Additional Exercises for *Convex Optimization*

Stephen Boyd Lieven Vandenberghe

April 22, 2010

This is a collection of additional exercises, meant to supplement those found in the book *Convex Optimization*, by Stephen Boyd and Lieven Vandenberghe. These exercises were used in several courses on convex optimization, EE364a (Stanford), EE236b (UCLA), or 6.975 (MIT), usually for homework, but sometimes as exam questions. Some of the exercises were originally written for the book, but were removed at some point. Many of them include a computational component using CVX, available at `www.stanford.edu/~boyd/cvx/`. Matlab files required for these exercises can be found at the book web site `www.stanford.edu/~boyd/cvxbook/`. Some of the exercises require a knowledge of elementary analysis.

You are free to use these exercises any way you like (for example in a course you teach), provided you acknowledge the source. In turn, we gratefully acknowledge the teaching assistants (and in some cases, students) who have helped us develop and debug these exercises. Pablo Parillo helped develop some of the exercises that were originally used in 6.975.

Course instructors can obtain solutions by request to `solutions@cambridge.org`, or by email to us. In either case please specify the course you are teaching and give its URL.

We'll update this document as new exercises become available, so the exercise numbers and sections will occasionally change. We have categorized the exercises into sections that follow the book chapters, as well as various additional application areas. Some exercises fit into more than one section, or don't fit well into any section, so we have just arbitrarily assigned these.

Stephen Boyd and Lieven Vandeberghe

Contents

1 Convex sets

1.1 Is the set $\{a \in \mathbf{R}^k \mid p(0) = 1,\ |p(t)| \leq 1 \text{ for } \alpha \leq t \leq \beta\}$, where

$$p(t) = a_1 + a_2 t + \cdots + a_k t^{k-1},$$

convex?

1.2 *Set distributive characterization of convexity.* [vT84, p21], [Roc70, Theorem 3.2] Show that $C \subseteq \mathbf{R}^n$ is convex if and only if $(\alpha + \beta)C = \alpha C + \beta C$ for all nonnegative α, β.

1.3 *Composition of linear-fractional functions.* Suppose $\phi : \mathbf{R}^n \to \mathbf{R}^m$ and $\psi : \mathbf{R}^m \to \mathbf{R}^p$ are the linear-fractional functions

$$\phi(x) = \frac{Ax + b}{c^T x + d}, \qquad \psi(y) = \frac{Ey + f}{g^T y + h},$$

with domains $\mathbf{dom}\,\phi = \{x \mid c^T x + d > 0\}$, $\mathbf{dom}\,\psi = \{y \mid g^T x + h > 0\}$. We associate with ϕ and ψ the matrices

$$\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}, \qquad \begin{bmatrix} E & f \\ g^T & h \end{bmatrix},$$

respectively.

Now consider the composition Γ of ψ and ϕ, *i.e.*, $\Gamma(x) = \psi(\phi(x))$, with domain

$$\mathbf{dom}\,\Gamma = \{x \in \mathbf{dom}\,\phi \mid \phi(x) \in \mathbf{dom}\,\psi\}.$$

Show that Γ is linear-fractional, and that the matrix associated with it is the product

$$\begin{bmatrix} E & f \\ g^T & h \end{bmatrix} \begin{bmatrix} A & b \\ c^T & d \end{bmatrix}.$$

1.4 *Dual of exponential cone.* The exponential cone $K_{\text{exp}} \subseteq \mathbf{R}^3$ is defined as

$$K_{\text{exp}} = \{(x, y, z) \mid y > 0,\ ye^{x/y} \leq z\}.$$

Find the dual cone K_{exp}^*.

We are not worried here about the fine details of what happens on the boundaries of these cones, so you really needn't worry about it. But we make some comments here for those who do care about such things.

The cone K_{exp} as defined above is not closed. To obtain its closure, we need to add the points

$$\{(x, y, z) \mid x \leq 0,\ y = 0,\ z \geq 0\}.$$

(This makes no difference, since the dual of a cone is equal to the dual of its closure.)

1.5 *Dual of intersection of cones.* Let C and D be closed convex cones in $\mathbf{R}^n$. In this problem we will show that

$$(C \cap D)^* = C^* + D^*.$$

Here, $+$ denotes set addition: $C^* + D^*$ is the set $\{u + v \mid u \in C^*,\ v \in D^*\}$. In other words, the dual of the intersection of two closed convex cones is the sum of the dual cones.

(a) Show that $C \cap D$ and $C^* + D^*$ are convex cones. (In fact, $C \cap D$ and $C^* + D^*$ are closed, but we won't ask you to show this.)

(b) Show that $(C \cap D)^* \supseteq C^* + D^*$.

(c) Now let's show $(C \cap D)^* \subseteq C^* + D^*$. You can do this by first showing

$$(C \cap D)^* \subseteq C^* + D^* \iff C \cap D \supseteq (C^* + D^*)^*.$$

You can use the following result:

If K is a closed convex cone, then $K^{**} = K$.

Next, show that $C \cap D \supseteq (C^* + D^*)^*$ and conclude $(C \cap D)^* = C^* + D^*$.

(d) Show that the dual of the polyhedral cone $V = \{x \mid Ax \succeq 0\}$ can be expressed as

$$V^* = \{A^T v \mid v \succeq 0\}.$$

1.6 *Polar of a set.* The *polar* of $C \subseteq \mathbf{R}^n$ is defined as the set

$$C^\circ = \{y \in \mathbf{R}^n \mid y^T x \leq 1 \text{ for all } x \in C\}.$$

(a) Show that C° is convex (even if C is not).

(b) What is the polar of a cone?

(c) What is the polar of the unit ball for a norm $\|\cdot\|$?

(d) Show that if C is closed and convex, with $0 \in \mathbf{int}\, C$, then $(C^\circ)^\circ = C$.

2 Convex functions

2.1 *Maximum of a convex function over a polyhedron.* Show that the maximum of a convex function f over the polyhedron $\mathcal{P} = \mathbf{conv}\{v_1, \ldots, v_k\}$ is achieved at one of its vertices, *i.e.*,

$$\sup_{x \in \mathcal{P}} f(x) = \max_{i=1,\ldots,k} f(v_i).$$

(A stronger statement is: the maximum of a convex function over a closed bound convex set is achieved at an extreme point, *i.e.*, a point in the set that is not a convex combination of any other points in the set.) *Hint.* Assume the statement is false, and use Jensen's inequality.

2.2 *A general vector composition rule.* Suppose

$$f(x) = h(g_1(x), g_2(x), \ldots, g_k(x))$$

where $h : \mathbf{R}^k \to \mathbf{R}$ is convex, and $g_i : \mathbf{R}^n \to \mathbf{R}$. Suppose that for each i, one of the following holds:

- h is nondecreasing in the ith argument, and g_i is convex
- h is nonincreasing in the ith argument, and g_i is concave
- g_i is affine.

Show that f is convex. (This composition rule subsumes all the ones given in the book, and is the one used in software systems such as CVX.)

2.3 *Logarithmic barrier for the second-order cone.* The function $f(x,t) = -\log(t^2 - x^T x)$, with $\mathbf{dom}\, f = \{(x,t) \in \mathbf{R}^n \times \mathbf{R} \mid t > \|x\|_2\}$ (*i.e.*, the second-order cone), is convex. (The function f is called the logarithmic barrier function for the second-order cone.) This can be shown many ways, for example by evaluating the Hessian and demonstrating that it is positive semidefinite. In this exercise you establish convexity of f using a relatively painless method, leveraging some composition rules and known convexity of a few other functions.

(a) Explain why $t - (1/t)u^T u$ is a concave function on $\mathbf{dom}\, f$. *Hint.* Use convexity of the quadratic over linear function.

(b) From this, show that $-\log(t - (1/t)u^T u)$ is a convex function on $\mathbf{dom}\, f$.

(c) From this, show that f is convex.

2.4 *A quadratic-over-linear composition theorem.* Suppose that $f : \mathbf{R}^n \to \mathbf{R}$ is nonnegative and convex, and $g : \mathbf{R}^n \to \mathbf{R}$ is positive and concave. Show that the function f^2/g, with domain $\mathbf{dom}\, f \cap \mathbf{dom}\, g$, is convex.

2.5 *A perspective composition rule.* [Mar05] Let $f : \mathbf{R}^n \to \mathbf{R}$ be a convex function with $f(0) \leq 0$.

(a) Show that the perspective $tf(x/t)$, with domain $\{(x,t) \mid t > 0,\ x/t \in \mathbf{dom}\, f\}$, is nonincreasing as a function of t.

(b) Let g be concave and positive on its domain. Show that the function

$$h(x) = g(x) f(x/g(x)), \qquad \mathbf{dom}\, h = \{x \in \mathbf{dom}\, g \mid x/g(x) \in \mathbf{dom}\, f\}$$

is convex.

(c) As an example, show that

$$h(x) = \frac{x^T x}{(\prod_{k=1}^n x_k)^{1/n}}, \qquad \mathbf{dom}\, h = \mathbf{R}^n_{++}$$

is convex.

2.6 *Perspective of log determinant.* Show that $f(X,t) = nt \log t - t \log \det X$, with $\mathbf{dom}\, f = \mathbf{S}^n_{++} \times \mathbf{R}_{++}$, is convex in (X,t). Use this to show that

$$\begin{aligned} g(X) &= n(\mathbf{tr}\, X) \log(\mathbf{tr}\, X) - (\mathbf{tr}\, X)(\log \det X) \\ &= n \left(\sum_{i=1}^n \lambda_i \right) \left(\log \sum_{i=1}^n \lambda_i - \sum_{i=1}^n \log \lambda_i \right), \end{aligned}$$

where λ_i are the eigenvalues of X, is convex on $\mathbf{S}^n_{++}$.

2.7 *Pre-composition with a linear fractional mapping.* Suppose $f : \mathbf{R}^m \to \mathbf{R}$ is convex, and $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and $d \in \mathbf{R}$. Show that $g : \mathbf{R}^n \to \mathbf{R}$, defined by

$$g(x) = (c^T x + d) f((Ax + b)/(c^T x + d)), \qquad \mathbf{dom}\, g = \{x \mid c^T x + d > 0\},$$

is convex.

2.8 *Scalar valued linear fractional functions.* A function $f : \mathbf{R}^n \to \mathbf{R}$ is called *linear fractional* if it has the form $f(x) = (a^T x + b)/(c^T x + d)$, with $\mathbf{dom}\, f = \{x \mid c^T x + d > 0\}$. When is a linear fractional function convex? When is a linear fractional function quasiconvex?

2.9 Show that the function

$$f(x) = \frac{\|Ax - b\|_2^2}{1 - x^T x}$$

is convex on $\{x \mid \|x\|_2 < 1\}$.

2.10 *Weighted geometric mean.* The geometric mean $f(x) = (\prod_k x_k)^{1/n}$ with $\mathbf{dom}\, f = \mathbf{R}^n_{++}$ is concave, as shown on page 74. Extend the proof to show that

$$f(x) = \prod_{k=1}^n x_k^{\alpha_k}, \qquad \mathbf{dom}\, f = \mathbf{R}^n_{++}$$

is concave, where α_k are nonnegative numbers with $\sum_k \alpha_k = 1$.

2.11 *Continued fraction function.* Show that the function

$$f(x) = \cfrac{1}{x_1 - \cfrac{1}{x_2 - \cfrac{1}{x_3 - \cfrac{1}{x_4}}}}$$

defined where every denominator is positive, is convex and decreasing. (There is nothing special about $n = 4$ here; the same holds for any number of variables.)

2.12 *Circularly symmetric Huber function.* The scalar Huber function is defined as

$$f_{\rm hub}(x) = \begin{cases} (1/2)x^2 & |x| \leq 1 \\ |x| - 1/2 & |x| > 1. \end{cases}$$

This convex function comes up in several applications, including robust estimation. This problem concerns generalizations of the Huber function to $\mathbf{R}^n$. One generalization to $\mathbf{R}^n$ is given by $f_{\rm hub}(x_1) + \cdots + f_{\rm hub}(x_n)$, but this function is not circularly symmetric, *i.e.*, invariant under transformation of x by an orthogonal matrix. A generalization to $\mathbf{R}^n$ that *is* circularly symmetric is

$$f_{\rm cshub}(x) = f_{\rm hub}(\|x\|) = \begin{cases} (1/2)\|x\|_2^2 & \|x\|_2 \leq 1 \\ \|x\|_2 - 1/2 & \|x\|_2 > 1. \end{cases}$$

(The subscript stands for 'circularly symmetric Huber function'.) Show that $f_{\rm cshub}$ is convex. Find the conjugate function $f^*_{\rm cshub}$.

2.13 *Reverse Jensen inequality.* Suppose f is convex, $\lambda_1 > 0$, $\lambda_i \leq 0$, $i = 2, \ldots, k$, and $\lambda_1 + \cdots + \lambda_n = 1$, and let $x_1, \ldots, x_n \in \mathbf{dom}\, f$. Show that the inequality

$$f(\lambda_1 x_1 + \cdots + \lambda_n x_n) \geq \lambda_1 f(x_1) + \cdots + \lambda_n f(x_n)$$

always holds. *Hints.* Draw a picture for the $n = 2$ case first. For the general case, express x_1 as a convex combination of $\lambda_1 x_1 + \cdots + \lambda_n x_n$ and $x_2, \ldots, x_n$, and use Jensen's inequality.

2.14 *Monotone extension of a convex function.* Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is convex. Recall that a function $h : \mathbf{R}^n \to \mathbf{R}$ is monotone nondecreasing if $h(x) \geq h(y)$ whenever $x \succeq y$. The *monotone extension* of f is defined as

$$g(x) = \inf_{z \succeq 0} f(x + z).$$

(We will assume that $g(x) > -\infty$.) Show that g is convex and monotone nondecreasing, and satisfies $g(x) \leq f(x)$ for all x. Show that if h is any other convex function that satisfies these properties, then $h(x) \leq g(x)$ for all x. Thus, g is the maximum convex monotone underestimator of f.

Remark. For simple functions (say, on $\mathbf{R}$) it is easy to work out what g is, given f. On $\mathbf{R}^n$, it can be very difficult to work out an explicit expression for g. However, systems such as CVX can immediately handle functions such as g, defined by partial minimization.

2.15 *Circularly symmetric convex functions.* Suppose $f : \mathbf{R}^n \to \mathbf{R}$ is symmetric with respect to rotations, *i.e.*, $f(x)$ depends only on $\|x\|_2$. Show that f must have the form $f(x) = \phi(\|x\|_2)$, where $\phi : \mathbf{R} \to \mathbf{R}$ is nondecreasing and convex, with $\mathbf{dom}\, f = \mathbf{R}$. (Conversely, any function of this form is symmetric and convex, so this form characterizes such functions.)

2.16 *Infimal convolution.* Let $f_1, \ldots, f_m$ be convex functions on $\mathbf{R}^n$. Their *infimal convolution*, denoted $g = f_1 \diamond \cdots \diamond f_m$ (several other notations are also used), is defined as

$$g(x) = \inf\{f_1(x_1) + \cdots + f_m(x_m) \mid x_1 + \cdots + x_m = x\},$$

with the natural domain (*i.e.*, defined by $g(x) < \infty$). In one simple interpretation, $f_i(x_i)$ is the cost for the ith firm to produce a mix of products given by x_i; $g(x)$ is then the optimal cost obtained

if the firms can freely exchange products to produce, all together, the mix given by x. (The name 'convolution' presumably comes from the observation that if we replace the sum above with the product, and the infimum above with integration, then we obtain the normal convolution.)

(a) Show that g is convex.

(b) Show that $g^* = f_1^* + \cdots + f_m^*$. In other words, the conjugate of the infimal convolution is the sum of the conjugates.

2.17 *Conjugate of composition of convex and linear function.* Suppose $A \in \mathbf{R}^{m\times n}$ with $\mathbf{rank}\, A = m$, and g is defined as $g(x) = f(Ax)$, where $f : \mathbf{R}^m \to \mathbf{R}$ is convex. Show that

$$g^*(y) = f^*((A^\dagger)^T y), \qquad \mathbf{dom}(g^*) = A^T \mathbf{dom}(f^*),$$

where $A^\dagger = (AA^T)^{-1}A$ is the pseudo-inverse of A. (This generalizes the formula given on page 95 for the case when A is square and invertible.)

2.18 [RV73, p104] Suppose $\lambda_1, \ldots, \lambda_n$ are positive. Show that the function $f : \mathbf{R}^n \to \mathbf{R}$, given by

$$f(x) = \prod_{i=1}^n (1 - e^{-x_i})^{\lambda_i},$$

is concave on

$$\mathbf{dom}\, f = \left\{ x \in \mathbf{R}^n_{++} \;\middle|\; \sum_{i=1}^n \lambda_i e^{-x_i} \leq 1 \right\}.$$

Hint. The Hessian is given by

$$\nabla^2 f(x) = f(x)\left(-yy^T + \mathbf{diag}(z)\right)$$

where $y_i = \lambda_i e^{-x_i}/(1 - e^{-x_i})$ and $z_i = y_i/(1 - e^{-x_i})$.

2.19 *Majorization and symmetric functions of eigenvalue.* We use $x_{[k]}$ to denote the kth largest element of a vector $x \in \mathbf{R}^n$: $x_{[1]}$, $x_{[2]}$, …, $x_{[n]}$ are the elements of x sorted in decreasing order. We say that a vector $y \in \mathbf{R}^n$ *majorizes* a vector $x \in \mathbf{R}^n$ if

$$\begin{aligned}
x_{[1]} &\leq y_{[1]} \\
x_{[1]} + x_{[2]} &\leq y_{[1]} + y_{[2]} \\
x_{[1]} + x_{[2]} + x_{[3]} &\leq y_{[1]} + y_{[2]} + y_{[3]} \\
&\vdots \\
x_{[1]} + x_{[2]} + \cdots + x_{[n-1]} &\leq y_{[1]} + y_{[2]} + \cdots + y_{[n-1]} \\
x_{[1]} + x_{[2]} + \cdots + x_{[n]} &= y_{[1]} + y_{[2]} + \cdots + y_{[n]}.
\end{aligned}$$

(Roughly speaking, the largest entry in y exceeds the largest entry in x, the sum of the two largest entries in y exceeds the sum of the largest two entries in x, etc.)

(a) It can be shown that y majorizes x if and only if there exists a *doubly stochastic matrix* P such that $x = Py$. A doubly stochastic matrix is a matrix with nonnegative elements and columns and rows that add up to one:

$$P_{ij} \geq 0, \quad i,j = 1,\ldots,n, \qquad P\mathbf{1} = \mathbf{1}, \qquad P^T\mathbf{1} = \mathbf{1}.$$

Use this characterization to show the following: If $f : \mathbf{R} \to \mathbf{R}$ is a convex function and y majorizes x, then

$$\sum_{i=1}^{n} f(x_i) \leq \sum_{i=1}^{n} f(y_i).$$

(b) We use the notation $\lambda_k(X)$ for the kth largest eigenvalue of a matrix $X \in \mathbf{S}^n$, so $\lambda_1(X)$, $\ldots$, $\lambda_n(X)$ are the eigenvalues of X sorted in decreasing order. Let r be an integer in $\{1, 2, \ldots, n\}$. Show that

$$\lambda_1(X) + \cdots + \lambda_r(X) = \sup \{\mathbf{tr}(XZ) \mid Z \in \mathbf{S}^n,\, 0 \preceq Z \preceq I,\, \mathbf{tr}\, Z = r\}. \tag{1}$$

What does this tell us about the convexity properties of the function $g(X) = \lambda_1(X) + \cdots + \lambda_r(X)$ (the sum of the largest r eigenvalues of X)?

Hint. Use the eigenvalue decomposition of X to reduce the maximization in (1) to a simple linear program.

(c) Let $X = \theta U + (1 - \theta)V$ be a convex combination of two matrices U, $V \in \mathbf{S}^n$. Use the results of part (b) to show that the vector

$$\theta \begin{bmatrix} \lambda_1(U) \\ \lambda_2(U) \\ \vdots \\ \lambda_n(U) \end{bmatrix} + (1 - \theta) \begin{bmatrix} \lambda_1(V) \\ \lambda_2(V) \\ \vdots \\ \lambda_n(V) \end{bmatrix}$$

majorizes the vector $(\lambda_1(X), \lambda_2(X), \ldots, \lambda_n(X))$.

(d) Combine the results of parts (a) and (c) to show that if $f : \mathbf{R} \to \mathbf{R}$ is convex, then the function $h : \mathbf{S}^n \to \mathbf{R}$ defined as

$$h(X) = \sum_{i=1}^{n} f(\lambda_i(X))$$

is convex.

For example, by taking $f(x) = x \log x$, we can conclude that the function $h(X) = \sum_{i=1}^{n} \lambda_i(X) \log(\lambda_i(X))$ is convex on $\mathbf{S}^n_{++}$. This function arises in quantum information theory where it is known as the (negative) Von Neumann entropy. For diagonal $X = \mathbf{diag}(x)$, it reduces to the negative Shannon entropy $\sum_i x_i \log x_i$.

3 Convex optimization problems

3.1 *Minimizing a function over the probability simplex.* Find simple necessary and sufficient conditions for $x \in \mathbf{R}^n$ to minimize a differentiable convex function f over the probability simplex, $\{x \mid \mathbf{1}^T x = 1,\ x \succeq 0\}$.

3.2 *'Hello World' in CVX.* Use CVX to verify the optimal values you obtained (analytically) for exercise 4.1 in *Convex Optimization*.

3.3 *Reformulating constraints in CVX.* Each of the following CVX code fragments describes a convex constraint on the scalar variables `x`, `y`, and `z`, but violates the CVX rule set, and so is invalid. Briefly explain why each fragment is invalid. Then, rewrite each one in an equivalent form that conforms to the CVX rule set. In your reformulations, you can use linear equality and inequality constraints, and inequalities constructed using CVX functions. You can also introduce additional variables, or use LMIs. Be sure to explain (briefly) why your reformulation is equivalent to the original constraint, if it is not obvious.

Check your reformulations by creating a small problem that includes these constraints, and solving it using CVX. Your test problem doesn't have to be feasible; it's enough to verify that CVX processes your constraints without error.

Remark. This *looks* like a problem about 'how to use CVX software', or 'tricks for using CVX'. But it really checks whether you understand the various composition rules, convex analysis, and constraint reformulation rules.

