $i : i$ helps Julia keep variable types straight. What conclusions do you draw from your computer runs?

(2) Verify our contention that it takes about $\frac{2}{3}n^3$ arithmetic operations to form the LU decomposition of an $n \times n$ matrix.

(3) Prove that (a) the product of two upper-triangular matrices is upper triangular, (b) the inverse of an upper-triangular matrix is upper triangular, (c) if the diagonal entries of an upper-triangular matrix are positive, then the diagonal entries of its inverse are positive, and (d) if the diagonal entries of an upper-triangular matrix are unity, then the diagonal entries of its inverse are unity. Similar statements apply to lower-triangular matrices.

(4)* Demonstrate that an orthogonal upper-triangular matrix is diagonal.

(5) Prove that the set of permutation matrices $\boldsymbol{P}$ forms a finite group closed under the formation of products and inverses. How many $n \times n$ permutation matrices exist? Recall that a permutation σ is a one-to-one map of the set $\{1, 2, \ldots, n\}$ onto itself.

(6) The entries of an $n \times n$ tridiagonal matrix $\boldsymbol{A} = (a_{ij})$ satisfy $a_{ij} = 0$ whenever $|i - j| > 1$. Write a Julia function that solves the equation $\boldsymbol{Ax} = \boldsymbol{b}$ by Gaussian elimination in $O(n)$ arithmetic operations.

(7) Find by hand the Cholesky decomposition of the matrix

$$\boldsymbol{A} = \begin{pmatrix} 2 & -2 \\ -2 & 5 \end{pmatrix}.$$

(8) Show that the matrices

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 3 & 2 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 0 & 0 \\ 2 & 0 & \sqrt{13} \end{pmatrix}$$

are both valid Cholesky-like decompositions of the positive semidefinite matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & 17 \end{pmatrix}.$$

(9) Suppose that the matrix $\boldsymbol{A} = (a_{ij})$ is banded in the sense that $a_{ij} = 0$ when $|i-j| > d$. Prove that the Cholesky decomposition $\boldsymbol{L} = (\ell_{ij})$ also satisfies the band condition $\ell_{ij} = 0$ when $|i-j| > d$.

(10) Given the Cholesky decomposition $\boldsymbol{L}$ of a positive definite matrix $\boldsymbol{A}$, describe a simple algorithm for computing $\boldsymbol{A}^{-1}$. Reduce this algorithm to Julia code.

(11) Show that inversion of an arbitrary square matrix $\boldsymbol{B}$ can be reduced to inversion of a positive definite matrix via the identity $\boldsymbol{B}^{-1} = \boldsymbol{B}^*(\boldsymbol{B}\boldsymbol{B}^*)^{-1}$. Since multiplication of two $n \times n$ matrices has computational complexity $O(n^3)$, matrix inversion also has computational complexity $O(n^3)$. For the record, this is definitely not the preferred method of matrix inversion.

(12) Write a Julia function to carry out the Cholesky least squares algorithm sketched in Example 5.2. Your output should include $\hat{\boldsymbol{\beta}}$, the predicted values $\hat{\boldsymbol{y}}$, the residual vector $\boldsymbol{r}$, and the residual sum of squares.

(13) Find the QR decomposition of the matrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & 3 & 3 \\ 1 & 3 & 1 \\ 1 & 1 & 5 \\ 1 & 1 & 3 \end{pmatrix}$$

by the Gram–Schmidt process.

(14) In linear regression find $\hat{\boldsymbol{\beta}}$, the predicted values $\hat{\boldsymbol{y}}$, the residual vector $\boldsymbol{r}$, and the residual sum of squares $\|\boldsymbol{r}\|^2$ in terms of the extended QR decomposition

$$(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{Q}, \boldsymbol{q}) \begin{pmatrix} \boldsymbol{R} & \boldsymbol{r} \\ \boldsymbol{0} & d \end{pmatrix}.$$

(15) If $\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$ is the QR decomposition of $\boldsymbol{X}$, then show that the projection matrix

$$\boldsymbol{X}(\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^* = \boldsymbol{Q}\boldsymbol{Q}^*.$$

Also show that $|\det \boldsymbol{X}| = |\det \boldsymbol{R}|$ when X is square and in general that

$$\det(\boldsymbol{X}^*\boldsymbol{X}) = (\det \boldsymbol{R})^2.$$

(16) Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ be n conjugate vectors for the $n \times n$ positive definite matrix $\boldsymbol{A}$. Describe how you can use the expansion $\boldsymbol{x} = \sum_{i=1}^n c_i \boldsymbol{v}_i$ to solve the linear equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

(17) Suppose that $\boldsymbol{A}$ is an $n \times n$ positive definite matrix and that the nontrivial vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ satisfy

$$\boldsymbol{u}_i^* \boldsymbol{A} \boldsymbol{u}_j = 0 \quad \text{and} \quad \boldsymbol{u}_i^* \boldsymbol{u}_j = 0$$

for all $i \neq j$. Demonstrate that the $\boldsymbol{u}_i$ are eigenvectors of $\boldsymbol{A}$.

(18) Suppose that the $n \times n$ symmetric matrix $\boldsymbol{A}$ satisfies $\boldsymbol{v}^*\boldsymbol{A}\boldsymbol{v} \neq 0$ for all $\boldsymbol{v} \neq \boldsymbol{0}$ and that $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$ is a basis of $\mathbb{R}^n$. If one defines $\boldsymbol{v}_1 = \boldsymbol{u}_1$ and inductively

$$\boldsymbol{v}_k = \boldsymbol{u}_k - \sum_{j=1}^{k-1} \frac{\boldsymbol{u}_k^* \boldsymbol{A} \boldsymbol{v}_j}{\boldsymbol{v}_j^* \boldsymbol{A} \boldsymbol{v}_j} \boldsymbol{v}_j$$

for $k = 2, \ldots, n$, then show that the vectors $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ are conjugate and provide a basis of $\mathbb{R}^n$. Note that $\boldsymbol{A}$ need not be positive definite.

(19)* Prove the second expression for s_i in equation (5.9).

(20) Consider matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{C}$ of dimensions $m \times n$, $n \times p$, and $p \times q$, respectively. The product $\boldsymbol{ABC}$ can be formed by two multiplications via either $(\boldsymbol{AB})\boldsymbol{C}$ or $\boldsymbol{A}(\boldsymbol{BC})$. From the standpoint of computational efficiency, show that the first order is preferred to the second order if and only if $\frac{1}{q} + \frac{1}{n} < \frac{1}{m} + \frac{1}{p}$.

(21)* Demonstrate that $\det(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*) = 1 + \boldsymbol{v}^*\boldsymbol{u}$ for $\boldsymbol{u}$ and $\boldsymbol{v}$ column vectors of the same length.

(22)* For $\rho > 0$ prove the matrix identity

$$(\boldsymbol{A}^*\boldsymbol{A} + \rho\boldsymbol{I})^{-1}\boldsymbol{A}^* \;=\; \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^* + \rho\boldsymbol{I})^{-1}.$$

(23) For an arbitrary matrix $\boldsymbol{A}$ verify the inequalities

$$1 \;\leq\; \frac{\|\boldsymbol{A}\|_F^2}{\|\boldsymbol{A}\|^2} \;\leq\; \text{rank}(\boldsymbol{A})$$

pertinent to the Frobenius and spectral norms.

(24) Let $\boldsymbol{A}$ be a 2×2 symmetric matrix. Prove that $\boldsymbol{A}$ is positive semidefinite if and only if $\text{tr}(\boldsymbol{A}) \geq 0$ and $\det(\boldsymbol{A}) \geq 0$. Produce a 3×3 symmetric matrix that satisfies these two conditions but fails to be positive semidefinite.

(25) Demonstrate that a strictly diagonally dominant matrix $\boldsymbol{A}$ is nonsingular. (Hint: Express $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ in coordinates.)

(26)* Show that the invertible lower triangular matrix

$$\boldsymbol{L} \;=\; \begin{pmatrix} \boldsymbol{L}_1 & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{L}_2 \end{pmatrix}$$

has inverse

$$\boldsymbol{L}^{-1} \;=\; \begin{pmatrix} \boldsymbol{L}_1^{-1} & \boldsymbol{0} \\ -\boldsymbol{L}_2^{-1}\boldsymbol{C}\boldsymbol{L}_1^{-1} & \boldsymbol{L}_2^{-1} \end{pmatrix}.$$

Implement this in recursive Julia code and compare it to ordinary inversion.

(27)* Consider the linear subspaces S of $\mathbb{R}^p$. The projection operators $P_S(\boldsymbol{x})$ are matrices, and we can define the distance between two such subspaces by $\|P_S - P_T\|$ using any matrix norm. Show that this defines a metric on the linear subspaces. In the case of two nondegenerate lines $S = \{c\boldsymbol{x} : c \in \mathbb{R}\}$ and $T = \{c\boldsymbol{y} : c \in \mathbb{R}\}$, show that $\|P_S - P_T\| = \|\frac{\boldsymbol{x}\boldsymbol{x}^*}{\|\boldsymbol{x}\|^2} - \frac{\boldsymbol{y}\boldsymbol{y}^*}{\|\boldsymbol{y}\|^2}\|$.

(28)* Show that $\|\boldsymbol{x}\|_* = \max_{i_1<i_2<\cdots<i_k}(|x_{i_1}| + |x_{i_2}| + \cdots |x_{i_k}|)$ is a vector norm on $\mathbb{R}^n$ for any integer k between 1 and n.

(29)* Let $f(x)$ be a continuous strictly monotonic function with functional inverse $f^{-1}(x)$. Consider the generalized mean

$$M_f(\boldsymbol{x}) \;=\; f^{-1}\Big[\frac{1}{n}\sum_{i=1}^{n} f(x_i)\Big]$$

of a vector $\boldsymbol{x}$ with n components x_i. Show that $M_f(\boldsymbol{x})$ possesses the following properties:

(a) $M_f(\boldsymbol{x})$ is continuous.
(b) $M_f(\boldsymbol{x})$ is strictly increasing in each entry x_i.
(c) $M_f(\boldsymbol{x}) = M_f(\boldsymbol{P}\boldsymbol{x})$ for every permutation matrix $\boldsymbol{P}$.
(d) $\min_i x_i \leq M_f(\boldsymbol{x}) \leq \max_i x_i$.
(e) $M_f(c\boldsymbol{1}) = c$.
(f) $M_{af+b}(\boldsymbol{x}) = M_f(\boldsymbol{x})$ whenever $a \neq 0$.

Further prove that the choice $f(x) = x^p$ for $x > 0$ leads to the power mean. The special values $p = 1$, $p = -1$, and $p = 2$ correspond to the arithmetic, harmonic, and quadratic means. Finally, show that the function $\sum_{i=1}^n [f(x_i) - f(\mu)]^2$ is minimized by taking $\mu = M_f(\boldsymbol{x})$.

(30)* Suppose you have to minimize the strictly convex quadratic

$$h(\boldsymbol{\beta}) \;=\; \frac{1}{2}\boldsymbol{\beta}^*\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{v}^*\boldsymbol{\beta}.$$

How can you leverage the QR decomposition of $\boldsymbol{C}$ to solve the problem? Note that $\boldsymbol{C}$ is not necessarily square, but it should have full column rank.

(31)* For a square matrix $\boldsymbol{A}$, show that the equation $(d\boldsymbol{I} + \boldsymbol{A})\boldsymbol{x} = \boldsymbol{y}$ has solution $\boldsymbol{x} = d^{-1}\sum_{n=0}^\infty (-1)^n d^{-n}\boldsymbol{A}^n\boldsymbol{y}$ whenever $\|\boldsymbol{A}\|_\sharp < d$ for some matrix norm $\|\cdot\|_\sharp$. Implement this method in code and test.

(32) Show that the pseudo-inverse $(\boldsymbol{A}^*\boldsymbol{A})^{-1}\boldsymbol{A}^*$ of a matrix $\boldsymbol{A}$ with linearly independent columns and QR decomposition $\boldsymbol{Q}\boldsymbol{R}$ reduces to $\boldsymbol{R}^{-1}\boldsymbol{Q}^*$.

(33)* Code and test a version of the conjugate gradient algorithm for least squares. Your code should never form the Gram matrix $\boldsymbol{X}^*\boldsymbol{X}$.

(34) Demonstrate the product $\boldsymbol{A}\boldsymbol{B}$ of two symmetric matrices is symmetric if and only if $\boldsymbol{A}$ and $\boldsymbol{B}$ commute.

(35) Show that the Frobenius inner product $\langle \boldsymbol{M}, \boldsymbol{N}\rangle = \mathrm{tr}(\boldsymbol{M}^*\boldsymbol{N})$ enjoys the properties

$$\begin{aligned}
\langle \boldsymbol{M}, \boldsymbol{N}\rangle &= \langle \boldsymbol{M}^*, \boldsymbol{N}^*\rangle \\
\langle \boldsymbol{M}\boldsymbol{A}, \boldsymbol{N}\rangle &= \langle \boldsymbol{M}, \boldsymbol{N}\boldsymbol{A}^*\rangle \\
\langle \boldsymbol{B}\boldsymbol{M}, \boldsymbol{N}\rangle &= \langle \boldsymbol{M}, \boldsymbol{B}^*\boldsymbol{N}\rangle \\
\langle \boldsymbol{M} \odot \boldsymbol{O}, \boldsymbol{N}\rangle &= \langle \boldsymbol{M}, \boldsymbol{N} \odot \boldsymbol{O}\rangle,
\end{aligned}$$

where $\boldsymbol{A} \odot \boldsymbol{B}$ denotes the denotes entry-wise multiplication (Hadamard product).

(36)* Cyclic coordinate descent algorithms update one parameter at a time in optimization. Consider minimizing the least squares criterion $f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to the bound constraints $l_j \le \beta_j \le u_j$. For instance, in nonnegative regression $l_j = 0$ and $u_j = \infty$. Demonstrate that the cyclic coordinate descent updates β_j by

$$\hat{\beta}_j \;=\; P_{[l_j,u_j]}\Big[\beta_j + \frac{\sum_i (y_i - \sum_k x_{ik}\beta_k)x_{ij}}{\sum_i x_{ij}^2}\Big],$$

where $\boldsymbol{X} = (x_{ij})$ and $P_{[l,u]}(z)$ projects z onto the interval $[l, u]$. Program and text this algorithm.

5.9. Solutions to Selected Problems

5.4 Suppose that $\boldsymbol{M}$ is $p \times p$. The eigenvalues λ_i of $\boldsymbol{M}$ occur on its diagonal. Each eigenvalue must have absolute value 1. Because $\|\boldsymbol{M}\|_F^2 = p = \sum_{i=1}^p |\lambda_i|^2$, all off-diagonal entries of $\boldsymbol{M}$ must be 0.

6.5. Problems

(1) On a one-dimensional problem of your choice, implement Newton's method. Check your result using the *fzero* function of the Julia package Roots.jl.

(2) Write a Julia program to solve Lambert's equation $we^w = x$ by Newton's method for $x > 0$. Prove that the iterates are defined by

$$w_{n+1} = w_n \frac{w_n + \frac{x}{w_n e^{w_n}}}{w_n + 1}.$$

Make the argument that $w_{n+1} > w_n$ when $w_n e^{w_n} < x$ and that $w_{n+1} < w_n$ when $w_n e^{w_n} > x$.

(3) In solving Lambert's equation $we^w = x$ by Newton's method, one must seed the algorithm with an initial guess. Argue that $w_0 = \ln x - \ln(\ln x)$ is good for moderate to large x. Show that it is exact for $x = e$. For small x, argue that the guess $w_0 = \frac{x}{1+cx}$ is good. What value of the constant c solves Lambert's equation when $x = e$? Plot these approximations against the solution curve of Lambert's equation.

(4) Halley's method for finding a root of the equation $f(x) = 0$ approximates $f(x)$ around x_n by the quadratic

$$q(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2} f''(x_n)(x - x_n)^2.$$

It then rearranges the equation $q(x) = 0$ in the form

$$x - x_n = -\frac{f(x_n)}{f'(x_n) + \frac{1}{2} f''(x_n)(x - x_n)}$$

and then approximates $x - x_n$ on the right-hand side by the Newton increment $-f(x_n)/f'(x_n)$. Show that these maneuvers yield the Halley update

$$x_{n+1} = x_n - \frac{2f(x_n)f'(x_n)}{2f'(x_n)^2 - f(x_n)f''(x_n)}.$$

Halley's method has a cubic rate of convergence. Compare its practical performance to Newton's method in solving Lambert's equation of the previous two problems.

(5) Consider the function

$$f(x) = \begin{cases} 0 & \text{if } x = 0 \\ x + x^2 \sin\left(\frac{2}{x}\right) & \text{if } x \neq 0. \end{cases}$$

Calculate its derivative $f'(x)$ for all x, and argue that Newton's method tends to be repelled by its root $x = 0$. This failure occurs despite the fact that $f(x)$ possesses a bounded derivative in a neighborhood of 0.

(6) The function $f(x) = x + x^{4/3}$ has $x = 0$ as a root. Derive the Newton updates

$$x_{n+1} = \frac{\frac{1}{3} x_n^{\frac{4}{3}}}{1 + \frac{4}{3} x_n^{\frac{1}{3}}},$$

and show that

$$\lim_{n \to \infty} \frac{x_{n+1}}{x_n^{\frac{4}{3}}} = \frac{1}{3}$$

for x_0 close to 0. This subquadratic rate of convergence occurs because $f(x)$ is not twice differentiable at 0.

(7) Characterize the behavior of Newton's method in minimizing the function $f(x) = \sqrt{x^2+1}$. When the method converges, what is its order of convergence?

(8) For any positive number x, show that

$$\log_2 x \quad = \quad m \pm \log_2(1+w) \quad = \quad m \pm \frac{\ln(1+w)}{\ln 2} \tag{6.7}$$

for an integer m and a real $w \in [0, \frac{1}{2})$ [**239**]. (Hints: Write $x = 2^n + b$ with $0 \leq b < 2^n$. Then $\log_2 x = n + \log_2(1+r)$ for $r = b2^{-n}$. When $r > \frac{1}{2}$, write

$$n + \log_2(1+r) \quad = \quad n + 1 - \log_2\Big(1 + \frac{1-r}{1+r}\Big).$$

Discuss the relevance of representation (6.7) in extracting $\ln x$.)

(9) For y positive the positive root of the equation $f(x) = \frac{1}{x^2} - y = 0$ is $\frac{1}{\sqrt{y}}$. Show that the Newton iterates for finding the root are

$$x_{n+1} \quad = \quad \frac{x_n(3 - yx_n^2)}{2}.$$

Alternatively, $x = \frac{1}{\sqrt{y}}$ solves the equation $g(x) = yx^2 - 1 = 0$. Demonstrate that this formulation gives rise to the Newton updates

$$x_{n+1} \quad = \quad \frac{1}{2}\Big(x_n + \frac{1}{yx_n}\Big).$$

The first scheme involves no reciprocals, but the second scheme has better convergence guarantees. Show that the second scheme satisfies $x_{n+1} \geq \frac{1}{\sqrt{y}}$ regardless of the value of $x_n > 0$. Also show that $x_{n+1} \leq x_n$ whenever $x_n \geq \frac{1}{\sqrt{y}}$. Hence, global convergence is assured.

(10) The binomial theorem states that for r real and $|x| < 1$

$$(1+x)^r \quad = \quad \sum_{k=0}^{\infty} \binom{r}{k} x^k \quad = \quad \sum_{k=0}^{\infty} \frac{r(r-1)\cdots(r-k+1)}{k!} x^k.$$

One can adapt this recipe to calculate matrix roots by defining

$$(\boldsymbol{I} + \boldsymbol{X})^r \quad = \quad \sum_{k=0}^{\infty} \binom{r}{k} \boldsymbol{X}^k$$

for square matrices $\boldsymbol{X} = (x_{ij})$ with norm $\|\boldsymbol{X}\| < 1$ [**114**]. Here the norm is generic except for the requirement that $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$. The most convenient choice is the Frobenius norm $\|\boldsymbol{X}\|_F^2$. One can extend the series method to matrices $\boldsymbol{Y}$ with larger norms by writing $\boldsymbol{Y}^r = c^r(\boldsymbol{I}+\boldsymbol{X})^r$, where $\boldsymbol{X} = \frac{1}{c}\boldsymbol{Y} - \boldsymbol{I}$ for come constant c. Show that under the Frobenius norm the optimal choice of c satisfies

$$c \quad = \quad \frac{\|\boldsymbol{Y}\|_F^2}{\mathrm{tr}(\boldsymbol{Y})} \quad \text{and} \quad \|\boldsymbol{X}\|_F^2 \quad = \quad \|\boldsymbol{I}\|_F^2 - \frac{\mathrm{tr}(\boldsymbol{Y})^2}{\|\boldsymbol{Y}\|_F^2}.$$

When $\mathrm{tr}(\boldsymbol{Y})$ is negative, the factor c^r may be complex.

(11) For the weighted least squares criterion

$$f(\boldsymbol{\beta}) \quad = \quad \frac{1}{2}\sum_{i=1}^{n} w_i\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2,$$

prove that the minimum is achieved when $\boldsymbol{\beta} = (\boldsymbol{X}^*\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{W}\boldsymbol{y}$, where $\boldsymbol{W}$ is a diagonal matrix with ith diagonal entry $w_i > 0$.

(12) Demonstrate Woodbury's generalization

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^*)^{-1} \quad = \quad \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1}+\boldsymbol{V}^*\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^*\boldsymbol{A}^{-1}$$

of the Sherman–Morrison matrix inversion formula for compatible matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{U}$, and $\boldsymbol{V}$.

(13) To explore whether the performance of Newton's method can be improved by a linear change of variables, consider solving the two problems $f(\boldsymbol{x}) = \boldsymbol{0}$ and $\boldsymbol{A}f(\boldsymbol{B}\boldsymbol{x}) = \boldsymbol{0}$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are invertible matrices of the right dimensions. Show that the two methods lead to basically the same iterates when started at $\boldsymbol{x}_0$ and $\boldsymbol{B}^{-1}\boldsymbol{x}_0$, respectively.

(14) To solve the vector-valued equation $g(\boldsymbol{x}) = \boldsymbol{0}$, one can minimize the function $f(\boldsymbol{x}) = \frac{1}{2}\|g(\boldsymbol{x})\|^2$. Show that $f(\boldsymbol{x})$ has gradient

$$\nabla f(\boldsymbol{x}) \quad = \quad dg(\boldsymbol{x})^* g(\boldsymbol{x})$$

and second differential

$$d^2 f(\boldsymbol{x}) \quad = \quad dg(\boldsymbol{x})^* dg(\boldsymbol{x}) + d^2 g(\boldsymbol{x})^* g(\boldsymbol{x}) \quad \approx \quad dg(\boldsymbol{x})^* dg(\boldsymbol{x}).$$

The approximation to the second differential is good when $\boldsymbol{x}$ is close to a root. When the number of components of $g(\boldsymbol{x})$ equals the dimension of $\boldsymbol{x}$, prove that the combination of this approximation and Newton's method of minimization leads to the standard Newton update

$$\boldsymbol{x}_{n+1} \quad = \quad \boldsymbol{x}_n - dg(\boldsymbol{x}_n)^{-1} g(\boldsymbol{x}_n).$$

Finally, argue in this case that the Newton increment is a descent direction for the objective function $f(\boldsymbol{x})$

(15)* If you apply Newton's method to solve $f(y) = \frac{1}{y^{1/2}} - xy^{3/2} = 0$ for $x > 0$, then show that the iterates are

$$y_{n+1} \quad = \quad y_n \left(\frac{3 + x y_n^2}{1 + 3x y_n^2} \right).$$

What is the root of the equation $f(y) = 0$? Observe that the quantity $c = xy_n^2$ should be calculated only once per iteration. Show that the factor $\frac{3+xy_n^2}{1+3xy_n^2} < 1$ if and only if $\frac{1}{\sqrt{x}} < y_n$, and $\frac{3+xy_n^2}{1+3xy_n^2} > 1$ if and only if $\frac{1}{\sqrt{x}} > y_n$. Why is this desirable? Finally, show that

$$y_{n+1} - \frac{1}{\sqrt{x}} \quad = \quad \frac{x(y_n - \frac{1}{\sqrt{x}})^3}{1 + 3x y_n^2}.$$

Now argue that the algorithm converges globally on $(0, \infty)$ and locally at a cubic rate.

(16)* If $f(x)$ has a root of multiplicity m at y, then $f(x)$ can be expressed in the form $f(x) = (x-y)^m g(x)$ with $g(y) \neq 0$. Show that the function $h(x) = \frac{f(x)}{f'(x)}$ has a simple root (multiplicity 1) at y and that Newton's method for finding a root of $h(x)$ iterates according to

$$x_{n+1} \quad = \quad x_n - \frac{f(x_n) f'(x_n)}{f'(x_n)^2 - f(x_n) f''(x_n)}$$

(17)* Suppose the real-valued $f(x)$ satisfies $f'(x) > 0$ and $f''(x) > 0$ for all x in its domain (d, ∞). If $f(x) = 0$ has a root r, then demonstrate that r is unique and that Newton's method converges to r regardless of its starting point. Further, prove that x_n converges monotonically to r from above when $x_0 > r$ and that $x_1 > r$ when $x_0 < r$.

(18)* Consider the map

$$f(\boldsymbol{x}) \;=\; \begin{pmatrix} x_1^2 + x_2^2 - 2 \\ x_1 - x_2 \end{pmatrix}$$

of the plane into itself. Show that $f(\boldsymbol{x}) = \boldsymbol{0}$ has the roots $-\boldsymbol{1}$ and $\boldsymbol{1}$ and no other roots. Prove that Newton's method iterates according to

$$x_{n+1,1} \;=\; x_{n+1,2} \;=\; \frac{x_{n1}^2 + x_{n2}^2 + 2}{2(x_{n1} + x_{n2})}$$

and that these iterates converge to the root $-\boldsymbol{1}$ if $x_{01} + x_{02}$ is negative and to the root $\boldsymbol{1}$ if $x_{01} + x_{02}$ is positive. If $x_{01} + x_{02} = 0$, then the first iterate is undefined. Finally, prove that

$$\lim_{n\to\infty} \frac{|x_{n+1,1} - y_1|}{|x_{n1} - y_1|^2} \;=\; \lim_{n\to\infty} \frac{|x_{n+1,2} - y_2|}{|x_{n2} - y_2|^2} \;=\; \frac{1}{2},$$

where $\boldsymbol{y}$ is the root relevant to the initial point $\boldsymbol{x}_0$.

(19)* Consider the quadratic function $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^*\boldsymbol{x}$, where $\boldsymbol{A}$ is symmetric with a negative eigenvalue. Show that $f(\boldsymbol{x})$ is unbounded below.

(20) Write, document, and test code for the Gauss-Newton algorithm. Include step halving and a convergence test. Apply the code to a typical nonlinear regression problem from the literature.

(21)* Show that the equation $y \ln y = ay + b$ with $b > 0$ and the equation $y = a + b\, e^{cy}$ with $b > 0$ and $c < 0$ can be solved in terms of Lambert's function introduced in Problems 2 and 3.

(22)* Suppose we write Lambert's equation $we^w = x$ as

$$f(w) \;=\; \log w + w - \log x = 0.$$

Show that Newton's method for $w > 0$ has the iterates

$$w_{n+1} \;=\; w_n \frac{1 - \log w_n + \log x}{w_n + 1}.$$

If the initial point $w_0 > 0$ satisfies $w_0 e^{w_0} < x$, then prove that the Newton iterates converge monotonically to the solution from below. (Hint: It might help to sketch the curve $f(w)$.)

(23)* Consider the problem of computing e^x given a fast algorithm for taking logarithms. Why can we restrict x to positive values? Show Newton's method iterates according to $y_{n+1} = y_n - y_n(\ln y_n - x)$. Also show that if $y_n > e^x$, then $y_{n+1} < e^x$, and if $y_n < e^x$, then $y_{n+1} < e^x$ as well. In the latter case prove that y_{n+1} is closer to e^x than y_n is. Prove that convergence is guaranteed whenever $y_0 \in (0, e^{1+x})$. (Hints: Apply the mean value theorem to $y_{n+1} - e^x$. Under what condition does $y_{n+1} > 0$?)

(24)* Let $x_1, \ldots, x_m$ be a random sample from the gamma density

$$f(x) \;=\; \Gamma(\alpha)^{-1}\beta^\alpha x^{\alpha-1}e^{-\beta x}$$

on $(0, \infty)$. Let $\overline{x} = \frac{1}{m}\sum_{i=1}^m x_i$ and $\overline{\ln x} = \frac{1}{m}\sum_{i=1}^m \ln x_i$. Setting the score function equal to 0 identifies $\beta = \alpha/\overline{x}$ as the maximum of the loglikelihood

$L(\alpha,\beta)$ of the sample for α fixed. Substituting this value of β in the loglikelihood reduces maximum likelihood estimation to optimization of the profile loglikelihood

$$L(\alpha) = m\alpha\ln\alpha - m\alpha\ln\overline{x} - m\ln\Gamma(\alpha) + m(\alpha-1)\overline{\ln x} - m\alpha.$$

As an alternative to Newton's method [**178**], we approximate $L(\alpha)$ by the surrogate function $g(\alpha) = c_0 + c_1\alpha + c_2\ln\alpha$. (No majorization is implied.) The coefficients here are generated by matching the derivatives $g^{(k)}(\alpha_n)$ and $L^{(k)}(\alpha_n)$ at the current iterate α_n for $k = 0, 1, 2$. Show that maximizing the surrogate leads to the update

$$\frac{1}{\alpha_{n+1}} = \frac{1}{\alpha_n} + \frac{\overline{\ln x} - \ln\overline{x} + \ln\alpha_n - \Psi(\alpha_n)}{\alpha_n - \alpha_n^2\Psi'(\alpha_n)},$$

where $\Psi(\alpha)$ is the digamma function (derivative of the log gamma function). Convergence is quick with a good starting value. (Hint: First maximize the surrogate. Note that the coefficient c_0 is irrelevant so it suffices to match first and second derivatives.)

6.6. Solutions to Selected Problems

6.15 Given $f(y) = \frac{1}{y^{1/2}} - xy^{3/2} = 0$, it follows that $f'(y) = -\frac{1}{2y^{3/2}} - \frac{3}{2}xy^{1/2}$. Hence, Newton's method gives

$$\begin{aligned} y_{n+1} &= y_n + \frac{\frac{1}{y_n^{1/2}} - xy_n^{3/2}}{\frac{1}{2y_n^{3/2}} + \frac{3}{2}xy_n^{1/2}} \\ &= y_n + \frac{2y_n - 2xy_n^3}{1 + 3xy_n^2} \\ &= y_n\left(\frac{3 + xy_n^2}{1 + 3xy_n^2}\right). \end{aligned}$$

Furthermore, $f(y) = 0$ if and only if $1 = xy^2$, so $y = x^{-1/2}$ is the solution. Now $\frac{3+xy_n^2}{1+3xy_n^2} > 1$ if and only if $3 + xy_n^2 > 1 + 3xy_n^2$, or $1 > xy_n^2$, which is equivalent to $y_n < \frac{1}{\sqrt{x}}$. As a consequence, we have $y_{n+1} > y_n$ when $y_n < \frac{1}{\sqrt{x}}$, and similarly $y_{n+1} < y_n$ when $y_n > \frac{1}{\sqrt{x}}$. The equalities

$$\begin{aligned} y_{n+1} - \frac{1}{\sqrt{x}} &= y_n\left(\frac{3 + xy_n^2}{1 + 3xy_n^2}\right) - \frac{1}{\sqrt{x}} \\ &= \frac{\sqrt{x}(3y_n + xy_n^3) - 1 - 3xy_n^2}{\sqrt{x}(1 + 3xy_n^2)} \\ &= \frac{x(y_n - \frac{1}{\sqrt{x}})^3}{1 + 3xy_n^2} \end{aligned}$$

demonstrate that $y_n - x^{-1/2}$ maintains constant sign and that the algorithm converges locally at a cubic rate. Overall, the algorithm converges globally because the iterates converge monotonically to the fixed point.

6.16 The first claim follows from the representation

$$h(x) = \frac{f(x)}{f'(x)} = \frac{(x-y)g(x)}{mg(x) + (x-y)g'(x)},$$

7.6. Problems

(1) Consider the linear program of minimizing $x_1 + x_2$ subject to the constraints $x_1 + 2x_2 \geq 3$, $2x_1 + x_2 \geq 5$, and $x_2 \geq 0$. Graph the feasible region, and solve the program by hand or by our Julia code.

(2) Consider the linear program of maximizing $x_1 + 2x_2 + 3x_3 + 4x_4 + 5$ subject to the constraints

$$\begin{aligned} 4x_1 + 3x_2 + 2x_3 + x_4 &\leq 10 \\ x_1 - x_3 + 2x_4 &= 2 \\ x_1 + x_2 + x_3 + x_4 &\geq 1 \end{aligned}$$

and $x_1 \geq 0$, $x_3 \geq 0$, $x_4 \geq 0$. Put this program into canonical form and solve.

(3) Convert the problem of minimizing $|x_1 + x_2 + x_3|$ subject to $x_1 - x_2 = 5$, $x_2 - x_3 = 7$, and $x_1 \geq 0$, and $x_3 \geq 2$ into a linear program and solve.

(4)* The simplest ℓ_1 regression problem consists of minimizing the function

$$f(\mu) = \sum_{i=1}^{n} |x_i - \mu|.$$

By taking one-sided derivatives, prove that a sample median of $x_1, \ldots, x_n$ solves the problem.

(5) Find an upper bound on the number of basic feasible points of a linear program.

(6) A set C is said to be convex if whenever $\boldsymbol{u}$ and $\boldsymbol{v}$ belong to C, then the entire line segment $[\boldsymbol{u}, \boldsymbol{v}] = \{t\boldsymbol{u} + (1-t)\boldsymbol{v} : t \in [0,1]\}$ belongs to C. Show that the feasible region of a linear program is convex. A set C is said to be closed if whenever a sequence $\boldsymbol{x}_n$ from C converges to a limit $\boldsymbol{x}$, then $\boldsymbol{x}$ also belongs to C. Show that the feasible region of a linear program is closed.

(7) Give an example of a linear program whose feasible region R is unbounded. Construct an objective $\boldsymbol{c}^*\boldsymbol{x}$ that is bounded below on R and an objective $\boldsymbol{c}^*\boldsymbol{x}$ that is unbounded below on R.

(8) A point $\boldsymbol{x}$ of a convex set C is called extreme if it cannot be expressed as a nontrivial convex combination $\boldsymbol{x} = t\boldsymbol{y} + (1-t)\boldsymbol{z}$ of two distinct points $\boldsymbol{y}$ and $\boldsymbol{z}$ from C. Here nontrivial means that t occurs in the open interval $(0,1)$. Prove that a point $\boldsymbol{x}$ in the feasible region $\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, ; \boldsymbol{x} \geq \boldsymbol{0}\}$ is extreme if and only if the columns $\boldsymbol{A}_B$ of $\boldsymbol{A}$ associated with its support set $B = \{i : x_i > 0\}$ are linearly independent.

(9)* The dual function of a linear program is defined by

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}),$$

where $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is the Lagrangian (7.1). Prove that the dual equals

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{cases} -\boldsymbol{b}^*\boldsymbol{\lambda} & \boldsymbol{c} - \boldsymbol{A}^*\boldsymbol{\lambda} - \boldsymbol{\mu} = \boldsymbol{0} \\ -\infty & \text{otherwise.} \end{cases}$$

(10)* In the notation of the previous problem, demonstrate the duality result

$$\max_{\{(\boldsymbol{\lambda}, \boldsymbol{\mu}) : \boldsymbol{\mu} \geq \boldsymbol{0}\}} \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}} \boldsymbol{c}^*\boldsymbol{x}$$

when the linear program has a solution. (Hints: For $\boldsymbol{x}$ feasible show that

$$\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \boldsymbol{c}^*\boldsymbol{x}.$$

At the constrained minimum $\boldsymbol{y}$ with multipliers $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\mu}}$, also show

$$\boldsymbol{c}^*\boldsymbol{y} \;=\; \mathcal{L}(\boldsymbol{y},\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\mu}}) \;=\; \mathcal{D}(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\mu}}).\ )$$

(11) Let $\boldsymbol{A}$ be a matrix with full row rank and

$$S \;=\; \{\boldsymbol{x}\in\mathbb{R}^p : \boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}\}$$

be an affine subspace. Show that the closest point to $\boldsymbol{y}$ in S is

$$P_S(\boldsymbol{y}) \;=\; \boldsymbol{y}-\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}(\boldsymbol{A}\boldsymbol{y}-\boldsymbol{b})$$

by minimizing the function $f(\boldsymbol{x})=\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$ subject to the constraint $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$. The matrix $\boldsymbol{P}=\boldsymbol{I}-\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}\boldsymbol{A}$ is an orthogonal projection. Check the properties $\boldsymbol{P}^2=\boldsymbol{P}$, $\boldsymbol{P}^*=\boldsymbol{P}$, and $\boldsymbol{P}\boldsymbol{x}=\boldsymbol{x}$ for $\boldsymbol{x}\in S$.

(12) Design and implement appropriate numerical examples to compare the performance of the revised simplex method and Karmarkar's algorithm.

(13) As an alternative to the revised simplex method and Karmarkar's algorithm, one can derive a path-following algorithm to solve the linear programming problem [**45**]. Consider minimizing $\boldsymbol{c}^*\boldsymbol{x}$ subject to the standard linear programming constraints $\boldsymbol{x}\geq\boldsymbol{0}$ and $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$. Given an initial feasible point $\boldsymbol{x}_0>\boldsymbol{0}$, one can devise a differential equation $\frac{d}{dt}\boldsymbol{x}(t)=G(\boldsymbol{x})$ whose solution $\boldsymbol{x}(t)$ is likely to converge to the optimal point. Simply take $G(\boldsymbol{x})=\boldsymbol{D}(\boldsymbol{x})\boldsymbol{P}(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x})$, where $\boldsymbol{D}(\boldsymbol{x})=\operatorname{diag}(\boldsymbol{x})$ is a diagonal matrix with diagonal entries given by the vector $\boldsymbol{x}$ and $\boldsymbol{P}(\boldsymbol{x})$ is orthogonal projection onto the null space of $\boldsymbol{A}\boldsymbol{D}(\boldsymbol{x})$. (See the previous problem.) The matrix $\boldsymbol{D}(\boldsymbol{x})$ slows the trajectory down as it approaches a boundary $x_i=0$. The matrix $\boldsymbol{P}(\boldsymbol{x})$ ensures that the value of $\boldsymbol{A}\boldsymbol{x}(t)$ remains fixed at the constant $\boldsymbol{b}$. Check this fact. Show that the choice $\boldsymbol{v}(\boldsymbol{x})=-\boldsymbol{P}(\boldsymbol{x})\boldsymbol{D}(\boldsymbol{x})\boldsymbol{c}$ yields $\frac{d}{dt}\boldsymbol{c}^*\boldsymbol{x}(t)\leq 0$. In other words, $\boldsymbol{c}^*\boldsymbol{x}(t)$ is a Liapunov function for the solution path.

(14) Reduce the path-following algorithm of the previous problem to Julia code. This can be accomplished by adopting Euler's method for solving the differential equation $\frac{d}{dt}\boldsymbol{x}(t)=G(\boldsymbol{x})$. In Euler's method we approximate

$$\boldsymbol{x}(t+\delta)\approx\boldsymbol{x}(t)+\delta\frac{d}{dt}\boldsymbol{x}(t)=\boldsymbol{x}(t)+\delta G(\boldsymbol{x})$$

for $\delta>0$ small. Note that Julia has a simple command to compute the pseudoinverse $\boldsymbol{M}^*(\boldsymbol{M}\boldsymbol{M}^*)^{-1}$ of a matrix $\boldsymbol{M}$ with full row rank. Your code must start with an $\boldsymbol{x}$ having all entries positive and satisfying $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$.

(15)* Rigorously prove Proposition 7.5.1.

(16) As a generalization of the Dinkelbach maneuver, consider $h(\boldsymbol{x})=\min_{i=1}^{r}\frac{f_i(\boldsymbol{x})}{g_i(\boldsymbol{x})}$ with all denominators $g_i(\boldsymbol{x})$ positive. To reduce $h(\boldsymbol{x})$ by iteration, let S be the index set $\operatorname{argmin}_i\frac{f_i(\boldsymbol{x}_n)}{g_i(\boldsymbol{x}_n)}$, and define for some $k\in S$

$$\boldsymbol{x}_{n+1} \;=\; \operatorname{argmin}_{\boldsymbol{x}}\,[f_k(\boldsymbol{x})-h(\boldsymbol{x}_n)g_k(\boldsymbol{x})].$$

Show that $h(\boldsymbol{x}_{n+1})\leq h(\boldsymbol{x}_n)$ and that strict inequality can be guaranteed unless $\frac{f_k(\boldsymbol{x}_{n+1})}{g_k(\boldsymbol{x}_{n+1})}=\frac{f_k(\boldsymbol{x}_n)}{g_k(\boldsymbol{x}_n)}$.

(17)* Given points $\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n$ in $\mathbb{R}^p$, design a linear program to decide whether another point $\boldsymbol{y}$ belongs to their convex hull.

(18)* Show that the function $f(\boldsymbol{x})=|3x_1+4x_2|+|2x_1+x_2|$ is convex and attains its minimum of 0 at $\boldsymbol{0}$. Furthermore, show that the slices $f(x_1,3)$ and $f(-4,x_2)$

achieve their minimal value of 5 at $x_1 = -4$ and $x_2 = 3$. Why do these facts not contradict Fermat's principle?

(19)* Suppose $f(\boldsymbol{x})$ is convex on $\mathbb{R}^p$ and achieves its minimum at $\boldsymbol{y}$. If in addition $f(\boldsymbol{x})$ is symmetric in the sense that $f(\boldsymbol{P}\boldsymbol{x}) = f(\boldsymbol{x})$ for all permutation matrices $\boldsymbol{P}$, then demonstrate that $f(\boldsymbol{x})$ possesses a symmetric minimum point $\boldsymbol{z}$ with $\boldsymbol{P}\boldsymbol{z} = \boldsymbol{z}$ for all $\boldsymbol{P}$. In other words, all components of $\boldsymbol{z}$ are equal.