(a) `norm([x + 2*y, x - y]) == 0`

(b) `square(square(x + y)) <= x - y`

(c) `1/x + 1/y <= 1; x >= 0; y >= 0`

(d) `norm([max(x,1), max(y,2)]) <= 3*x + y`

(e) `x*y >= 1; x >= 0; y >= 0`

(f) `(x + y)^2/sqrt(y) <= x - y + 5`

(g) `x^3 + y^3 <= 1; x >= 0; y >= 0`

(h) `x + z <= 1 + sqrt(x*y - z^2); x >= 0; y >= 0`

3.4 *Optimal activity levels.* Solve the optimal activity level problem described in exercise 4.17 in *Convex Optimization*, for the instance with problem data

$$A = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 0 & 0 & 3 & 1 \\ 0 & 3 & 1 & 1 \\ 2 & 1 & 2 & 5 \\ 1 & 0 & 3 & 2 \end{bmatrix}, \qquad c^{\max} = \begin{bmatrix} 100 \\ 100 \\ 100 \\ 100 \\ 100 \end{bmatrix}, \qquad p = \begin{bmatrix} 3 \\ 2 \\ 7 \\ 6 \end{bmatrix}, \qquad p^{\text{disc}} = \begin{bmatrix} 2 \\ 1 \\ 4 \\ 2 \end{bmatrix}, \qquad q = \begin{bmatrix} 4 \\ 10 \\ 5 \\ 10 \end{bmatrix}.$$

You can do this by forming the LP you found in your solution of exercise 4.17, or more directly, using CVX. Give the optimal activity levels, the revenue generated by each one, and the total revenue generated by the optimal solution. Also, give the average price per unit for each activity level, *i.e.*, the ratio of the revenue associated with an activity, to the activity level. (These numbers should

be between the basic and discounted prices for each activity.) Give a *very brief* story explaining, or at least commenting on, the solution you find.

3.5 *Minimizing the ratio of convex and concave piecewise-linear functions.* We consider the problem

$$\begin{array}{ll} \text{minimize} & \dfrac{\max_{i=1,\ldots,m}(a_i^T x + b_i)}{\min_{i=1,\ldots,p}(c_i^T x + d_i)} \\ \text{subject to} & Fx \preceq g, \end{array}$$

with variable $x \in \mathbf{R}^n$. We assume that $c_i^T x + d_i > 0$ for all x satisfying $Fx \preceq g$, and that the problem is feasible. This problem is quasiconvex, and can be solved using bisection, with each iteration involving a feasibility LP. Show how this problem can be solved by solving *one* LP, using a trick similar to one described in §4.3.2.

3.6 *Two problems involving two norms.* We consider the problem

$$\text{minimize} \quad \frac{\|Ax - b\|_1}{1 - \|x\|_\infty}, \tag{2}$$

and the very closely related problem

$$\text{minimize} \quad \frac{\|Ax - b\|_1^2}{1 - \|x\|_\infty}. \tag{3}$$

In both problems, the variable is $x \in \mathbf{R}^n$, and the data are $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$. Note that the only difference between problem (2) and (3) is the square in the numerator. In both problems, the constraint $\|x\|_\infty < 1$ is implicit. You can assume that $b \notin \mathcal{R}(A)$, in which case the constraint $\|x\|_\infty < 1$ can be replaced with $\|x\|_\infty \leq 1$.

Answer the following two questions, for each of the two problems. (So you will answer four questions all together.)

(a) Is the problem, exactly as stated (and for all problem data), convex? If not, is it quasiconvex? Justify your answer.

(b) Explain how to solve the problem. Your method can involve an SDP solver, an SOCP solver, an LP solver, or any combination. You can include a one-parameter bisection, if necessary. (For example, you can solve the problem by bisection on a parameter, where each iteration consists of solving an SOCP feasibility problem.)

Give the best method you can. In judging best, we use the following rules:

- *Bisection methods are worse than 'one-shot' methods.* Any method that solves the problem above by solving *one* LP, SOCP, or SDP problem is better than any method that uses a one-parameter bisection. In other words, use a bisection method only if you cannot find a 'one-shot' method.
- *Use the simplest solver needed to solve the problem.* We consider an LP solver to be simpler than an SOCP solver, which is considered simpler than an SDP solver. Thus, a method that uses an LP solver is better than a method that uses an SOCP solver, which in turn is better than a method that uses an SDP solver.

3.7 *The illumination problem.* In lecture 1 we encountered the function

$$f(p) = \max_{i=1,\ldots,n} |\log a_i^T p - \log I_{\rm des}|$$

where $a_i \in \mathbf{R}^m$, and $I_{\rm des} > 0$ are given, and $p \in \mathbf{R}^m_+$.

(a) Show that $\exp f$ is convex on $\{p \mid a_i^T p > 0,\ i = 1, \ldots, n\ \}$.

(b) Show that the constraint 'no more than half of the total power is in any 10 lamps' is convex (*i.e.*, the set of vectors p that satisfy the constraint is convex).

(c) Show that the constraint 'no more than half of the lamps are on' is (in general) *not* convex.

3.8 *Schur complements and LMI representation.* Recognizing Schur complements (see §A5.5) often helps to represent nonlinear convex constraints as linear matrix inequalities (LMIs). Consider the function

$$f(x) = (Ax + b)^T (P_0 + x_1 P_1 + \cdots + x_n P_n)^{-1} (Ax + b)$$

where $A \in \mathbf{R}^{m\times n}$, $b \in \mathbf{R}^m$, and $P_i = P_i^T \in \mathbf{R}^{m\times m}$, with domain

$$\mathbf{dom}\, f = \{x \in \mathbf{R}^n \mid P_0 + x_1 P_1 + \cdots + x_n P_n \succ 0\}.$$

This is the composition of the matrix fractional function and an affine mapping, and so is convex. Give an LMI representation of $\mathbf{epi}\, f$. That is, find a symmetric matrix $F(x,t)$, affine in (x,t), for which

$$x \in \mathbf{dom}\, f, \quad f(x) \le t \qquad \Longleftrightarrow \qquad F(x,t) \succeq 0.$$

Remark. LMI representations, such as the one you found in this exercise, can be directly used in software systems such as CVX.

3.9 *Complex least-norm problem.* We consider the complex least ℓ_p-norm problem

$$\begin{array}{ll} \text{minimize} & \|x\|_p \\ \text{subject to} & Ax = b, \end{array}$$

where $A \in \mathbf{C}^{m\times n}$, $b \in \mathbf{C}^m$, and the variable is $x \in \mathbf{C}^n$. Here $\|\cdot\|_p$ denotes the ℓ_p-norm on $\mathbf{C}^n$, defined as

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$

for $p \ge 1$, and $\|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$. We assume A is full rank, and $m < n$.

(a) Formulate the complex least ℓ_2-norm problem as a least ℓ_2-norm problem with real problem data and variable. *Hint.* Use $z = (\Re x, \Im x) \in \mathbf{R}^{2n}$ as the variable.

(b) Formulate the complex least ℓ_∞-norm problem as an SOCP.

(c) Solve a random instance of both problems with $m = 30$ and $n = 100$. To generate the matrix A, you can use the Matlab command `A = randn(m,n) + i*randn(m,n)`. Similarly, use `b = randn(m,1) + i*randn(m,1)` to generate the vector b. Use the Matlab command `scatter` to plot the optimal solutions of the two problems on the complex plane, and comment (briefly) on what you observe. You can solve the problems using the CVX functions `norm(x,2)` and `norm(x,inf)`, which are overloaded to handle complex arguments. To utilize this feature, you will need to declare variables to be `complex` in the `variable` statement. (In particular, you do not have to manually form or solve the SOCP from part (b).)

3.10 *Linear programming with random cost vector.* We consider the linear program

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b. \end{array}$$

Here, however, the cost vector c is random, normally distributed with mean $\mathbf{E}\, c = c_0$ and covariance $\mathbf{E}(c - c_0)(c - c_0)^T = \Sigma$. ($A$, b, and x are deterministic.) Thus, for a given $x \in \mathbf{R}^n$, the cost $c^T x$ is a (scalar) Gaussian variable.

We can attach several different meanings to the goal 'minimize $c^T x$'; we explore some of these below.

(a) How would you minimize the expected cost $\mathbf{E}\, c^T x$ subject to $Ax \preceq b$?

(b) In general there is a tradeoff between small expected cost and small cost variance. One way to take variance into account is to minimize a linear combination

$$\mathbf{E}\, c^T x + \gamma\, \mathbf{var}(c^T x) \tag{4}$$

of the expected value $\mathbf{E}\, c^T x$ and the variance $\mathbf{var}(c^T x) = \mathbf{E}(c^T x)^2 - (\mathbf{E}\, c^T x)^2$. This is called the 'risk-sensitive cost', and the parameter $\gamma \geq 0$ is called the *risk-aversion parameter*, since it sets the relative values of cost variance and expected value. (For $\gamma > 0$, we are willing to tradeoff an increase in expected cost for a decrease in cost variance). How would you minimize the risk-sensitive cost? Is this problem a convex optimization problem? Be as specific as you can.

(c) We can also minimize the risk-sensitive cost, but with $\gamma < 0$. This is called 'risk-seeking'. Is this problem a convex optimization problem?

(d) Another way to deal with the randomness in the cost $c^T x$ is to formulate the problem as

$$\begin{array}{ll} \text{minimize} & \beta \\ \text{subject to} & \mathbf{prob}(c^T x \geq \beta) \leq \alpha \\ & Ax \preceq b. \end{array}$$

Here, α is a fixed parameter, which corresponds roughly to the reliability we require, and might typically have a value of 0.01. Is this problem a convex optimization problem? Be as specific as you can. Can you obtain risk-seeking by choice of α? Explain.

3.11 Formulate the following optimization problems as semidefinite programs. The variable is $x \in \mathbf{R}^n$; $F(x)$ is defined as

$$F(x) = F_0 + x_1 F_1 + x_2 F_2 + \cdots + x_n F_n$$

with $F_i \in \mathbf{S}^m$. The domain of f in each subproblem is $\mathbf{dom}\, f = \{x \in \mathbf{R}^n \mid F(x) \succ 0\}$.

(a) Minimize $f(x) = c^T F(x)^{-1} c$ where $c \in \mathbf{R}^m$.

(b) Minimize $f(x) = \max_{i=1,\ldots,K} c_i^T F(x)^{-1} c_i$ where $c_i \in \mathbf{R}^m$, $i = 1, \ldots, K$.

(c) Minimize $f(x) = \sup_{\|c\|_2 \leq 1} c^T F(x)^{-1} c$.

(d) Minimize $f(x) = \mathbf{E}(c^T F(x)^{-1} c)$ where c is a random vector with mean $\mathbf{E}\, c = \bar{c}$ and covariance $\mathbf{E}(c - \bar{c})(c - \bar{c})^T = S$.

3.12 *A matrix fractional function.*[And79] Show that $X = B^T A^{-1} B$ solves the SDP

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, X \\ \text{subject to} & \left[\begin{array}{cc} A & B \\ B^T & X \end{array}\right] \succeq 0, \end{array}$$

with variable $X \in \mathbf{S}^n$, where $A \in \mathbf{S}^m_{++}$ and $B \in \mathbf{R}^{m\times n}$ are given.

Conclude that $\mathbf{tr}(B^T A^{-1} B)$ is a convex function of (A, B), for A positive definite.

3.13 *Trace of harmonic mean of matrices.* [And79] The matrix $H(A, B) = 2(A^{-1} + B^{-1})^{-1}$ is known as the *harmonic mean* of positive definite matrices A and B. Show that $X = (1/2)H(A, B)$ solves the SDP

$$\begin{array}{ll} \text{maximize} & \mathbf{tr}\, X \\ \text{subject to} & \left[\begin{array}{cc} X & X \\ X & X \end{array}\right] \preceq \left[\begin{array}{cc} A & 0 \\ 0 & B \end{array}\right], \end{array}$$

with variable $X \in \mathbf{S}^n$. The matrices $A \in \mathbf{S}^n_{++}$ and $B \in \mathbf{S}^n_{++}$ are given. Conclude that the function $\mathbf{tr}\left((A^{-1} + B^{-1})^{-1}\right)$, with domain $\mathbf{S}^n_{++} \times \mathbf{S}^n_{++}$, is concave.

Hint. Verify that the matrix

$$R = \left[\begin{array}{cc} A^{-1} & I \\ B^{-1} & -I \end{array}\right]$$

is nonsingular. Then apply the congruence transformation defined by R to the two sides of matrix inequality in the SDP, to obtain an equivalent inequality

$$R^T \left[\begin{array}{cc} X & X \\ X & X \end{array}\right] R \preceq R^T \left[\begin{array}{cc} A & 0 \\ 0 & B \end{array}\right] R.$$

3.14 *Trace of geometric mean of matrices.* [And79]

$$G(A, B) = A^{1/2} \left(A^{-1/2} B A^{-1/2}\right)^{1/2} A^{1/2}$$

is known as the *geometric mean* of positive definite matrices A and B. Show that $X = G(A, B)$ solves the SDP

$$\begin{array}{ll} \text{maximize} & \mathbf{tr}\, X \\ \text{subject to} & \left[\begin{array}{cc} A & X \\ X & B \end{array}\right] \succeq 0. \end{array}$$

The variable is $X \in \mathbf{S}^n$. The matrices $A \in \mathbf{S}^n_{++}$ and $B \in \mathbf{S}^n_{++}$ are given.

Conclude that the function $\mathbf{tr}\, G(A, B)$ is concave, for $A,\ B$ positive definite.

Hint. The symmetric matrix square root is monotone: if U and V are positive semidefinite with $U \preceq V$ then $U^{1/2} \preceq V^{1/2}$.

3.15 *Transforming a standard form convex problem to conic form.* In this problem we show that any convex problem can be cast in conic form, provided some technical conditions hold. We start with a standard form convex problem with linear objective (without loss of generality):

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b, \end{array}$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ are convex, and $x \in \mathbf{R}^n$ is the variable. For simplicity, we will assume that $\mathbf{dom}\, f_i = \mathbf{R}^n$ for each i.

Now introduce a new scalar variable $t \in \mathbf{R}$ and form the convex problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & tf_i(x/t) \leq 0, \quad i = 1, \ldots, m, \\ & Ax = b, \quad t = 1. \end{array}$$

Define

$$K = \mathbf{cl}\{(x,t) \in \mathbf{R}^{n+1} \mid tf_i(x/t) \leq 0,\ i = 1, \ldots, m,\ t > 0\}.$$

Then our original problem can be expressed as

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & (x,t) \in K, \\ & Ax = b, \quad t = 1. \end{array}$$

This is a conic problem when K is proper.

You will relate some properties of the original problem to K.

(a) Show that K is a convex cone. (It is closed by definition, since we take the closure.)

(b) Suppose the original problem is strictly feasible, *i.e.*, there exists a point $\bar{x}$ with $f_i(x) < 0$, $i = 1, \ldots, m$. (This is called Slater's condition.) Show that K has nonempty interior.

(c) Suppose that the inequalities define a bounded set, *i.e.*, $\{x \mid f_i(x) \leq 0,\ i = 1, \ldots, m\}$ is bounded. Show that K is pointed.

3.16 *Exploring nearly optimal points.* An optimization algorithm will find *an* optimal point for a problem, provided the problem is feasible. It is often useful to explore the set of nearly optimal points. When a problem has a 'strong minimum', the set of nearly optimal points is small; all such points are close to the original optimal point found. At the other extreme, a problem can have a 'soft minimum', which means that there are many points, some quite far from the original optimal point found, that are feasible and have nearly optimal objective value. In this problem you will use a typical method to explore the set of nearly optimal points.

We start by finding the optimal value $p^\star$ of the given problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array}$$

as well as an optimal point $x^\star \in \mathbf{R}^n$. We then pick a small positive number ϵ, and a vector $c \in \mathbf{R}^n$, and solve the problem

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \\ & f_0(x) \leq p^\star + \epsilon. \end{array}$$

Note that any feasible point for this problem is ϵ-suboptimal for the original problem. Solving this problem multiple times, with different c's, will generate (perhaps different) ϵ-suboptimal points. If

the problem has a strong minimum, these points will all be close to each other; if the problem has a weak minimum, they can be quite different.

There are different strategies for choosing c in these experiments. The simplest is to choose the c's randomly; another method is to choose c to have the form $\pm e_i$, for $i = 1, \ldots, n$. (This method gives the 'range' of each component of x, over the ϵ-suboptimal set.)

You will carry out this method for the following problem, to determine whether it has a strong minimum or a weak minimum. You can generate the vectors c randomly, with enough samples for you to come to your conclusion. You can pick $\epsilon = 0.01p^\star$, which means that we are considering the set of 1% suboptimal points.

The problem is a minimum fuel optimal control problem for a vehicle moving in $\mathbf{R}^2$. The position at time kh is given by $p(k) \in \mathbf{R}^2$, and the velocity by $v(k) \in \mathbf{R}^2$, for $k = 1, \ldots, K$. Here $h > 0$ is the sampling period. These are related by the equations

$$p(k+1) = p(k) + hv(k), \quad v(k+1) = (1-\alpha)v(k) + (h/m)f(k), \quad k = 1, \ldots, K-1,$$

where $f(k) \in \mathbf{R}^2$ is the force applied to the vehicle at time kh, $m > 0$ is the vehicle mass, and $\alpha \in (0, 1)$ models drag on the vehicle; in the absense of any other force, the vehicle velocity decreases by the factor $1 - \alpha$ in each discretized time interval. (These formulas are approximations of more accurate formulas that involve matrix exponentials.)

The force comes from two thrusters, and from gravity:

$$f(k) = \begin{bmatrix} \cos\theta_1 \\ \sin\theta_1 \end{bmatrix} u_1(k) + \begin{bmatrix} \cos\theta_2 \\ \sin\theta_2 \end{bmatrix} u_2(k) + \begin{bmatrix} 0 \\ -mg \end{bmatrix}, \quad k = 1, \ldots, K-1.$$

Here $u_1(k) \in \mathbf{R}$ and $u_2(k) \in \mathbf{R}$ are the (nonnegative) thruster force magnitudes, θ_1 and θ_2 are the directions of the thrust forces, and $g = 10$ is the constant acceleration due to gravity.

The total fuel use is

$$F = \sum_{k=1}^{K-1} \left(u_1(k) + u_2(k) \right).$$

(Recall that $u_1(k) \geq 0$, $u_2(k) \geq 0$.)

The problem is to minimize fuel use subject to the initial condition $p(1) = 0$, $v(1) = 0$, and the way-point constraints

$$p(k_i) = w_i, \quad i = 1, \ldots, M.$$

(These state that at the time hk_i, the vehicle must pass through the location $w_i \in \mathbf{R}^2$.) In addition, we require that the vehicle should remain in a square operating region,

$$\|p(k)\|_\infty \leq P^{\max}, \quad k = 1, \ldots, K.$$

Both parts of this problem concern the specific problem instance with data given in `thrusters_data.m`.

(a) Find an optimal trajectory, and the associated minimum fuel use $p^\star$. Plot the trajectory $p(k)$ in $\mathbf{R}^2$ (*i.e.*, in the p_1, p_2 plane). Verify that it passes through the way-points.

(b) Generate several 1% suboptimal trajectories using the general method described above, and plot the associated trajectories in $\mathbf{R}^2$. Would you say this problem has a strong minimum, or a weak minimum?

3.17 *Minimum fuel optimal control.* Solve the minimum fuel optimal control problem described in exercise 4.16 of *Convex Optimization*, for the instance with problem data

$$A = \begin{bmatrix} -1 & 0.4 & 0.8 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ 0.3 \end{bmatrix}, \qquad x_{\rm des} = \begin{bmatrix} 7 \\ 2 \\ -6 \end{bmatrix}, \qquad N = 30.$$

You can do this by forming the LP you found in your solution of exercise 4.16, or more directly using CVX. Plot the actuator signal $u(t)$ as a function of time t.

3.18 *Heuristic suboptimal solution for Boolean LP.* This exercise builds on exercises 4.15 and 5.13 in *Convex Optimization*, which involve the Boolean LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \\ & x_i \in \{0,1\}, \quad i = 1, \ldots, n, \end{array}$$

with optimal value $p^\star$. Let $x^{\rm rlx}$ be a solution of the LP relaxation

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \preceq b \\ & 0 \preceq x \preceq \mathbf{1}, \end{array}$$

so $L = c^T x^{\rm rlx}$ is a lower bound on $p^\star$. The relaxed solution $x^{\rm rlx}$ can also be used to guess a Boolean point $\hat{x}$, by rounding its entries, based on a threshold $t \in [0,1]$:

$$\hat{x}_i = \begin{cases} 1 & x_i^{\rm rlx} \geq t \\ 0 & \text{otherwise}, \end{cases}$$

for $i = 1, \ldots, n$. Evidently $\hat{x}$ is Boolean (*i.e.*, has entries in $\{0,1\}$). If it is feasible for the Boolean LP, *i.e.*, if $A\hat{x} \preceq b$, then it can be considered a guess at a good, if not optimal, point for the Boolean LP. Its objective value, $U = c^T \hat{x}$, is an upper bound on $p^\star$. If U and L are close, then $\hat{x}$ is nearly optimal; specifically, $\hat{x}$ cannot be more than $(U - L)$-suboptimal for the Boolean LP.

This rounding need not work; indeed, it can happen that for all threshold values, $\hat{x}$ is infeasible. But for some problem instances, it can work well.

Of course, there are many variations on this simple scheme for (possibly) constructing a feasible, good point from $x^{\rm rlx}$.

Finally, we get to the problem. Generate problem data using

```
rand('state',0);
n=100;
m=300;
A=rand(m,n);
b=A*ones(n,1)/2;
c=-rand(n,1);
```

You can think of x_i as a job we either accept or decline, and $-c_i$ as the (positive) revenue we generate if we accept job i. We can think of $Ax \preceq b$ as a set of limits on m resources. A_{ij}, which is positive, is the amount of resource i consumed if we accept job j; b_i, which is positive, is the amount of resource i available.

Find a solution of the relaxed LP and examine its entries. Note the associated lower bound L. Carry out threshold rounding for (say) 100 values of t, uniformly spaced over $[0, 1]$. For each value of t, note the objective value $c^T\hat{x}$ and the maximum constraint violation $\max_i(A\hat{x} - b)_i$. Plot the objective value and the maximum violation versus t. Be sure to indicate on the plot the values of t for which $\hat{x}$ is feasible, and those for which it is not.

Find a value of t for which $\hat{x}$ is feasible, and gives minimum objective value, and note the associated upper bound U. Give the gap $U - L$ between the upper bound on $p^\star$ and the lower bound on $p^\star$. If you define vectors `obj` and `maxviol`, you can find the upper bound as `U=min(obj(find(maxviol<=0)))`.

3.19 *Optimal operation of a hybrid vehicle.* Solve the instance of the hybrid vehicle operation problem described in exercise 4.65 in *Convex Optimization*, with problem data given in the file `hybrid_veh_data.m`, and fuel use function $F(p) = p + \gamma p^2$ (for $p \geq 0$).

Hint. You will actually formulate and solve a *relaxation* of the original problem. You may find that some of the equality constraints you relaxed to inequality constraints do not hold for the solution found. This is not an error: it just means that there is no incentive (in terms of the objective) for the inequality to be tight. You can fix this in (at least) two ways. One is to go back and adjust certain variables, without affecting the objective and maintaining feasibility, so that the relaxed constraints hold with equality. Another simple method is to add to the objective a term of the form

$$\epsilon \sum_{t=1}^{T} \max\{0, -P_{\mathrm{mg}}(t)\},$$

where ϵ is small and positive. This makes it more attractive to use the brakes to extract power from the wheels, even when the battery is (or will be) full (which removes any fuel incentive).