7.7. Solutions to Selected Problems

7.4 The one-sided derivatives of $f(\mu)$ are

$$\begin{aligned} f'_+(\mu) &= \sum_{x_i<\mu} 1 - \sum_{x_i \geq \mu} 1 \\ f'_-(\mu) &= \sum_{x_i \leq \mu} 1 - \sum_{x_i > \mu} 1. \end{aligned}$$

The minimum is achieved when $f'_-(\mu) \leq 0$ and $f'_+(\mu) \geq 0$. These are precisely the conditions characterizing a median.

7.9 The Lagrangian is

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \boldsymbol{c}^*\boldsymbol{x} - \boldsymbol{\lambda}^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\mu}^*\boldsymbol{x} \\ &= \boldsymbol{\lambda}^*\boldsymbol{b} + (\boldsymbol{c} - \boldsymbol{A}^*\boldsymbol{\lambda} - \boldsymbol{\mu})^*\boldsymbol{x} \end{aligned}$$

If we send the components of $\boldsymbol{x}$ separately to $\pm\infty$, then it clear that the dual $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ equals $-\infty$ unless

$$\boldsymbol{c} - \boldsymbol{A}^*\boldsymbol{\lambda} - \boldsymbol{\mu} = \boldsymbol{0}. \tag{7.4}$$

If this condition holds, then $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{\lambda}^*\boldsymbol{b}$.

7.10 The dual function $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu})$ equals $-\infty$ unless equation (B.1) is valid, in which case it equals $\boldsymbol{\lambda}^*\boldsymbol{b}$. At a feasible point $\boldsymbol{x}$, $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and $\boldsymbol{\mu}^*\boldsymbol{x} \geq 0$. Therefore,

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{c}^*\boldsymbol{x} - \boldsymbol{\lambda}^*(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\mu}^*\boldsymbol{x} \leq \boldsymbol{c}^*\boldsymbol{x}$$

for all $\boldsymbol{\lambda}$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$. In particular, if $\boldsymbol{y}$ is the solution of the primal problem, then we have $\mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \boldsymbol{c}^*\boldsymbol{y}$. This inequality continues to hold if we now maximize over $\boldsymbol{\lambda}$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$. The inequality $\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \leq \boldsymbol{c}^*\boldsymbol{y}$ is called weak duality. In the primal problem, the Lagrange multiplier condition is

$$\boldsymbol{0} = \boldsymbol{c} - \boldsymbol{A}^*\hat{\boldsymbol{\lambda}} - \hat{\boldsymbol{\mu}},$$

which is precisely equation (B.1). Owing to feasibility and complementary slackness, the equalities

$$\boldsymbol{c}^*\boldsymbol{y} = \mathcal{L}(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}) = \boldsymbol{c}^*\boldsymbol{y} - \hat{\boldsymbol{\lambda}}^*(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{b}) - \hat{\boldsymbol{\mu}}^*\boldsymbol{y} = \hat{\boldsymbol{\lambda}}^*\boldsymbol{b}$$

hold at a solution $\boldsymbol{y}$ of the primal problem . Hence,

$$\max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \mathcal{D}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \geq \hat{\boldsymbol{\lambda}}^*\boldsymbol{b} = \boldsymbol{c}^*\boldsymbol{y}.$$

This inequality is called strong duality. Together with weak duality it implies the claim of the problem.

7.15 Suppose $\boldsymbol{x} \in \mathbb{R}^n$ is a boundary point. Without loss of generality take $x_n = 0$. An optimal point subject to this condition minimizes

$$f(\boldsymbol{x}) = \left(\frac{1}{n} - 0\right)^2 + \sum_{i=1}^{n-1} \left(\frac{1}{n} - x_i\right)^2$$

8.10. Problems

(1) Describe the behavior of the power method applied to the matrix

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Explain your empirical findings by invoking the *eigen* command of Julia to reveal $\boldsymbol{A}$'s eigenvalues and eigenvectors.

(2) Find the eigenvalues and eigenvectors of the matrix

$$\boldsymbol{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}$$

of RJ Wilson by divide and conquer and Jacobi's method.

(3) Find the eigenvalues and eigenvectors of the rotation matrix

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Note that the eigenvalues are complex conjugates.

(4) Find the eigenvalues and eigenvectors of the reflection matrix

$$\begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

(5) Suppose λ is an eigenvalue of the orthogonal matrix $\boldsymbol{O}$ with corresponding eigenvector $\boldsymbol{v}$. Show that $\boldsymbol{v}$ has real entries only if $\lambda = \pm 1$.

(6) A matrix $\boldsymbol{A}$ with real entries is said to be skew-symmetric if $\boldsymbol{A}^* = -\boldsymbol{A}$. Show that all eigenvalues of a skew-symmetric matrix are imaginary or 0.

(7) Continuing the previous problem, suppose that $\boldsymbol{A}$ is skew-symmetric. Prove that $\boldsymbol{I} - \boldsymbol{A}$ is invertible and that $(\boldsymbol{I} - \boldsymbol{A})^{-1}(\boldsymbol{I} + \boldsymbol{A})$ is orthogonal. The latter matrix is called the Cayley transform of $\boldsymbol{A}$.

(8) Consider an $n \times n$ upper triangular matrix $\boldsymbol{U}$ with distinct nonzero diagonal entries. Let λ be its mth diagonal entry, and write

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_{11} & \boldsymbol{U}_{12} & \boldsymbol{U}_{13} \\ \boldsymbol{0} & \lambda & \boldsymbol{U}_{23} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{U}_{33} \end{pmatrix}$$

in block form. Verify that λ is an eigenvalue of $\boldsymbol{U}$ with eigenvector

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{v} \\ -1 \\ \boldsymbol{0} \end{pmatrix}, \quad \boldsymbol{v} = (\boldsymbol{U}_{11} - \lambda \boldsymbol{I}_{m-1})^{-1} \boldsymbol{U}_{12},$$

where $\boldsymbol{I}_{m-1}$ is the $(m-1) \times (m-1)$ identity matrix.

(9) Suppose the $m \times m$ symmetric matrix $\boldsymbol{A}$ has eigenvalues

$$\lambda_1 < \lambda_2 \leq \cdots \leq \lambda_{m-1} < \lambda_m.$$

The iterative scheme $\boldsymbol{x}_{n+1} = (\boldsymbol{A} - \eta_n \boldsymbol{I})\boldsymbol{x}_n$ with normalization after each step can be used to approximate either λ_1 or λ_m [**113**]. Consider the criterion

$$\sigma_n \quad = \quad \frac{\boldsymbol{x}_{n+1}^* \boldsymbol{A} \boldsymbol{x}_{n+1}}{\boldsymbol{x}_{n+1}^* \boldsymbol{x}_{n+1}}.$$

Choosing η_n to maximize σ_n causes $\lim_{n\to\infty} \sigma_n = \lambda_m$, while choosing η_n to minimize σ_n causes $\lim_{n\to\infty} \sigma_n = \lambda_1$. If $\tau_k = \boldsymbol{x}_n^* \boldsymbol{A}^k \boldsymbol{x}_n$, then show that the extrema of σ_n as a function of η are given by the roots of the quadratic equation

$$0 \quad = \quad \det \begin{pmatrix} 1 & \eta & \eta^2 \\ \tau_0 & \tau_1 & \tau_2 \\ \tau_1 & \tau_2 & \tau_3 \end{pmatrix}.$$

(10) Apply the algorithm of the previous problem to find the largest and smallest eigenvalues of the matrix in problem 2.

(11)* Show that

$$\boldsymbol{v}^*(\boldsymbol{I} + \boldsymbol{F}\boldsymbol{F}^*)^{-1}\boldsymbol{v} \quad \leq \quad \|\boldsymbol{v}\|^2$$

for all $\boldsymbol{v}$. (Hint: Apply the inequality $(1+x)^{-1} \leq 1$ for $x \geq 0$.)

(12) Denote the smallest and largest eigenvalues of an $m \times m$ symmetric matrix $\boldsymbol{C}$ by $\lambda_1[\boldsymbol{C}]$ and $\lambda_m[\boldsymbol{C}]$. For any two $m \times m$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and any $\alpha \in [0,1]$, demonstrate that

$$\begin{aligned} \lambda_1[\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}] &\geq \alpha\lambda_1[\boldsymbol{A}] + (1-\alpha)\lambda_1[\boldsymbol{B}] \\ \lambda_m[\alpha\boldsymbol{A} + (1-\alpha)\boldsymbol{B}] &\leq \alpha\lambda_m[\boldsymbol{A}] + (1-\alpha)\lambda_m[\boldsymbol{B}]. \end{aligned}$$

(13) Given the assumptions of the previous problem, show that the smallest and largest eigenvalues satisfy

$$\begin{aligned} \lambda_1[\boldsymbol{A} + \boldsymbol{B}] &\geq \lambda_1[\boldsymbol{A}] + \lambda_1[\boldsymbol{B}] \\ \lambda_m[\boldsymbol{A} + \boldsymbol{B}] &\leq \lambda_m[\boldsymbol{A}] + \lambda_m[\boldsymbol{B}]. \end{aligned}$$

(14) For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, define $\boldsymbol{A} \trianglerighteq \boldsymbol{0}$ to mean that $\boldsymbol{A}$ is positive semidefinite and $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$ to mean that $\boldsymbol{A} - \boldsymbol{B} \trianglerighteq \boldsymbol{0}$. Show that $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$ and $\boldsymbol{B} \trianglerighteq \boldsymbol{C}$ imply $\boldsymbol{A} \trianglerighteq \boldsymbol{C}$. Also show that $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$ and $\boldsymbol{B} \trianglerighteq \boldsymbol{A}$ imply $\boldsymbol{A} = \boldsymbol{B}$. Thus, $\trianglerighteq$ induces a partial order on the set of symmetric matrices.

(15) In the notation of the previous problem, show that two positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$ if and only if they satisfy $\boldsymbol{B}^{-1} \trianglerighteq \boldsymbol{A}^{-1}$. If $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$, then prove that $\det \boldsymbol{A} \geq \det \boldsymbol{B}$ and $\operatorname{tr} \boldsymbol{A} \geq \operatorname{tr} \boldsymbol{B}$.

(16) Suppose the symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfy $\boldsymbol{A} \trianglerighteq \boldsymbol{B}$ in the notation of the previous two problems. If in addition $\operatorname{tr}(\boldsymbol{A}) = \operatorname{tr}(\boldsymbol{B})$, then demonstrate that $\boldsymbol{A} = \boldsymbol{B}$. (Hint: Consider the matrix $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{B}$.)

(17) Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be positive semidefinite matrices of the same dimension. Prove that the matrix $a\boldsymbol{A} + b\boldsymbol{B}$ is positive semidefinite for every pair of nonnegative scalars a and b. Thus, the set of positive semidefinite matrices is a convex cone.

(18) One of the simplest ways of showing that a symmetric matrix is positive semidefinite is to show that it is the covariance matrix of a random vector. Use this insight to prove that if the symmetric matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are positive semidefinite, then the matrix $\boldsymbol{C} = (c_{ij})$ with entries $c_{ij} = a_{ij}b_{ij}$ is also positive semidefinite [**190**]. (Hint: Take independent random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ with covariance matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and form the random vector $\boldsymbol{z}$ with components $z_i = x_i y_i$.)

(19) Continuing the previous problem, suppose that the $n \times n$ symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ have entries $a_{ij} = i(n-j+1)$ and $b_{ij} = \sum_{k=1}^{i} \sigma_k^2$ for $j \geq i$ and all $\sigma_k^2 \geq 0$. Show that $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite [**190**]. (Hint: For $\boldsymbol{A}$, consider the order statistics from a random sample of the uniform distribution on $[0,1]$.)

(20) Find the svds of a symmetric matrix $\boldsymbol{S}$, an orthogonal matrix $\boldsymbol{O}$, and a vector outer product $\boldsymbol{u}\boldsymbol{v}^*$.

(21)* Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding orthonormal eigenvectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. Show that the modified matrix $(\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^*)\boldsymbol{A}$ has the eigenvalues $0, \lambda_2, \ldots, \lambda_n$ and the eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$, respectively.

(22)* Use the svd to prove that the full rank $m \times n$ matrices are dense in the set of $m \times n$ matrices.

(23)* Every invertible matrix $\boldsymbol{M}$ has a polar decomposition $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{O}$ into a product of a positive definite matrix $\boldsymbol{A}$ and an orthogonal matrix $\boldsymbol{O}$. Find the polar decomposition in terms of the svd of $\boldsymbol{M}$. Show that $\boldsymbol{O}$ is the closest orthogonal matrix to $\boldsymbol{M}$ in the Frobenius norm. (Hint: Apply Proposition A.7.5.)

(24) Find the svd of the matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}.$$

Express the singular vectors in terms of sines and cosines.

(25)* Consider a weighted graph with n nodes and edge weights $w_{ij} \geq 0$. The Laplacian $\boldsymbol{L}$ of the graph is defined to be the difference $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $\boldsymbol{D}$ is diagonal with diagonal entries $d_{ii} = \sum_{j \neq i} w_{ij}$ and $\boldsymbol{W}$ has entries w_{ij}. Show that $\boldsymbol{L}$ is symmetric and positive semidefinite. If the graph is connected, then show that the $\boldsymbol{L}$ has pseudo-inverse

$$\boldsymbol{L}^- \;=\; \Big(\boldsymbol{L} + \frac{\boldsymbol{1}\boldsymbol{1}^*}{n}\Big)^{-1} - \frac{\boldsymbol{1}\boldsymbol{1}^*}{n}.$$

(Hints: The associated quadratic form is $\boldsymbol{v}^*\boldsymbol{L}\boldsymbol{v} = \sum_i \sum_j w_{ij}(v_i - v_j)^2$, and $\boldsymbol{L}\boldsymbol{1} = \boldsymbol{0}$.)

(26) Show that the polynomial $p(x) = x^n + c_{n-1}x^{n-1} + \cdots + c_1 x + c_0$ can be expressed as

$$\begin{aligned} p(x) &= \det(x\boldsymbol{I}_n - \boldsymbol{C}) \\ \boldsymbol{C} &= \begin{pmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{pmatrix}. \end{aligned}$$

Hence, extraction of the roots of $p(x)$ reduces to the extraction of the eigenvalues of the companion matrix $\boldsymbol{C}$. (Hint: Prove the formula by induction on n.)

(27)* Demonstrate that the $m \times m$ matrix $\boldsymbol{A} = a\boldsymbol{I}_m + b\mathbf{1}\mathbf{1}^*$ has the eigenvector $\mathbf{1}$ with eigenvalue $a + mb$ and $m - 1$ orthogonal eigenvectors

$$\boldsymbol{u}_i \;=\; \frac{1}{i-1}\sum_{j=1}^{i-1}\boldsymbol{e}_j - \boldsymbol{e}_i,$$

$i = 2, \ldots, m$, with eigenvalue a. Under what circumstances is $\boldsymbol{A}$ positive definite.

(28) In Jacobi's eigen-decomposition algorithm, let $T = \tan(\theta)$, $C = \cos(\theta)$, and $S = \sin(\theta)$. Based on equation (8.2) argue that $r = \frac{c-a}{2b} = \frac{C^2-S^2}{2CS} = \frac{1-T^2}{2T}$ and therefore that T is determined by the quadratic $T^2 + 2rT - 1 = 0$. Show that this fact implies $T = \frac{1}{r+\sqrt{1+r^2}}$, $C = \frac{1}{\sqrt{1+t^2}}$, and $S = TC$ when $r > 0$.

(29)* Consider the set G of real $2n \times 2n$ matrices $\boldsymbol{M}$ satisfying $\boldsymbol{M}^*\boldsymbol{A}\boldsymbol{M} = \boldsymbol{A}$, where

$$\boldsymbol{A} \;=\; \begin{pmatrix} \mathbf{0} & \boldsymbol{I}_n \\ -\boldsymbol{I}_n & \mathbf{0} \end{pmatrix}.$$

Show that (a) $\boldsymbol{A}$ is invertible with $\det \boldsymbol{A} = 1$, (b) G is a group under matrix multiplication, and (c) all $\boldsymbol{M} \in G$ have $\det \boldsymbol{M} = \pm 1$. Actually, $\det \boldsymbol{M} = 1$, but this is harder to prove.

(30)* Prove that the positive definite square root of a positive definite matrix $\boldsymbol{A}$ is unique.

(31)* Prove that the map $\boldsymbol{A} \mapsto \sqrt{\boldsymbol{A}}$ is continuous on positive semidefinite matrices. Now consider two $m \times m$ positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with spectra contained in $[\delta, \infty)$. Show that the positive definite square roots of these matrices satisfy

$$\|\sqrt{\boldsymbol{A}} - \sqrt{\boldsymbol{B}}\| \;\leq\; \frac{1}{2\sqrt{\delta}}\|\boldsymbol{A} - \boldsymbol{B}\|.$$

(32)* A normal matrix $\boldsymbol{M}$ commutes with its transpose. For instance, every symmetric matrix is normal. If $\boldsymbol{M}$ is normal with polar decomposition $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{O}$, then demonstrate that $\boldsymbol{A}\boldsymbol{O} = \boldsymbol{O}\boldsymbol{A}$.

(33) Problem 3 of Chapter 5 demonstrates that the set of invertible $m \times m$ upper-triangular matrices forms a mathematical group G. The subset of G with positive diagonal entries constitutes a subgroup H, which can be used to decompose G into disjoint subsets called cosets. There are left cosets and right cosets. The left cosets $gH = \{gh : h \in H\}$ can be constructed by considering the diagonal matrices with entries of ± 1. (These also form a subgroup.) Let g_S denote such a matrix with -1 in precisely the positions determined by the index set S. Show that the coset $g_S H$ consists of those matrices in G with negative diagonal entries determined by S and positive diagonal entries elsewhere. Also show that $g_S H$ coincides with the right coset Hg_S. Multiplying a left coset $g_S H$ by an element g_T on the left produces the left coset $(g_T g_S)H$. Likewise, multiplying a right coset Hg_S by an element g_T on the right produces the right coset $H(g_S g_T)$. The proof of Proposition 8.7.2 claims that the matrix $\boldsymbol{T}_k = \boldsymbol{\Delta}^k(\boldsymbol{R}_k \cdots \boldsymbol{R}_2\boldsymbol{R}_1)(\boldsymbol{\Lambda}^k\boldsymbol{U})^{-1}\boldsymbol{\Delta}$ has positive diagonal entries. Given the current discussion, verify this fact.

(34)* The m vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$ in $\mathbb{R}^n$ define the parallelpiped

$$P \;=\; \Big\{\sum_{i=1}^{m} t_i\boldsymbol{w}_i : \; t_i \in [0,1] \; \forall \, i\Big\}$$

and the $n \times m$ matrix $\boldsymbol{W}$ whose ith column is $\boldsymbol{w}_i$. Note that $P = \boldsymbol{W}C$ is the image of the unit cube C in $\mathbb{R}^m$ under the linear map $\boldsymbol{x} \mapsto \boldsymbol{W}\boldsymbol{x}$. Assuming $m \leq n$, show that the volume of P within the subspace S spanned by the $\boldsymbol{w}_i$ satisfies $\text{vol}(P) = \sqrt{\det(\boldsymbol{W}^*\boldsymbol{W})}$. This result is pertinent to defining surface area.

(35)* Consider two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Prove that any eigenvalue λ of $\boldsymbol{A}\boldsymbol{B}$ is an eigenvalue of $\boldsymbol{B}\boldsymbol{A}$ and vice versa. (Hint: Consider the two cases $\lambda \neq 0$ and $\lambda = 0$ separately.)

8.11. Solutions to Selected Problems

8.11 If the positive semidefinite matrix $\boldsymbol{F}\boldsymbol{F}^*$ has spectral decomposition $\boldsymbol{O}\boldsymbol{\Lambda}\boldsymbol{O}^*$, then

$$\begin{aligned} \boldsymbol{v}^*(\boldsymbol{I} + \boldsymbol{F}\boldsymbol{F}^*)^{-1}\boldsymbol{v} &= \boldsymbol{v}^*\boldsymbol{O}(\boldsymbol{I} + \boldsymbol{\Lambda})^{-1}\boldsymbol{O}^*\boldsymbol{v} \\ \|\boldsymbol{v}\|^2 &= \|\boldsymbol{O}^*\boldsymbol{v}\|^2. \end{aligned}$$

Once we replace $\boldsymbol{O}^*\boldsymbol{v}$ by $\boldsymbol{u}$ and assume $\boldsymbol{\Lambda}$ has diagonal entries λ_i, then the claim follows from the hint with $x = \lambda_i$.

8.21 Note that

$$\begin{aligned} (\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^*)\boldsymbol{A}\boldsymbol{v}_1 &= \lambda_1(\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^*)\boldsymbol{v}_1 = \lambda_1(\boldsymbol{v}_1 - \boldsymbol{v}_1) = \boldsymbol{0} \\ (\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^*)\boldsymbol{A}\boldsymbol{v}_j &= \lambda_j(\boldsymbol{I} - \boldsymbol{v}_1\boldsymbol{v}_1^*)\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j \quad j > 1. \end{aligned}$$

8.22 Suppose the matrix $\boldsymbol{M}$ has full svd $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. Let $\boldsymbol{J}$ be a diagonal matrix with 1's along its diagonal and dimensionally matched to $\boldsymbol{\Sigma}$. Then $\boldsymbol{M}_\epsilon = \boldsymbol{U}(\boldsymbol{\Sigma} + \epsilon\boldsymbol{J})\boldsymbol{V}^*$ for $\epsilon > 0$ has full rank, and $\|\boldsymbol{M}_\epsilon - \boldsymbol{M}\|_F = \epsilon\|\boldsymbol{J}\|_F$, which can be made arbitrarily small by taking ϵ small.

8.23 Take the full svd $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ and put $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^*$ and $\boldsymbol{O} = \boldsymbol{U}\boldsymbol{V}^*$. Then $\boldsymbol{A}$ is positive definite and $\boldsymbol{O}$ is orthogonal. Assuming $\boldsymbol{M}$ is $n \times n$ and $\boldsymbol{W}$ is any $n \times n$ orthogonal matrix, then

$$\begin{aligned} \|\boldsymbol{M} - \boldsymbol{W}\|_F^2 &= \|\boldsymbol{M}\|_F^2 - 2\,\text{tr}(\boldsymbol{M}\boldsymbol{W}^*) + \|\boldsymbol{W}\|_F^2 \\ &= \|\boldsymbol{M}\|_F^2 - 2\,\text{tr}(\boldsymbol{M}\boldsymbol{W}^*) + n. \end{aligned}$$

The distance is minimized by maximizing $\text{tr}(\boldsymbol{M}\boldsymbol{W}^*)$. Because all singular values of $\boldsymbol{W}$ equal 1, Proposition A.7.5 implies $\text{tr}(\boldsymbol{M}\boldsymbol{W}^*) \leq \sum_i \sigma_i$, where σ_i is the ith singular value of $\boldsymbol{M}$. This upper bound is attained by the choice $\boldsymbol{W} = \boldsymbol{O} = \boldsymbol{U}\boldsymbol{V}^*$. Indeed, then

$$\text{tr}(\boldsymbol{M}\boldsymbol{W}^*) = \text{tr}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{V}\boldsymbol{U}^*) = \text{tr}(\boldsymbol{\Sigma}).$$

8.25 $\boldsymbol{L}$ is clearly symmetric. Because the associated quadratic

$$\boldsymbol{v}^*\boldsymbol{L}\boldsymbol{v} = \sum_i \sum_j w_{ij}(v_i - v_j)^2 \geq 0,$$

$\boldsymbol{L}$ is positive semidefinite. Furthermore, $\boldsymbol{L}\boldsymbol{1} = \boldsymbol{0}$ by definition. Suppose $\boldsymbol{L}\boldsymbol{v} = \boldsymbol{0}$ for some $\boldsymbol{v} = (v_i)$ not a multiple of $\boldsymbol{1}$. Let S be the set $\{j : v_j = v_1\}$. Since $0 = \boldsymbol{v}^*\boldsymbol{L}\boldsymbol{v} = \sum_i \sum_j w_{ij}(v_i - v_j)^2$, all neighbors of node 1 belong to S. All neighbors of those neighbors also belong to S, and so forth. Given that the graph is connected, it follows that $S = \{1, \ldots, n\}$. But this gives the contradiction $\boldsymbol{v} = v_1\boldsymbol{1}$. Thus, the svd $\boldsymbol{L} = \sum_{i=1}^n \lambda_i \boldsymbol{v}_i\boldsymbol{v}_i^*$ has all singular values $\lambda_i > 0$ except $\lambda_n = 0$. The pseudo-inverse of $\boldsymbol{L}$ is

$$\boldsymbol{L}^- = \sum_{i<n} \lambda_i^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^*.$$

9.4. Problems

(1) In a majorization-minimization algorithm with objective function $f(\boldsymbol{x})$ and surrogate function $g(\boldsymbol{x} \mid \boldsymbol{x}_n)$, show that the sequence $g(\boldsymbol{x}_{n+1} \mid \boldsymbol{x}_n)$ decreases.

(2) In an MM algorithm having a differentiable objective $f(\boldsymbol{x})$, a differentiable surrogate $g(\boldsymbol{x} \mid \boldsymbol{y})$, and no constraints, prove that the gradient identity $\nabla f(\boldsymbol{x}) = \nabla g(\boldsymbol{x} \mid \boldsymbol{x})$ when $\boldsymbol{y} = \boldsymbol{x}$. Here the gradient of $g(\boldsymbol{x} \mid \boldsymbol{y})$ is taken with respect to its left argument $\boldsymbol{x}$.

(3) In the previous problem assume that $f(\boldsymbol{x})$ and $g(\boldsymbol{x} \mid \boldsymbol{y})$ are twice differentiable in $\boldsymbol{x}$ for each $\boldsymbol{y}$. Demonstrate that the difference matrix $d^2 g(\boldsymbol{x} \mid \boldsymbol{x}) - d^2 f(\boldsymbol{x})$ is positive semidefinite.

(4)* Prove that the function $\ln \Gamma(t)$ is convex.

(5) Derive the balanced ANOVA estimates (9.8) by the method of Lagrange multipliers.

(6) Prove the majorization

$$\begin{aligned}(x+y-z)^2 &\leq -(x_n+y_n-z_n)^2 + 2(x_n+y_n-z_n)(x+y-z) \\ &\quad +3[(x-x_n)^2+(y-y_n)^2+(z-z_n)^2]\end{aligned}$$

separating the variables x, y, and z. In Example 9.7 this would facilitate penalizing parameter curvature rather than changes in parameter values.

(7) Find a quadratic upper bound majorizing the function e^{-x^2} around the point x_n.

(8) For the function $f(x) = \ln(1+e^x)$, derive the majorization

$$f(x) \leq f(x_n) + f'(x_n)(x-x_n) + \frac{1}{8}(x-x_n)^2$$

by the quadratic upper bound principle.

(9) Demonstrate the majorizations

$$\begin{aligned}xy &\leq \frac{1}{2}(x^2+y^2) + \frac{1}{2}(x_n-y_n)^2 - (x_n-y_n)(x-y) \\ -xy &\leq \frac{1}{2}(x^2+y^2) + \frac{1}{2}(x_n+y_n)^2 - (x_n+y_n)(x+y).\end{aligned} \tag{9.12}$$

(10) Based on problem 9, consider minimizing Rosenbrock's function

$$f(\boldsymbol{x}) = 100(x_1^2-x_2)^2 + (x_1-1)^2.$$

Show that up to an irrelevant constant $f(\boldsymbol{x})$ is majorized by the sum of the two functions

$$\begin{aligned}g_1(x_1 \mid \boldsymbol{x}_n) &= 200x_1^4 - [200(x_{n1}^2+x_{n2})-1]x_1^2 - 2x_1 \\ g_2(x_2 \mid \boldsymbol{x}_n) &= 200x_2^2 - 200(x_{n1}^2+x_{n2})x_2.\end{aligned}$$

Hence, to implement the corresponding MM algorithm, one must minimize a quartic in x_1 and a quadratic in x_2 at each iteration. Program the MM algorithm, and check whether it converges to the global minimum of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \mathbf{1}$.

(11) Verify the majorization $-\ln x \leq -\ln x_n + \frac{x_n}{x} - 1$ for x and x_n positive. Use this to design an MM algorithm for minimizing the convex function $f(x) = ax - \ln x$ for $a > 0$. Why do the iterates converge to a^{-1}?

(12) A number μ_q is said to be a q quantile of the m numbers $x_1, \ldots, x_m$ if it satisfies

$$\frac{1}{m}\sum_{x_i \leq \mu_q} 1 \geq q \quad \text{and} \quad \frac{1}{m}\sum_{x_i \geq \mu_q} 1 \geq 1-q.$$

If we define

$$\rho_q(r) = \begin{cases} qr & r \geq 0 \\ -(1-q)r & r < 0 \end{cases},$$

then it turns out that μ is a q quantile if and only if μ minimizes the function $f_q(\mu) = \sum_{i=1}^m \rho_q(x_i - \mu)$. Medians correspond to the case $q = 1/2$. Show that $\rho_q(r)$ is majorized by the quadratic

$$\zeta_q(r \mid r_n) = \frac{1}{4}\left[\frac{r^2}{|r_n|} + (4q-2)r + |r_n|\right].$$

Deduce from this majorization the MM algorithm

$$\begin{aligned} \mu_{n+1} &= \frac{m(2q-1) + \sum_{i=1}^m w_{ni}x_i}{\sum_{i=1}^m w_{ni}} \\ w_{ni} &= \frac{1}{|x_i - \mu_n|} \end{aligned}$$

for finding a q quantile. This interesting algorithm involves no sorting, only arithmetic operations [**123**].

(13) Implement the MM algorithm of the previous problem in Julia.

(14) The ABO system is typed in matching blood donors and recipients. The ABO locus has three alleles, A, B, and O, and four phenotypes, A, B, AB, and O. Because alleles A and B are dominant to allele O, phenotype A corresponds to the genotypes A/A and A/O, and phenotype B corresponds to genotypes B/B and B/O. Dominance complicates the calculation of the allele frequencies p_A, p_B, and p_O. Suppose in a random sample of unrelated individuals there are x_A type A individuals, x_B type B individuals, x_{AB} type AB individuals, and x_O type O individuals. Under genetic equilibrium, the loglikelihood of the data can be written as

$$\begin{aligned} L(\boldsymbol{p}) &= x_A \ln(p_A^2 + 2p_Ap_O) + x_B \ln(p_B^2 + 2p_Bp_O) \\ &\quad + x_{AB} \ln(2p_Ap_B) + x_O \ln p_O^2 \end{aligned}$$

after omitting an irrelevant multinomial coefficient. By parameter splitting, derive the following MM algorithm:

$$\begin{aligned} p_{n+1,A} &= \frac{2x_{nA/A} + x_{nA/O} + x_{AB}}{2x} \\ p_{n+1,B} &= \frac{2x_{nB/B} + x_{nB/O} + x_{AB}}{2x} \\ p_{n+1,O} &= \frac{x_{nA/O} + x_{nB/O} + 2x_O}{2x}, \end{aligned}$$

where

$$\begin{aligned} x &= x_A + x_B + x_{AB} + x_O \\ x_{nA/A} &= x_A \frac{p_{nA}^2}{p_{nA}^2 + 2p_{nA}p_{nO}} \\ x_{nA/O} &= x_A \frac{2p_{nA}p_{nO}}{p_{nA}^2 + 2p_{nA}p_{nO}} \\ x_{nB/B} &= x_B \frac{p_{nB}^2}{p_{nB}^2 + 2p_{nB}p_{nO}} \\ x_{nB/O} &= x_B \frac{2p_{nB}p_{nO}}{p_{nB}^2 + 2p_{nB}p_{nO}}. \end{aligned}$$

Interpret this algorithm as gene counting, and apply it to the real data $x_A = 186$, $x_B = 38$, $x_{AB} = 13$, and $x_O = 284$ of Clarke et al. [**48**]. You should arrive at the estimates $p_A = 0.2136$, $p_B = 0.0501$, and $p_O = 0.7363$.

(15) A random variable X concentrated on the nonnegative integers is said to have a power series distribution if

$$\Pr(X = k) = \frac{c_k \theta^k}{q(\theta)}. \tag{9.13}$$

In equation (9.13), θ is a positive parameter, the coefficients c_k are nonnegative, and $q(\theta) = \sum_{k=0}^{\infty} c_k \theta^k$ is the appropriate normalizing constant [**204**]. Examples include the binomial, negative binomial, Poisson, and logarithmic families and versions of these families truncated at 0. If $x_1, \ldots, x_m$ is a random sample from the discrete density (9.13) and $q(\theta)$ is log-concave, then devise an MM algorithm with updates

$$\theta_{n+1} = \frac{\bar{x} q(\theta_n)}{q'(\theta_n)}, \tag{9.14}$$

where $\bar{x}$ is the sample average of the observations x_i.

(16) Continuing problem 15, consider the truncated Poisson density with normalizing function $q(\theta) = e^\theta - 1$. Prove that $q(\theta)$ is log-concave, and program the MM algorithm (9.14) for estimating θ. Reproduce the MM iterates shown in Table 9.4 for the choices $\bar{x} = 2$ and $m = 10$. In the table $L(\theta)$ denotes the loglikelihood.

TABLE 9.4. Performance of Algorithm (9.14) on Truncated Poisson Data

n	θ_n	$L(\theta_n)$	n	θ_n	$L(\theta_n)$
0	1.00000	-5.41325	7	1.59161	-4.34467
1	1.26424	-4.63379	8	1.59280	-4.34466
2	1.43509	-4.40703	9	1.59329	-4.34466
3	1.52381	-4.35635	10	1.59349	-4.34466
4	1.56424	-4.34670	11	1.59357	-4.34466
5	1.58151	-4.34501	12	1.59360	-4.34466
6	1.58867	-4.34472	13	1.59362	-4.34466

(17) The gamma-Poisson distribution occurs in a model for the frequency of words in a document [**190**]. A gamma-Poisson random variable X is assumed to be Poisson given its intensity λ, which is taken to be gamma distributed with parameters α and β. Show that this definition generates the negative binomial density

$$p_x = \int_0^\infty e^{-\lambda}\frac{\lambda^x}{x!}\frac{\beta^\alpha\lambda^{\alpha-1}}{\Gamma(\alpha)}e^{-\beta\lambda}d\lambda = \frac{\beta^\alpha\Gamma(x+\alpha)}{\Gamma(\alpha)(1+\beta)^{x+\alpha}x!}$$

on the set of nonnegative integers. Demonstrate that the gamma-Poisson distribution has Laplace transform $\mathrm{E}(e^{-\theta X}) = \beta^\alpha(1+\beta-e^{-\theta})^{-\alpha}$, mean $\alpha\beta^{-1}$, and variance $\alpha(\beta+1)\beta^{-2}$. Now consider d independent documents with document i having w_i total words. It is reasonable to assume that the number of occurrences X_{ij} of word j in document i is Poisson with mean $\lambda_j w_i$, where λ_j is gamma with parameters α_j and β_j. For word counts x_{ij} and v vocabulary words, show that this translates into the likelihood

$$L(\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{i=1}^d\prod_{j=1}^v \frac{\beta_j^{\alpha_j}\Gamma(x_{ij}+\alpha_j)w_i^{x_{ij}}}{\Gamma(\alpha_j)(w_i+\beta_j)^{x_{ij}+\alpha_j}x_{ij}!}.$$

One can estimate $(\boldsymbol{\alpha},\boldsymbol{\beta})$ by block descent. Show that an MM update of β_j is

$$\beta_{n+1,j} = \frac{d\alpha_{nj}}{\sum_{i=1}^d \frac{x_{ij}+\alpha_{nj}}{w_i+\beta_{nj}}} \approx \frac{d\alpha_{nj}}{\sum_{i=1}^d \frac{x_{ij}+\alpha_{nj}}{w_i}}$$

and an MM update of α_j is

$$\alpha_{n+1,j} = \frac{\sum_{i=1}^d\sum_{k=0}^{x_{ij}-1}\frac{\alpha_{nj}}{\alpha_{nj}+k}}{\sum_{i=1}^d \ln\frac{w_i+\beta_{nj}}{\beta_{nj}}} \approx \frac{\alpha_{nj}\sum_{i=1}^d \ln\frac{x_{ij}+\alpha_{nj}-1}{\alpha_{nj}}}{\sum_{i=1}^d \ln\frac{w_i+\beta_{nj}}{\beta_{nj}}}.$$

(Hint: Mimic the reasoning in Example 9.3. Better approximations can be constructed by taking more terms in certain Euler–Maclaurin summation formulas.)

(18) Consider a random sample $x_1,\ldots,x_m$ from a two-parameter Weibull density

$$f(x) = \frac{\kappa}{\lambda}\left(\frac{x}{\lambda}\right)^{\kappa-1}e^{-(x/\lambda)^\kappa}$$

over the interval $(0,\infty)$. Show that the maximum of the loglikelihood with respect to $\lambda > 0$ for $\kappa > 0$ fixed satisfies $\lambda^\kappa = \frac{1}{m}\sum_{i=1}^m x_i^\kappa$ [**11**]. Derive the profile loglikelihood

$$m\ln\kappa - m\ln\left(\sum_{i=1}^m e^{\kappa\ln x_i}\right) + (\kappa-1)\sum_{i=1}^m \ln x_i + \text{constant}$$

by substituting this λ into the ordinary loglikelihood. Since the second term of the profile loglikelihood is concave in κ, one can apply the majorization (9.3). Prove that this leads to the MM algorithm

$$\frac{1}{\kappa_{n+1}} = \frac{\sum_{i=1}^m x_i^{\kappa_n}\ln x_i}{\sum_{i=1}^m x_i^{\kappa_n}} - \frac{1}{m}\sum_{i=1}^m \ln x_i.$$

(19) The Rasch model of test taking says that person i gives a correct response to question j of an exam with probability

$$\frac{e^{\alpha_i+\beta_j}}{1+e^{\alpha_i+\beta_j}}.$$

Justify the loglikelihood

$$L(\boldsymbol{\alpha},\boldsymbol{\beta}) \;=\; \sum_{i=1}^{p} x_i\alpha_i + \sum_{j=1}^{s} y_j\beta_j - \sum_{i=1}^{p}\sum_{j=1}^{s} \ln(1+e^{\alpha_i+\beta_j}),$$

where x_i is the number of correct answers given by person i, and y_j is the number of correct answers of question j given by all test takers. Based on the previous problem, majorize $-L(\boldsymbol{\alpha},\boldsymbol{\beta})$ by a quadratic. Minimizing the quadratic can be accomplished by one step of Newton's method. Thus, the Rasch model succumbs to a straightforward MM algorithm.

(20)* For $\boldsymbol{x}$ in $\mathbb{R}^2$, show that the function $f(\boldsymbol{x}) = \frac{x_1^2}{x_2}$ is convex on the domain $x_2 > 0$.

(21) Prove the minorization

$$\begin{aligned}\ln[1-\tanh(x)^2] \;&\geq\; \ln[1-\tanh(x_n)^2] - 2\tanh(x_n)(x-x_n)\\ &\quad -(x-x_n)^2\end{aligned}$$

by the quadratic upper bound principle.

(22)* Suppose $\boldsymbol{A} = (a_{ij})$ is a symmetric matrix and $\boldsymbol{B} = (b_{ij})$ is the diagonal matrix with ith diagonal $b_{ii} = \sum_j |a_{ij}|$. Prove that $\boldsymbol{x}^*\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{x}^*\boldsymbol{B}\boldsymbol{x}$ for all $\boldsymbol{x}$.

(23) Show that the linear function $f(\boldsymbol{x}) = \boldsymbol{v}^*\boldsymbol{x}$ is majorized by the quadratic

$$g(\boldsymbol{x} \mid \boldsymbol{x}_n) \;=\; \frac{1}{2}(\boldsymbol{v}^*\boldsymbol{x} - \boldsymbol{v}^*\boldsymbol{x}_n + 1)^2 - \frac{1}{2} + \boldsymbol{v}^*\boldsymbol{x}_n.$$

(24)* For all variables positive, demonstrate the majorization

$$\frac{1}{\sum_{i=1}^m a_i x_i} \;\leq\; \frac{1}{(\sum_{j=1}^m a_j x_{nj})^2}\sum_{i=1}^{m}\frac{a_i x_{ni}^2}{x_i}.$$

(25) Show that any norm $\|\boldsymbol{x}\|_*$ satisfies the majorization

$$\|\boldsymbol{x}\|_* \;\leq\; \frac{1}{2}\left(\|\boldsymbol{x}_n\|_* + \frac{\|\boldsymbol{x}\|_*^2}{\|\boldsymbol{x}_n\|_*}\right)$$

for $\boldsymbol{x}_n \neq \boldsymbol{0}$.

(26) Suppose $f(\boldsymbol{x})$ is a convex function with domain $\mathbb{R}^p$. Derive the majorization

$$f\Big(\sum_{i=1}^{m}\boldsymbol{x}_i\Big) \;\leq\; \frac{1}{m}\sum_{i=1}^{m} f[m(\boldsymbol{x}_i - \boldsymbol{x}_{ni} + \bar{\boldsymbol{x}}_n)]$$

involving vectors $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m$, where $\bar{\boldsymbol{x}}_n = \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_{ni}$ and $\boldsymbol{x}_{ni}$ represents the value of $\boldsymbol{x}_i$ at iteration n in some optimization scheme. This majorization splits the vectors. The choices $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_*$ and $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_*^2$ involving an arbitrary norm are important in practice.

(27)* Suppose $h(x)$ is nondecreasing, $r(x)$ is nonincreasing, and $f'(x)=r(x)h'(x)$. Prove that

$$g(x \mid x_n) \;=\; f(x_n) + r(x_n)h(x) - r(x_n)h(x_n)$$

majorizes $f(x)$ around x_n. Construct an example where this majorization applies.