Find the optimal fuel consumption, and compare to the fuel consumption with a non-hybrid version of the same vehicle (*i.e.*, one without a battery). Plot the braking power, engine power, motor/generator power, and battery energy versus time.

How would you use optimal dual variables for this problem to find $\partial F_{\mathrm{total}}/\partial E_{\mathrm{batt}}^{\max}$, *i.e.*, the partial derivative of optimal fuel consumption with respect to battery capacity? (You can just assume that this partial derivative exists.) You do not have to give a long derivation or proof; you can just state how you would fine this derivative from optimal dual variables for the problem. Verify your method numerically, by changing the battery capacity a small amount and re-running the optimization, and comparing this to the prediction made using dual variables.

3.20 *Optimal vehicle speed scheduling.* A vehicle (say, an airplane) travels along a fixed path of n segments, between $n + 1$ waypoints labeled $0, \ldots, n$. Segment i starts at waypoint $i - 1$ and terminates at waypoint i. The vehicle starts at time $t = 0$ at waypoint 0. It travels over each segment at a constant (nonnegative) speed; s_i is the speed on segment i. We have lower and upper limits on the speeds: $s^{\min} \preceq s \preceq s^{\max}$. The vehicle does not stop at the waypoints; it simply proceeds to the next segment. The travel distance of segment i is d_i (which is positive), so the travel time over segment i is d_i/s_i. We let τ_i, $i = 1, \ldots, n$, denote the time at which the vehicle

arrives at waypoint i. The vehicle is required to arrive at waypoint i, for $i = 1, \ldots, n$, between times $\tau_i^{\min}$ and $\tau_i^{\max}$, which are given. The vehicle consumes fuel over segment i at a rate that depends on its speed, $\Phi(s_i)$, where Φ is positive, increasing, and convex, and has units of kg/s.

You are given the data d (segment travel distances), $s^{\min}$ and $s^{\max}$ (speed bounds), $\tau^{\min}$ and $\tau^{\max}$ (waypoint arrival time bounds), and the fuel use function $\Phi : \mathbf{R} \to \mathbf{R}$. You are to choose the speeds $s_1, \ldots, s_n$ so as to minimize the total fuel consumed in kg.

(a) Show how to pose this as a convex optimization problem. If you introduce new variables, or change variables, you must explain how to recover the optimal speeds from the solution of your problem. If convexity of the objective or any constraint function in your formulation is not obvious, explain why it is convex.

(b) Carry out the method of part (a) on the problem instance with data in `veh_speed_sched_data.m`. Use the fuel use function $\Phi(s_i) = as_i^2 + bs_i + c$ (the parameters a, b, and c are defined in the data file). What is the optimal fuel consumption? Plot the optimal speed versus segment, using the matlab command `stairs` to better show constant speed over the segments.

3.21 *Norm approximation via SOCP, for ℓ_p-norms with rational p.*

(a) Use the observation at the beginning of exercise 4.26 in *Convex Optimization* to express the constraint

$$y \leq \sqrt{z_1 z_2}, \qquad y, z_1, z_2 \geq 0,$$

with variables y, z_1, z_2, as a second-order cone constraint. Then extend your result to the constraint

$$y \leq (z_1 z_2 \cdots z_n)^{1/n}, \qquad y \geq 0, \qquad z \succeq 0,$$

where n is a positive integer, and the variables are $y \in \mathbf{R}$ and $z \in \mathbf{R}^n$. First assume that n is a power of two, and then generalize your formulation to arbitrary positive integers.

(b) Express the constraint

$$f(x) \leq t$$

as a second-order cone constraint, for the following two convex functions f:

$$f(x) = \begin{cases} x^\alpha & x \geq 0 \\ 0 & x < 0, \end{cases}$$

where α is rational and nonnegative, and

$$f(x) = x^\alpha, \qquad \mathbf{dom}\, f = \mathbf{R}_{++},$$

where α is rational and negative.

(c) Formulate the norm approximation problem

$$\text{minimize} \quad \|Ax - b\|_p$$

as a second-order cone program, where p is a rational number greater than or equal to one. The variable in the optimization problem is $x \in \mathbf{R}^n$. The matrix $A \in \mathbf{R}^{m \times n}$ and the vector $b \in \mathbf{R}^m$ are given. For an m-vector y, the norm $\|y\|_p$ is defined as

$$\|y\|_p = \left(\sum_{k=1}^{m} |y_k|^p \right)^{1/p}$$

when $p \geq 1$.

4 Duality

4.1 *Numerical perturbation analysis example.* Consider the quadratic program

$$\begin{array}{ll} \text{minimize} & x_1^2 + 2x_2^2 - x_1x_2 - x_1 \\ \text{subject to} & x_1 + 2x_2 \leq u_1 \\ & x_1 - 4x_2 \leq u_2, \\ & 5x_1 + 76x_2 \leq 1, \end{array}$$

with variables x_1, x_2, and parameters u_1, u_2.

(a) Solve this QP, for parameter values $u_1 = -2$, $u_2 = -3$, to find optimal primal variable values $x_1^\star$ and $x_2^\star$, and optimal dual variable values $\lambda_1^\star$, $\lambda_2^\star$ and $\lambda_3^\star$. Let $p^\star$ denote the optimal objective value. Verify that the KKT conditions hold for the optimal primal and dual variables you found (within reasonable numerical accuracy).

Hint: See §3.6 of the CVX users' guide to find out how to retrieve optimal dual variables. To specify the quadratic objective, use `quad_form()`.

(b) We will now solve some perturbed versions of the QP, with

$$u_1 = -2 + \delta_1, \qquad u_2 = -3 + \delta_2,$$

where δ_1 and δ_2 each take values from $\{-0.1, 0, 0.1\}$. (There are a total of nine such combinations, including the original problem with $\delta_1 = \delta_2 = 0$.) For each combination of δ_1 and δ_2, make a prediction $p^\star_{\text{pred}}$ of the optimal value of the perturbed QP, and compare it to $p^\star_{\text{exact}}$, the exact optimal value of the perturbed QP (obtained by solving the perturbed QP). Put your results in the two righthand columns in a table with the form shown below. Check that the inequality $p^\star_{\text{pred}} \leq p^\star_{\text{exact}}$ holds.

δ_1	δ_2	$p^\star_{\text{pred}}$	$p^\star_{\text{exact}}$
0	0		
0	−0.1		
0	0.1		
−0.1	0		
−0.1	−0.1		
−0.1	0.1		
0.1	0		
0.1	−0.1		
0.1	0.1		

4.2 *A determinant maximization problem.* We consider the problem

$$\begin{array}{ll} \text{minimize} & \log\det X^{-1} \\ \text{subject to} & A_i^T X A_i \preceq B_i, \quad i = 1, \ldots, m, \end{array}$$

with variable $X \in \mathbf{S}^n$, and problem data $A_i \in \mathbf{R}^{n \times k_i}$, $B_i \in \mathbf{S}^{k_i}_{++}$, $i = 1, \ldots, m$. The constraint $X \succ 0$ is implicit.

We can give several interpretations of this problem. Here is one, from statistics. Let z be a random variable in $\mathbf{R}^n$, with covariance matrix X, which is unknown. However, we do have (matrix) upper

bounds on the covariance of the random variables $y_i = A_i^T z \in \mathbf{R}^{k_i}$, which is $A_i^T X A_i$. The problem is to find the covariance matrix for z, that is consistent with the known upper bounds on the covariance of y_i, that has the largest volume confidence ellipsoid.

Derive the Lagrange dual of this problem. Be sure to state what the dual variables are (*e.g.*, vectors, scalars, matrices), any constraints they must satisfy, and what the dual function is. If the dual function has any implicit equality constraints, make them explicit. You can assume that $\sum_{i=1}^m A_i A_i^T \succ 0$, which implies the feasible set of the original problem is bounded.

What can you say about the optimal duality gap for this problem?

4.3 The relative entropy between two vectors x, $y \in \mathbf{R}^n_{++}$ is defined as

$$\sum_{k=1}^n x_k \log(x_k/y_k).$$

This is a convex function, jointly in x and y. In the following problem we calculate the vector x that minimizes the relative entropy with a given vector y, subject to equality constraints on x:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{k=1}^n x_k \log(x_k/y_k) \\ \text{subject to} & Ax = b \\ & \mathbf{1}^T x = 1 \end{array}$$

The optimization variable is $x \in \mathbf{R}^n$. The domain of the objective function is $\mathbf{R}^n_{++}$. The parameters $y \in \mathbf{R}^n_{++}$, $A \in \mathbf{R}^{m\times n}$, and $b \in \mathbf{R}^m$ are given.

Derive the Lagrange dual of this problem and simplify it to get

$$\text{maximize} \quad b^T z - \log \textstyle\sum_{k=1}^n y_k e^{a_k^T z}$$

(a_k is the kth column of A).

4.4 *Source localization from range measurements.* [BSL08] A signal emitted by a source at an unknown position $x \in \mathbf{R}^n$ ($n = 2$ or $n = 3$) is received by m sensors at known positions $y_1, \ldots, y_m \in \mathbf{R}^n$. From the strength of the received signals, we can obtain noisy estimates d_k of the distances $\|x - y_k\|_2$. We are interested in estimating the source position x based on the measured distances d_k.

In the following problem the error between the squares of the actual and observed distances is minimized:

$$\text{minimize} \quad f_0(x) = \sum_{k=1}^m \left(\|x - y_k\|_2^2 - d_k^2 \right)^2.$$

Introducing a new variable $t = x^T x$, we can express this as

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{k=1}^m \left(t - 2y_k^T x + \|y_k\|_2^2 - d_k^2 \right)^2 \\ \text{subject to} & x^T x - t = 0. \end{array} \tag{5}$$

The variables are $x \in \mathbf{R}^n$, $t \in \mathbf{R}$. Although this problem is not convex, it can be shown that strong duality holds. (It is a variation on the problem discussed on page 229 and in exercise 5.29 of *Convex Optimization*.)

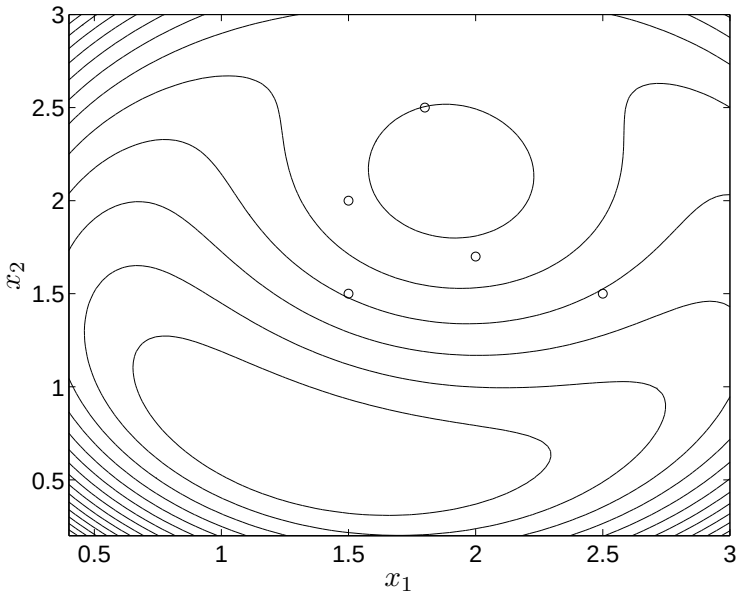
Solve (5) for an example with $m = 5$,

$$y_1 = \begin{bmatrix} 1.8 \\ 2.5 \end{bmatrix}, \qquad y_2 = \begin{bmatrix} 2.0 \\ 1.7 \end{bmatrix}, \qquad y_3 = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}, \qquad y_4 = \begin{bmatrix} 1.5 \\ 2.0 \end{bmatrix}, \qquad y_5 = \begin{bmatrix} 2.5 \\ 1.5 \end{bmatrix},$$

and

$$d = (2.00,\ 1.24,\ 0.59,\ 1.31,\ 1.44).$$

The figure shows some contour lines of the cost function f_0, with the positions y_k indicated by circles.



To solve the problem, you can note that $x^\star$ is easily obtained from the KKT conditions for (5) if the optimal multiplier $\nu^\star$ for the equality constraint is known. You can use one of the following two methods to find $\nu^\star$.

- Derive the dual problem, express it as an SDP, and solve it using CVX.
- Reduce the KKT conditions to a nonlinear equation in ν, and pick the correct solution (similarly as in exercise 5.29 of *Convex Optimization*).

4.5 *Projection on the ℓ_1 ball.* Consider the problem of projecting a point $a \in \mathbf{R}^n$ on the unit ball in ℓ_1-norm:

$$\begin{array}{ll} \text{minimize} & (1/2)\|x - a\|_2^2 \\ \text{subject to} & \|x\|_1 \leq 1. \end{array}$$

Derive the dual problem and describe an efficient method for solving it. Explain how you can obtain the optimal x from the solution of the dual problem.

4.6 *A nonconvex problem with strong duality.* On page 229 of *Convex Optimization*, we consider the problem

$$\begin{array}{ll} \text{minimize} & f(x) = x^T A x + 2b^T x \\ \text{subject to} & x^T x \leq 1 \end{array} \tag{6}$$

with variable $x \in \mathbf{R}^n$, and data $A \in \mathbf{S}^n$, $b \in \mathbf{R}^n$. We do not assume that A is positive semidefinite, and therefore the problem is not necessarily convex. In this exercise we show that x is (globally) optimal if and only if there exists a λ such that

$$\|x\|_2 \leq 1, \qquad \lambda \geq 0, \qquad A + \lambda I \succeq 0, \qquad (A + \lambda I)x = -b, \qquad \lambda(1 - \|x\|_2^2) = 0. \tag{7}$$

From this we will develop an efficient method for finding the global solution. The conditions (7) are the KKT conditions for (6) with the inequality $A + \lambda I \succeq 0$ added.

(a) Show that if x and λ satisfy (7), then $f(x) = \inf_{\tilde{x}} L(\tilde{x}, \lambda) = g(\lambda)$, where L is the Lagrangian of the problem and g is the dual function. Therefore strong duality holds, and x is globally optimal.

(b) Next we show that the conditions (7) are also necessary. Assume that x is globally optimal for (6). We distinguish two cases.

(i) $\|x\|_2 < 1$. Show that (7) holds with $\lambda = 0$.

(ii) $\|x\|_2 = 1$. First prove that $(A + \lambda I)x = -b$ for some $\lambda \geq 0$. (In other words, the negative gradient $-(Ax + b)$ of the objective function is normal to the unit sphere at x, and point away from the origin.) You can show this by contradiction: if the condition does not hold, then there exists a direction v with $v^T x < 0$ and $v^T(Ax + b) < 0$. Show that $f(x + tv) < f(x)$ for small positive t.
It remains to show that $A + \lambda I \succeq 0$. If not, there exists a w with $w^T(A + \lambda I)w < 0$, and without loss of generality we can assume that $w^T x \neq 0$. Show that the point $y = x + tw$ with $t = -2w^T x / w^T w$ satisfies $\|y\|_2 = 1$ and $f(y) < f(x)$.

(c) The optimality conditions (7) can be used to derive a simple algorithm for (6). Using the eigenvalue decomposition $A = \sum_{i=1}^n \alpha_i q_i q_i^T$, of A, we make a change of variables $y_i = q_i^T x$, and write (6) as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \alpha_i y_i^2 + 2\sum_{i=1}^n \beta_i y_i \\ \text{subject to} & y^T y \leq 1 \end{array}$$

where $\beta_i = q_i^T b$. The transformed optimality conditions (7) are

$$\|y\|_2 \leq 1, \qquad \lambda \geq -\alpha_n, \qquad (\alpha_i + \lambda)y_i = -\beta_i, \quad i = 1, \ldots, n, \qquad \lambda(1 - \|y\|_2) = 0,$$

if we assume that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n$. Give an algorithm for computing the solution y and λ.

4.7 *Connection between perturbed optimal cost and Lagrange dual functions.* In this exercise we explore the connection between the optimal cost, as a function of perturbations to the righthand sides of the constraints,

$$p^\star(u) = \inf\{f_0(x) \mid \exists x \in \mathcal{D},\ f_i(x) \leq u_i,\ i = 1, \ldots, m\},$$

(as in §5.6), and the Lagrange dual function

$$g(\lambda) = \inf_x \left(f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)\right),$$

with domain restricted to $\lambda \succeq 0$. We assume the problem is convex. We consider a problem with inequality constraints only, for simplicity.

We have seen several connections between $p^\star$ and g:

- *Slater's condition and strong duality.* Slater's condition is: there exists $u \prec 0$ for which $p^\star(u) < \infty$. Strong duality (which follows) is: $p^\star(0) = \sup_\lambda g(\lambda)$. (Note that we include the condition $\lambda \succeq 0$ in the domain of g.)
- *A global inequality.* We have $p^\star(u) \geq p^\star(0) - \lambda^{\star T} u$, for any u, where $\lambda^\star$ maximizes g.
- *Local sensitivity analysis.* If $p^\star$ is differentiable at 0, then we have $\nabla p^\star(0) = -\lambda^\star$, where $\lambda^\star$ maximizes g.

In fact the two functions are closely related by conjugation. Show that

$$p^\star(u) = (-g)^*(-u).$$

Here $(-g)^*$ is the conjugate of the function $-g$. You can show this for $u \in \mathbf{int}\,\mathbf{dom}\, p^\star$.

Hint. Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & \tilde{f}_i(x) = f_i(x) - u_i \leq 0, \quad i = 1, \ldots, m. \end{array}$$

Verify that Slater's condition holds for this problem, for $u \in \mathbf{int}\,\mathbf{dom}\, p^\star$.

4.8 *Exact penalty method for SDP.* Consider the pair of primal and dual SDPs

$$\begin{array}{llll} \text{(P)} & \text{minimize} & c^T x & \\ & \text{subject to} & F(x) \preceq 0 & \end{array} \qquad \begin{array}{llll} \text{(D)} & \text{maximize} & \mathbf{tr}(F_0 Z) & \\ & \text{subject to} & \mathbf{tr}(F_i Z) + c_i = 0, & i = 1, \ldots, m \\ & & Z \succeq 0, & \end{array}$$

where $F(x) = F_0 + x_1 F_1 + \cdots + x_n F_n$ and $F_i \in \mathbf{S}^p$ for $i = 0, \ldots, n$. Let $Z^\star$ be a solution of (D). Show that every solution $x^\star$ of the unconstrained problem

$$\text{minimize} \quad c^T x + M \max\{0, \lambda_{\max}(F(x))\},$$

where $M > \mathbf{tr}\, Z^\star$, is a solution of (P).

4.9 *Quadratic penalty.* Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{8}$$

where the functions $f_i : \mathbf{R}^n \to \mathbf{R}$ are differentiable and convex.

Show that

$$\phi(x) = f_0(x) + \alpha \sum_{i=1}^m \max\{0, f_i(x)\}^2,$$

where $\alpha > 0$, is convex. Suppose $\tilde{x}$ minimizes ϕ. Show how to find from $\tilde{x}$ a feasible point for the dual of (8). Find the corresponding lower bound on the optimal value of (8).

4.10 *Binary least-squares.* We consider the non-convex least-squares approximation problem with binary constraints

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & x_k^2 = 1, \quad k = 1, \ldots, n, \end{array} \tag{9}$$

where $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$. We assume that $\mathbf{rank}(A) = n$, *i.e.*, A^TA is nonsingular.

One possible application of this problem is as follows. A signal $\hat{x} \in \{-1, 1\}^n$ is sent over a noisy channel, and received as $b = A\hat{x} + v$ where $v \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian noise. The solution of (9) is the maximum likelihood estimate of the input signal $\hat{x}$, based on the received signal b.

(a) Derive the Lagrange dual of (9) and express it as an SDP.

(b) Derive the dual of the SDP in part (a) and show that it is equivalent to

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(A^TAZ) - 2b^TAz + b^Tb \\ \text{subject to} & \mathbf{diag}(Z) = \mathbf{1} \\ & \begin{bmatrix} Z & z \\ z^T & 1 \end{bmatrix} \succeq 0. \end{array} \tag{10}$$

Interpret this problem as a relaxation of (9). Show that if

$$\mathbf{rank}(\begin{bmatrix} Z & z \\ z^T & 1 \end{bmatrix}) = 1 \tag{11}$$

at the optimum of (10), then the relaxation is exact, *i.e.*, the optimal values of problems (9) and (10) are equal, and the optimal solution z of (10) is optimal for (9). This suggests a heuristic for rounding the solution of the SDP (10) to a feasible solution of (9), if (11) does not hold. We compute the eigenvalue decomposition

$$\begin{bmatrix} Z & z \\ z^T & 1 \end{bmatrix} = \sum_{i=1}^{n+1} \lambda_i \begin{bmatrix} v_i \\ t_i \end{bmatrix} \begin{bmatrix} v_i \\ t_i \end{bmatrix}^T,$$

where $v_i \in \mathbf{R}^n$ and $t_i \in \mathbf{R}$, and approximate the matrix by a rank-one matrix

$$\begin{bmatrix} Z & z \\ z^T & 1 \end{bmatrix} \approx \lambda_1 \begin{bmatrix} v_1 \\ t_1 \end{bmatrix} \begin{bmatrix} v_1 \\ t_1 \end{bmatrix}^T.$$

(Here we assume the eigenvalues are sorted in decreasing order). Then we take $x = \mathbf{sign}(v_1)$ as our guess of good solution of (9).

(c) We can also give a probabilistic interpretation of the relaxation (10). Suppose we interpret z and Z as the first and second moments of a random vector $v \in \mathbf{R}^n$ (*i.e.*, $z = \mathbf{E}\, v$, $Z = \mathbf{E}\, vv^T$). Show that (10) is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\, \|Av - b\|_2^2 \\ \text{subject to} & \mathbf{E}\, v_k^2 = 1, \quad k = 1, \ldots, n, \end{array}$$

where we minimize over all possible probability distributions of v.

This interpretation suggests another heuristic method for computing suboptimal solutions of (9) based on the result of (10). We choose a distribution with first and second moments $\mathbf{E}\, v = z$, $\mathbf{E}\, vv^T = Z$ (for example, the Gaussian distribution $\mathcal{N}(z, Z - zz^T)$). We generate a number of samples $\tilde{v}$ from the distribution and round them to feasible solutions $x = \mathbf{sign}(\tilde{v})$. We keep the solution with the lowest objective value as our guess of the optimal solution of (9).

(d) Solve the dual problem (10) using CVX. Generate problem instances using the Matlab code

```
randn('state',0)
m = 50;
n = 40;
A = randn(m,n);
xhat = sign(randn(n,1));
b = A*xhat + s*randn(m,1);
```

for four values of the noise level s: $s = 0.5$, $s = 1$, $s = 2$, $s = 3$. For each problem instance, compute suboptimal feasible solutions x using the the following heuristics and compare the results.