(28) Consider a graph with edge set E and edge weights w_{ij}. The Laplacian of the graph is the quadratic form

$$\mathcal{L}(\boldsymbol{x}) \;=\; \frac{1}{2}\sum_{(i,j)\in E} w_{ij}(x_i - x_j)^2.$$

Show that $\mathcal{L}(\boldsymbol{x})$ is majorized by the quadratic function

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}) &\leq \sum_{(i,j)\in E} w_{ij}\Big[\Big(x_i - \frac{1}{2}x_{ni} - \frac{1}{2}x_{nj}\Big)^2 + \Big(x_j - \frac{1}{2}x_{ni} - \frac{1}{2}x_{nj}\Big)^2\Big] \\ &= \sum_{(i,j\in E)} w_{ij}[x_i^2 + x_j^2 - (x_i + x_j)(x_{nj} + x_{ni})] + c_n\end{aligned}$$

with variables separated, where c_n is an irrelevant constant.

(29) Let $f(x)$ be nondecreasing and $g(\boldsymbol{x})$ be concave. Demonstrate the majorization

$$f[g(\boldsymbol{x})] \leq f[g(\boldsymbol{x}_n) + dg(\boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{x}_n)].$$

(30) Suppose $f(x)$ is concave and differentiable. Verify the majorization

$$f[g(\boldsymbol{x})] \leq f[g(\boldsymbol{x}_n)] + f'[g(\boldsymbol{x}_n)][g(\boldsymbol{x}) - g(\boldsymbol{x}_n)].$$

(31) Suppose $f(x)$ is real and a-Lipschitz and $g(\boldsymbol{x})$ is b-Lipschitz. Verify the majorization

$$f[g(\boldsymbol{x})] \leq f[g(\boldsymbol{x}_n)] + ab\|\boldsymbol{x} - \boldsymbol{x}_n\|.$$

(32) Assume that the real-valued function satisfies the L-smooth majorization

$$f(\boldsymbol{x}) \leq f(\boldsymbol{x}_n) + df(\boldsymbol{x}_n)(\boldsymbol{x} - \boldsymbol{x}_n) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_n\|^2.$$

If $|\boldsymbol{x}|$ denotes the component-wise absolute value function, then demonstrate the majorization

$$f(|\boldsymbol{x}|) \leq f(|\boldsymbol{x}_n|) + df(|\boldsymbol{x}_n|)(|\boldsymbol{x}| - |\boldsymbol{x}_n|) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{x}_n\|^2.$$

This majorization is separable and has a simple proximal operator. Thus, $f(|\boldsymbol{x}|)$ can be routinely minimized, which is equivalent to minimization of $f(\boldsymbol{x})$ over the nonnegative orthant.

(33)* Prove the inequality $(x+y)^p \leq x^p + y^p$ for $p \in (0,1)$ and x and y nonnegative. (Hint: First reduce to a one-dimensional problem by dividing.)

(34)* Prove the squared hinge function $f(u) = \max\{1-u, 0\}^2$ is majorized by the surrogate function $g(u \mid u_n) = (u-u_n)^2$ for $u_n \geq 1$ and by $g(u \mid u_n) = (1-u)^2$ for $u_n < 1$. Draw a crude graph to visualize the problem.

(35)* Show that the function $f(x) = x^2 e^{-x^2}$ satisfies the quadratic upper bound majorization

$$f(x) \leq f(x_n) + f'(x_n)(x - x_n) + (x - x_n)^2.$$

See the article [**88**] for a potential application to edge enhancement in imaging.

(36)* The Cauchy probability density with location μ and scale σ can be written as

$$f(x) = \frac{1}{\pi\sigma\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]}.$$

The usual approach to maximum likelihood estimation of μ and σ involves finding the roots of polynomials of degree $2m-1$ and $2m$, respectively, for m sample points $x_1, \ldots, x_m$. However, this process tends to be complicated by the existence multiple local maxima. Derive the MM updates

$$\mu_{n+1} = \frac{\sum_{i=1}^m w_{ni}x_i}{\sum_{i=1}^m w_{ni}}, \quad \text{and} \quad \sigma_{n+1} = \sqrt{\frac{2\sum_{i=1}^m w_{ni}(x_i - \mu_{n+1})^2}{m}}.$$

for estimating μ and σ using positive weights w_{ni}. (Hint: Invoke the supporting hyperplane minorization for the function $-\log(1+y)$.)

(37)* Implement the MM algorithm of the previous problem in code. Verify that the loglikelihood is driven uphill.

(38) For negative binomial data $x_1, \ldots, x_m$ with parameters (p, r), show that Newton's method for updating r for fixed p relies on the derivatives

$$\begin{aligned}\frac{\partial}{\partial r}\ln \mathcal{L}(p,r) &= \sum_{i=1}^m \Big(\sum_{j=0}^{x_i-1}\frac{1}{r+j}+\ln p\Big)\\ \frac{\partial^2}{\partial r^2}\ln \mathcal{L}(p,r) &= -\sum_{i=1}^m\sum_{j=0}^{x_i-1}\frac{1}{(r+j)^2}.\end{aligned}$$

The Newton iterates are

$$r_{n+1} = r_n - \frac{\frac{\partial}{\partial r}\ln \mathcal{L}(p,r_n)}{\frac{\partial^2}{\partial r^2}\ln \mathcal{L}(p,r_n)}.$$

Program, document, and test this Newton's method.

(39)* Show that the function $f(x) = (|x|-w)^2$ for $w > 0$ is majorized by the convex quadratics

$$g(x \mid x_n) = \begin{cases}(x-w)^2 & x_n > 0\\ x^2+w^2 & x_n = 0\\ (x+w)^2 & x_n < 0.\end{cases}$$

When $w < 0$, show that $f(x)$ is majorized by

$$\frac{|x_n|-w}{|x_n|}(x^2-x_n^2)+f(x_n) = \frac{|x_n|-w}{|x_n|}(x^2-x_n^2)+(|x_n|-w)^2.$$

Note that no quadratic majorization exists when $x_n = 0$ and $w < 0$. Why is the case $w = 0$ trivial? (Hint: Graph $f(x)$.)

9.5. Solutions to Selected Problems

9.4 The derivatives

$$\begin{aligned}\frac{d}{dx}\Gamma(x) &= \int_0^\infty \log(y)y^{x-1}e^{-y}dy\\ \frac{d^2}{dx^2}\Gamma(x) &= \int_0^\infty \log(y)^2y^{x-1}e^{-y}dy\end{aligned}$$

emerge after differentiating under the integral sign. It follows that

$$\begin{aligned}\frac{d}{dx}\log\Gamma(x) &= \frac{\int_0^\infty \log(y)y^{x-1}e^{-y}dy}{\int_0^\infty y^{x-1}e^{-y}dy}\\ \frac{d^2}{dx^2}\log\Gamma(x) &= \frac{\int_0^\infty \log(y)^2y^{x-1}e^{-y}dy}{\int_0^\infty y^{x-1}e^{-y}dy}-\left[\frac{\int_0^\infty \log(y)y^{x-1}e^{-y}dy}{\int_0^\infty y^{x-1}e^{-y}dy}\right]^2\\ &= \mathrm{E}[\log(Y)^2]-\mathrm{E}[\log(Y)]^2\\ &= \mathrm{Var}[\log(Y)]\\ &\geq 0\end{aligned}$$

10.10. Problems

(1) Rewrite and test either the k-means or k-nearest neighbors algorithm with ℓ_1 distances substituted for ℓ_2 distances.

(2) In k-nearest neighbor regression, one predicts the value of a test response by taking a weighted average of the responses of the k-nearest training points. Plausible weights are the inverses of the distances in feature space between the test point and the neighboring training points. Write a Julia program implementing this algorithm. Note that responses are not considered features.

(3) In the EM clustering model, suppose we relax the assumption that the cluster distributions share a common covariance matrix $\boldsymbol{\Omega}$. This can cause the likelihood function to be unbounded. Imposing a prior stabilizes estimation [**41**]. It is mathematically convenient to impose independent inverse Wishart priors on the different covariance matrices $\boldsymbol{\Omega}_j$. This amounts to adding the logprior

$$-\sum_{j=1}^{k}\left[\frac{a}{2}\ln\det\boldsymbol{\Omega}_j+\frac{b}{2}\operatorname{tr}(\boldsymbol{\Omega}_j^{-1}\boldsymbol{S}_j)\right]$$

to the loglikelihood, where the positive constants a and b and the positive definite matrices $\boldsymbol{S}_j$ must be determined. Show that imposition of this prior does not change the MM updates of the fractions π_j or the cluster centers $\boldsymbol{\mu}_j$. The most natural choice is to take all $\boldsymbol{S}_j$ equal to the sample variance matrix

$$\boldsymbol{S} \;=\; \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{y}_i-\bar{\boldsymbol{y}})(\boldsymbol{y}_i-\bar{\boldsymbol{y}})^*.$$

Show that the MM updates of the $\boldsymbol{\Omega}_j$ are now

$$\begin{aligned}\boldsymbol{\Omega}_{n+1,j} &= \frac{a}{a+\sum_{i=1}^{m}w_{nij}}\left(\frac{b}{a}\boldsymbol{S}\right)+\frac{\sum_{i=1}^{m}w_{nij}}{a+\sum_{i=1}^{m}w_{nij}}\tilde{\boldsymbol{\Omega}}_{n+1,j}\\ \tilde{\boldsymbol{\Omega}}_{n+1,j} &= \frac{1}{\sum_{i=1}^{m}w_{nij}}\sum_{i=1}^{m}w_{nij}(\boldsymbol{y}_i-\boldsymbol{\mu}_{n+1,j})(\boldsymbol{y}_i-\boldsymbol{\mu}_{n+1,j})^*.\end{aligned}$$

In other words, the penalized EM update is a convex combination of the standard EM update and the mode $\frac{b}{a}\boldsymbol{S}$ of the prior. Chen and Tan [**41**] tentatively recommend the choice $a=b=2/\sqrt{m}$.

(4) The mean shift algorithm is a technique for finding the local maxima of an empirically constructed probability density [**79, 86**]. The algorithm has applications in computer vision and image processing. Suppose $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_m$ is a random sample of points from some unknown density function. One can approximate the density by a weighted sum

$$f(\boldsymbol{x}) \;=\; \sum_i w_i h(\|\boldsymbol{x}-\boldsymbol{y}_i\|^2)$$

of shifted versions of a density $h(\|\boldsymbol{x}\|^2)$. Let us assume that $h(s)$ is nonnegative, decreasing, convex, and differentiable on $[0,\infty)$. The mean shift algorithm looks for a local maximum of $f(\boldsymbol{x})$ in a neighborhood of an initial point $\boldsymbol{x}_0$ and iterates according to

$$\boldsymbol{x}_{n+1} \;=\; \frac{\sum_i w_i h'(\|\boldsymbol{x}_n-\boldsymbol{y}_i\|^2)\boldsymbol{y}_i}{\sum_i w_i h'(\|\boldsymbol{x}_n-\boldsymbol{y}_i\|^2)}.$$

Prove that the iterates constitute an MM algorithm for maximizing $f(\boldsymbol{x})$. What are the iterates in the Gaussian setting $h(s) = e^{-cs}$?

(5) Describe and program a naive Bayes classification algorithm for Gaussian distributed features. Assume that the features are independently distributed.

(6) In MCDA there are an infinity of regular simplexes. One appealing simplex has vertices

$$\boldsymbol{v}_j = \begin{cases} (c-1)^{-1/2}\mathbf{1} & \text{if } j = 1 \\ r\mathbf{1} + s\boldsymbol{e}_{j-1} & \text{if } 2 \le j \le c, \end{cases}$$

where

$$r = -\frac{1+\sqrt{c}}{(c-1)^{3/2}}, \qquad s = \sqrt{\frac{c}{c-1}},$$

and $\boldsymbol{e}_j$ is the jth standard coordinate vector in $\mathbb{R}^{c-1}$. Show that these choices place the c vertices on the surface of the unit sphere in $\mathbb{R}^{c-1}$ and that all vertex pairs are equidistant. Any rotation, dilation, or translation of these vectors preserves the equidistant property.

(7) Demonstrate that it is impossible to situate $c+1$ points in $\mathbb{R}^{c-1}$ so that all pairs of points are equidistant under the Euclidean norm [**237**].

(8) If the matrix $\boldsymbol{A}$ has full svd $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, then the best rank r approximation of $\boldsymbol{A}$ in the Frobenius norm is

$$\boldsymbol{B} = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*$$

based on the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ and diagonal entries of $\boldsymbol{\Sigma}$. This is the content of the Eckart–Young theorem [**76**] as proved in Proposition A.7.6. Given this result, use existing utilities in Julia to write a function for matrix completion.

(9) In the Kullback–Leibler divergence version of NMF, suppose the entries y_{ij} of $\boldsymbol{Y}$ are integers. They can then be viewed as realizations of independent Poisson random variables with means $\sum_k u_{ik}v_{kj}$. In this setting the loglikelihood is

$$L(\boldsymbol{U}, \boldsymbol{V}) = \sum_i \sum_j \Big[y_{ij} \ln \Big(\sum_k u_{ik} v_{kj} \Big) - \sum_k u_{ik} v_{kj} \Big].$$

Maximization with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ should lead to a good factorization. Verify the minorization

$$\ln \Big(\sum_k u_{ik} v_{kj} \Big) \ge \sum_k \frac{a_{nikj}}{s_{nij}} \ln \Big(\frac{s_{nij}}{a_{nikj}} u_{ik} v_{kj} \Big),$$

where

$$a_{nikj} = u_{nik} v_{nkj}, \quad s_{nij} = \sum_k a_{nikj},$$

and n indicates the current iteration. Given this minorization, derive the alternating multiplicative updates

$$u_{n+1,ik} = u_{nik} \frac{\sum_j y_{ij} \frac{v_{nkj}}{s_{nij}}}{\sum_j v_{nkj}}$$

and

$$v_{n+1,kj} = v_{nkj} \frac{\sum_i y_{ij} \frac{u_{nik}}{s_{nij}}}{\sum_i v_{nik}}.$$

(10) In problem 9 calculate the partial derivative

$$\frac{\partial}{\partial u_{il}} L(\boldsymbol{U}, \boldsymbol{V}) \;=\; \sum_j v_{lj} \Big(\frac{y_{ij}}{\sum_k u_{ik} v_{kj}} - 1 \Big).$$

Show that the conditions $\min\{u_{il}, -\frac{\partial}{\partial u_{il}} L(\boldsymbol{U}, \boldsymbol{V})\} = 0$ for all pairs (i, l) are both necessary and sufficient for $\boldsymbol{U}$ to maximize $L(\boldsymbol{U}, \boldsymbol{V})$ when $\boldsymbol{V}$ is fixed. The same conditions apply in minimizing the Frobenius criterion $\|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}\|_F^2$ with different partial derivatives.

(11) In the matrix factorization, it may be worthwhile shrinking the estimates of the entries of $\boldsymbol{U}$ and $\boldsymbol{V}$ toward 0 [**194**]. Let λ and μ be positive constants, and consider the penalized objective functions

$$\begin{aligned} l(\boldsymbol{U}, \boldsymbol{V}) &= L(\boldsymbol{U}, \boldsymbol{V}) - \lambda \sum_i \sum_k u_{ik} - \mu \sum_k \sum_j v_{kj} \\ r(\boldsymbol{V}, \boldsymbol{W}) &= \|\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{V}\|_F^2 + \lambda \sum_i \sum_k u_{ik}^2 + \mu \sum_k \sum_j v_{kj}^2 \end{aligned}$$

with lasso and ridge penalties, respectively. Here the Poisson loglikelihood $L(\boldsymbol{U}, \boldsymbol{V})$ is studied in problem 9. Derive block optimization updates for these objective functions. Explain why the updates maintain positivity and why shrinkage is obvious, with stronger shrinkage for the lasso penalty with small parameters.

(12) Find the Euclidean projection operators for the four following closed convex sets in $\mathbb{R}^p$: (a) closed ball $\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{z}\| \leq r\}$, (b) closed rectangle $\{\boldsymbol{x} : \boldsymbol{a} \leq \boldsymbol{x} \leq \boldsymbol{b}\}$, (c) hyperplane $\{\boldsymbol{x} : \boldsymbol{a}^* \boldsymbol{x} = b\}$ for $\boldsymbol{a} \neq \boldsymbol{0}$, and (d) closed halfspace $\{\boldsymbol{x} : \boldsymbol{a}^* \boldsymbol{x} \leq b\}$. Write Julia functions implementing these projections.

(13) Let Sym_n denote the subspace of $n \times n$ symmetric matrices. Find the Euclidean projection mapping an arbitrary $n \times n$ matrix $\boldsymbol{X}$ onto Sym_n.

(14) Let Skew_n denote the subspace of $n \times n$ skew-symmetric matrices. Show that the map $\boldsymbol{X} \mapsto \frac{1}{2}(\boldsymbol{X} - \boldsymbol{X}^*)$ constitutes projection onto Skew_n and that the subspaces Sym_n and Skew_n are orthogonal complements under the Frobenius inner product.

(15) Let Pos_n denote the set of $n \times n$ positive semidefinite matrices. Find the Euclidean projection mapping an $n \times n$ symmetric matrix $\boldsymbol{X}$ onto Pos_n. (Hint: The Frobenius norm is invariant under multiplication of its argument by an orthogonal matrix.)

(16) Let S_r be the set whose points have at most r nonzero coordinates. Show that projecting $\boldsymbol{y}$ onto S_r amounts to sorting the coordinates of $\boldsymbol{y}$ by magnitude, saving the r largest, and sending the remaining $n - r$ coordinates to zero. For what $\boldsymbol{y}$ is the projection multivalued?

(17) The set $U_k^n = \{0, 1\}^n \cap \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}^* \boldsymbol{1} = k\}$ is the collection of $\binom{n}{k}$ vertices of the unit cube whose coordinates sum to k. Show that projection of $\boldsymbol{y}$ onto U_k^n replaces the k largest entries of $\boldsymbol{y}$ by 1 and the remaining entries by 0.

(18)* Find the Euclidean projection of a point $\boldsymbol{x}$ onto a line segment $[\boldsymbol{a}, \boldsymbol{b}] \subset \mathbb{R}^n$. (Hint: This is a purely one-dimensional problem.)

(19) Consider the piecewise linear function $f(\mu) = c\mu + \sum_{i=1}^n w_i |y_i - \mu|$, where the points y_i satisfy $y_1 < y_2 < \cdots < y_n$ and the positive weights satisfy $\sum_{i=1}^n w_i = 1$. Show that $f(\mu)$ has no minimum when $|c| > 1$. What happens when $c = 1$ or $c = -1$? This leaves the case $|c| < 1$. Show that a minimum

occurs when

$$\sum_{y_i>\mu} w_i - \sum_{y_i\leq\mu} w_i \;\leq\; c \text{ and } \sum_{y_i\geq\mu} w_i - \sum_{y_i<\mu} w_i \;\geq\; c.$$

The case $c = 0$ produces the weighted median. (Hints: A crude plot of $f(\mu)$ might help. What conditions on the right-hand and left-hand derivatives of $f(\mu)$ characterize a minimum?)

(20)* Registration of two images is one of the vexing problems of medical imaging. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ represent two sets of matched anatomical landmarks in $\mathbb{R}^2$. Registration can be achieved by mapping $\boldsymbol{x}_j$ into $\boldsymbol{y}_j$ for each j. Affine maps $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ are the simplest relevant maps. To estimate $\boldsymbol{A}$ and $\boldsymbol{b}$, one can minimize the objective

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_F^2,$$

where the matrix $\boldsymbol{B}$ has all columns equal to $\boldsymbol{b}$, and the matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ have jth columns $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$, respectively. One can estimate parameters by block relaxation. Show that the minimum of the objective with respect to $\boldsymbol{b}$ for $\boldsymbol{A}$ fixed is achieved by taking $\boldsymbol{b}$ equal to the average of the columns of the matrix $\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}$. For $\boldsymbol{B}$ fixed, show that the optimal $\boldsymbol{A}$ is

$$\boldsymbol{A} \;=\; (\boldsymbol{Y} - \boldsymbol{B})\boldsymbol{X}^*(\boldsymbol{X}\boldsymbol{X}^*)^{-1},$$

assuming $\boldsymbol{X}$ has full rank. Why do these solutions represent minimum points?

(21) Program the algorithm for k-means clustering with missing data described in the text.

(22)* Devise and code projected gradient descent for minimizing the least squares criterion $f(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$ subject to $\|\boldsymbol{\beta}\| \leq r$.

(23)* The Minkowski sum $A+B$ of two sets A and B is defined to be the set of all sums $\boldsymbol{a}+\boldsymbol{b}$ for $\boldsymbol{a} \in A$ and $\boldsymbol{b} \in B$. One can easily prove that $A+B$ is convex whenever A and B are both convex. Unfortunately, $A + B$ is not necessarily closed when A and B are both closed; it is closed if one of the two sets is compact and the other is closed. It is also closed when A and B are both polyhedral. A natural question is how to project a point $\boldsymbol{x}$ onto $A+B$ given both convexity and closure. Suppose that the projection operators $P_A(\boldsymbol{y})$ and $P_B(\boldsymbol{z})$ are both known. Show that the alternating algorithm

$$\begin{aligned} \boldsymbol{y}_{n+1} &= P_A(\boldsymbol{x} - \boldsymbol{z}_n) \\ \boldsymbol{z}_{n+1} &= P_B(\boldsymbol{x} - \boldsymbol{y}_{n+1}) \end{aligned}$$

is just block relaxation for solving this problem. One can argue that $\boldsymbol{y}_n + \boldsymbol{z}_n$ converges to the solution. Describe this block relaxation algorithm in more detail when B is the sphere $\{\boldsymbol{x} : \|\boldsymbol{x}\| \leq r\}$. In this setting $A + B$ is a kind of halo of the set A.

(24)* Show that Lloyd's algorithm is an MM algorithm.

(25) In multidimensional scaling assume uniform weights. If we impose the constraint $\sum_{i=1}^q \boldsymbol{x}_i = \boldsymbol{0}$, then show that the Lagrangian for the SMACOF surrogate

(one majorization, not two) becomes up to a constant

$$\begin{aligned}\mathcal{L}(\boldsymbol{X},\boldsymbol{\lambda}) &= \sum_{1\le i<j\le q}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2-2\sum_{i=1}^q\boldsymbol{v}_{ni}^*\boldsymbol{x}_i+\boldsymbol{\lambda}^*\sum_{i=1}^q\boldsymbol{x}_i\\ \boldsymbol{v}_{ni} &= \sum_{j\neq i}d_{ij}\frac{\boldsymbol{x}_{ni}-\boldsymbol{x}_{nj}}{\|\boldsymbol{x}_{ni}-\boldsymbol{x}_{nj}\|}.\end{aligned}$$

Demonstrate that solving the stationary equations leads to the simple MM updates

$$\boldsymbol{x}_{n+1,i} = \frac{1}{q}\sum_{j=1}^q\boldsymbol{x}_{n+1,j}+\frac{\boldsymbol{v}_{ni}}{q} = \frac{\boldsymbol{v}_{ni}}{q}.$$

(26)* Apply k-means clustering to the data $(x_1,x_2,x_3)=(0,2,3)$, and show that the initial clusters $\{0,2\}$ and $\{3\}$ are stable and preserved by the algorithm. What is the optimal clustering into to 2 groups? Thus, k-means can converge to an inferior solution.

(27)* The trimmed mean is a robust estimator of location and a device for outlier detection. To find a trimmed mean for multivariate data $\boldsymbol{X}=(x_{ij})$, a reasonable procedure is to minimize the negative multivariate Gaussian loglikelihood

$$f(\boldsymbol{w},\boldsymbol{\mu},\boldsymbol{\Omega}) = \frac{1}{2}\sum_{i=1}^m w_i\Big[\ln\det\boldsymbol{\Omega}+(\boldsymbol{x}_i-\boldsymbol{\mu})^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\Big],$$

where $\boldsymbol{\mu}$ is the mean of the features, $\boldsymbol{\Omega}$ is the covariance matrix, and i is the case index. Here all $w_i\in\{0,1\}$ and $\sum_i w_i=k$ is the number of retained cases. The updates of $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ are

$$\begin{aligned}\hat{\boldsymbol{\mu}} &= \frac{\sum_i w_i\boldsymbol{x}_i}{\sum_i w_{ij}}\\ \hat{\boldsymbol{\Omega}} &= \frac{1}{k}\sum_{i=1}^m w_i(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^*.\end{aligned}$$

The weights w_i can be calculated by finding the k smallest values of the quantities $(\boldsymbol{x}_i-\boldsymbol{\mu})^*\boldsymbol{\Omega}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$ and setting the corresponding $w_i=1$. The two kinds of updates are alternated until convergence. This model requires $k\ge d$, where d is the number of features. Once again the possibility of converging to an inferior minimum cannot be ruled out. Code and test this procedure.

10.11. Solutions to Selected Problems

10.18 A point on the line through $\boldsymbol{a}$ and $\boldsymbol{b}$ can be expressed as an affine combination $t\boldsymbol{a}+(1-t)\boldsymbol{b}$ for t a scalar. Points on the convex segment correspond to the values $t\in[0,1]$. Thus, we must minimize the quadratic

$$\begin{aligned}f(t) &= \frac{1}{2}\|\boldsymbol{x}-t\boldsymbol{a}-(1-t)\boldsymbol{b}\|^2\\ &= \frac{t^2}{2}\|\boldsymbol{b}-\boldsymbol{a}\|^2+t(\boldsymbol{x}-\boldsymbol{b})^*(\boldsymbol{b}-\boldsymbol{a})+\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{b}\|^2\end{aligned}$$

subject to $t\in[0,1]$. The unconstrained minimum occurs at $t_{\min}=\frac{(\boldsymbol{b}-\boldsymbol{x})^*(\boldsymbol{b}-\boldsymbol{a})}{\|\boldsymbol{b}-\boldsymbol{a}\|^2}$. If $t_{\min}<0$, then $f'(0)>0$, and the projected point is $\boldsymbol{a}$. If $t_{\min}>1$, then the projected point is $\boldsymbol{b}$. Otherwise, the projected point is $t_{\min}\boldsymbol{a}+(1-t_{\min})\boldsymbol{b}$.

11.10. Problems

(1) Write and test a fast Julia function to multiply two large positive integers.
(2) Verify the explicit numerical results mentioned in Example 11.1.
(3) Explicitly calculate discrete Fourier transforms for the sequences $c_j = 1$, $c_j = 1_{\{0\}}$, $c_j = (-1)^j$, and $c_j = 1_{\{0,1,\ldots,n/2-1\}}$ defined on $\{0, 1, \ldots, n-1\}$. For the last two sequences assume that n is even.
(4) Show that the sequence $c_j = j$ on $\{0, 1, \ldots, n-1\}$ has discrete Fourier transform

$$\hat{c}_k = \begin{cases} \frac{n-1}{2} & k = 0 \\ -\frac{1}{2} + \frac{i}{2}\cot\frac{k\pi}{n} & k \neq 0. \end{cases}$$

(5) For $0 \leq r < n/2$, define the rectangular and triangular smoothing sequences

$$\begin{aligned} c_j &= \frac{1}{2r+1} 1_{\{-r \leq j \leq r\}} \\ d_j &= \frac{1}{r} 1_{\{-r \leq j \leq r\}} \left(1 - \frac{|j|}{r}\right) \end{aligned}$$

and extend them to have period n. Show that

$$\begin{aligned} \hat{c}_k &= \frac{1}{n(2r+1)} \frac{\sin\frac{(2r+1)k\pi}{n}}{\sin} \frac{k\pi}{n} \\ \hat{d}_k &= \frac{1}{nr^2} \left(\frac{\sin\frac{rk\pi}{n}}{\sin\frac{k\pi}{n}}\right)^2. \end{aligned}$$

(6) Prove parts (b) and (c) of Proposition 11.2.1.
(7) For $0 \leq m \leq n-1$ and a periodic function $f(x)$ on [0,1], define the sequence $b_m = f(m/n)$. If $\hat{b}_k$ is the discrete Fourier transform of the sequence b_m, then we can approximate $f(x)$ by $\sum_{k=-\lfloor n/2 \rfloor}^{\lfloor n/2 \rfloor} \hat{b}_k e^{2\pi i k x}$. Show that this approximation is exact when $f(x)$ is equal to $e^{2\pi i j x}$, $\cos(2\pi j x)$, or $\sin(2\pi j x)$ for j satisfying $0 \leq |j| < \lfloor n/2 \rfloor$.
(8) Continuing problem 7, let c_k be the kth Fourier series coefficient of a general periodic function $f(x)$. If $|c_k| \leq a r^{|k|}$ for constants $a \geq 0$ and $0 \leq r < 1$, then verify using equation (11.9) that

$$|\hat{b}_k - c_k| \leq a r^n \frac{r^k + r^{-k}}{1 - r^n}$$

for $|k| < n$. Functions analytic around 0 automatically possess Fourier coefficients satisfying the bound $|c_k| \leq a r^{|k|}$.
(9) Continuing problems (7) and (8), suppose a constant $a \geq 0$ and positive integer p exist such that

$$|c_k| \leq \frac{a}{|k|^{p+1}}$$

for all $k \neq 0$. (One can show that this criterion holds if $f^{(p+1)}(x)$ is piecewise continuous.) Verify the inequality

$$|\hat{b}_k - c_k| \leq \frac{a}{n^{p+1}} \sum_{j=1}^{\infty} \left[\frac{1}{\left(j + \frac{k}{n}\right)^{p+1}} + \frac{1}{\left(j - \frac{k}{n}\right)^{p+1}} \right]$$

when $|k| < n/2$. To simplify this inequality, demonstrate that

$$\begin{aligned}\sum_{j=1}^{\infty} \frac{1}{(j+\alpha)^{p+1}} &< \int_{\frac{1}{2}}^{\infty} (x+\alpha)^{-p-1} dx \\ &= \frac{1}{p\Big(\frac{1}{2}+\alpha\Big)^p}\end{aligned}$$

for $\alpha > -1/2$. Finally, conclude that

$$|\hat{b}_k - c_k| < \frac{a}{pn^{p+1}}\left[\frac{1}{\Big(\frac{1}{2}+\frac{k}{n}\Big)^p} + \frac{1}{\Big(\frac{1}{2}-\frac{k}{n}\Big)^p}\right].$$

(10) For a complex number c with $|c| > 1$, show that the periodic function $f(x) = (c - e^{2\pi i x})^{-1}$ has the Fourier series coefficients $c_k = c^{-k-1}1_{\{k\geq 0\}}$. Argue from equation (11.9) that the discrete Fourier transform approximation $\hat{b}_k$ to c_k is

$$\hat{b}_k = \begin{cases} c^{-k-1}\frac{1}{1-c^{-n}} & 0 \leq k \leq \frac{n}{2}-1 \\ c^{-n-k-1}\frac{1}{1-c^{-n}} & -\frac{n}{2} \leq k \leq 0. \end{cases}$$

(11) For some purposes it is preferable to have a purely real transform. If $c_1, \ldots, c_{n-1}$ is a finite sequence of real numbers, then we define its discrete sine transform by

$$\hat{c}_k = \frac{2}{n}\sum_{j=1}^{n-1} c_j \sin\Big(\frac{\pi k j}{n}\Big).$$

Show that this transform has inverse

$$\breve{d}_j = \sum_{k=1}^{n-1} d_k \sin\Big(\frac{\pi k j}{n}\Big).$$

(Hint: It is helpful to consider $c_1, \ldots, c_{n-1}$ as part of a sequence of period $2n$ that is odd about n.)

(12) From a real sequence c_k of period $2n$ we can concoct a complex sequence of period n according to the recipe $d_k = c_{2k} + ic_{2k+1}$. Because it is quicker to take the discrete Fourier transform of the sequence d_k than that of the sequence c_k, it is desirable to have a simple method of constructing $\hat{c}_k$ from $\hat{d}_k$. Show that

$$\hat{c}_k = \frac{1}{4}\Big(\hat{d}_k + \hat{d}^*_{n-k}\Big) - \frac{i}{4}\Big(\hat{d}_k - \hat{d}^*_{n-k}\Big)e^{-\frac{\pi i k}{n}},$$

where z^* denotes the complex conjugate of the complex number z.

(13) Let $F(s) = \sum_{n=1}^{\infty} f_n s^n$ be a probability generating function. Show that the equation $F(s) = 1$ has only the solution $s = 1$ on $|s| = 1$ if and only if the set $\{n\colon f_n > 0\}$ has greatest common divisor 1.

(14) Let W be the waiting time until the first run of r heads in a coin-tossing experiment. If heads occur with probability p, and tails occur with probability $q = 1 - p$ per trial, then show that W has the generating function displayed in equation (11.14). (Hint: Argue that either $W = r$ or $W = k + 1 + W_k$, where $0 \leq k \leq r - 1$ is the initial number of heads and W_k is a probabilistic replica of W.)

(15) Consider a power series $f(x) = \sum_{m=0}^{\infty} c_m x^m$ with radius of convergence $r > 0$. Prove that

$$\sum_{m=k \bmod n}^{\infty} c_m x^m = \frac{1}{n} \sum_{j=0}^{n-1} u_n^{-jk} f(u_n^j x)$$

for any x with $|x| < r$. As a special case, verify the identity

$$\sum_{m=k \bmod n}^{\infty} \binom{p}{m} = \frac{2^p}{n} \sum_{j=0}^{n-1} \cos\left[\frac{(p-2k)j\pi}{n}\right] \cos^p\left[\frac{j\pi}{n}\right]$$

for any positive integer p.

(16) For a fixed positive integer n, we define the segmental functions ${}_n\alpha_j(x)$ of x as the discrete Fourier transform coefficients

$${}_n\alpha_j(x) = \frac{1}{n} \sum_{k=0}^{n-1} e^{x u_n^k} u_n^{-jk}.$$

The segmental functions generalize the hyperbolic trig functions $\cosh(x)$ and $\sinh(x)$. Prove the following assertions:

(a) ${}_n\alpha_j(x) = {}_n\alpha_{j+n}(x)$.
(b) ${}_n\alpha_j(x+y) = \sum_{k=0}^{n-1} {}_n\alpha_k(x) {}_n\alpha_{j-k}(y)$.
(c) ${}_n\alpha_j(x) = \sum_{k=0}^{\infty} x^{j+kn}/(j+kn)!$ for $0 \le j \le n-1$.
(d) $\frac{d}{dx}\left[{}_n\alpha_j(x)\right] = {}_n\alpha_{j-1}(x)$.
(e) Consider the differential equation $\frac{d^n}{dx^n} f(x) = k f(x)$ with initial conditions $\frac{d^j}{dx^j} f(0) = c_j$ for $0 \le j \le n-1$, where k and the c_j are constants. Show that

$$f(x) = \sum_{j=0}^{n-1} c_j k^{-\frac{j}{n}} {}_n\alpha_j(k^{\frac{1}{n}} x).$$

(f) The differential equation $\frac{d^n}{dx^n} f(x) = k f(x) + g(x)$ with initial conditions $\frac{d^j}{dx^j} f(0) = c_j$ for $0 \le j \le n-1$ has solution

$$\begin{aligned} f(x) &= \int_0^x k^{-\frac{n-1}{n}} {}_n\alpha_{n-1}[k^{\frac{1}{n}}(x-y)] g(y) dy \\ &\quad + \sum_{j=0}^{n-1} c_j k^{-\frac{j}{n}} {}_n\alpha_j(k^{\frac{1}{n}} x). \end{aligned}$$

(g) $\lim_{x\to\infty} e^{-x} {}_n\alpha_j(x) = 1/n$.
(h) In a Poisson process of intensity 1, $e^{-x} {}_n\alpha_j(x)$ is the probability that the number of random points on $[0, x]$ equals j modulo n.
(i) Relative to this Poisson process, let N_x count every nth random point on $[0, x]$. Then N_x has probability generating function

$$P(s) = e^{-x} \sum_{j=0}^{n-1} s^{-\frac{j}{n}} {}_n\alpha_j(s^{\frac{1}{n}} x).$$

(j) Furthermore, N_x has mean

$$\mathrm{E}(N_x) = \frac{x}{n} - \frac{e^{-x}}{n} \sum_{j=0}^{n-1} j {}_n\alpha_j(x).$$

(k) $\lim_{x\to\infty}\Big[\mathrm{E}(N_x) - x/n\Big] = -(n-1)/(2n)$.

(17)* The Hadamard transform $\boldsymbol{x} \mapsto \boldsymbol{H}_n\boldsymbol{x}$ turns a vector $\boldsymbol{x}$ with $n = 2^d$ real components into a vector $\hat{\boldsymbol{x}}$ with n real components. The Hadamard matrices are defined recursively by the formula

$$\boldsymbol{H}_n \;=\; \frac{1}{\sqrt{2}}\begin{pmatrix} \boldsymbol{H}_{\frac{n}{2}} & \boldsymbol{H}_{\frac{n}{2}} \\ \boldsymbol{H}_{\frac{n}{2}} & -\boldsymbol{H}_{\frac{n}{2}} \end{pmatrix}$$

starting from $\boldsymbol{H}_1 = 1$. Prove that $\boldsymbol{H}_n$ satisfies $\boldsymbol{H}_n^* = \boldsymbol{H}_n$ and $\boldsymbol{H}_n^2 = \boldsymbol{I}$. Thus, the Hadamard transform is its own inverse. Note that all entries of $\boldsymbol{H}_n$ are multiples of $\frac{\pm 1}{\sqrt{n}}$.

(18) Show that the transpose of a circulant matrix is circulant. Under what conditions is a circulant matrix symmetric? What if we replace transpose by adjoint (conjugate transpose) and symmetric by Hermitian (self-adjoint)?

(19) The $n \times n$ permutation matrix $\boldsymbol{S}_n$ satisfying $\boldsymbol{S}_n\boldsymbol{e}_k = \boldsymbol{e}_{k+1}$ for $k = 1, \ldots, n-1$ and $\boldsymbol{S}_n\boldsymbol{e}_n = \boldsymbol{e}_1$ is called a circular shift. Prove that the $n \times n$ matrix $\boldsymbol{C}$ is circulant if and only if $\boldsymbol{S}_n\boldsymbol{C} = \boldsymbol{C}\boldsymbol{S}_n$.

(20) Prove that the sum or product of two circulant matrices of the same size is circulant. Also prove that the inverse of an invertible circulant matrix C is circulant. Thus, the collection of invertible circulant matrices forms a group. Why is this group commutative?

(21) Show that the nth root of unity u_n satisfies

$$\prod_{j=0}^{n-1} u_n^j \;=\; \begin{cases} -1 & n \text{ is even} \\ 1 & n \text{ is odd} \end{cases}$$

for all $n \geq 1$.

(22) Derive the cubic polynomial equations (11.8) by substitution. The identity

$$\begin{aligned} \hat{r}_0\hat{r}_1 &= \hat{r}_0^2 + \hat{r}_1^2 + \hat{r}_2^2 + (\hat{r}_0\hat{r}_1 + \hat{r}_0\hat{r}_2 + \hat{r}_1\hat{r}_2)(u_3 + u_3^2) \\ &= \hat{r}_0^2 + \hat{r}_1^2 + \hat{r}_2^2 - (\hat{r}_0\hat{r}_1 + \hat{r}_0\hat{r}_2 + \hat{r}_1\hat{r}_2) \end{aligned}$$

is representative is of the needed steps.

(23)* Consider the symmetric $n \times n$ matrix

$$\boldsymbol{C} \;=\; \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-2} & \rho^{n-3} & \rho^{n-4} & \cdots & \rho \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

for $|\rho| \leq 1$ arising in the autoregressive AR(1) model of statistics and signal processing. Show that $\boldsymbol{C}$ has rank 1 for $\rho = \pm 1$. Otherwise, show that $\boldsymbol{C}$ is invertible with the symmetric tridiagonal inverse

$$\boldsymbol{C}^{-1} \;=\; \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 \\ -\rho & 1+\rho^2 & -\rho & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & -\rho \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Use this representation to prove that $\boldsymbol{C}$ is positive definite when $|\rho| < 1$.

(24)* Given two finite sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ of nonnegative integers, show how to find all possible sums $a_i + b_j$ efficiently and for each sum count how often it appears. For instance, if $\boldsymbol{a} = (1, 2, 3)$ and $\boldsymbol{b} = (2, 4)$, then the sum 3 can be obtained in 1 way, 4 in 1 way, 5 in 2 ways, 6 in 1 way, and 7 in 1 way.

(25)* Consider two strings drawn from an alphabet $\mathcal{A}$ with d letters. Suppose we assign a complex number to each letter. This transforms the strings into two numeric sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ of length m and $n > m$. It is of interest to count the number of occurrences of $\boldsymbol{a}$ in $\boldsymbol{b}$. We can measure the difference between $\boldsymbol{a}$ and the block $(b_k, \ldots, b_{k+m-1})$ of $\boldsymbol{b}$ by

$$\begin{aligned} c_k &= \sum_{j=0}^{m-1} |a_j - b_{j+k}|^2 \\ &= \sum_{j=0}^{m-1} |a_j|^2 - 2 \sum_{j=0}^{m-1} \text{Real}(a_j b_{j+k}^*) + \sum_{j=0}^{m-1} |b_{j+k}|^2, \end{aligned}$$

where z^* denotes the complex conjugate of z. Show how we can choose the numbers assigned to the letters to reveal which blocks of $\boldsymbol{b}$ perfectly match $\boldsymbol{a}$. This assignment should support fast matching.

(26)* An upper triangular $n \times n$ Toeplitz matrices can be written as

$$\boldsymbol{T} = \sum_{j=0}^{n-1} t_j \boldsymbol{N}^j, \quad \boldsymbol{N} = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \tag{11.18}$$

where $\boldsymbol{N}$ is an upper triangular matrix with a superdiagonal of 1's, $\boldsymbol{N}^0 = \boldsymbol{I}_n$, and $\boldsymbol{N}^n = \boldsymbol{0}$. Show that that the product of two upper triangular Toeplitz matrices is an upper triangular Toeplitz, upper triangular Toeplitz matrices commute, and $\boldsymbol{T}^{-1}$ exists as an upper triangular Toeplitz matrix whenever $\boldsymbol{T}$ is nonsingular.