(i) $x^{(a)} = \mathbf{sign}(x_{\rm ls})$ where $x_{\rm ls}$ is the solution of the least-squares problem

$$\text{minimize} \quad \|Ax - b\|_2^2.$$

(ii) $x^{(b)} = \mathbf{sign}(z)$ where z is the optimal value of the variable z in the SDP (10).

(iii) $x^{(c)}$ is computed from a rank-one approximation of the optimal solution of (10), as explained in part (b) above.

(iv) $x^{(d)}$ is computed by rounding 100 samples of $\mathcal{N}(z, Z - zz^T)$, as explained in part (c) above.

4.11 *Monotone transformation of the objective.* Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m. \end{array} \tag{12}$$

where $f_i : \mathbf{R}^n \to \mathbf{R}$ for $i = 0, 1, \ldots, m$ are convex. Suppose $\phi : \mathbf{R} \to \mathbf{R}$ is increasing and convex. Then the problem

$$\begin{array}{ll} \text{minimize} & \tilde{f}_0(x) = \phi(f_0(x)) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \end{array} \tag{13}$$

is convex and equivalent to it; in fact, it has the same optimal set as (12).

In this problem we explore the connections between the duals of the two problems (12) and (13). We assume f_i are differentiable, and to make things specific, we take $\phi(a) = \exp a$.

(a) Suppose λ is feasible for the dual of (12), and $\bar{x}$ minimizes

$$f_0(x) + \sum_{i=1}^m \lambda_i f_i(x).$$

Show that $\bar{x}$ also minimizes

$$\exp f_0(x) + \sum_{i=1}^m \tilde{\lambda}_i f_i(x)$$

for appropriate choice of $\tilde{\lambda}$. Thus, $\tilde{\lambda}$ is dual feasible for (13).

(b) Let $p^\star$ denote the optimal value of (12) (so the optimal value of (12) is $\exp p^\star$). From λ we obtain the bound

$$p^\star \geq g(\lambda),$$

where g is the dual function for (12). From $\tilde{\lambda}$ we obtain the bound $\exp p^\star \geq \tilde{g}(\tilde{\lambda})$, where $\tilde{g}$ is the dual function for (13). This can be expressed as

$$p^\star \geq \log \tilde{g}(\tilde{\lambda}).$$

How do these bounds compare? Are they the same, or is one better than the other?

4.12 *Variable bounds and dual feasibility.* In many problems the constraints include *variable bounds*, as in

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & l_i \leq x_i \leq u_i, \quad i = 1, \ldots, n. \end{array} \tag{14}$$

Let $\mu \in \mathbf{R}^n_+$ be the Lagrange multipliers associated with the constraints $x_i \leq u_i$, and let $\nu \in \mathbf{R}^n_+$ be the Lagrange multipliers associated with the constraints $l_i \geq x_i$. Thus the Lagrangian is

$$L(x, \lambda, \mu, \nu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \mu^T(x - u) + \nu^T(l - x).$$

(a) Show that for any $x \in \mathbf{R}^n$ and any λ, we can choose $\mu \succeq 0$ and $\nu \succeq 0$ so that x minimizes $L(x, \lambda, \mu, \nu)$. In particular, it is very easy to find dual feasible points.

(b) Construct a dual feasible point (λ, μ, ν) by applying the method you found in part (a) with $x = (l + u)/2$ and $\lambda = 0$. From this dual feasible point you get a lower bound on $f^\star$. Show that this lower bound can be expressed as

$$f^\star \geq f_0((l + u)/2) - ((u - l)/2)^T \, |\nabla f_0((l + u)/2)|$$

where $|\cdot|$ means componentwise. Can you prove this bound directly?

4.13 *Deducing costs from samples of optimal decision.* A system (such as a firm or an organism) chooses a vector of values x as a solution of the LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \succeq b, \end{array}$$

with variable $x \in \mathbf{R}^n$. You can think of $x \in \mathbf{R}^n$ as a vector of activity levels, $b \in \mathbf{R}^m$ as a vector of requirements, and $c \in \mathbf{R}^n$ as a vector of costs or prices for the activities. With this interpretation, the LP above finds the cheapest set of activity levels that meet all requirements. (This interpretation is not needed to solve the problem.)

We suppose that A is known, along with a set of data

$$(b^{(1)}, x^{(1)}), \quad \ldots, \quad (b^{(r)}, x^{(r)}),$$

where $x^{(j)}$ is an optimal point for the LP, with $b = b^{(j)}$. (The solution of an LP need not be unique; all we say here is that $x^{(j)}$ is *an* optimal solution.) Roughly speaking, we have samples of optimal decisions, for different values of requirements.

You *do not* know the cost vector c. Your job is to compute the tightest possible bounds on the costs c_i from the given data. More specifically, you are to find $c_i^{\max}$ and $c_i^{\min}$, the maximum and minimum possible values for c_i, consistent with the given data.

Note that if x is optimal for the LP for a given c, then it is also optimal if c is scaled by any positive factor. To normalize c, then, we will assume that $c_1 = 1$. Thus, we can interpret c_i as the relative cost of activity i, compared to activity 1.

(a) Explain how to find $c_i^{\max}$ and $c_i^{\min}$. Your method can involve the solution of a reasonable number (not exponential in n, m or r) of convex or quasiconvex optimization problems.

(b) Carry out your method using the data found in `deducing_costs_data.m`. You may need to determine whether individual inequality constraints are tight; to do so, use a tolerance threshold of $\epsilon = 10^{-3}$. (In other words: if $a_k^T x - b_k \leq 10^{-3}$, you can consider this inequality as tight.)
Give the values of $c_i^{\max}$ and $c_i^{\min}$, and make a very brief comment on the results.

4.14 *Kantorovich inequality.*

(a) Suppose $a \in \mathbf{R}^n$ with $a_1 \geq a_2 \geq \cdots \geq a_n > 0$, and $b \in \mathbf{R}^n$ with $b_k = 1/a_k$.
Derive the KKT conditions for the convex optimization problem

$$\begin{array}{ll} \text{minimize} & -\log(a^T x) - \log(b^T x) \\ \text{subject to} & x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

Show that $x = (1/2, 0, \ldots, 0, 1/2)$ is optimal.

(b) Suppose $A \in \mathbf{S}^n_{++}$ with eigenvalues λ_k sorted in decreasing order. Apply the result of part (a), with $a_k = \lambda_k$, to prove the *Kantorovich inequality:*

$$2\left(u^T A u\right)^{1/2}\left(u^T A^{-1} u\right)^{1/2} \leq \sqrt{\frac{\lambda_1}{\lambda_n}} + \sqrt{\frac{\lambda_n}{\lambda_1}}$$

for all u with $\|u\|_2 = 1$.

4.15 State and solve the optimality conditions for the problem

$$\begin{array}{ll} \text{minimize} & \log\det\left(\left[\begin{array}{cc} X_1 & X_2 \\ X_2^T & X_3 \end{array}\right]^{-1}\right) \\ \text{subject to} & \mathbf{tr}\, X_1 = \alpha \\ & \mathbf{tr}\, X_2 = \beta \\ & \mathbf{tr}\, X_3 = \gamma. \end{array}$$

The optimization variable is

$$X = \left[\begin{array}{cc} X_1 & X_2 \\ X_2^T & X_3 \end{array}\right],$$

with $X_1 \in \mathbf{S}^n$, $X_2 \in \mathbf{R}^{n\times n}$, $X_3 \in \mathbf{S}^n$. The domain of the objective function is $\mathbf{S}^{2n}_{++}$. We assume $\alpha > 0$, and $\alpha\gamma > \beta^2$.

4.16 Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & -\log\det X + \mathbf{tr}(SX) \\ \text{subject to} & X \text{ is tridiagonal} \end{array}$$

with domain $\mathbf{S}^n_{++}$ and variable $X \in \mathbf{S}^n$. The matrix $S \in \mathbf{S}^n$ is given. Show that the optimal X_{opt} satisfies

$$(X_{\text{opt}}^{-1})_{ij} = S_{ij}, \quad |i-j| \leq 1.$$

4.17 We denote by $f(A)$ the sum of the largest r eigenvalues of a symmetric matrix $A \in \mathbf{S}^n$ (with $1 \leq r \leq n$), *i.e.*,

$$f(A) = \sum_{k=1}^{r} \lambda_k(A),$$

where $\lambda_1(A), \ldots, \lambda_n(A)$ are the eigenvalues of A sorted in decreasing order.

(a) Show that the optimal value of the SDP

$$\begin{array}{ll} \text{maximize} & \mathbf{tr}(AX) \\ \text{subject to} & \mathbf{tr}\, X = r \\ & 0 \preceq X \preceq I, \end{array}$$

with variable $X \in \mathbf{S}^n$, is equal to $f(A)$.

(b) Show that f is a convex function.

(c) Assume $A(x) = A_0 + x_1 A_1 + \cdots + x_m A_m$, with $A_k \in \mathbf{S}^n$. Use the observation in part (a) to formulate the optimization problem

$$\text{minimize} \quad f(A(x)),$$

with variable $x \in \mathbf{R}^m$, as an SDP.

4.18 *An exact penalty function.* Suppose we are given a convex problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \; i = 1, \ldots, m \end{array} \tag{15}$$

with dual

$$\begin{array}{ll} \text{maximize} & g(\lambda) \\ \text{subject to} & \lambda \succeq 0. \end{array} \tag{16}$$

We assume that Slater's condition holds, so we have strong duality and the dual optimum is attained. For simplicity we will assume that there is a unique dual optimal solution $\lambda^\star$.

For fixed $t > 0$, consider the unconstrained minimization problem

$$\text{minimize } f_0(x) + t \max_{i=1,\ldots,m} f_i(x)^+, \tag{17}$$

where $f_i(x)^+ = \max\{f_i(x), 0\}$.

(a) Show that the objective function in (17) is convex.

(b) We can express (17) as

$$\begin{array}{lll} \text{minimize} & f_0(x) + ty & \\ \text{subject to} & f_i(x) \leq y, & i = 1. \ldots, m \\ & 0 \leq y & \end{array} \tag{18}$$

where the variables are x and $y \in \mathbf{R}$.

Find the Lagrange dual problem of (18) and express it in terms of the Lagrange dual function g for problem (15).

(c) Use the result in (b) to prove the following property. If $t > \mathbf{1}^T\lambda^\star$, then any minimizer of (17) is also an optimal solution of (15).

(The second term in (17) is called a *penalty function* for the constraints in (15). It is zero if x is feasible, and adds a penalty to the cost function when x is infeasible. The penalty function is called *exact* because for t large enough, the solution of the unconstrained problem (17) is also a solution of (15).)

4.19 *Infimal convolution.* Let $f_1, \ldots, f_m$ be convex functions on $\mathbf{R}^n$. Their *infimal convolution*, denoted $g = f_1 \diamond \cdots \diamond f_m$ (several other notations are also used), is defined as

$$g(x) = \inf\{f_1(x_1) + \cdots + f_m(x_m) \mid x_1 + \cdots + x_m = x\},$$

with the natural domain (*i.e.*, defined by $g(x) < \infty$). In one simple interpretation, $f_i(x_i)$ is the cost for the ith firm to produce a mix of products given by x_i; $g(x)$ is then the optimal cost obtained if the firms can freely exchange products to produce, all together, the mix given by x. (The name 'convolution' presumably comes from the observation that if we replace the sum above with the product, and the infimum above with integration, then we obtain the normal convolution.)

(a) Show that g is convex.

(b) Show that $g^* = f_1^* + \cdots + f_m^*$. In other words, the conjugate of the infimal convolution is the sum of the conjugates.

(c) Verify the identity in part (b) for the specific case of two strictly convex quadratic functions, $f_i(x) = (1/2)x^T P_i x$, with $P_i \in \mathbf{S}^n_{++}$, $i = 1, 2$.
Hint: Depending on how you work out the conjugates, you might find the matrix identity $(X + Y)^{-1}Y = X^{-1}(X^{-1} + Y^{-1})^{-1}$ useful.

5 Approximation and fitting

5.1 *Three measures of the spread of a group of numbers.* For $x \in \mathbf{R}^n$, we define three functions that measure the spread or width of the set of its elements (or coefficients). The first function is the *spread*, defined as

$$\phi_{\mathrm{sprd}}(x) = \max_{i=1,\ldots,n} x_i - \min_{i=1,\ldots,n} x_i.$$

This is the width of the smallest interval that contains all the elements of x.

The second function is the *standard deviation*, defined as

$$\phi_{\mathrm{stdev}}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \right)^{1/2}.$$

This is the statistical standard deviation of a random variable that takes the values $x_1, \ldots, x_n$, each with probability $1/n$.

The third function is the average absolute deviation from the median of the values:

$$\phi_{\mathrm{aamd}}(x) = (1/n) \sum_{i=1}^{n} |x_i - \mathrm{med}(x)|,$$

where $\mathrm{med}(x)$ denotes the median of the components of x, defined as follows. If $n = 2k-1$ is odd, then the median is defined as the value of middle entry when the components are sorted, *i.e.*, $\mathrm{med}(x) = x_{[k]}$, the kth largest element among the values $x_1, \ldots, x_n$. If $n = 2k$ is even, we define the median as the average of the two middle values, *i.e.*, $\mathrm{med}(x) = (x_{[k]} + x_{[k+1]})/2$.

Each of these functions measures the spread of the values of the entries of x; for example, each function is zero if and only if all components of x are equal, and each function is unaffected if a constant is added to each component of x.

Which of these three functions is convex? For each one, either show that it is convex, or give a counterexample showing it is not convex. By a counterexample, we mean a specific x and y such that Jensen's inequality fails, *i.e.*, $\phi((x+y)/2) > (\phi(x) + \phi(y))/2$.

5.2 *Minimax rational fit to the exponential.* (See exercise 6.9 of *Convex Optimization*.) We consider the specific problem instance with data

$$t_i = -3 + 6(i-1)/(k-1), \quad y_i = e^{t_i}, \quad i = 1, \ldots, k,$$

where $k = 201$. (In other words, the data are obtained by uniformly sampling the exponential function over the interval $[-3, 3]$.) Find a function of the form

$$f(t) = \frac{a_0 + a_1 t + a_2 t^2}{1 + b_1 t + b_2 t^2}$$

that minimizes $\max_{i=1,\ldots,k} |f(t_i) - y_i|$. (We require that $1 + b_1 t_i + b_2 t_i^2 > 0$ for $i = 1, \ldots, k$.)

Find optimal values of a_0, a_1, a_2, b_1, b_2, and give the optimal objective value, computed to an accuracy of 0.001. Plot the data and the optimal rational function fit on the same plot. On a different plot, give the fitting error, *i.e.*, $f(t_i) - y_i$.

Hint. You can use `strcmp(cvx_status,'Solved')`, after `cvx_end`, to check if a feasibility problem is feasible.

5.3 *Approximation with trigonometric polynomials.* Suppose $y : \mathbf{R} \to \mathbf{R}$ is a 2π-periodic function. We will approximate y with the trigonometric polynomial

$$f(t) = \sum_{k=0}^{K} a_k \cos(kt) + \sum_{k=1}^{K} b_k \sin(kt).$$

We consider two approximations: one that minimizes the L_2-norm of the error, defined as

$$\|f - y\|_2 = \left(\int_{-\pi}^{\pi} (f(t) - y(t))^2 \, dt \right)^{1/2},$$

and one that minimizes the L_1-norm of the error, defined as

$$\|f - y\|_1 = \int_{-\pi}^{\pi} |f(t) - y(t)| \, dt.$$

The L_2 approximation is of course given by the (truncated) Fourier expansion of y.

To find an L_1 approximation, we discretize t at $2N$ points,

$$t_i = -\pi + i\pi/N, \quad i = 1, \ldots, 2N,$$

and approximate the L_1 norm as

$$\|f - y\|_1 \approx (\pi/N) \sum_{i=1}^{2N} |f(t_i) - y(t_i)|.$$

(A standard rule of thumb is to take N at least 10 times larger than K.) The L_1 approximation (or really, an approximation of the L_1 approximation) can now be found using linear programming.

We consider a specific case, where y is a 2π-periodic square-wave, defined for $-\pi \leq t \leq \pi$ as

$$y(t) = \begin{cases} 1 & |t| \leq \pi/2 \\ 0 & \text{otherwise.} \end{cases}$$

(The graph of y over a few cycles explains the name 'square-wave'.)

Find the optimal L_2 approximation and (discretized) L_1 optimal approximation for $K = 10$. You can find the L_2 optimal approximation analytically, or by solving a least-squares problem associated with the discretized version of the problem. Since y is even, you can take the sine coefficients in your approximations to be zero. Show y and the two approximations on a single plot.

In addition, plot a histogram of the residuals (*i.e.*, the numbers $f(t_i) - y(t_i)$) for the two approximations. Use the same horizontal axis range, so the two residual distributions can easily be compared. (Matlab command `hist` might be helpful here.) Make some brief comments about what you see.

5.4 *Penalty function approximation.* We consider the approximation problem

$$\text{minimize} \quad \phi(Ax - b)$$

where $A \in \mathbf{R}^{m \times n}$ and $b \in \mathbf{R}^m$, the variable is $x \in \mathbf{R}^n$, and $\phi : \mathbf{R}^m \to \mathbf{R}$ is a convex penalty function that measures the quality of the approximation $Ax \approx b$. We will consider the following choices of penalty function:

(a) *Euclidean norm.*

$$\phi(y) = \|y\|_2 = (\sum_{k=1}^{m} y_k^2)^{1/2}.$$

(b) *ℓ_1-norm.*

$$\phi(y) = \|y\|_1 = \sum_{k=1}^{m} |y_k|.$$

(c) *Sum of the largest $m/2$ absolute values.*

$$\phi(y) = \sum_{k=1}^{\lfloor m/2 \rfloor} |y|_{[k]}$$

where $|y|_{[1]}$, $|y|_{[2]}$, $|y|_{[3]}$, ..., denote the absolute values of the components of y sorted in decreasing order.

(d) *A piecewise-linear penalty.*

$$\phi(y) = \sum_{k=1}^{m} h(y_k), \qquad h(u) = \begin{cases} 0 & |u| \leq 0.2 \\ |u| - 0.2 & 0.2 \leq |u| \leq 0.3 \\ 2|u| - 0.5 & |u| \geq 0.3. \end{cases}$$

(e) *Huber penalty.*

$$\phi(y) = \sum_{k=1}^{m} h(y_k), \qquad h(u) = \begin{cases} u^2 & |u| \leq M \\ M(2|u| - M) & |u| \geq M \end{cases}$$

with $M = 0.2$.

(f) *Log-barrier penalty.*

$$\phi(y) = \sum_{k=1}^{m} h(y_k), \qquad h(u) = -\log(1 - u^2), \quad \mathbf{dom}\, h = \{u \mid |u| < 1\}.$$

Here is the problem. Generate data A and b as follows:

```
m = 200;
n = 100;
A = randn(m,n);
b = randn(m,1);
b = b/(1.01*max(abs(b)));
```

(The normalization of b ensures that the domain of $\phi(Ax - b)$ is nonempty if we use the log-barrier penalty.) To compare the results, plot a histogram of the vector of residuals $y = Ax - b$, for each of the solutions x, using the Matlab command

```
hist(A*x-b,m/2);
```

Some additional hints and remarks for the individual problems:

(a) This problem can be solved using least-squares (`x=A\b`).

(b) Use the CVX function `norm(y,1)`.

(c) Use the CVX function `norm_largest()`.

(d) Use CVX, with the overloaded `max()`, `abs()`, and `sum()` functions.

(e) Use the CVX function `huber()`.

(f) The current version of CVX does not directly handle the logarithm. However, you can reformulate this problem as

$$\text{maximize} \quad \left(\prod_{k=1}^{m} \left((1-(Ax-b)_k)(1+(Ax-b)_k)\right)\right)^{1/2m},$$

and use the CVX function `geomean()`.

5.5 *$\ell_{1.5}$ optimization.* Optimization and approximation methods that use both an ℓ_2-norm (or its square) and an ℓ_1-norm are currently very popular in statistics, machine learning, and signal and image processing. Examples include Huber estimation, LASSO, basis pursuit, SVM, various ℓ_1-regularized classification methods, total variation de-noising, etc. Very roughly, an ℓ_2-norm corresponds to Euclidean distance (squared), or the negative log-likelihood function for a Gaussian; in contrast the ℓ_1-norm gives 'robust' approximation, *i.e.*, reduced sensitivity to outliers, and also tends to yield sparse solutions (of whatever the argument of the norm is). (All of this is just background; you don't need to know any of this to solve the problem.)

In this problem we study a natural method for blending the two norms, by using the $\ell_{1.5}$-norm, defined as

$$\|z\|_{1.5} = \left(\sum_{i=1}^{k} |z_i|^{3/2}\right)^{2/3}$$

for $z \in \mathbf{R}^k$. We will consider the simplest approximation or regression problem:

$$\text{minimize} \quad \|Ax-b\|_{1.5},$$

with variable $x \in \mathbf{R}^n$, and problem data $A \in \mathbf{R}^{m\times n}$ and $b \in \mathbf{R}^m$. We will assume that $m > n$ and the A is full rank (*i.e.*, rank n). The hope is that this $\ell_{1.5}$-optimal approximation problem should share some of the good features of ℓ_2 and ℓ_1 approximation.

(a) Give optimality conditions for this problem. Try to make these as simple as possible. Your solution should have the form 'x is optimal for the $\ell_{1.5}$-norm approximation problem if and only if ...'.

(b) Explain how to formulate the $\ell_{1.5}$-norm approximation problem as an SDP. (Your SDP can include linear equality and inequality constraints.)

(c) Solve the specific numerical instance generated by the following code:

```
randn('state',0);
A=randn(100,30);
b=randn(100,1);
```

Numerically verify the optimality conditions. Give a histogram of the residuals, and repeat for the ℓ_2-norm and ℓ_1-norm approximations. You can use any method you like to solve the problem (but of course you must explain how you did it); in particular, you do not need to use the SDP formulation found in part (b).

5.6 *Total variation image interpolation.* A grayscale image is represented as an $m \times n$ matrix of intensities U^{orig}. You are given the values U_{ij}^{orig}, for $(i,j) \in \mathcal{K}$, where $\mathcal{K} \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$. Your job is to *interpolate* the image, by guessing the missing values. The reconstructed image will be represented by $U \in \mathbf{R}^{m \times n}$, where U satisfies the interpolation conditions $U_{ij} = U_{ij}^{\text{orig}}$ for $(i,j) \in \mathcal{K}$.

The reconstruction is found by minimizing a roughness measure subject to the interpolation conditions. One common roughness measure is the ℓ_2 variation (squared),

$$\sum_{i=2}^{m}\sum_{j=2}^{n}\left((U_{ij}-U_{i-1,j})^2+(U_{ij}-U_{i,j-1})^2\right).$$

Another method minimizes instead the *total variation*,

$$\sum_{i=2}^{m}\sum_{j=2}^{n}\left(|U_{ij}-U_{i-1,j}|+|U_{ij}-U_{i,j-1}|\right).$$

Evidently both methods lead to convex optimization problems.

Carry out ℓ_2 and total variation interpolation on the problem instance with data given in `tv_img_interp.m`. This will define `m`, `n`, and matrices `Uorig` and `Known`. The matrix `Known` is $m \times n$, with (i,j) entry one if $(i,j) \in \mathcal{K}$, and zero otherwise. The mfile also has skeleton plotting code. (We give you the entire original image so you can compare your reconstruction to the original; obviously your solution cannot access U_{ij}^{orig} for $(i,j) \notin \mathcal{K}$.)

5.7 *Piecewise-linear fitting.* In many applications some function in the model is not given by a formula, but instead as tabulated data. The tabulated data could come from empirical measurements, historical data, numerically evaluating some complex expression or solving some problem, for a set of values of the argument. For use in a convex optimization model, we then have to fit these data with a convex function that is compatible with the solver or other system that we use. In this problem we explore a very simple problem of this general type.

Suppose we are given the data (x_i, y_i), $i = 1,\ldots,m$, with $x_i,\ y_i \in \mathbf{R}$. We will assume that x_i are sorted, *i.e.*, $x_1 < x_2 < \cdots < x_m$. Let $a_0 < a_1 < a_2 < \cdots < a_K$ be a set of fixed knot points, with $a_0 \le x_1$ and $a_K \ge x_m$. Explain how to find the convex piecewise linear function f, defined over $[a_0, a_K]$, with knot points a_i, that minimizes the least-squares fitting criterion

$$\sum_{i=1}^{m}(f(x_i)-y_i)^2.$$

You must explain what the variables are and how they parametrize f, and how you ensure convexity of f.

Hints. One method to solve this problem is based on the Lagrange basis, $f_0,\ldots,f_K$, which are the piecewise linear functions that satisfy

$$f_j(a_i) = \delta_{ij}, \quad i,j = 0,\ldots,K.$$

Another method is based on defining $f(x) = \alpha_i x + \beta_i$, for $x \in (a_{i-1}, a_i]$. You then have to add conditions on the parameters α_i and β_i to ensure that f is continuous and convex.

Apply your method to the data in the file `pwl_fit_data.m`, which contains data with $x_j \in [0,1]$. Find the best affine fit (which corresponds to $a = (0,1)$), and the best piecewise-linear convex function fit for 1, 2, and 3 internal knot points, evenly spaced in $[0,1]$. (For example, for 3 internal knot points we have $a_0 = 0,\ a_1 = 0.25,\ a_2 = 0.50,\ a_3 = 0.75,\ a_4 = 1$.) Give the least-squares fitting cost for each one. Plot the data and the piecewise-linear fits found. Express each function in the form

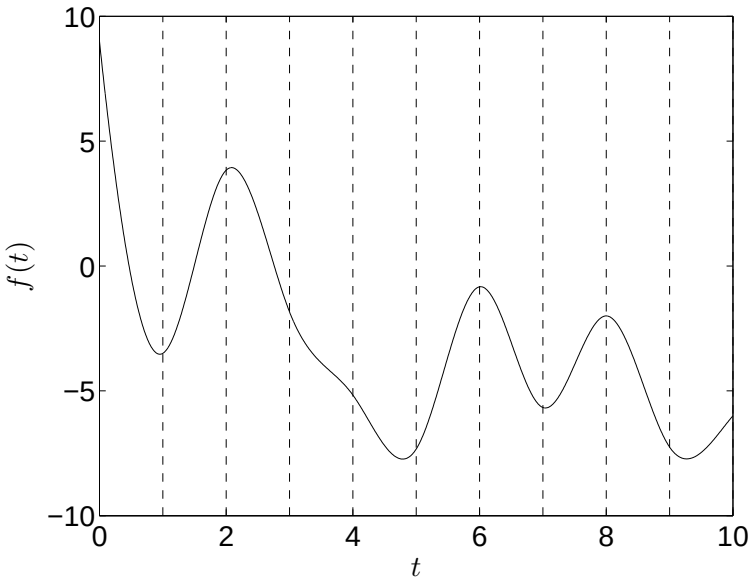
$$f(x) = \max_{i=1\ldots,K}(\alpha_i x + \beta_i).$$

(In this form the function is easily incorporated into an optimization problem.)

5.8 *Least-squares fitting with convex splines.* A *cubic spline* (or *fourth-order spline*) with breakpoints α_0, α_1, ..., α_M (that satisfy $\alpha_0 < \alpha_1 < \cdots < \alpha_M$) is a piecewise-polynomial function with the following properties:

- the function is a cubic polynomial on each interval $[\alpha_i, \alpha_{i+1}]$
- the function values, and the first and second derivatives are continuous on the interval (α_0, α_M).

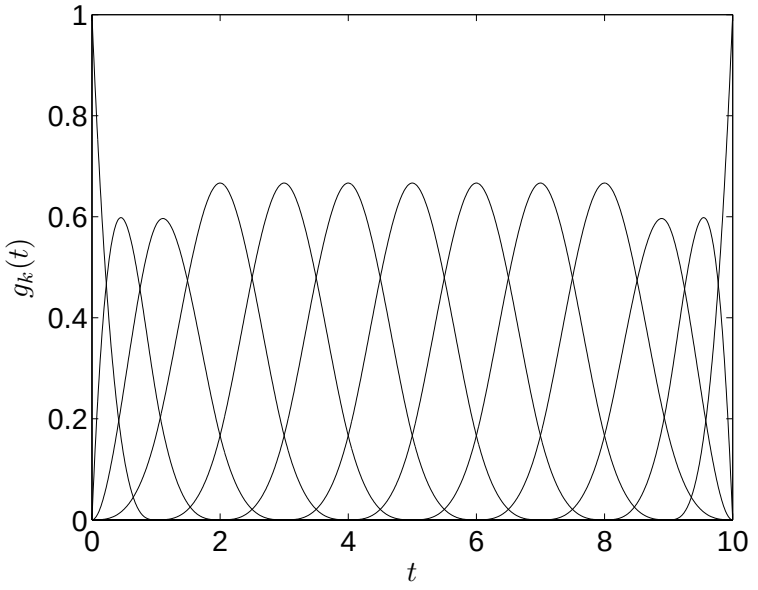
The figure shows an example of a cubic spline $f(t)$ with $M = 10$ segments and breakpoints $\alpha_0 = 0$, $\alpha_1 = 1$, ..., $\alpha_{10} = 10$.



In approximation problems with splines it is convenient to parametrize a spline as a linear combination of basis functions, called *B-splines*. The precise definition of B-splines is not important for our purposes; it is sufficient to know that every cubic spline can be written as a linear combination of $M + 3$ cubic B-splines $g_k(t)$, *i.e.*, in the form

$$f(t) = x_1 g_1(t) + \cdots + x_{M+3} g_{M+3}(t) = x^T g(t),$$

and that there exist efficient algorithms for computing $g(t) = (g_1(t), \ldots, g_{M+3}(t))$. The next figure shows the 13 B-splines for the breakpoints 0, 1, ..., 10.



In this exercise we study the problem of fitting a cubic spline to a set of data points, subject to the constraint that the spline is a convex function. Specifically, the breakpoints $\alpha_0, \ldots, \alpha_M$ are fixed, and we are given N data points (t_k, y_k) with $t_k \in [\alpha_0, \alpha_M]$. We are asked to find the convex cubic spline $f(t)$ that minimizes the least-squares criterion

$$\sum_{k=1}^{N} (f(t_k) - y_k)^2.$$

We will use B-splines to parametrize f, so the variables in the problem are the coefficients x in $f(t) = x^T g(t)$. The problem can then be written as

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{k=1}^{N} \left(x^T g(t_k) - y_k \right)^2 \\ \text{subject to} & x^T g(t) \text{ is convex in } t \text{ on } [\alpha_0, \alpha_M]. \end{array} \tag{19}$$

(a) Express problem (19) as a convex optimization problem of the form

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|_2^2 \\ \text{subject to} & Gx \preceq h. \end{array}$$

(b) Use CVX to solve a specific instance of the optimization problem in part (a). As in the figures above, we take $M = 10$ and $\alpha_0 = 0$, $\alpha_1 = 1$, $\ldots$, $\alpha_{10} = 10$.

Download the Matlab files `spline_data.m` and `bsplines.m`. The first m-file is used to generate the problem data. The command `[t, y] = spline_data` will generate two vectors t, y of length $N = 51$, with the data points t_k, y_k.

The second function can be used to compute the B-splines, and their first and second derivatives, at any given point $u \in [0, 10]$. The command `[g, gp, gpp] = bsplines(u)` returns three vectors of length 13 with elements $g_k(u)$, $g_k'(u)$, and $g_k''(u)$. (The right derivatives are returned for $u = 0$, and the left derivatives for $u = 10$.)

Solve the convex spline fitting problem (19) for this example, and plot the optimal spline.

5.9 *Robust least-squares with interval coefficient matrix.* An *interval matrix* in $\mathbf{R}^{m\times n}$ is a matrix whose entries are intervals:

$$\mathcal{A} = \{A \in \mathbf{R}^{m\times n} \mid |A_{ij} - \bar{A}_{ij}| \leq R_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,n\}.$$

The matrix $\bar{A} \in \mathbf{R}^{m\times n}$ is called the *nominal value* or *center value*, and $R \in \mathbf{R}^{m\times n}$, which is elementwise nonnegative, is called the *radius*.

The robust least-squares problem, with interval matrix, is

$$\text{minimize} \quad \sup_{A\in\mathcal{A}} \|Ax - b\|_2,$$

with optimization variable $x \in \mathbf{R}^n$. The problem data are $\mathcal{A}$ (*i.e.*, $\bar{A}$ and R) and $b \in \mathbf{R}^m$. The objective, as a function of x, is called the *worst-case residual norm*. The robust least-squares problem is evidently a convex optimization problem.

(a) Formulate the interval matrix robust least-squares problem as a standard optimization problem, *e.g.*, a QP, SOCP, or SDP. You can introduce new variables if needed. Your reformulation should have a number of variables and constraints that grows linearly with m and n, and not exponentially.

(b) Consider the specific problem instance with $m = 4$, $n = 3$,

$$\mathcal{A} = \begin{bmatrix} 60 \pm 0.05 & 45 \pm 0.05 & -8 \pm 0.05 \\ 90 \pm 0.05 & 30 \pm 0.05 & -30 \pm 0.05 \\ 0 \pm 0.05 & -8 \pm 0.05 & -4 \pm 0.05 \\ 30 \pm 0.05 & 10 \pm 0.05 & -10 \pm 0.05 \end{bmatrix}, \qquad b = \begin{bmatrix} -6 \\ -3 \\ 18 \\ -9 \end{bmatrix}.$$

(The first part of each entry in $\mathcal{A}$ gives $\bar{A}_{ij}$; the second gives R_{ij}, which are all 0.05 here.) Find the solution x_{ls} of the nominal problem (*i.e.*, minimize $\|\bar{A}x - b\|_2$), and robust least-squares solution x_{rls}. For each of these, find the nominal residual norm, and also the worst-case residual norm. Make sure the results make sense.

5.10 *Identifying a sparse linear dynamical system.* A linear dynamical system has the form

$$x(t+1) = Ax(t) + Bu(t) + w(t), \quad t = 1,\ldots,T-1,$$

where $x(t) \in \mathbf{R}^n$ is the state, $u(t) \in \mathbf{R}^m$ is the input signal, and $w(t) \in \mathbf{R}^n$ is the process noise, at time t. We assume the process noises are IID $\mathcal{N}(0, W)$, where $W \succ 0$ is the covariance matrix. The matrix $A \in \mathbf{R}^{n\times n}$ is called the dynamics matrix or the state transition matrix, and the matrix $B \in \mathbf{R}^{n\times m}$ is called the input matrix.

You are given accurate measurements of the state and input signal, *i.e.*, $x(1),\ldots,x(T)$, $u(1),\ldots,u(T-1)$, and W is known. Your job is to find a state transition matrix $\hat{A}$ and input matrix $\hat{B}$ from these data, that are plausible, and in addition are sparse, *i.e.*, have many zero entries. (The sparser the better.)

By doing this, you are effectively estimating the structure of the dynamical system, *i.e.*, you are determining which components of $x(t)$ and $u(t)$ affect which components of $x(t+1)$. In some applications, this structure might be more interesting than the actual values of the (nonzero) coefficients in $\hat{A}$ and $\hat{B}$.

By plausible, we mean that

$$\sum_{t=1}^{T-1} \left\| W^{-1/2} \left(x(t+1) - \hat{A}x(t) - \hat{B}u(t) \right) \right\|_2^2 \in n(T-1) \pm 2\sqrt{2n(T-1)},$$

where $a \pm b$ means the interval $[a-b, a+b]$. (You can just take this as our definition of plausible. But to explain this choice, we note that when $\hat{A} = A$ and $\hat{B} = B$, the left-hand side is χ^2, with $n(T-1)$ degrees of freedom, and so has mean $n(T-1)$ and standard deviation $\sqrt{2n(T-1)}$.)

(a) Describe a method for finding $\hat{A}$ and $\hat{B}$, based on convex optimization.

We are looking for a *very simple* method, that involves solving *one* convex optimization problem. (There are many extensions of this basic method, that would improve the simple method, *i.e.*, yield sparser $\hat{A}$ and $\hat{B}$ that are still plausible. We're not asking you to describe or implement any of these.)

(b) Carry out your method on the data found in `sparse_lds_data.m`. Give the values of $\hat{A}$ and $\hat{B}$ that you find, and verify that they are plausible.

In the data file, we give you the true values of A and B, so you can evaluate the performance of your method. (Needless to say, you are not allowed to use these values when forming $\hat{A}$ and $\hat{B}$.) Using these true values, give the number of false positives and false negatives in both $\hat{A}$ and $\hat{B}$. A false positive in $\hat{A}$, for example, is an entry that is nonzero, while the corresponding entry in A is zero. A false negative is an entry of $\hat{A}$ that is zero, while the corresponding entry of A is nonzero. To judge whether an entry of $\hat{A}$ (or $\hat{B}$) is nonzero, you can use the test $|\hat{A}_{ij}| \geq 0.01$ (or $|\hat{B}_{ij}| \geq 0.01$).

5.11 *Measurement with bounded errors.* A series of K measurements $y_1, \ldots, y_K \in \mathbf{R}^p$, are taken in order to estimate an unknown vector $x \in \mathbf{R}^q$. The measurements are related to the unknown vector x by $y_i = Ax + v_i$, where v_i is a measurement noise that satisfies $\|v_i\|_\infty \leq \alpha$ but is otherwise unknown. (In other words, the entries of $v_1, \ldots, v_K$ are no larger than α.) The matrix A and the measurement noise norm bound α are known. Let X denote the set of vectors x that are consistent with the observations $y_1, \ldots, y_K$, *i.e.*, the set of x that could have resulted in the measurements made. Is X convex?

Now we will examine what happens when the measurements are occasionally in error, *i.e.*, for a few i we have no relation between x and y_i. More precisely suppose that I_{fault} is a subset of $\{1, \ldots, K\}$, and that $y_i = Ax + v_i$ with $\|v_i\|_\infty \leq \alpha$ (as above) for $i \notin I_{\text{fault}}$, but for $i \in I_{\text{fault}}$, there is no relation between x and y_i. The set I_{fault} is the set of times of the faulty measurements.

Suppose you know that I_{fault} has at most J elements, *i.e.*, out of K measurements, at most J are faulty. You do not know I_{fault}; you know only a bound on its cardinality (size). For what values of J is X, the set of x consistent with the measurements, convex?

5.12 *Least-squares with a few permuted measurements.* We want to estimate a vector $x \in \mathbf{R}^n$, given some linear measurements of x corrupted with Gaussian noise. Here's the catch: some of the measurements have been *permuted*.

More precisely, our measurement vector $y \in \mathbf{R}^m$ has the form

$$y = P(Ax + v),$$

where v_i are IID $\mathcal{N}(0, I)$ measurement noises, $x \in \mathbf{R}^n$ is the vector of parameters we wish to estimate, and $P \in \mathbf{R}^{n\times n}$ is a permutation matrix. (This means that each row and column of P has exactly one entry equal to one, and the remaining $n-1$ entries zero.) We know that fewer than k of the measurements are permuted; *i.e.*, $Pe_i \neq e_i$ for at most k indices i. We wish to guess what x is, and also what P is. There are $\begin{pmatrix} m \\ k \end{pmatrix}$ possible values of P, which is very large in the cases we are interested in, such as $m = 100$ and $k = 5$.

Once we make a guess $\hat{P}$ for P, we can get the maximum likelihood estimate of x by minimizing $\|Ax - P^T y\|_2$. The residual $A\hat{x} - P^T y$ is then our guess of what $P^T v$ is, and should be consistent with being a sample of a $\mathcal{N}(0, I)$ vector.

In principle, we can find the maximum likelihood estimate of x and P by solving a set of $\begin{pmatrix} m \\ k \end{pmatrix}$ least-squares problems, and choosing one that has minimum residual norm. But this is not practical unless m and k are both very small.

Describe a *heuristic* method for approximately solving this problem, using convex optimization.

5.13 *Fitting with censored data.* In some experiments there are two kinds of measurements or data available: The usual ones, in which you get a number (say), and *censored data*, in which you don't get the specific number, but are told something about it, such as a lower bound. A classic example is a study of lifetimes of a set of subjects (say, laboratory mice). For those who have died by the end of data collection, we get the lifetime. For those who have not died by the end of data collection, we do not have the lifetime, but we do have a lower bound, *i.e.*, the length of the study. These are the censored data values.

We wish to fit a set of data points,

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(K)}, y^{(K)}),$$

with $x^{(k)} \in \mathbf{R}^n$ and $y^{(k)} \in \mathbf{R}$, with a linear model of the form $y \approx c^T x$. The vector $c \in \mathbf{R}^n$ is the model parameter, which we want to choose. We will use a least-squares criterion, *i.e.*, choose c to minimize

$$J = \sum_{k=1}^{K} \left(y^{(k)} - c^T x^{(k)}\right)^2.$$

Here is the tricky part: some of the values of $y^{(k)}$ are censored; for these entries, we have only a (given) lower bound. We will re-order the data so that $y^{(1)}, \ldots, y^{(M)}$ are given (*i.e.*, uncensored), while $y^{(M+1)}, \ldots, y^{(K)}$ are all censored, *i.e.*, unknown, but larger than D, a given number. All the values of $x^{(k)}$ are known.

(a) Explain how to find c (the model parameter) and $y^{(M+1)}, \ldots, y^{(K)}$ (the censored data values) that minimize J.

(b) Carry out the method of part (a) on the data values in `cens_fit_data.m`. Report $\hat{c}$, the value of c found using this method.

Also find $\hat{c}_{\text{ls}}$, the least-squares estimate of c obtained by simply ignoring the censored data samples, *i.e.*, the least-squares estimate based on the data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(M)}, y^{(M)}).$$

The data file contains c_{true}, the true value of c, in the vector `c_true`. Use this to give the two relative errors

$$\frac{\|c_{\text{true}} - \hat{c}\|_2}{\|c_{\text{true}}\|_2}, \qquad \frac{\|c_{\text{true}} - \hat{c}_{\text{ls}}\|_2}{\|c_{\text{true}}\|_2}.$$

5.14 *Spectrum analysis with quantized measurements.* A sample is made up of n compounds, in quantities $q_i \geq 0$, for $i = 1, \ldots, n$. Each compound has a (nonnegative) spectrum, which we represent as a vector $s^{(i)} \in \mathbf{R}^m_+$, for $i = 1, \ldots, n$. (Precisely what $s^{(i)}$ means won't matter to us.) The spectrum of the sample is given by $s = \sum_{i=1}^n q_i s^{(i)}$. We can write this more compactly as $s = Sq$, where $S \in \mathbf{R}^{m \times n}$ is a matrix whose columns are $s^{(1)}, \ldots, s^{(n)}$.

Measurement of the spectrum of the sample gives us an interval for each spectrum value, *i.e.*, $l, u \in \mathbf{R}^m_+$ for which

$$l_i \leq s_i \leq u_i, \quad i = 1, \ldots, m.$$

(We don't directly get s.) This occurs, for example, if our measurements are quantized.

Given l and u (and S), we cannot in general deduce q exactly. Instead, we ask you to do the following. For each compound i, find the range of possible values for q_i consistent with the spectrum measurements. We will denote these ranges as $q_i \in [q_i^{\min}, q_i^{\max}]$. Your job is to find $q_i^{\min}$ and $q_i^{\max}$.

Note that if $q_i^{\min}$ is large, we can confidently conclude that there is a significant amount of compound i in the sample. If $q_i^{\max}$ is small, we can confidently conclude that there is not much of compound i in the sample.

(a) Explain how to find $q_i^{\min}$ and $q_i^{\max}$, given S, l, and u.

(b) Carry out the method of part (a) for the problem instance given in `spectrum_data.m`. (Executing this file defines the problem data, and plots the compound spectra and measurement bounds.) Plot the minimum and maximum values versus i, using the commented out code in the data file. Report your values for $q_4^{\min}$ and $q_4^{\max}$.

6 Statistical estimation

6.1 *Maximum likelihood estimation of x and noise mean and covariance.* Consider the maximum likelihood estimation problem with the linear measurement model

$$y_i = a_i^T x + v_i, \quad i = 1, \ldots, m.$$

The vector $x \in \mathbf{R}^n$ is a vector of unknown parameters, y_i are the measurement values, and v_i are independent and identically distributed measurement errors.

In this problem we make the assumption that the *normalized* probability density function of the errors is given (normalized to have zero mean and unit variance), but not their mean and variance. In other words, the density of the measurement errors v_i is

$$p(z) = \frac{1}{\sigma} f(\frac{z-\mu}{\sigma}),$$

where f is a given, normalized density. The parameters μ and σ are the mean and standard deviation of the distribution p, and are not known.

The maximum likelihood estimates of x, μ, σ are the maximizers of the log-likelihood function

$$\sum_{i=1}^m \log p(y_i - a_i^T x) = -m \log \sigma + \sum_{i=1}^m \log f(\frac{y_i - a_i^T x - \mu}{\sigma}),$$

where y is the observed value. Show that if f is log-concave, then the maximum likelihood estimates of x, μ, σ can be determined by solving a convex optimization problem.

6.2 *Mean and covariance estimation with conditional independence constraints.* Let $X \in \mathbf{R}^n$ be a Gaussian random variable with density

$$p(x) = \frac{1}{(2\pi)^{n/2}(\det S)^{1/2}} \exp(-(x-a)^T S^{-1}(x-a)/2).$$

The conditional density of a subvector $(X_i, X_j) \in \mathbf{R}^2$ of X, given the remaining variables, is also Gaussian, and its covariance matrix R_{ij} is equal to the Schur complement of the 2×2 submatrix

$$\begin{bmatrix} S_{ii} & S_{ij} \\ S_{ij} & S_{jj} \end{bmatrix}$$

in the covariance matrix S. The variables X_i, X_j are called *conditionally independent* if the covariance matrix R_{ij} of their conditional distribution is diagonal.