11.11. Solutions to Selected Problems

11.17 The claim is obviously true for $m = 1$. Assume it is true for an arbitrary m. Symmetry is clear by inspection. Induction now gives

$$\begin{aligned} \boldsymbol{H}_{m+1}^2 &= \frac{1}{\sqrt{2}} \begin{pmatrix} \boldsymbol{H}_m & \boldsymbol{H}_m \\ \boldsymbol{H}_m & -\boldsymbol{H}_m \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} \boldsymbol{H}_m & \boldsymbol{H}_m \\ \boldsymbol{H}_m & -\boldsymbol{H}_m \end{pmatrix} \\ &= \frac{1}{2} \begin{pmatrix} 2\boldsymbol{H}_m^2 & \boldsymbol{H}_m^2 - \boldsymbol{H}_m^2 \\ \boldsymbol{H}_m^2 - \boldsymbol{H}_m^2 & 2\boldsymbol{H}_m^2 \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{I}_{2^m} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{2^m} \end{pmatrix}. \end{aligned}$$

11.23 If $\rho = 1$, then $\boldsymbol{C} = \boldsymbol{1}\boldsymbol{1}^*$, which has rank 1. If $\rho = -1$, then let $\boldsymbol{v}$ be the vector with jth entry $(-1)^j$. In this case $\boldsymbol{C} = \boldsymbol{v}\boldsymbol{v}^*$ also has rank 1. Otherwise, $\boldsymbol{C}^{-1}$ is symmetric and strictly diagonally dominant because $1 + \rho^2 - 2\rho = (1 - \rho)^2 > 0$. Thus, $\boldsymbol{C}$ is positive definite with positive definite inverse. To verify the suggested

12.8. Problems

(1) Discuss how you would use the inverse method to generate a random variable with (a) the continuous logistic density

$$f(x|\mu,\sigma) = \frac{e^{-\frac{x-\mu}{\sigma}}}{\sigma[1+e^{-\frac{x-\mu}{\sigma}}]^2},$$

(b) the Pareto density

$$f(x|\alpha,\beta) = \frac{\beta\alpha^\beta}{x^{\beta+1}}1_{(\alpha,\infty)}(x),$$

and (c) the Weibull density

$$f(x|\delta,\gamma) = \frac{\gamma}{\delta}x^{\gamma-1}e^{-\frac{x^\gamma}{\delta}}1_{(0,\infty)}(x),$$

where α, β, γ, δ, and σ are taken positive.

(2) Continuing problem 1, discuss how the inverse method applies to (d) the Gumbel density

$$f(x) = e^{-x}e^{-e^{-x}},$$

(e) the arcsine density

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}1_{(0,1)}(x),$$

and (f) the slash density

$$f(x) = \alpha x^{\alpha-1}1_{(0,1)}(x),$$

where $\alpha > 0$.

(3) Suppose the random variable X has distribution function $F(x)$. If U is uniformly distributed on $[0,1]$, then show that

$$Y = F^{[-1]}[UF(t)]$$

is distributed as X conditional on $X \leq t$ and that

$$Z = F^{[-1]}\{F(t)+U[1-F(t]\}$$

is distributed as X conditional on $X > t$.

(4) One can implement the inverse method even when the quantile function $F^{[-1]}(y)$ is not explicitly available. Show that Newton's method for solving $F(x) = y$ has the update

$$x_{n+1} = x_n - \frac{F(x_n)-y}{f(x_n)}$$

when $F(x)$ has density $f(x) = F'(x)$. Observe that if $F(x_n) > y$, then $x_{n+1} < x_n$, and if $F(x_n) < y$, then $x_{n+1} > x_n$. Prove that x_n approaches the solution from above if $F(x)$ is convex and from below if $F(x)$ is concave. Implement Newton's method for the standard normal distribution, and describe its behavior. Convexity or concavity is determined by $F''(x) = f'(x)$. Many distributions possess a unique mode m such that $f'(x) > 0$ for $x < m$ and $f'(x) < 0$ for $x > m$. The mode serves as a safe point for starting the Newton iterations.

(5) Describe one method for generating independent Poisson deviates and implement it in code.

(6) Describe and implement acceptance-rejection sampling for the beta distribution with parameters $\alpha > 1$ and $\beta > 1$. In this case $x^{\alpha-1}(1-x)^{\beta-1} \leq c$ for all $x \in [0,1]$ and some constant c. First, calculate the least value of c.
(7) Implement the sampling method sketched in Example 5.3 for generating multivariate normal random deviates.
(8) Based on random normal deviates, discuss how one can generates random deviates from the log normal, chi-square, F, and Student's t distributions.
(9) The transition matrix $\boldsymbol{P}$ of a finite Markov chain is said to be doubly stochastic if each of its column sums equals 1. Find an equilibrium distribution in this setting. Prove that symmetric transition matrices are doubly stochastic.
(10) The random variable Y stochastically dominates the random variable X provided $\Pr(Y \leq u) \leq \Pr(X \leq u)$ for all real u. Using quantile coupling, we can construct on a common probability space probabilistic copies X_c of X and Y_c of Y such that $X_c \leq Y_c$ with probability 1. If X has distribution function $F(x)$ and Y has distribution function $G(y)$, then define $F^{[-1]}(u)$ and $G^{[-1]}(u)$. If U is uniformly distributed on $[0,1]$, demonstrate that $X_c = F^{[-1]}(U)$ and $Y_c = G^{[-1]}(U)$ yield quantile couplings with the property $X_c \leq Y_c$.
(11) Let $Z_0, Z_1, Z_2, \ldots$ be a realization of a finite-state ergodic chain. If we sample every kth epoch, then show (a) that the sampled chain $Z_0, Z_k, Z_{2k}, \ldots$ is ergodic, (b) that it possesses the same equilibrium distribution as the original chain, and (c) that it is reversible if the original chain is. Thus, based on the ergodic theorem, we can estimate theoretical means by sample averages using only every kth epoch of the original chain.
(12) An acceptance function $a : (0,\infty) \mapsto [0,1]$ satisfies the functional identity $a(x) = xa(1/x)$. Prove that the detailed balance condition

$$\pi_i q_{ij} a_{ij} \quad = \quad \pi_j q_{ji} a_{ji}$$

holds if the acceptance probability a_{ij} is defined by

$$a_{ij} \quad = \quad a\left(\frac{\pi_j q_{ji}}{\pi_i q_{ij}}\right)$$

in terms of an acceptance function $a(x)$. Check that the Barker function $a(x) = x/(1+x)$ qualifies as an acceptance function and that any acceptance function is dominated by the Metropolis acceptance function in the sense that $a(x) \leq \min\{x, 1\}$ for all x.
(13) Consider the Cartesian product space $\{0,1\} \times \{0,1\}$ equipped with the probability distribution

$$(\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}) \quad = \quad \left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right).$$

Demonstrate that sequential Gibbs sampling does not satisfy detailed balance by showing that $\pi_{00} s_{00,11} \neq \pi_{11} s_{11,00}$, where $s_{00,11}$ and $s_{11,00}$ are entries of the matrix S for first resampling component one and then resampling component two.
(14) Implement a Metropolis-driven random walk to generate binomial deviates with n trials and success probability p. If the random walk is in state x, then it should propose states $x-1$ and $x+1$ with equal probabilities. Check that the visited states have approximate mean np and approximate variance $np(1-p)$.
(15) Implement a Metropolis-driven random walk to generate standard normal deviates. If the random walk is in state x, then it should propose to move to state

$x + c(U - \frac{1}{2})$, where $c > 0$ and U is a uniform deviate from $[0, 1]$. Check that the walk increment has mean and variance 0 and $\frac{c^2}{12}$, respectively. What value of c would you suggest?

(16) Design and implement a Gibbs sampler for generating bivariate normal deviates.

(17) It is known that every planar graph can be colored by four colors [**34**]. Design, program, and test a simulated annealing algorithm to find a four coloring of any planar graph. (Suggestions: Represent the graph by a list of nodes and a list of edges. Assign to each node a color represented by a number between 1 and 4. The cost of a coloring is the number of edges with incident nodes of the same color. In the proposal stage of the simulated annealing solution, randomly choose a node, randomly reassign its color, and recalculate the cost. If successful, simulated annealing will find a coloring with the minimum cost of 0.)

(18) A Sudoku puzzle is a 9×9 matrix, with some entries containing predefined digits. The goal is to completely fill in the matrix, using the digits 1 through 9, in such a way that each row, column, and symmetrically placed 3×3 submatrix displays each digit exactly once. In mathematical language, a completed Sudoku matrix is a Latin square subject to further constraints on the 3×3 submatrices. The initial partially filled in matrix is assumed to have a unique completion. Design, program, and test a simulated annealing algorithm to solve a Sudoku puzzle.

(19) Given a finite set of integers, the partition problem determines whether these integers can be partitioned into two subsets with the same sum. For example, if the set is $\{1, 2, 3, 5, 7\}$, then the answer is yes because $1+3+5 = 2+7$. Design, program, and test a simulated annealing algorithm to solve the partition problem.

(20) Devise a simulated annealing algorithm to approximate the maximum number of 1's that can occur in the hardcore model.

(21) Consider a graph with $2n$ nodes. The graph bisection problem involves dividing the nodes into two disjoint subsets A and B of size n such that the number of edges extending between a node in A and a node in B is as small as possible. Design and implement a simulated annealing algorithm to find a best pair of subsets. This problem has implications for the design of microchips.

(22)* Suppose X is a random variable symmetrically distributed around 0. If X has quantile function $Q(u)$, then show that the folded random variable $Y = |X|$ has quantile function $-Q[(1-u)/2]$. You may assume X is continuously distributed. In practice, this exercise extends the inverse sampling method to $|X|$ and related random variables such as X^2.

(23)* Given $\delta > 0$, generate a sequence of independent standard normal deviates X_n and define $Y_0 = X_0$ and inductively

$$Y_{n+1} \quad = \quad e^{-\delta} Y_n + \sqrt{1 - e^{-2\delta}} X_{n+1}.$$

Show that the Y_n are correlated standard normal deviates with covariances

$$\mathrm{Cov}(Y_m, Y_n) \quad = \quad e^{-|n-m|\delta}.$$

(24)* Describe and implement an algorithm for generating beta random deviates that makes use of gamma random deviates.

(25)* Given a vector $\boldsymbol{w} \in \mathbb{R}^p$ of positive weights and a total capacity $c > 0$, the knapsack problem asks for the number of vectors $\boldsymbol{x}$ with 0/1 entries satisfying $\boldsymbol{x}^*\boldsymbol{w} \leq c$. Suggest a Markov chain method for solving the knapsack problem.

(26)* An asymmetric triangular distribution has triangular-shaped probability density function concentrated on the interval $[a, b]$ with mode at $m \in (a, b)$. Find the corresponding distribution function $F(x)$, and show that inverse method generates the random deviate

$$X = \begin{cases} a + \sqrt{U(b-a)(m-a)} & 0 < U < F(m) \\ b - \sqrt{(1-U)(b-a)(b-m)} & F(m) \le U, \end{cases}$$

where U is uniform on $[0, 1]$. (Hint: The total mass of the density determines the value at the mode.)

(27)* Suppose X follows a standard normal distribution. Devise an acceptance-rejection method for sampling deviates from $Y = |X|$ based on an exponentially distributed instrumental random variable Z. What is the density $f(x)$ of Y? Find the least constant c with $ce^{-x} \ge e^{-\frac{x^2}{2}}$ for all $x \ge 0$.

(28)* Devise and program an algorithm that generates random uniform deviates from the ball $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{c}\| \le r\}$.

12.9. Solutions to Selected Problems

12.22 The quantile equation is $F(x) - F(-x) = u$. By symmetry this amounts to $1 - 2F(-x) = u$, or $F(-x) = (1-u)/2$. It now follows that

$$-x = Q[F(-x)] = Q[(1-u)/2],$$

or $x = -Q[(1-u)/2]$.

12.23 Because $Y_{n+1} = e^{-\delta} Y_n + \sqrt{1 - e^{-2\delta}} X_{n+1}$ is the sum of two independent Gaussian deviates, it is Gaussian. By induction Y_{n+1} has mean 0 and variance $e^{-2\delta} + 1 - e^{-2\delta} = 1$. Without loss of generality fix $m \le n$. By induction on $n \ge m$

$$\begin{aligned} \operatorname{Cov}(Y_m, Y_{n+1}) &= e^{-\delta} \operatorname{Cov}(Y_m, Y_n) + \sqrt{1 - e^{-2\delta}} \operatorname{Cov}(Y_m, X_{n+1}) \\ &= e^{-\delta} \operatorname{Cov}(Y_m, Y_n) \\ &= e^{-(n+1-m)\delta}. \end{aligned}$$

12.24 If X and Y are independent gamma variates with parameters $(a, 1)$ and $(b, 1)$, respectively, then the ratio $\frac{X}{X+Y}$ is beta(a, b). The code

```
using Statistics

function beta_deviate(a::T, b::T, n::Int) where T <: Real
  x = gamma_deviate(a, 1.0, n)
  y = gamma_deviate(b, 1.0, n)
  z = zeros(T, n)
  @. z = x / (x + y)
  return z
end

(a, b, n) = (3.0, 2.0, 10000);
z = beta_deviate(a, b, n);
mu = a / (a + b);
sigma2 = (a * b) / ((a + b)^2 * (a + b + 1));
println(mean(z),"  ",mu,"  ", std(z)^2,"  ", sigma2)
```

13.6. Elaborations of Neural Networks

Fitting of massive data sets has driven innovation in neural net modeling. Two of the most notable advances are stochastic gradient descent and convolutional neural nets. The former tactic implements gradient descent on a small random subset of the data at each iteration. The latter tactic addresses the issue of parameter parsimony in image analysis. Convolutional neural nets were inspired by the connectivity patterns among neurons of the animal visual cortex. Let us briefly comment on both tactics.

In stochastic gradient descent, denote the objective for case i by $f_i(\boldsymbol{\theta})$. The overall objective is often scaled by the sample size m and written $f(\boldsymbol{\theta}) = \frac{1}{m}\sum_i^m f_i(\boldsymbol{\theta})$. At iteration n one chooses a random subset (batch) of cases S_n with fixed size $|S_n| = b$. Stochastic gradient descent updates $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \frac{t_n}{b}\sum_{i\in S_i}\nabla f_i(\boldsymbol{\theta}_n)$. On average

$$\mathrm{E}\left[\frac{1}{b}\sum_{i\in S_i}\nabla f_i(\boldsymbol{\theta}_n)\right] \;=\; \nabla f(\boldsymbol{\theta}_n),$$

so the updates are unbiased relative to ordinary gradient descent. To achieve computational savings, b is taken so that $b \ll m$. These updates can be quite noisy, so the sample average iterate $\bar{\boldsymbol{\theta}}_n = \frac{1}{n}\sum_{k=1}^n \boldsymbol{\theta}_k$ is usually tracked. Noisy updates have the advantage of allowing the estimates $\boldsymbol{\theta}_n$ to escape the basins of attraction of poor local minima. The disadvantage of stochastic gradient descent is slow local convergence of $\bar{\boldsymbol{\theta}}_n$ to the optimal point. This disadvantage has prompted early stopping as a remedy.

Convolutional neural nets include the input layer where the inputs $z_{\boldsymbol{i}}$ are sent to the outputs

$$u_{\boldsymbol{i}} \;=\; \sum_{\boldsymbol{j}\in N_{\boldsymbol{i}}} w_{\boldsymbol{j}} z_{\boldsymbol{i}-\boldsymbol{j}} + b_{\boldsymbol{i}}$$

prior to passing through the activation function attached to the level. Here the neighborhood $N_{\boldsymbol{i}}$ of pixel $\boldsymbol{i} = (i_1, i_2)$ in dimension 2 is typically a small square of nearby pixels. The constant b_i is the intercept. Parameter parsimony is achieved by limiting the neighborhood size and by sharing the weights $w_{\boldsymbol{j}}$ and intercepts $b_{\boldsymbol{i}}$ across an entire square of pixels $\boldsymbol{i}$. Edge effects where $\boldsymbol{i} - \boldsymbol{j}$ falls outside the defined set of pixels are usually handled by setting $z_{\boldsymbol{i}-\boldsymbol{j}} = 0$. Pooling is another device commonly employed in convolutional neural nets. Here an entire square S of input values $z_{\boldsymbol{i}}$ is replaced by its average value $\frac{1}{|S|}\sum_{\boldsymbol{i}\in S} z_{\boldsymbol{i}}$ or its maximum value $\max_{\boldsymbol{i}\in S} z_{\boldsymbol{i}}$ within the square. Pooling of congruent non-overlapping squares obviously reduces the dimensionality of the next layer.

13.7. Problems

(1) Apply the central difference formulas (13.2) and (13.3) to approximate $f''(1)$, where $f(x) = x^{-1}$ and $\delta = 10^{-k}$ for $k = 1, \ldots, 6$. Find the approximation errors and check if the predicted error estimates are accurate.

(2) The formula $f^{(n)}(x) = \lim_{h\to 0}\frac{1}{h^n}\sum_{k=0}^n (-1)^{n-k}\binom{n}{k} f(x+kh)$ generalizes formula (13.2) to $n > 2$ and $f(x)$ sufficiently smooth. Implement this generalization in code and test it.

(3) Continuing the previous problem, demonstrate the vector identity

$$\begin{bmatrix} f(x) \\ \vdots \\ f(x+nh) \end{bmatrix} \;=\; \boldsymbol{M}_n \begin{bmatrix} f(x)\frac{h^0}{0!} \\ \vdots \\ f^{(n)}(x)\frac{h^n}{n!} \end{bmatrix} + O(h^{n+1}),$$

where the matrix $\boldsymbol{M}_n$ has entries $m_{kj} = k^j$. Here 0^j is interpreted as 1 when $j = 0$ and the row index $k = 0$. Prove that $\det \boldsymbol{M}_n = \prod_{i=1}^n i!$, implying that $\boldsymbol{M}_n$ is invertible. To verify the formula for $f^{(n)}(x)$, prove the identity $1_{\{k=n\}} = \sum_{j=0}^n (-1)^{n-j} \binom{n}{j} j^k$. This nails down the bottom row of $\boldsymbol{M}_n^{-1}$, which is the pertinent row. (Hints: See https://nhigham.com/2021/06/15/what-is-a-vandermonde-matrix/ for the determinant formula. The bottom row of $\boldsymbol{M}_n$ involves the inverse of the matrix constructed by truncating Pascal's triangle.)

(4) If the function $f(x)$ has a power series expansion $\sum_{j=0}^\infty a_j x^j$ converging in a disc $\{x : |x| < r\}$ centered at 0 in the complex plane, then we can approximate the derivatives $f^{(j)}(0) = j! a_j$ by evaluating $f(x)$ on the boundary of a small circle of radius $h < r$. This is done by expanding the function $t \to f(he^{2\pi it})$ in the Fourier series

$$f(he^{2\pi it}) \quad = \quad \sum_{j=0}^\infty a_j h^j e^{2\pi ijt}.$$

Thus, if we take the discrete Fourier transform $\hat{b}_k$ of the sequence $b_j = f(hu_n^j)$, then equation (11.9) mutates into

$$\hat{b}_k - a_k h^k \quad = \quad \sum_{l=1}^\infty a_{ln+k} h^{ln+k} \quad = \quad O\Big(h^{n+k}\Big)$$

for $0 \le k \le n-1$ under fairly mild conditions on the coefficients a_j. Rearranging this equation gives the derivative approximation

$$f^{(k)}(0) \quad = \quad k! a_k \quad = \quad \frac{k! \hat{b}_k}{h^k} + O\Big(h^n\Big)$$

highlighted by Henrici [**112**]. Implement this method in computer code and test its accuracy.

(5) Numerical differentiation can be improved by Romberg's acceleration technique. If $f(x)$ has $2k$ continuous derivatives, then the central difference formula

$$\begin{aligned} D_0(h) &= \frac{f(x+h) - f(x-h)}{2h} \\ &= \sum_{j=0}^{k-1} f^{(2j+1)}(x) \frac{h^{2j}}{(2j+1)!} + O(h^{2k-1}) \end{aligned}$$

follows from an application of Taylor's theorem. Show that the inductively defined quantities

$$\begin{aligned} D_j(h) &= \frac{4^j D_{j-1}(\frac{1}{2}h) - D_{j-1}(h)}{4^j - 1} \\ &= D_{j-1}\Big(\frac{1}{2}h\Big) - \frac{1}{4^j - 1}\Big[D_{j-1}(h) - D_{j-1}\Big(\frac{1}{2}h\Big)\Big] \end{aligned}$$

satisfy

$$f'(x) \quad = \quad D_j(h) + O(h^{2j+2})$$

for $j = 0, \ldots, k-1$. Verify that

$$D_1(h) \quad = \quad \frac{1}{6}\Big[8f\Big(x + \frac{1}{2}h\Big) - 8f\Big(x - \frac{1}{2}h\Big) - f(x+h) + f(x-h)\Big].$$

Finally, try this improvement of central differencing on a few representative functions such as $\sin(x)$, e^x, and $\ln \Gamma(x)$.