Formulate the following problem as a convex optimization problem. We are given N independent samples $y_1, \ldots, y_N \in \mathbf{R}^n$ of X. We are also given a list $\mathcal{N} \in \{1, \ldots, n\} \times \{1, \ldots, n\}$ of pairs of conditionally independent variables: $(i, j) \in \mathcal{N}$ means X_i and X_j are conditionally independent. The problem is to compute the maximum likelihood estimate of the mean a and the covariance matrix S, subject to the constraint that X_i and X_j are conditionally independent for $(i, j) \in \mathcal{N}$.

6.3 *Maximum likelihood estimation for exponential family.* A probability distribution on $\mathcal{D} \subseteq \mathbf{R}^n$, parametrized by $\theta \in \mathbf{R}^m$, is called an *exponential family* if it has the form

$$p_\theta(x) = a(\theta) \exp(\theta^T c(x))$$

for $x \in \mathcal{D}$, where $c : \mathbf{R}^n \to \mathbf{R}^m$, and

$$a(\theta) = \left(\int_{\mathcal{D}} \exp(\theta^T c(x))\, dx \right)^{-1}.$$

(We consider only values of θ for which the integral above is finite.) Many families of distributions have this form, for appropriate choice of the parameter θ.

(a) When $c(x) = x$ and $\mathcal{D} = \mathbf{R}^n_+$, what is the associated family of distributions? What is the set of valid values of θ?

(b) Explain how to represent the normal family $\mathcal{N}(\mu, \Sigma)$ as an exponential family. *Hint.* Use parameter $(z, Y) = (\Sigma^{-1}\mu, \Sigma^{-1})$. With this parameter, $\theta^T c(x)$ has the form $z^T c_1(x) + \mathbf{tr}\, Y C_2(x)$, where $C_2(x) \in \mathbf{S}^n$.

(c) Show that for any $x \in \mathcal{D}$, the log-likelihood function $\log p_\theta(x)$ is concave in θ. This means that maximum-likelihood estimation for an exponential family leads to a convex optimization problem. You don't have to give a formal proof of concavity of $\log p_\theta(x)$; if you like, you can approximate the integral appearing in the expression as a (finite) Riemann sum, show concavity of this approxmation, and then just state 'take the limit'.

6.4 *Maximum likelihood prediction of team ability.* A set of n teams compete in a tournament. We model each team's ability by a number $a_j \in [0, 1]$, $j = 1, \ldots, n$. When teams j and k play each other, the probability that team j wins is equal to $\mathbf{prob}(a_j - a_k + v > 0)$, where $v \sim \mathcal{N}(0, \sigma^2)$.

You are given the outcome of m past games. These are organized as

$$(j^{(i)}, k^{(i)}, y^{(i)}), \quad i = 1, \ldots, m,$$

meaning that game i was played between teams $j^{(i)}$ and $k^{(i)}$; $y^{(i)} = 1$ means that team $j^{(i)}$ won, while $y^{(i)} = -1$ means that team $k^{(i)}$ won. (We assume there are no ties.)

(a) Formulate the problem of finding the maximum likelihood estimate of team abilities, $\hat{a} \in \mathbf{R}^n$, given the outcomes, as a convex optimization problem. You will find the *game incidence matrix* $A \in \mathbf{R}^{m \times n}$, defined as

$$A_{il} = \begin{cases} y^{(i)} & l = j^{(i)} \\ -y^{(i)} & l = k^{(i)} \\ 0 & \text{otherwise,} \end{cases}$$

useful.

The prior constraints $\hat{a}_i \in [0, 1]$ should be included in the problem formulation. Also, we note that if a constant is added to all team abilities, there is no change in the probabilities of game outcomes. This means that $\hat{a}$ is determined only up to a constant, like a potential. But this doesn't affect the ML estimation problem, or any subsequent predictions made using the estimated parameters.

(b) Find $\hat{a}$ for the team data given in `team_data.m`, in the matrix `train`. (This matrix gives the outcomes for a tournament in which each team plays each other team once.) You may find the `cvx` function `log_normcdf` helpful for this problem.

You can form A using the commands

```
A = sparse(1:m,train(:,1),train(:,3),m,n) + ...
    sparse(1:m,train(:,2),-train(:,3),m,n);
```

(c) Use the maximum likelihood estimate $\hat{a}$ found in part (b) to predict the outcomes of next year's tournament games, given in the matrix `test`, using $\hat{y}^{(i)} = \mathbf{sign}(\hat{a}_{j^{(i)}} - \hat{a}_{k^{(i)}})$. Compare these predictions with the actual outcomes, given in the third column of `test`. Given the fraction of correctly predicted outcomes.

The games played in `train` and `test` are the same, so another, simpler method for predicting the outcomes in `test` it to just assume the team that won last year's match will also win this year's match. Give the percentage of correctly predicted outcomes using this simple method.

6.5 *Estimating a vector with unknown measurement nonlinearity.* (A specific instance of exercise 7.9 in *Convex Optimization*.) We want to estimate a vector $x \in \mathbf{R}^n$, given some measurements

$$y_i = \phi(a_i^T x + v_i), \quad i = 1, \ldots, m.$$

Here $a_i \in \mathbf{R}^n$ are known, v_i are IID $\mathcal{N}(0, \sigma^2)$ random noises, and $\phi : \mathbf{R} \to \mathbf{R}$ is an unknown monotonic increasing function, known to satisfy

$$\alpha \leq \phi'(u) \leq \beta,$$

for all u. (Here α and β are known positive constants, with $\alpha < \beta$.) We want to find a maximum likelihood estimate of x and ϕ, given y_i. (We also know a_i, σ, α, and β.)

This sounds like an infinite-dimensional problem, since one of the parameters we are estimating is a function. In fact, we only need to know the m numbers $z_i = \phi^{-1}(y_i)$, $i = 1, \ldots, m$. So by estimating ϕ we really mean estimating the m numbers $z_1, \ldots, z_m$. (These numbers are not arbitrary; they must be consistent with the prior information $\alpha \leq \phi'(u) \leq \beta$ for all u.)

(a) Explain how to find a maximum likelihood estimate of x and ϕ (*i.e.*, $z_1, \ldots, z_m$) using convex optimization.

(b) Carry out your method on the data given in `nonlin_meas_data.m`, which includes a matrix $A \in \mathbf{R}^{m \times n}$, with rows $a_1^T, \ldots, a_m^T$. Give $\hat{x}_{\mathrm{ml}}$, the maximum likelihood estimate of x. Plot your estimated function $\hat{\phi}_{\mathrm{ml}}$. (You can do this by plotting $(\hat{z}_{\mathrm{ml}})_i$ versus y_i, with y_i on the vertical axis and $(\hat{z}_{\mathrm{ml}})_i$ on the horizontal axis.)

Hint. You can assume the measurements are numbered so that y_i are sorted in nondecreasing order, *i.e.*, $y_1 \leq y_2 \leq \cdots \leq y_m$. (The data given in the problem instance for part (b) is given in this order.)

6.6 *Maximum likelihood estimation of an increasing nonnegative signal.* We wish to estimate a scalar signal $x(t)$, for $t = 1, 2, \ldots, N$, which is known to be nonnegative and monotonically nondecreasing:

$$0 \leq x(1) \leq x(2) \leq \cdots \leq x(N).$$

This occurs in many practical problems. For example, $x(t)$ might be a measure of wear or deterioration, that can only get worse, or stay the same, as time t increases. We are also given that $x(t) = 0$ for $t \leq 0$.

We are given a noise-corrupted moving average of x, given by

$$y(t) = \sum_{\tau=1}^{k} h(\tau)x(t-\tau) + v(t), \quad t = 2, \ldots, N+1,$$

where $v(t)$ are independent $\mathcal{N}(0,1)$ random variables.

(a) Show how to formulate the problem of finding the maximum likelihood estimate of x, given y, taking into account the prior assumption that x is nonnegative and monotonically nondecreasing, as a convex optimization problem. Be sure to indicate what the problem variables are, and what the problem data are.

(b) We now consider a specific instance of the problem, with problem data (*i.e.*, N, k, h, and y) given in the file `l_estim_incr_signal_data.m`. Find the maximum likelihood estimate $\hat{x}_{\mathrm{ml}}$, and plot it. Also find the maximum likelihood estimate $\hat{x}_{\mathrm{ml,free}}$ *not taking into account the signal nonnegativity and monotonicity*, and plot it as well. Be sure to explain how you solve the problem, and to give your source code for the solution.

(*Note:* We will reveal the true signal x used to generate the data in the solutions.)

6.7 *Relaxed and discrete A-optimal experiment design.* This problem concerns the A-optimal experiment design problem, described on page 387, with data generated as follows.

```
n = 5; % dimension of parameters to be estimated
p = 20; % number of available types of measurements
m = 30; % total number of measurements to be carried out
randn('state', 0);
V=randn(n,p); % columns are vi, the possible measurement vectors
```

Solve the relaxed A-optimal experiment design problem,

$$\begin{array}{ll} \text{minimize} & (1/m)\,\mathbf{tr}\left(\sum_{i=1}^{p} \lambda_i v_i v_i^T\right)^{-1} \\ \text{subject to} & \mathbf{1}^T\lambda = 1, \quad \lambda \succeq 0, \end{array}$$

with variable $\lambda \in \mathbf{R}^p$. Find the optimal point $\lambda^\star$ and the associated optimal value of the relaxed problem. This optimal value is a lower bound on the optimal value of the discrete A-optimal experiment design problem,

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\left(\sum_{i=1}^{p} m_i v_i v_i^T\right)^{-1} \\ \text{subject to} & m_1 + \cdots + m_p = m, \quad m_i \in \{0, \ldots, m\}, \quad i = 1, \ldots, p, \end{array}$$

with variables $m_1, \ldots, m_p$. To get a suboptimal point for this discrete problem, round the entries in $m\lambda^\star$ to obtain integers $\hat{m}_i$. If needed, adjust these by hand or some other method to ensure that they sum to m, and compute the objective value obtained. This is, of course, an upper bound on the optimal value of the discrete problem. Give the gap between this upper bound and the lower bound obtained from the relaxed problem. Note that the two objective values can be interpreted as mean-square estimation error $\mathbf{E}\,\|\hat{x} - x\|_2^2$.

6.8 *Optimal detector design.* We adopt here the notation of §7.3 of the book. Explain how to design a (possibly randomized) detector that minimizes the worst-case probability of our estimate being off by more than one,

$$P_{\text{wc}} = \max_{\theta} \mathbf{prob}(|\hat{\theta} - \theta| \geq 2).$$

(The probability above is under the distribution associated with θ.)

Carry out your method for the problem instance with data in `off_by_one_det_data.m`. Give the optimal detection probability matrix D. Compare the optimal worst-case probability $P_{\text{wc}}^{\star}$ with the worst-case probability $P_{\text{wc}}^{\text{ml}}$ obtained using a maximum-likelihood detector.

6.9 *Experiment design with condition number objective.* Explain how to solve the experiment design problem (§7.5) with the condition number $\mathbf{cond}(E)$ of E (the error covariance matrix) as the objective to be minimized.

7 Geometry

7.1 *Efficiency of maximum volume inscribed ellipsoid.* In this problem we prove the following geometrical result. Suppose C is a polyhedron in $\mathbf{R}^n$, symmetric about the origin, and described as

$$C = \{x \mid -1 \leq a_i^T x \leq 1,\ i = 1, \ldots, p\}.$$

Let

$$\mathcal{E} = \{x \mid x^T Q^{-1} x \leq 1\},$$

with $Q \in \mathbf{S}^n_{++} ucc0$, be the maximum volume ellipsoid with center at the origin, inscribed in C. Then the ellipsoid

$$\sqrt{n}\mathcal{E} = \{x \mid x^T Q^{-1} x \leq n\}$$

(*i.e.*, the ellipsoid $\mathcal{E}$, scaled by a factor $\sqrt{n}$ about the origin) *contains* C.

(a) Show that the condition $\mathcal{E} \subseteq C$ is equivalent to $a_i^T Q a_i \leq 1$ for $i = 1, \ldots, p$.

(b) The volume of $\mathcal{E}$ is proportional to $(\det Q)^{1/2}$, so we can find the maximum volume ellipsoid $\mathcal{E}$ inside C by solving the convex problem

$$\begin{array}{ll} \text{minimize} & \log\det Q^{-1} \\ \text{subject to} & a_i^T Q a_i \leq 1, \quad i = 1. \ldots, p. \end{array} \tag{20}$$

The variable is the matrix $Q \in \mathbf{S}^n$ and the domain of the objective function is $\mathbf{S}^n_{++}$.
Derive the Lagrange dual of problem (20).

(c) Note that Slater's condition for (20) holds ($a_i^T Q a_i < 1$ for $Q = \epsilon I$ and $\epsilon > 0$ small enough), so we have strong duality, and the KKT conditions are necessary and sufficient for optimality. What are the KKT conditions for (20)?
Suppose Q is optimal. Use the KKT conditions to show that

$$x \in C \quad \Longrightarrow \quad x^T Q^{-1} x \leq n.$$

In other words $C \subseteq \sqrt{n}\mathcal{E}$, which is the desired result.

7.2 *Euclidean distance matrices.* A matrix $X \in \mathbf{S}^n$ is a *Euclidean distance matrix* if its elements x_{ij} can be expressed as

$$x_{ij} = \|p_i - p_j\|_2^2, \quad i, j = 1, \ldots, n,$$

for some vectors $p_1, \ldots, p_n$ (of arbitrary dimension). In this exercise we prove several classical characterizations of Euclidean distance matrices, derived by I. Schoenberg in the 1930s.

(a) Show that X is a Euclidean distance matrix if and only if

$$X = \mathbf{diag}(Y)\mathbf{1}^T + \mathbf{1}\,\mathbf{diag}(Y)^T - 2Y \tag{21}$$

for some matrix $Y \in \mathbf{S}^n_+$ (the symmetric positive semidefinite matrices of order n). Here, $\mathbf{diag}(Y)$ is the n-vector formed from the diagonal elements of Y, and $\mathbf{1}$ is the n-vector with all its elements equal to one. The equality (21) is therefore equivalent to

$$x_{ij} = y_{ii} + y_{jj} - 2y_{ij}, \quad i, j = 1, \ldots, n.$$

Hint. Y is the Gram matrix associated with the vectors $p_1, \ldots, p_n$, *i.e.*, the matrix with elements $y_{ij} = p_i^T p_j$.

(b) Show that the set of Euclidean distance matrices is a convex cone.

(c) Show that X is a Euclidean distance matrix if and only if

$$\mathbf{diag}(X) = 0, \qquad X_{22} - X_{21}\mathbf{1}^T - \mathbf{1}X_{21}^T \preceq 0. \tag{22}$$

The subscripts refer to the partitioning

$$X = \begin{bmatrix} x_{11} & X_{21}^T \\ X_{21} & X_{22} \end{bmatrix}$$

with $X_{21} \in \mathbf{R}^{n-1}$, and $X_{22} \in \mathbf{S}^{n-1}$.

Hint. The definition of Euclidean distance matrix involves only the distances $\|p_i - p_j\|_2$, so the origin can be chosen arbitrarily. For example, it can be assumed without loss of generality that $p_1 = 0$. With this assumption there is a unique Gram matrix Y for a given Euclidean distance matrix X. Find Y from (21), and relate it to the lefthand side of the inequality (22).

(d) Show that X is a Euclidean distance matrix if and only if

$$\mathbf{diag}(X) = 0, \qquad (I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T) \preceq 0. \tag{23}$$

Hint. Use the same argument as in part (c), but take the mean of the vectors p_k at the origin, *i.e.*, impose the condition that $p_1 + p_2 + \cdots + p_n = 0$.

(e) Suppose X is a Euclidean distance matrix. Show that the matrix $W \in \mathbf{S}^n$ with elements

$$w_{ij} = e^{-x_{ij}}, \quad i, j = 1, \ldots, n,$$

is positive semidefinite.

Hint. Use the following identity from probability theory. Define $z \sim \mathcal{N}(0, I)$ (the normal distribution with zero mean and covariance I). Then

$$\mathbf{E}\, e^{izx} = e^{-\frac{1}{2}\|x\|_2^2}$$

for all x, where $i = \sqrt{-1}$ and $\mathbf{E}$ denotes expectation with respect to z. (This is the characteristic function of a multivariate normal distribution.)

7.3 *Minimum total covering ball volume.* We consider a collection of n points with locations $x_1, \ldots, x_n \in \mathbf{R}^k$. We are also given a set of m groups or subsets of these points, $G_1, \ldots, G_m \subseteq \{1, \ldots, n\}$. For each group, let V_i be the volume of the smallest Euclidean ball that contains the points in group G_i. (The volume of a Euclidean ball of radius r in $\mathbf{R}^k$ is $a_k r^k$, where a_k is known constant that is positive but otherwise irrelevant here.) We let $V = V_1 + \cdots + V_m$ be the total volume of these minimal covering balls.

The points $x_{k+1}, \ldots, x_n$ are fixed (*i.e.*, they are problem data). The variables to be chosen are $x_1, \ldots, x_k$. Formulate the problem of choosing $x_1, \ldots, x_k$, in order to minimize the total minimal covering ball volume V, as a convex optimization problem. Be sure to explain any new variables you introduce, and to justify the convexity of your objective and inequality constraint functions.

7.4 *Maximum-margin multiclass classification.* In an m-category pattern classification problem, we are given m sets $C_i \subseteq \mathbf{R}^n$. Set C_i contains N_i examples of feature vectors in class i. The learning problem is to find a decision function $f : \mathbf{R}^n \to \{1, 2, \ldots, m\}$ that maps each training example to its class, and also generalizes reliably to feature vectors that are not included in the training sets C_i.

(a) A common type of decision function for two-way classification is

$$f(x) = \begin{cases} 1 & \text{if } a^T x + b > 0 \\ 2 & \text{if } a^T x + b < 0. \end{cases}$$

In the simplest form, finding f is equivalent to solving a feasibility problem: find a and b such that

$$\begin{array}{ll} a^T x + b > 0 & \text{if } x \in C_1 \\ a^T x + b < 0 & \text{if } x \in C_2. \end{array}$$

Since these strict inequalities are homogeneous in a and b, they are feasible if and only if the nonstrict inequalities

$$\begin{array}{ll} a^T x + b \geq 1 & \text{if } x \in C_1 \\ a^T x + b \leq -1 & \text{if } x \in C_2 \end{array}$$

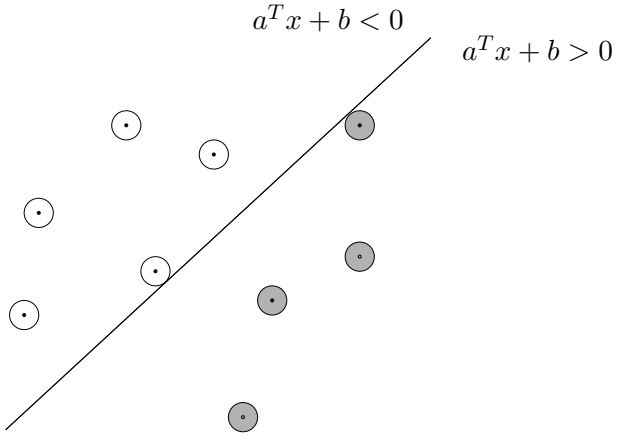
are feasible. This is a feasibility problem with $N_1 + N_2$ linear inequalities in $n + 1$ variables a, b.

As an extension that improves the robustness (*i.e.*, generalization capability) of the classifier, we can impose the condition that the decision function f classifies all points in a neighborhood of C_1 and C_2 correctly, and we can maximize the size of the neighborhood. This problem can be expressed as

$$\begin{array}{lll} \text{maximize} & t & \\ \text{subject to} & a^T x + b > 0 & \text{if } \mathbf{dist}(x, C_1) \leq t, \\ & a^T x + b < 0 & \text{if } \mathbf{dist}(x, C_2) \leq t, \end{array}$$

where $\mathbf{dist}(x, C) = \min_{y \in C} \|x - y\|_2$.

This is illustrated in the figure. The centers of the shaded disks form the set C_1. The centers of the other disks form the set C_2. The set of points at a distance less than t from C_i is the union of disks with radius t and center in C_i. The hyperplane in the figure separates the two expanded sets. We are interested in expanding the circles as much as possible, until the two expanded sets are no longer separable by a hyperplane.



Since the constraints are homogeneous in a, b, we can again replace them with nonstrict inequalities

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & a^T x + b \geq 1 \ \text{ if } \mathbf{dist}(x, C_1) \leq t, \\ & a^T x + b \leq -1 \ \text{ if } \mathbf{dist}(x, C_2) \leq t. \end{array} \tag{24}$$

The variables are a, b, and t.

(b) Next we consider an extension to more than two classes. If $m > 2$ we can use a decision function

$$f(x) = \operatorname*{argmax}_{i=1,\ldots,m} (a_i^T x + b_i),$$

parameterized by m vectors $a_i \in \mathbf{R}^n$ and m scalars b_i. To find f, we can solve a feasibility problem: find a_i, b_i, such that

$$a_i^T x + b_i > \max_{j \neq i} (a_j^T x + b_j) \quad \text{if } x \in C_i, \quad i = 1, \ldots, m,$$

or, equivalently,

$$a_i^T x + b_i \geq 1 + \max_{j \neq i} (a_j^T x + b_j) \quad \text{if } x \in C_i, \quad i = 1, \ldots, m.$$

Similarly as in part (a), we consider a robust version of this problem:

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & a_i^T x + b_i \geq 1 + \max_{j \neq i} (a_j^T x + b_j) \text{ if } \mathbf{dist}(x, C_i) \leq t, \\ & \qquad i = 1, \ldots, m. \end{array} \tag{25}$$

The variables in the problem are $a_i \in \mathbf{R}^n$, $b_i \in \mathbf{R}$, $i = 1, \ldots, m$, and t.

Formulate the optimization problems (24) and (25) as SOCPs (if possible), or as quasiconvex optimization problems involving SOCP feasibility problems (otherwise).

7.5 *Three-way linear classification.* We are given data

$$x^{(1)}, \ldots, x^{(N)}, \quad y^{(1)}, \ldots, y^{(M)}, \quad z^{(1)}, \ldots, z^{(P)},$$

three nonempty sets of vectors in $\mathbf{R}^n$. We wish to find three affine functions on $\mathbf{R}^n$,

$$f_i(z) = a_i^T z - b_i, \quad i = 1, 2, 3,$$

that satisfy the following properties:

$$\begin{array}{l} f_1(x^{(j)}) > \max\{f_2(x^{(j)}), f_3(x^{(j)})\}, \quad j = 1, \ldots, N, \\ f_2(y^{(j)}) > \max\{f_1(y^{(j)}), f_3(y^{(j)})\}, \quad j = 1, \ldots, M, \\ f_3(z^{(j)}) > \max\{f_1(z^{(j)}), f_2(z^{(j)})\}, \quad j = 1, \ldots, P. \end{array}$$

In words: f_1 is the largest of the three functions on the x data points, f_2 is the largest of the three functions on the y data points, f_3 is the largest of the three functions on the z data points. We

can give a simple geometric interpretation: The functions f_1, f_2, and f_3 partition $\mathbf{R}^n$ into three regions,

$$\begin{aligned} R_1 &= \{z \mid f_1(z) > \max\{f_2(z), f_3(z)\}\}, \\ R_2 &= \{z \mid f_2(z) > \max\{f_1(z), f_3(z)\}\}, \\ R_3 &= \{z \mid f_3(z) > \max\{f_1(z), f_2(z)\}\}, \end{aligned}$$

defined by where each function is the largest of the three. Our goal is to find functions with $x^{(j)} \in R_1$, $y^{(j)} \in R_2$, and $z^{(j)} \in R_3$.

Pose this as a convex optimization problem. You may not use strict inequalities in your formulation.

Solve the specific instance of the 3-way separation problem given in `sep3way_data.m`, with the columns of the matrices `X`, `Y` and `Z` giving the $x^{(j)}$, $j = 1, \ldots, N$, $y^{(j)}$, $j = 1, \ldots, M$ and $z^{(j)}$, $j = 1, \ldots, P$. To save you the trouble of plotting data points and separation boundaries, we have included the plotting code in `sep3way_data.m`. (Note that `a1`, `a2`, `a3`, `b1` and `b2` contain arbitrary numbers; you should compute the correct values using CVX.)

7.6 *Feature selection and sparse linear separation.* Suppose $x^{(1)}, \ldots, x^{(N)}$ and $y^{(1)}, \ldots, y^{(M)}$ are two given nonempty collections or classes of vectors in $\mathbf{R}^n$ that can be (strictly) separated by a hyperplane, *i.e.*, there exists $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$ such that

$$a^T x^{(i)} - b \geq 1, \quad i = 1, \ldots, N, \qquad a^T y^{(i)} - b \leq -1, \quad i = 1, \ldots, M.$$

This means the two classes are (weakly) separated by the slab

$$S = \{z \mid |a^T z - b| \leq 1\},$$

which has thickness $2/\|a\|_2$. You can think of the components of $x^{(i)}$ and $y^{(i)}$ as *features*; a and b define an affine function that combines the features and allows us to distinguish the two classes.

To find the thickest slab that separates the two classes, we can solve the QP

$$\begin{array}{ll} \text{minimize} & \|a\|_2 \\ \text{subject to} & a^T x^{(i)} - b \geq 1, \quad i = 1, \ldots, N \\ & a^T y^{(i)} - b \leq -1, \quad i = 1, \ldots, M, \end{array}$$

with variables $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$. (This is equivalent to the problem given in (8.23), p424, §8.6.1; see also exercise 8.23.)

In this problem we seek (a, b) that separate the two classes with a thick slab, and also has a sparse, *i.e.*, there are many j with $a_j = 0$. Note that if $a_j = 0$, the affine function $a^T z - b$ does not depend on z_j, *i.e.*, the jth feature is not used to carry out classification. So a sparse a corresponds to a classification function that is parsimonious; it depends on just a few features. So our goal is to find an affine classification function that gives a thick separating slab, and also uses as few features as possible to carry out the classification.

This is in general a hard combinatorial (bi-criterion) optimization problem, so we use the standard heuristic of solving

$$\begin{array}{ll} \text{minimize} & \|a\|_2 + \lambda\|a\|_1 \\ \text{subject to} & a^T x^{(i)} - b \geq 1, \quad i = 1, \ldots, N \\ & a^T y^{(i)} - b \leq -1, \quad i = 1, \ldots, M, \end{array}$$

where $\lambda \geq 0$ is a weight vector that controls the trade-off between separating slab thickness and (indirectly, through the ℓ_1 norm) sparsity of a.

Get the data in `sp_ln_sp_data.m`, which gives $x^{(i)}$ and $y^{(i)}$ as the columns of matrices `X` and `Y`, respectively. Find the thickness of the maximum thickness separating slab. Solve the problem above for 100 or so values of λ over an appropriate range (we recommend log spacing). For each value, record the separation slab thickness $2/\|a\|_2$ and $\mathbf{card}(a)$, the cardinality of a (*i.e.*, the number of nonzero entries). In computing the cardinality, you can count an entry a_j of a as zero if it satisfies $|a_j| \leq 10^{-4}$. Plot these data with slab thickness on the vertical axis and cardinality on the horizontal axis.

Use this data to choose a set of 10 features out of the 50 in the data. Give the indices of the features you choose. You may have several choices of sets of features here; you can just choose one. Then find the maximum thickness separating slab that uses only the chosen features. (This is standard practice: once you've chosen the features you're going to use, you optimize again, using only those features, and without the ℓ_1 regularization.

7.7 *Thickest slab separating two sets.* We are given two sets in $\mathbf{R}^n$: a polyhedron

$$C_1 = \{x \mid Cx \preceq d\},$$

defined by a matrix $C \in \mathbf{R}^{m\times n}$ and a vector $d \in \mathbf{R}^m$, and an ellipsoid

$$C_2 = \{Pu + q \mid \|u\|_2 \leq 1\},$$

defined by a matrix $P \in \mathbf{R}^{n\times n}$ and a vector $q \in \mathbf{R}^n$. We assume that the sets are nonempty and that they do not intersect. We are interested in the optimization problem

$$\begin{array}{ll} \text{maximize} & \inf_{x\in C_1} a^T x - \sup_{x\in C_2} a^T x \\ \text{subject to} & \|a\|_2 = 1. \end{array}$$

with variable $a \in \mathbf{R}^n$.

Explain how you would solve this problem. You can answer the question by reducing the problem to a standard problem class (LP, QP, SOCP, SDP, ...), or by describing an algorithm to solve it.

Remark. The geometrical interpretation is as follows. If we choose

$$b = \frac{1}{2}(\inf_{x\in C_1} a^T x + \sup_{x\in C_2} a^T x),$$

then the hyperplane $H = \{x \mid a^T x = b\}$ is the maximum margin separating hyperplane separating C_1 and C_2. Alternatively, a gives us the thickest slab that separates the two sets.

7.8 *Bounding object position from multiple camera views.* A small object is located at unknown position $x \in \mathbf{R}^3$, and viewed by a set of m cameras. Our goal is to find a box in $\mathbf{R}^3$,

$$\mathcal{B} = \{z \in \mathbf{R}^3 \mid l \preceq z \preceq u\},$$

for which we can guarantee $x \in \mathcal{B}$. We want the smallest possible such bounding box. (Although it doesn't matter, we can use volume to judge 'smallest' among boxes.)

Now we describe the cameras. The object at location $x \in \mathbf{R}^3$ creates an image on the image plane of camera i at location

$$v_i = \frac{1}{c_i^T x + d_i}(A_i x + b_i) \in \mathbf{R}^2.$$

The matrices $A_i \in \mathbf{R}^{2\times 3}$, vectors $b_i \in \mathbf{R}^2$ and $c_i \in \mathbf{R}^3$, and real numbers $d_i \in \mathbf{R}$ are known, and depend on the camera positions and orientations. We assume that $c_i^T x + d_i > 0$. The 3×4 matrix

$$P_i = \begin{bmatrix} A_i & b_i \\ c_i^T & d_i \end{bmatrix}$$

is called the *camera matrix* (for camera i). It is often (but not always) the case the that the first 3 columns of P_i (*i.e.*, A_i stacked above c_i^T) form an orthogonal matrix, in which case the camera is called *orthographic*.

We do not have direct access to the image point v_i; we only know the (square) pixel that it lies in. In other words, the camera gives us a measurement $\hat{v}_i$ (the center of the pixel that the image point lies in); we are guaranteed that

$$\|v_i - \hat{v}_i\|_\infty \leq \rho_i/2,$$

where ρ_i is the pixel width (and height) of camera i. (We know nothing else about v_i; it could be any point in this pixel.)

Given the data A_i, b_i, c_i, d_i, $\hat{v}_i$, ρ_i, we are to find the smallest box $\mathcal{B}$ (*i.e.*, find the vectors l and u) that is guaranteed to contain x. In other words, find the smallest box in $\mathbf{R}^3$ that contains all points consistent with the observations from the camera.

(a) Explain how to solve this using convex or quasiconvex optimization. You must explain any transformations you use, any new variables you introduce, etc. If the convexity or quasiconvexity of any function in your formulation isn't obvious, be sure justify it.

(b) Solve the specific problem instance given in the file `camera_data.m`. Be sure that your final numerical answer (*i.e.*, l and u) stands out.

7.9 *Triangulation from multiple camera views.* A projective camera can be described by a linear-fractional function $f : \mathbf{R}^3 \to \mathbf{R}^2$,

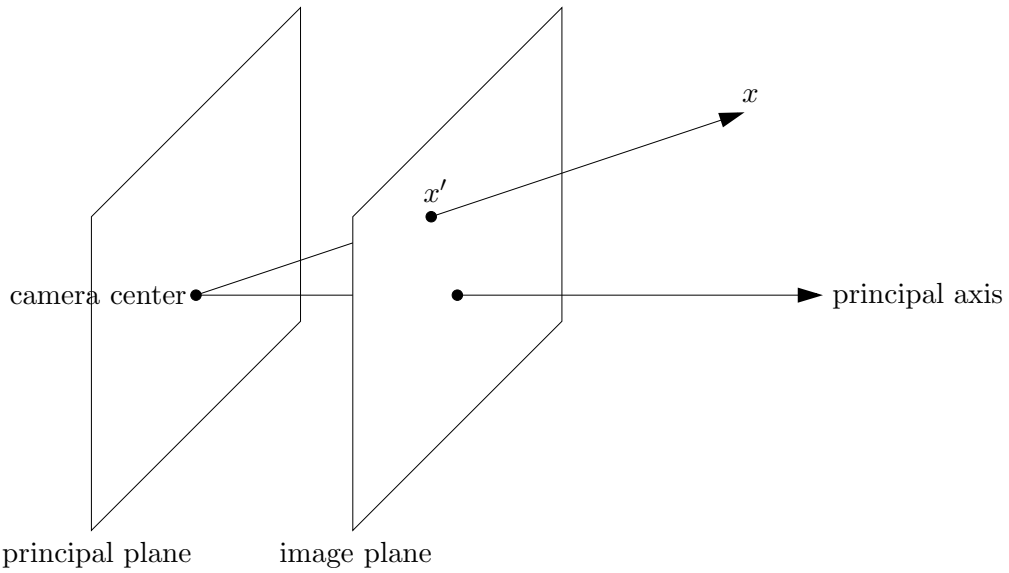
$$f(x) = \frac{1}{c^T x + d}(Ax + b), \qquad \mathbf{dom}\, f = \{x \mid c^T x + d > 0\},$$

with

$$\mathbf{rank}(\begin{bmatrix} A \\ c^T \end{bmatrix}) = 3.$$

The domain of f consists of the points in front of the camera.

Before stating the problem, we give some background and interpretation, most of which will not be needed for the actual problem.



The 3×4-matrix

$$P = \begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$$

is called the *camera matrix* and has rank 3. Since f is invariant with respect to a scaling of P, we can normalize the parameters and assume, for example, that $\|c\|_2 = 1$. The numerator $c^T x + d$ is then the distance of x to the plane $\{z \mid c^T z + d = 0\}$. This plane is called the *principal plane*. The point

$$x_{\rm c} = -\begin{bmatrix} A \\ c^T \end{bmatrix}^{-1} \begin{bmatrix} b \\ d \end{bmatrix}$$

lies in the principal plane and is called the *camera center*. The ray $\{x_{\rm c} + \theta c \mid \theta \geq 0\}$, which is perpendicular to the principal plane, is the *principal axis*. We will define the *image plane* as the plane parallel to the principal plane, at a unit distance from it along the principal axis.

The point x' in the figure is the intersection of the image plane and the line through the camera center and x, and is given by

$$x' = x_{\rm c} + \frac{1}{c^T(x - x_{\rm c})}(x - x_{\rm c}).$$

Using the definition of $x_{\rm c}$ we can write $f(x)$ as

$$f(x) = \frac{1}{c^T(x - x_{\rm c})} A(x - x_{\rm c}) = A(x' - x_{\rm c}) = Ax' + b.$$

This shows that the mapping $f(x)$ can be interpreted as a projection of x on the image plane to get x', followed by an affine transformation of x'. We can interpret $f(x)$ as the point x' expressed in some two-dimensional coordinate system attached to the image plane.

In this exercise we consider the problem of determining the position of a point $x \in \mathbf{R}^3$ from its image in N cameras. Each of the cameras is characterized by a known linear-fractional mapping f_k and camera matrix P_k:

$$f_k(x) = \frac{1}{c_k^T x + d_k}(A_k x + b_k), \qquad P_k = \begin{bmatrix} A_k & b_k \\ c_k^T & d_k \end{bmatrix}, \qquad k = 1, \ldots, N.$$

The image of the point x in camera k is denoted $y^{(k)} \in \mathbf{R}^2$. Due to camera imperfections and calibration errors, we do not expect the equations $f_k(x) = y^{(k)}$, $k = 1, \ldots, N$, to be exactly solvable. To estimate the point x we therefore minimize the maximum error in the N equations by solving

$$\text{minimize} \quad g(x) = \max_{k=1,\ldots,N} \|f_k(x) - y^{(k)}\|_2. \tag{26}$$

(a) Show that (26) is a quasiconvex optimization problem. The variable in the problem is $x \in \mathbf{R}^3$. The functions f_k (*i.e.*, the parameters A_k, b_k, c_k, d_k) and the vectors $y^{(k)}$ are given.

(b) Solve the following instance of (26) using CVX (and bisection): $N = 4$,

$$P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 10 \end{bmatrix},$$

$$P_3 = \begin{bmatrix} 1 & 1 & 1 & -10 \\ -1 & 1 & 1 & 0 \\ -1 & -1 & 1 & 10 \end{bmatrix}, \qquad P_4 = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 10 \end{bmatrix},$$

$$y^{(1)} = \begin{bmatrix} 0.98 \\ 0.93 \end{bmatrix}, \qquad y^{(2)} = \begin{bmatrix} 1.01 \\ 1.01 \end{bmatrix}, \qquad y^{(3)} = \begin{bmatrix} 0.95 \\ 1.05 \end{bmatrix}, \qquad y^{(4)} = \begin{bmatrix} 2.04 \\ 0.00 \end{bmatrix}.$$

You can terminate the bisection when a point is found with accuracy $g(x) - p^\star \leq 10^{-4}$, where $p^\star$ is the optimal value of (26).

7.10 *Ellipsoidal peeling.* In this exercise you will implement an outlier identification and removal technique called *ellipsoidal peeling*. We are given a set of points $x_1, \ldots, x_N \in \mathbf{R}^n$; our goal is to find a small (measured by volume) ellipsoid $\mathcal{E}$, for which $x_i \in \mathcal{E}$ for $i \notin \mathcal{O}$, where $\mathcal{O} \subset \{1, \ldots, N\}$ is the set of 'outliers'. Of course, there is a trade-off between $\mathbf{card}\,\mathcal{O}$ (the candinality of the set of outliers) and $\mathbf{vol}\,\mathcal{E}$. Of course, once we choose $\mathcal{O}$, we can find $\mathcal{E}$ as the minimum volume ellipsoid that contains x_i for $i \notin \mathcal{O}$.

Ellipsoidal peeling a heuristic for finding reasonable choices for $\mathcal{O}$. We start with $\mathcal{O} = \emptyset$, and find the minimum volume ellipsoid $\mathcal{E}$ containing all x_i for $i \notin \mathcal{O}$ (which, at this first step, is all x_i). Some of the points x_i will be on the surface of $\mathcal{E}$; we add these points to $\mathcal{O}$, and repeat. Roughly speaking, in each step we 'peel off' the points that lie on surface of the smallest volume enclosing ellipsoid. We then plot $\mathbf{vol}\,\mathcal{E}$ versus $\mathbf{card}\,\mathcal{O}$, and hope that we see a clear knee of the curve.

There are many variations on this approach. For example, instead of dropping all points on the surface of the current ellipsoid, we might drop only the one that corresponds to the largest Lagrange multiplier for the constraint that requires $x_i \in \mathcal{E}$.

Apply ellipsoidal peeling to the data given in `ellip_peel_data.m`. Plot $\mathbf{vol}\,E$ (on a log scale) versus $\mathbf{card}\,\mathcal{O}$. The data includes a list of the 'true' outliers. How did ellipsoidal peeling do?

Hint. In CVX, you should use `det_rootn` (which is handled exactly), rather than `log_det` (which is handled using an inefficient iterative procedure).

7.11 *Projection onto the probability simplex.* In this problem you will work out a simple method for finding the Euclidean projection y of $x \in \mathbf{R}^n$ onto the probability simplex $\mathcal{P} = \{z \mid z \succeq 0,\ \mathbf{1}^T z = 1\}$.

Hints. Consider the problem of minimizing $(1/2)\|y - x\|_2^2$ subject to $y \succeq 0$, $\mathbf{1}^T y = 1$. Form the partial Lagrangian

$$L(y, \nu) = (1/2)\|y - x\|_2^2 + \nu(\mathbf{1}^T y - 1),$$

leaving the constraint $y \succeq 0$ implicit. Show that $y = (x - \nu\mathbf{1})_+$ minimizes $L(y, \nu)$ over $y \succeq 0$.

7.12 *Minimum total covering ball volume.* We consider a collection of n points with locations $x_1, \ldots, x_n \in \mathbf{R}^k$. We are also given a set of m groups or subsets of these points, $G_1, \ldots, G_m \subseteq \{1, \ldots, n\}$. For each group, let V_i be the volume of the smallest Euclidean ball than contains the points in group G_i. (The volume of a Euclidean ball of radius r in $\mathbf{R}^k$ is $a_k r^k$, where a_k is known constant that is positive but otherwise irrelevant here.) We let $V = V_1 + \cdots + V_m$ be the total volume of these minimal covering balls.

The points $x_{k+1}, \ldots, x_n$ are fixed (*i.e.*, they are problem data). The variables to be chosen are $x_1, \ldots, x_k$. Formulate the problem of choosing $x_1, \ldots, x_k$, in order to minimize the total minimal covering ball volume V, as a convex optimization problem. Be sure to explain any new variables you introduce, and to justify the convexity of your objective and inequality constraint functions.

7.13 *Conformal mapping via convex optimization.* Suppose that Ω is a closed bounded region in $\mathbf{C}$ with no holes (*i.e.*, it is simply connected). The Riemann mapping theorem states that there exists a conformal mapping φ from Ω onto $D = \{z \in \mathbf{C} \mid |z| \leq 1\}$, the unit disk in the complex plane. (This means that φ is an analytic function, and maps Ω one-to-one onto D.)

One proof of the Riemann mapping theorem is based on an infinite dimensional optimization problem. We choose a point $a \in \mathbf{int}\,\Omega$ (the interior of Ω). Among all analytic functions that map $\partial\Omega$ (the boundary of Ω) into D, we choose one that maximizes the magnitude of the derivative at a. Amazingly, it can be shown that this function is a conformal mapping of Ω onto D.

We can use this theorem to construct an approximate conformal mapping, by sampling the boundary of Ω, and by restricting the optimization to a finite-dimensional subspace of analytic functions. Let $b_1, \ldots, b_N$ be a set of points in $\partial\Omega$ (meant to be a sampling of the boundary). We will search only over polynomials of degree up to n,

$$\hat{\varphi}(z) = \alpha_1 z^n + \alpha_2 z^{n-1} + \cdots + \alpha_n z + \alpha_{n+1},$$

where $\alpha_1, \ldots, \alpha_{n+1} \in \mathbf{C}$. With these approximations, we obtain the problem

$$\begin{array}{ll} \text{maximize} & |\hat{\varphi}'(a)| \\ \text{subject to} & |\hat{\varphi}(b_i)| \leq 1, \quad i = 1, \ldots, N, \end{array}$$

with variables $\alpha_1, \ldots, \alpha_{n+1} \in \mathbf{C}$. The problem data are $b_1, \ldots, b_N \in \partial\Omega$ and $a \in \mathbf{int}\,\Omega$.

(a) Explain how to solve the problem above via convex or quasiconvex optimization.

(b) Carry out your method on the problem instance given in `conf_map_data.m`. This file defines the boundary points b_i and plots them. It also contains code that will plot $\hat{\varphi}(b_i)$, the boundary of the mapped region, once you provide the values of α_j; these points should be very close to the boundary of the unit disk. (Please turn in this plot, and give us the values of α_j that you find.) The function `polyval` may be helpful.

Remarks.

- We've been a little informal in our mathematics here, but it won't matter.
- You do not need to know any complex analysis to solve this problem; we've told you everything you need to know.
- A basic result from complex analysis tells us that $\hat{\varphi}$ is one-to-one if and only if the image of the boundary does not 'loop over' itself. (We mention this just for fun; we're not asking you to verify that the $\hat{\varphi}$ you find is one-to-one.)

7.14 *Fitting a vector field to given directions.* This problem concerns a vector field on $\mathbf{R}^n$, *i.e.*, a function $F : \mathbf{R}^n \to \mathbf{R}^n$. We are given the *direction* of the vector field at points $x^{(1)}, \ldots, x^{(N)} \in \mathbf{R}^n$,

$$q^{(i)} = \frac{1}{\|F(x^{(i)})\|_2} F(x^{(i)}), \quad i = 1, \ldots, N.$$

(These directions might be obtained, for example, from samples of trajectories of the differential equation $\dot{z} = F(z)$.) The goal is to fit these samples with a vector field of the form

$$\hat{F} = \alpha_1 F_1 + \cdots + \alpha_m F_m,$$

where $F_1, \ldots, F_m : \mathbf{R}^n \to \mathbf{R}^n$ are given (basis) functions, and $\alpha \in \mathbf{R}^m$ is a set of coefficients that we will choose.

We will measure the fit using the maximum angle error,

$$J = \max_{i=1,\ldots,N} \left| \angle(q^{(i)}, \hat{F}(x^{(i)})) \right|,$$

where $\angle(z, w) = \cos^{-1}((z^T w)/\|z\|_2\|w\|_2)$ denotes the angle between nonzero vectors z and w. We are only interested in the case when J is smaller than $\pi/2$.

(a) Explain how to choose α so as to minimize J using convex optimization. Your method can involve solving multiple convex problems. Be sure to explain how you handle the constraints $\hat{F}(x^{(i)}) \neq 0$.

(b) Use your method to solve the problem instance with data given in `vfield_fit_data.m`, with an affine vector field fit, *i.e.*, $\hat{F}(z) = Az + b$. (The matrix A and vector b are the parameters α above.) Give your answer to the nearest degree, as in '$20^\circ < J^\star \leq 21^\circ$'.

This file also contains code that plots the vector field directions, and also (but commented out) the directions of the vector field fit, $\hat{F}(x^{(i)})/\|\hat{F}(x^{(i)})\|_2$. Create this plot, with your fitted vector field.