(6) Write code to evaluate $\sqrt{x}$ and its derivative based on its continued fraction expansion.

(7) The fractional linear transformations $\frac{\alpha s+\beta}{\gamma s+\delta}$ [**117**] play a role in continued fraction expansions. To avoid trivial cases where the fractional linear transformation is undefined or constant, we impose the condition that $\alpha\delta - \beta\gamma \neq 0$. The restricted set of fractional linear transformations (or Möbius functions) forms a group under functional composition. This group is the homomorphic image of the group of invertible 2×2 matrices under the correspondence

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \longrightarrow \frac{\alpha s + \beta}{\gamma s + \delta}. \tag{13.7}$$

A group homomorphism is a function between two groups that preserves the underlying algebraic operation. Show that the correspondence (13.7) qualifies as a group homomorphism in the sense that if $f_i(s) = \frac{\alpha_i s+\beta_i}{\gamma_i s+\delta_i}$ for $i = 1, 2$, then

$$\begin{pmatrix} \alpha_1 & \beta_1 \\ \gamma_1 & \delta_1 \end{pmatrix} \begin{pmatrix} \alpha_2 & \beta_2 \\ \gamma_2 & \delta_2 \end{pmatrix} \longrightarrow f_1 \circ f_2(s).$$

The homomorphism (13.7) correctly pairs the two identity elements $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $f(s) = s$ of the groups.

(8)* Show that the fractional linear transformations $\frac{\alpha s}{\gamma s+\delta}$ with $\alpha\delta = 1$ constitute a subgroup of the set of fractional linear transformations and send the numbers $r = \frac{\alpha-\delta}{\gamma}$ into themselvers. Furthermore, prove that the iterates $s_{k+1} = \frac{\alpha s_k}{\gamma s_k+\delta}$ are locally attracted to r when $|\delta| < |\alpha|$.

(9) Demonstrate that the fixed point $\sqrt{c} - r$ of the fractional linear transformation $f(y) = \frac{c-r^2}{2r+y}$ is locally attractive. This fact increases our confidence that the continued fraction expansion of $\sqrt{c} - r$ converges as expected.

(10) One can use ForwardDiff to extract second derivatives. If $f(x)$ is a real-valued function of a real variable, then the relevant Julia code is

```
y = 1.0
ForwardDiff.derivative(x -> ForwardDiff.derivative(f, x), y)
```

for the point $y = 1$. Try this on a few representative functions and points y. How accurate are the results? Generalize this code to a real-valued function of a multivariate argument.

(11) Program Search.jl to perform logistic regression. No derivatives need be input. Apply your code to the Titanic data, which can be extracted via the commands

```
using RDatasets
df = dataset("count", "titanic") # data frame
y = convert(Vector{Float64}, df[:, 1]); # responses
X = Tables.matrix(df[:, 2:4]) # cases by predictors
```

If necessary, convert the predictors to real numbers.

(12) Implement this chapter's neural net code on data of your choice with multiple hidden layers. The data should be available in a public repository reachable by Julia.

(13) Prove that the function $f(x) = \sqrt{x^2+\epsilon}$ for $\epsilon > 0$ small is majorized by the quadratic $g(x) = \sqrt{x_n^2+\epsilon} + \frac{1}{2\sqrt{x_n^2+\epsilon}}(x^2 - x_n^2)$. Plot $f(x)$ and note how well it approximates $|x|$. Finally, prove that $f(x)$ is convex.

(14) Based on the relu identity $x_+ = \frac{1}{2}(x+|x|)$, show that x_+ is well approximated by the function $f(x) = \frac{1}{2}(x + \sqrt{x^2+\epsilon})$. Furthermore, show that $f(x)$ is convex and majorized by the quadratic

$$g(x \mid x_n) \quad = \quad \frac{1}{2}\Big[x + \sqrt{x_n^2+\epsilon} + \frac{1}{2\sqrt{x_n^2+\epsilon}}(x^2 - x_n^2)\Big].$$

(15)* Continuing the two previous problems, demonstrate that

$$\begin{aligned} 0 &\leq \sqrt{x^2+\epsilon} - |x| \leq \sqrt{\epsilon} \\ 0 &\leq \frac{1}{2}(x + \sqrt{x^2+\epsilon}) - x_+ \leq \frac{1}{2}\sqrt{\epsilon} \end{aligned}$$

for all x.

(16) Decide whether the following functions are differentiable at the indicated points.

$$\begin{aligned} f(x) &= x^2 \sin\left(\frac{1}{x}\right) \quad \text{at } x = 0 \\ g(\boldsymbol{x}) &= \begin{cases} 0 & \boldsymbol{x} = \boldsymbol{0} \\ \frac{x_1^3}{x_1^2+x_2^2} & \boldsymbol{x} \neq \boldsymbol{0} \end{cases} \quad \text{at } \boldsymbol{x} = \boldsymbol{0} \\ h(\boldsymbol{x}) &= \|\boldsymbol{x}\| \quad \text{for all } \boldsymbol{x} \end{aligned}$$

If the differential exists, calculate it.

(17)* Suppose the $f(\boldsymbol{x})$ is a differentiable function satisfying the Lipschitz inequality $\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|$ for all $\boldsymbol{x}$ and $\boldsymbol{y}$. Prove that $\|df(\boldsymbol{x})\| \leq L$.

(18)* If $f(x)$ is a real-valued differentiable function on $[a, b]$, then demonstrate that $f(b) - f(a) = f'(x)(b-a)$ for some intermediate point x.

(19)* Suppose that $f(\boldsymbol{x})$ is a differentiable function from an open convex set U of $\mathbb{R}^p$ to an open set of $\mathbb{R}^n$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ belong to U, prove that there exists a scalar $s \in [0, 1]$ such that

$$\|f(\boldsymbol{y}) - f(\boldsymbol{x})\| \quad \leq \quad \|df[\boldsymbol{x} + s(\boldsymbol{y} - \boldsymbol{x})]\| \, \|\boldsymbol{y} - \boldsymbol{x}\|.$$

This is the mean-value inequality.

(20)* Program and test software that leverages automatic differentiation in projected gradient descent. Apply the software to constrained least squares.

13.8. Solutions to Selected Problems

13.8 These transformations correspond to the group of lower triangular matrices with determinant 1. This subset is the intersection of two subgroups and therefore itself constitutes a subgroup. The stated point r is a fixed point because

$$\frac{\alpha\frac{\alpha-\delta}{\gamma}}{\gamma\frac{\alpha-\delta}{\gamma} + \delta} \quad = \quad \frac{\alpha - \delta}{\gamma}.$$

Biomathematics Comprehensive Exam
Course 205 Questions

Kenneth Lange

Department of Biomathematics
UCLA School of Medicine
Los Angeles, CA 90095

August 2019

Questions

Work any combination of problems totaling 100 points.

1. (20 points) In linear regression with response vector $\boldsymbol{y}$, design matrix $\boldsymbol{X}$, and regression coefficient vector $\boldsymbol{\beta}$, find the estimate $\hat{\boldsymbol{\beta}}$, the predicted values $\hat{\boldsymbol{y}}$, the residual vector $\boldsymbol{r}$, and the residual sum of squares $\|\boldsymbol{r}\|^2$ in terms of the extended QR decomposition
$$(\boldsymbol{X}, \boldsymbol{y}) \;=\; (\boldsymbol{Q}, \boldsymbol{q}) \begin{pmatrix} \boldsymbol{R} & \boldsymbol{r} \\ \boldsymbol{0} & d \end{pmatrix}.$$

2. (20 points) Convert the problem of minimizing $|x_1| - |x_2|$ subject to $x_1 + x_2 = 5$, $2x_1 + 3x_2 - x_3 \leq 0$, and $x_3 \geq 4$ into a linear program in standard form.

3. (30 points) Show that the function
$$f(\boldsymbol{x}) \;=\; x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3 + (x_1^2 + x_2^2 + x_3^2)^2$$
is convex and possesses a global minimum at $\boldsymbol{x} = \boldsymbol{0}$.

4. (30 points) Suppose $f(\boldsymbol{x})$ is convex on $\mathbb{R}^p$ and achieves its minimum at $\boldsymbol{y}$. If $f(\boldsymbol{x})$ is also symmetric in the sense that $f(\boldsymbol{P}\boldsymbol{x}) = f(\boldsymbol{x})$ for all permutation matrices $\boldsymbol{P}$, then demonstrate that $f(\boldsymbol{x})$ possesses a symmetric minimum point $\boldsymbol{z}$ with $\boldsymbol{P}\boldsymbol{z} = \boldsymbol{z}$ for all such $\boldsymbol{P}$. (Hint: Symmetrize the point $\boldsymbol{y}$.)

5. (30 points) Suppose $f(\boldsymbol{x})$ is a convex function with domain $\mathbb{R}^p$. Derive the majorization
$$f\Big(\sum_{i=1}^m \boldsymbol{x}_i\Big) \;\leq\; \frac{1}{m}\sum_{i=1}^m f[m(\boldsymbol{x}_i - \boldsymbol{x}_{ni} + \bar{\boldsymbol{x}}_n)]$$

involving vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$, where $\bar{\boldsymbol{x}}_n = \frac{1}{m}\sum_{i=1}^m \boldsymbol{x}_{ni}$ and $\boldsymbol{x}_{ni}$ represents the value of $\boldsymbol{x}_i$ at iteration n in some optimization scheme. This majorization splits the vectors. The choices $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_*$ and $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_*^2$ involving an arbitrary norm are important in practice.

6. (30 points) Devise an acceptance-rejection method for generating beta deviates based on the inequality $x^{\alpha-1}(1-x)^{\beta-1} \leq x^{\alpha-1} + (1-x)^{\beta-1}$.

Biomathematics Comprehensive Exam
Course 205 Questions

Kenneth Lange

Department of Biomathematics
UCLA School of Medicine
Los Angeles, CA 90095

September 2020

Questions

Work any combination of problems totaling 100 points.

1. (10 points) Consider computing the number e^{-x} via the two expansions

$$\sum_{n=0}^{\infty} \frac{(-x)^n}{n!} = \frac{1}{\sum_{n=0}^{\infty} \frac{x^n}{n!}}$$

in single precision for $x > 0$. Which expansion is apt to be more accurate? Explain why?

2. (30 points) To find the square root of $c > 0$, consider the iteration scheme $x_{n+1} = f(x_n)$ with

$$f(x) = \frac{c + ax}{a + x}$$

and $a^2 > c$. Show that

$$f'(x) = \frac{a^2 - c}{(a + x)^2}$$

and that $|f'(x)| < 1$. Now explain in detail why $\sqrt{c}$ is fixed point and why x_n converges to $\sqrt{c}$ regardless of the choice of $x_0 \geq 0$. What is the local rate of convergence at the fixed point?

3. (25 points) Let $\boldsymbol{A}$ be a 2×2 symmetric matrix. Prove that $\boldsymbol{A}$ is positive semidefinite if and only if $\text{tr}(\boldsymbol{A}) \geq 0$ and $\det(\boldsymbol{A}) \geq 0$. Produce a 3×3 symmetric matrix that satisfies these two conditions but fails to be positive semidefinite.

4. (20 points) Demonstrate that $\det(\boldsymbol{I} + \boldsymbol{u}\boldsymbol{v}^*) = 1 + \boldsymbol{v}^*\boldsymbol{u}$ for $\boldsymbol{u}$ and $\boldsymbol{v}$ column vectors of the same length.

5. (25 points) Find the Euclidean projection of a point $\boldsymbol{x}$ onto a line segment $[\boldsymbol{a}, \boldsymbol{b}] \subset \mathbb{R}^n$. (Hint: This is a purely one-dimensional problem.)

6. (40 points) Registration of two images is one of the vexing problems of medical imaging. Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ represent two sets of matched anatomical landmarks in R^2. Registration can be achieved by mapping $\boldsymbol{x}_j$ into $\boldsymbol{y}_j$ for each j. Affine maps $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ are the simplest relevant maps. To estimate $\boldsymbol{A}$ and $\boldsymbol{b}$, one can minimize the objective

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X} - \boldsymbol{B}\|_F^2,$$

where the matrix $\boldsymbol{B}$ has all columns equal to $\boldsymbol{b}$, and the matrices $\boldsymbol{Y}$ and $\boldsymbol{X}$ have jth columns $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$, respectively. One can estimate parameters by block descent. Show that the minimum of the objective with respect to $\boldsymbol{b}$ for $\boldsymbol{A}$ fixed is achieved by taking $\boldsymbol{b}$ equal to the average of the columns of the matrix $\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}$. For $\boldsymbol{B}$ fixed, show that the optimal $\boldsymbol{A}$ is

$$\boldsymbol{A} \;=\; (\boldsymbol{Y} - \boldsymbol{B})\boldsymbol{X}^*(\boldsymbol{X}\boldsymbol{X}^*)^{-1},$$

assuming $\boldsymbol{X}$ has full rank. Why do these solutions represent minimum points?

Biomathematics Comprehensive Exam
Course 205 Questions

Kenneth Lange

Department of Biomathematics
UCLA School of Medicine
Los Angeles, CA 90095

August 2021

Questions

Work any combination of problems totaling 100 points.

1. (40 points) If you apply Newton's method to find a root of the function $f(y) = \frac{1}{y^{1/2}} - xy^{3/2}$ for $x > 0$, then show that the iterates are

$$y_{n+1} = y_n \left(\frac{3 + xy_n^2}{1 + 3xy_n^2} \right).$$

   What is the root of the equation $f(y) = 0$? Observe that the quantity $c = xy_n^2$ should be calculated only once per iteration. Show that the factor $\frac{3+xy_n^2}{1+3xy_n^2}$ is less than 1 if and only if $\frac{1}{\sqrt{x}} \leq y_n$. Why is this desirable? Finally, show that $\frac{1}{\sqrt{x}} \leq y_n$ if and only if $(\sqrt{x}y_n - 1)^3 \geq 0$. Use this equivalence to prove that $\frac{1}{\sqrt{x}} \leq y_n$ if and only if $\frac{1}{\sqrt{x}} \leq y_{n+1}$. Together these facts establish global convergence of the algorithm on $(0, \infty)$. The algorithm has a fast cubic rate of convergence. (Caution: This is a root finding problem, not an optimization problem.)

2. (25 points) Show that the $m \times m$ matrix $\boldsymbol{A} = a\boldsymbol{I}_m + b\mathbf{1}\mathbf{1}^*$ has the eigenvector $\mathbf{1}$ with eigenvalue $a + mb$ and $m - 1$ orthogonal eigenvectors

$$\boldsymbol{u}_i = \frac{1}{i-1} \sum_{j=1}^{i-1} (\boldsymbol{e}_j - \boldsymbol{e}_i),$$

   $i = 2, \ldots, m$, with eigenvalue a. Under what circumstances is $\boldsymbol{A}$ positive definite.

3. (25 points) Prove the squared hinge function $\max\{1-u, 0\}^2$ is majorized by $(u - u_n)^2$ at $u_n \geq 1$ and by $(1 - u)^2$ for $u_n < 1$. Draw a crude graph to visualize the problem.

4. (30 points) Prove the inequality $(x+y)^p \leq x^p + y^p$ for $p \in (0,1)$ and x and y nonnegative. (Hint: First reduce to a one-dimensional problem by dividing.)

5. (25 points) Suppose you have to minimize the convex quadratic

$$h(\boldsymbol{\beta}) \;=\; \frac{1}{2}\boldsymbol{\beta}^*\boldsymbol{C}^*\boldsymbol{C}\boldsymbol{\beta} + \boldsymbol{v}^*\boldsymbol{\beta}.$$

   How can you leverage the QR decomposition of $\boldsymbol{C}$ to solve the problem? Note that $\boldsymbol{C}$ is not necessarily square, but it should have full column rank.

6. (30 points) Suppose X is a random variable symmetrically distributed around 0. If X has quantile function $Q(u)$, then show that the folded random variable $Y = |X|$ has quantile function $-Q[(1-u)/2]$. You may assume X is continuously distributed. In practice, this exercise extends the inverse sampling method to $|X|$ and related random variables such as X^2.

Biomathematics Comprehensive Exam
Course 205 Questions

Kenneth Lange

Department of Biomathematics
UCLA School of Medicine
Los Angeles, CA 90095

August 2022

Questions

Work any combination of problems totaling 100 points.

1. (20 points) Consider the set G of real $2n \times 2n$ matrices $\boldsymbol{M}$ satisfying $\boldsymbol{M}^* \boldsymbol{A} \boldsymbol{M} = \boldsymbol{A}$, where

$$\boldsymbol{A} \;=\; \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_n \\ -\boldsymbol{I}_n & \boldsymbol{0} \end{pmatrix}.$$

   Show that G is a group under matrix multiplication and that all $\boldsymbol{M} \in G$ have $\det \boldsymbol{M} = \pm 1$. Actually, $\det \boldsymbol{M} = 1$, but this is harder to prove. (Hint: Show that $\det \boldsymbol{A} = 1$ when $n = 1$. Otherwise, assume this fact.)

2. (30 points) Let $x_1, \ldots, x_m$ be a random sample from the gamma density

$$f(x) \;=\; \Gamma(\alpha)^{-1} \beta^\alpha x^{\alpha-1} e^{-\beta x}$$

   on $(0, \infty)$. Let $\overline{x} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\overline{\ln x} = \frac{1}{m} \sum_{i=1}^m \ln x_i$. Setting the score function equal to 0 identifies $\beta = \alpha / \overline{x}$ as the maximum of the loglikelihood $L(\alpha, \beta)$ of the sample for α fixed. Substituting this value of β in the loglikelihood reduces maximum likelihood estimation to optimization of the profile loglikelihood

$$L(\alpha) \;=\; m\alpha \ln \alpha - m\alpha \ln \overline{x} - m \ln \Gamma(\alpha) + m(\alpha - 1)\overline{\ln x} - m\alpha.$$

   As an alternative to Newton's method, we approximate $L(\alpha)$ by the surrogate function $g(\alpha) = c_0 + c_1 \alpha + c_2 \ln \alpha$. (No majorization is implied.) The coefficients here are generated by matching the derivatives $g^{(k)}(\alpha_n)$ and $L^{(k)}(\alpha_n)$ at the current iterate α_n for $k = 0, 1, 2$. Show that maximizing the surrogate leads to the update

$$\frac{1}{\alpha_{n+1}} \;=\; \frac{1}{\alpha_n} + \frac{\overline{\ln x} - \ln \overline{x} + \ln \alpha_n - \Psi(\alpha_n)}{\alpha_n - \alpha_n^2 \Psi'(\alpha_n)},$$

where $\Psi(\alpha)$ is the digamma function (derivative of the log gamma function). Convergence is quick with a good starting value. (Hint: First maximize the surrogate. Note that the coefficient c_0 is irrelevant so it suffices to match first and second derivatives.)

3. (20 points) Show that the function $f(\boldsymbol{x}) = |3x_1 + 4x_2| + |2x_1 + x_2|$ is convex and attains its minimum of 0 at $\mathbf{0}$. Furthermore, show that the slices $f(x_1, 3)$ and $f(-4, x_2)$ achieve their minimal value of 5 at $x_1 = -4$ and $x_2 = 3$. Why are these facts not contradictory?

4. (25 points) Consider the function $f(x) = [x - \text{sgn}(x)w]^2 = (|x| - w)^2$ for $w > 0$. Show that it is majorized by the convex quadratics

$$f(x) \leq \begin{cases} (x-w)^2 & x_n > 0 \\ x^2 + w^2 & x_n = 0 \\ (x+w)^2 & x_n < 0. \end{cases}$$

(Hint: Graph $f(x)$.)

5. (20 points) Apply k-means clustering to the data $(x_1, x_2, x_3) = (0, 2, 3)$, and show that the initial clusters $\{0, 2\}$ and $\{3\}$ are stable and preserved by the algorithm. What is the optimal clustering into to 2 groups? Thus, k-means can converge to an inferior solution.

6. (30 points) An asymmetric triangular distribution has triangular-shaped probability density function concentrated on the interval $[a, b]$ with mode at $m \in (a, b)$. Find the corresponding distribution function $F(x)$, and show that inverse method generates the random deviate

$$X = \begin{cases} a + \sqrt{U(b-a)(m-a)} & 0 < U < F(m) \\ b - \sqrt{(1-U)(b-a)(b-m)} & F(m) \leq U, \end{cases}$$

where U is uniform on $[0, 1]$. (Hint: The total mass of the density determines the value at the mode.)

7. (30 points) Consider the problem of computing e^x given a fast algorithm for taking logarithms. Why can we restrict x to positive values? Show Newton's method iterates according to $y_{n+1} = y_n - y_n(\ln y_n - x)$. Also show that If $y_n > y_\infty$, then $y_{n+1} < y_\infty$, and if $y_n < y_\infty$, then $y_{n+1} < y_\infty$ as well. In the latter case prove that y_{n+1} is closer to y_∞ than y_n is. Prove that convergence is global whenever $y_0 \in (0, e^{1+x})$. (Hints: Apply the mean value theorem to $y_{n+1} - x$. Under what condition does $y_{n+1} > 0$?)