7.15 *Robust minimum volume covering ellipsoid.* Suppose z is a point in $\mathbf{R}^n$ and $\mathcal{E}$ is an ellipsoid in $\mathbf{R}^n$ with center c. The *Mahalanobis distance* of the point to the ellipsoid center is defined as

$$M(z, \mathcal{E}) = \inf\{t \geq 0 \mid z \in c + t(\mathcal{E} - c)\},$$

which is the factor by which we need to scale the ellipsoid about its center so that z is on its boundary. We have $z \in \mathcal{E}$ if and only if $M(z, \mathcal{E}) \leq 1$. We can use $(M(z, \mathcal{E}) - 1)_+$ as a measure of the Mahalanobis distance of the point z to the ellipsoid $\mathcal{E}$.

Now we can describe the problem. We are given m points $x_1, \ldots, x_m \in \mathbf{R}^n$. The goal is to find the optimal trade-off between the volume of the ellipsoid $\mathcal{E}$ and the total Mahalanobis distance of the points to the ellipsoid, *i.e.*,

$$\sum_{i=1}^{m} (M(z, \mathcal{E}) - 1)_+ .$$

Note that this can be considered a robust version of finding the smallest volume ellipsoid that covers a set of points, since here we allow one or more points to be outside the ellipsoid.

(a) Explain how to solve this problem. You must say clearly what your variables are, what problem you solve, and why the problem is convex.

(b) Carry out your method on the data given in `rob_min_vol_ellips_data.m`. Plot the optimal trade-off curve of ellipsoid volume versus total Mahalanobis distance. For some selected points on the trade-off curve, plot the ellipsoid and the points (which are in $\mathbf{R}^2$). We are only interested in the region of the curve where the ellipsoid volume is within a factor of ten (say) of the minimum volume ellipsoid that covers all the points.

Important. Depending on how you formulate the problem, you might encounter problems that are unbounded below, or where CVX encounters numerical difficulty. Just avoid these by appropriate choice of parameter.

Very important. If you use Matlab version 7.0 (which is filled with bugs) you might find that functions involving determinants don't work in CVX. If you use this version of Matlab, then you must download the file `blkdiag.m` on the course website and put it in your Matlab path before the default version (which has a bug).

8 Unconstrained and equality constrained minimization

8.1 *Suggestions for exercises 9.30 in* Convex Optimization. We recommend the following to generate a problem instance:

```
n = 100;
m = 200;
randn('state',1);
A=randn(m,n);
```

Of course, you should try out your code with different dimensions, and different data as well.

In all cases, be sure that your line search *first* finds a step length for which the tentative point is in $\mathbf{dom}\, f$; if you attempt to evaluate f outside its domain, you'll get complex numbers, and you'll never recover.

To find expressions for $\nabla f(x)$ and $\nabla^2 f(x)$, use the chain rule (see Appendix A.4); if you attempt to compute $\partial^2 f(x)/\partial x_i \partial x_j$, you will be sorry.

To compute the Newton step, you can use `vnt=-H\g`.

8.2 *Suggestions for exercise 9.31 in* Convex Optimization. For 9.31a, you should try out $N = 1$, $N = 15$, and $N = 30$. You might as well compute and store the Cholesky factorization of the Hessian, and then back solve to get the search directions, even though you won't really see any speedup in Matlab for such a small problem. After you evaluate the Hessian, you can find the Cholesky factorization as `L=chol(H,'lower')`. You can then compute a search step as `-L'\(L\g)`, where `g` is the gradient at the current point. Matlab will do the right thing, *i.e.*, it will first solve `L\g` using forward substitution, and then it will solve `-L'\(L\g)` using backward substitution. Each substitution is order n^2.

To fairly compare the convergence of the three methods (*i.e.*, $N = 1$, $N = 15$, $N = 30$), the horizontal axis should show the approximate total number of flops required, and not the number of iterations. You can compute the approximate number of flops using $n^3/3$ for each factorization, and $2n^2$ for each solve (where each 'solve' involves a forward substitution step and a backward substitution step).

8.3 *Efficient numerical method for a regularized least-squares problem.* We consider a regularized least squares problem with smoothing,

$$\text{minimize} \quad \sum_{i=1}^{k}(a_i^T x - b_i)^2 + \delta \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \eta \sum_{i=1}^{n} x_i^2,$$

where $x \in \mathbf{R}^n$ is the variable, and $\delta, \eta > 0$ are parameters.

(a) Express the optimality conditions for this problem as a set of linear equations involving x. (These are called the normal equations.)

(b) Now assume that $k \ll n$. Describe an efficient method to solve the normal equations found in (1). Give an approximate flop count for a general method that does not exploit structure, and also for your efficient method.

(c) *A numerical instance.* In this part you will try out your efficient method. We'll choose $k = 100$ and $n = 2000$, and $\delta = \eta = 1$. First, randomly generate A and b with these dimensions. Form the normal equations as in (1), and solve them using a generic method. Next, write (short) code implementing your efficient method, and run it on your problem instance. Verify that the solutions found by the two methods are nearly the same, and also that your efficient method is much faster than the generic one.

Note: You'll need to know some things about Matlab to be sure you get the speedup from the efficient method. Your method should involve solving linear equations with tridiagonal coefficient matrix. In this case, both the factorization and the back substitution can be carried out very efficiently. The Matlab documentation says that banded matrices are recognized and exploited, when solving equations, but we found this wasn't always the case. To be sure Matlab knows your matrix is tridiagonal, you can declare the matrix as sparse, using `spdiags`, which can be used to create a tridiagonal matrix. You could also create the tridiagonal matrix conventionally, and then convert the resulting matrix to a sparse one using `sparse`.

One other thing you need to know. Suppose you need to solve a group of linear equations with the same coefficient matrix, *i.e.*, you need to compute $F^{-1}a_1, ..., F^{-1}a_m$, where F is invertible and a_i are column vectors. By concatenating columns, this can be expressed as a single matrix

$$\left[F^{-1}a_1 \ \cdots \ F^{-1}a_m\right] = F^{-1}\left[a_1 \ \cdots \ a_m\right].$$

To compute this matrix using Matlab, you should collect the righthand sides into one matrix (as above) and use Matlab's backslash operator: `F\A`. This will do the right thing: factor the matrix F once, and carry out multiple back substitutions for the righthand sides.

8.4 *Newton method for approximate total variation de-noising.* Total variation de-noising is based on the bi-criterion problem with the two objectives

$$\|x - x^{\mathrm{cor}}\|_2, \qquad \phi_{\mathrm{tv}}(x) = \sum_{i=1}^{n-1} |x_{i+1} - x_i|.$$

Here $x^{\mathrm{cor}} \in \mathbf{R}^n$ is the (given) corrupted signal, $x \in \mathbf{R}^n$ is the de-noised signal to be computed, and ϕ_{tv} is the total variation function. This bi-criterion problem can be formulated as an SOCP, or, by squaring the first objective, as a QP. In this problem we consider a method used to approximately formulate the total variation de-noising problem as an unconstrained problem with twice differentiable objective, for which Newton's method can be used.

We first observe that the Pareto optimal points for the bi-criterion total variation de-noising problem can found as the minimizers of the function

$$\|x - x^{\mathrm{cor}}\|_2^2 + \mu\phi_{\mathrm{tv}}(x),$$

where $\mu \geq 0$ is parameter. (Note that the Euclidean norm term has been squared here, and so is twice differentiable.) In *approximate total variation de-noising*, we substitute a twice differentiable approximation of the total variation function,

$$\phi_{\mathrm{atv}}(x) = \sum_{i=1}^{n-1} \left(\sqrt{\epsilon^2 + (x_{i+1} - x_i)^2} - \epsilon\right),$$

for the total variation function $\phi_{\rm tv}$. Here $\epsilon > 0$ is parameter that controls the level of approximation. In approximate total variation de-noising, we use Newton's method to minimize

$$\psi(x) = \|x - x^{\rm cor}\|_2^2 + \mu\phi_{\rm atv}(x).$$

(The parameters $\mu > 0$ and $\epsilon > 0$ are given.)

(a) Find expressions for the gradient and Hessian of ψ.

(b) Explain how you would exploit the structure of the Hessian to compute the Newton direction for ψ efficiently. (Your explanation can be brief.) Compare the approximate flop count for your method with the flop count for a generic method that does not exploit any structure in the Hessian of ψ.

(c) Implement Newton's method for approximate total variation de-noising. Get the corrupted signal $x^{\rm cor}$ from the file `approx_tv_denoising_data.m`, and compute the de-noised signal $x^\star$, using parameters $\epsilon = 0.001$, $\mu = 50$ (which are also in the file). Use line search parameters $\alpha = 0.01$, $\beta = 0.5$, initial point $x^{(0)} = 0$, and stopping criterion $\lambda^2/2 \leq 10^{-8}$. Plot the Newton decrement versus iteration, to verify asymptotic quadratic convergence. Plot the final smoothed signal $x^\star$, along with the corrupted one $x^{\rm cor}$.

8.5 Derive the Newton equation for the unconstrained minimization problem

$$\text{minimize} \quad (1/2)x^T x + \log \textstyle\sum_{i=1}^m \exp(a_i^T x + b_i).$$

Give an efficient method for solving the Newton system, assuming the matrix $A \in \mathbf{R}^{m\times n}$ (with rows a_i^T) is dense with $m \ll n$. Give an approximate flop count of your method.

8.6 We consider the equality constrained problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}(CX) - \log\det X \\ \text{subject to} & \mathbf{diag}(X) = \mathbf{1}. \end{array}$$

The variable is the matrix $X \in \mathbf{S}^n$. The domain of the objective function is $\mathbf{S}^n_{++}$. The matrix $C \in \mathbf{S}^n$ is a problem parameter. This problem is similar to the analytic centering problem discussed in lecture 11 (p.18–19) and pages 553-555 of the textbook. The differences are the extra linear term $\mathbf{tr}(CX)$ in the objective, and the special form of the equality constraints. (Note that the equality constraints can be written as $\mathbf{tr}(A_i X) = 1$ with $A_i = e_i e_i^T$, a matrix of zeros except for the i, i element, which is equal to one.)

(a) Show that X is optimal if and only if

$$X \succ 0, \qquad X^{-1} - C \text{ is diagonal}, \qquad \mathbf{diag}(X) = \mathbf{1}.$$

(b) The Newton step ΔX at a feasible X is defined as the solution of the Newton equations

$$X^{-1}\Delta X X^{-1} + \mathbf{diag}(w) = -C + X^{-1}, \qquad \mathbf{diag}(\Delta X) = 0,$$

with variables $\Delta X \in \mathbf{S}^n$, $w \in \mathbf{R}^n$. (Note the two meanings of the **diag** function: $\mathbf{diag}(w)$ is the diagonal matrix with the vector w on its diagonal; $\mathbf{diag}(\Delta X)$ is the vector of the diagonal elements of ΔX.) Eliminating ΔX from the first equation gives an equation

$$\mathbf{diag}(X\,\mathbf{diag}(w)X) = \mathbf{1} - \mathbf{diag}(XCX).$$

This is a set of n linear equations in n variables, so it can be written as $Hw = g$. Give a simple expression for the coefficients of the matrix H.

(c) Implement the feasible Newton method in Matlab. You can use $X = I$ as starting point. The code should terminate when $\lambda(X)^2/2 \leq 10^{-6}$, where $\lambda(X)$ is the Newton decrement.

You can use the Cholesky factorization to evaluate the cost function: if $X = LL^T$ where L is triangular with positive diagonal then $\log\det X = 2\sum_i \log L_{ii}$.

To ensure that the iterates remain feasible, the line search has to consist of two phases. Starting at $t = 1$, you first need to backtrack until $X + t\Delta X \succ 0$. Then you continue the backtracking until the condition of sufficient decrease

$$f_0(X + t\Delta X) \leq f_0(X) + \alpha t\,\mathbf{tr}(\nabla f_0(X)\Delta X)$$

is satisfied. To check that a matrix $X + t\Delta X$ is positive definite, you can use the Cholesky factorization with two output arguments (`[R, p] = chol(A)` returns $p > 0$ if A is not positive definite).

Test your code on randomly generated problems of sizes $n = 10, \ldots, 100$ (for example, using `n = 100; C = randn(n); C = C + C'`).

8.7 *Estimation of a vector from one-bit measurements.* A system of m sensors is used to estimate an unknown parameter $x \in \mathbf{R}^n$. Each sensor makes a noisy measurement of some linear combination of the unknown parameters, and quantizes the measured value to one bit: it returns $+1$ if the measured value exceeds a certain threshold, and -1 otherwise. In other words, the output of sensor i is given by

$$y_i = \mathbf{sign}(a_i^T x + v_i - b_i) = \begin{cases} 1 & a_i^T x + v_i \geq b_i \\ -1 & a_i^T x + v_i < b_i, \end{cases}$$

where a_i and b_i are known, and v_i is measurement error. We assume that the measurement errors v_i are independent random variables with a zero-mean unit-variance Gaussian distribution (*i.e.*, with a probability density $\phi(v) = (1/\sqrt{2\pi})e^{-v^2/2}$). As a consequence, the sensor outputs y_i are random variables with possible values ± 1. We will denote $\mathbf{prob}(y_i = 1)$ as $P_i(x)$ to emphasize that it is a function of the unknown parameter x:

$$\begin{aligned} P_i(x) &= \mathbf{prob}(y_i = 1) = \mathbf{prob}(a_i^T x + v_i \geq b_i) = \frac{1}{\sqrt{2\pi}} \int_{b_i - a_i^T x}^{\infty} e^{-t^2/2} dt \\ 1 - P_i(x) &= \mathbf{prob}(y_i = -1) = \mathbf{prob}(a_i^T x + v_i < b_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{b_i - a_i^T x} e^{-t^2/2} dt. \end{aligned}$$

The problem is to estimate x, based on observed values $\bar{y}_1$, $\bar{y}_2$, $\ldots$, $\bar{y}_m$ of the m sensor outputs.

We will apply the maximum likelihood (ML) principle to determine an estimate $\hat{x}$. In maximum likelihood estimation, we calculate $\hat{x}$ by maximizing the *log-likelihood function*

$$l(x) = \log\left(\prod_{\bar{y}_i = 1} P_i(x) \prod_{\bar{y}_i = -1} (1 - P_i(x))\right) = \sum_{\bar{y}_i = 1} \log P_i(x) + \sum_{\bar{y}_i = -1} \log(1 - P_i(x)).$$

(a) Show that the maximum likelihood estimation problem

$$\text{maximize } l(x)$$

is a convex optimization problem. The variable is x. The measured vector $\bar{y}$, and the parameters a_i and b_i are given.

(b) Solve the ML estimation problem with data defined in `one_bit_meas_data.m`, using Newton's method with backtracking line search. This file will define a matrix A (with rows a_i^T), a vector b, and a vector $\bar{y}$ with elements ± 1.

Remark. The Matlab functions `erfc` and `erfcx` are useful to evaluate the following functions:

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-t^2/2}\,dt = \frac{1}{2}\mathbf{erfc}(-\frac{u}{\sqrt{2}}), \qquad \frac{1}{\sqrt{2\pi}}\int_{u}^{\infty} e^{-t^2/2}\,dt = \frac{1}{2}\mathbf{erfc}(\frac{u}{\sqrt{2}})$$

$$\frac{1}{\sqrt{2\pi}}e^{u^2/2}\int_{-\infty}^{u} e^{-t^2/2}\,dt = \frac{1}{2}\mathbf{erfcx}(-\frac{u}{\sqrt{2}}), \qquad \frac{1}{\sqrt{2\pi}}e^{u^2/2}\int_{u}^{\infty} e^{-t^2/2}\,dt = \frac{1}{2}\mathbf{erfcx}(\frac{u}{\sqrt{2}}).$$

9 Interior point methods

9.1 *Dual feasible point from analytic center.* We consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m, \end{array} \tag{27}$$

where the functions f_i are convex and differentiable. For $u > p^\star$, define $x_{\rm ac}(u)$ as the analytic center of the inequalities

$$f_0(x) \leq u, \qquad f_i(x) \leq 0, \quad i = 1, \ldots, m,$$

i.e.,

$$x_{\rm ac}(u) = \operatorname{argmin} \left(-\log(u - f_0(x)) - \sum_{i=1}^m \log(-f_i(x)) \right).$$

Show that $\lambda \in \mathbf{R}^m$, defined by

$$\lambda_i = \frac{u - f_0(x_{\rm ac}(u))}{-f_i(x_{\rm ac}(u))}, \quad i = 1, \ldots, m$$

is dual feasible for the problem above. Express the corresponding dual objective value in terms of u, $x_{\rm ac}(u)$ and the problem parameters.

9.2 *Efficient solution of Newton equations.* Explain how you would solve the Newton equations in the barrier method applied to the quadratic program

$$\begin{array}{ll} \text{minimize} & (1/2)x^T x + c^T x \\ \text{subject to} & Ax \preceq b \end{array}$$

where $A \in \mathbf{R}^{m \times n}$ is dense. Distinguish two cases, $m \gg n$ and $n \gg m$, and give the most efficient method in each case.

9.3 *Efficient solution of Newton equations.* Describe an efficient method for solving the Newton equation in the barrier method for the quadratic program

$$\begin{array}{ll} \text{minimize} & (1/2)(x - a)^T P^{-1}(x - a) \\ \text{subject to} & 0 \preceq x \preceq \mathbf{1}, \end{array}$$

with variable $x \in \mathbf{R}^n$. The matrix $P \in \mathbf{S}^n$ and the vector $a \in \mathbf{R}^n$ are given.

Assume that the matrix P is large, positive definite, and sparse, and that P^{-1} is dense. 'Efficient' means that the complexity of the method should be much less than $O(n^3)$.

9.4 *Dual feasible point from incomplete centering.* Consider the SDP

$$\begin{array}{ll} \text{minimize} & \mathbf{1}^T x \\ \text{subject to} & W + \mathbf{diag}(x) \succeq 0, \end{array}$$

with variable $x \in \mathbf{R}^n$, and its dual

$$\begin{array}{ll} \text{maximize} & -\mathbf{tr}\, WZ \\ \text{subject to} & Z_{ii} = 1, \quad i = 1, \ldots, n \\ & Z \succeq 0, \end{array}$$

with variable $X \in \mathbf{S}^n$. (These problems arise in a relaxation of the two-way partitioning problem, described on page 219; see also exercises 5.39 and 11.23.)

Standard results for the barrier method tell us that when x is on the central path, *i.e.*, minimizes the function

$$\phi(x) = t\mathbf{1}^T x + \log\det(W + \mathbf{diag}(x))^{-1}$$

for some parameter $t > 0$, the matrix

$$Z = \frac{1}{t}(W + \mathbf{diag}(x))^{-1}$$

is dual feasible, with objective value $-\mathbf{tr}\, WZ = \mathbf{1}^T x - n/t$.

Now suppose that x is strictly feasible, but not necessarily on the central path. (For example, x might be the result of using Newton's method to minimize ϕ, but with early termination.) Then the matrix Z defined above will not be dual feasible. In this problem we will show how to construct a dual feasible $\hat{Z}$ (which agrees with Z as given above when x is on the central path), from any point x that is *near* the central path. Define $X = W + \mathbf{diag}(x)$, and let $v = -\nabla^2\phi(x)^{-1}\nabla\phi(x)$ be the Newton step for the function ϕ defined above. Define

$$\hat{Z} = \frac{1}{t}\left(X^{-1} - X^{-1}\,\mathbf{diag}(v)X^{-1}\right).$$

(a) Verify that when x is on the central path, we have $\hat{Z} = Z$.

(b) Show that $\hat{Z}_{ii} = 1$, for $i = 1, \ldots, n$.

(c) Let $\lambda(x) = \nabla\phi(x)^T\nabla^2\phi(x)^{-1}\nabla\phi(x)$ be the Newton decrement at x. Show that

$$\lambda(x) = \mathbf{tr}(X^{-1}\,\mathbf{diag}(v)X^{-1}\,\mathbf{diag}(v)) = \mathbf{tr}(X^{-1/2}\,\mathbf{diag}(v)X^{-1/2})^2.$$

(d) Show that $\lambda(x) < 1$ implies that $\hat{Z} \succ 0$. Thus, when x is near the central path (meaning, $\lambda(x) < 1$), Z is dual feasible.

9.5 *Standard form LP barrier method.* In the following three parts of this exercise, you will implement a barrier method for solving the standard form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \quad x \succeq 0, \end{array}$$

with variable $x \in \mathbf{R}^n$, where $A \in \mathbf{R}^{m\times n}$, with $m < n$. Throughout these exercises we will assume that A is full rank, and the sublevel sets $\{x \mid Ax = b,\ x \succeq 0,\ c^T x \leq \gamma\}$ are all bounded. (If this is not the case, the centering problem is unbounded below.)

(a) *Centering step.* Implement Newton's method for solving the centering problem

$$\begin{array}{ll} \text{minimize} & c^T x - \sum_{i=1}^n \log x_i \\ \text{subject to} & Ax = b, \end{array}$$

with variable x, given a strictly feasible starting point x_0.

Your code should accept A, b, c, and x_0, and return $x^\star$, the primal optimal point, $\nu^\star$, a dual optimal point, and the number of Newton steps executed.

Use the block elimination method to compute the Newton step. (You can also compute the Newton step via the KKT system, and compare the result to the Newton step computed via block elimination. The two steps should be close, but if any x_i is very small, you might get a warning about the condition number of the KKT matrix.)

Plot $\lambda^2/2$ versus iteration k, for various problem data and initial points, to verify that your implementation gives asymptotic quadratic convergence. As stopping criterion, you can use $\lambda^2/2 \leq 10^{-6}$. Experiment with varying the algorithm parameters α and β, observing the effect on the total number of Newton steps required, for a fixed problem instance. Check that your computed $x^\star$ and $\nu^\star$ (nearly) satisfy the KKT conditions.

To generate some random problem data (*i.e.*, A, b, c, x_0), we recommend the following approach. First, generate A randomly. (You might want to check that it has full rank.) Then generate a random positive vector x_0, and take $b = Ax_0$. (This ensures that x_0 is strictly feasible.) The parameter c can be chosen randomly. To be sure the sublevel sets are bounded, you can add a row to A with all positive elements. If you want to be able to repeat a run with the same problem data, be sure to set the state for the uniform and normal random number generators.

Here are some hints that may be useful.

- We recommend computing λ^2 using the formula $\lambda^2 = -\Delta x_{\rm nt}^T \nabla f(x)$. You don't really need λ for anything; you can work with λ^2 instead. (This is important for reasons described below.)
- There can be small numerical errors in the Newton step $\Delta x_{\rm nt}$ that you compute. When x is nearly optimal, the computed value of λ^2, *i.e.*, $\lambda^2 = -\Delta x_{\rm nt}^T \nabla f(x)$, can actually be (slightly) negative. If you take the squareroot to get λ, you'll get a complex number, and you'll never recover. Moreover, your line search will never exit. However, this only happens when x is nearly optimal. So if you exit on the condition $\lambda^2/2 \leq 10^{-6}$, everything will be fine, even when the computed value of λ^2 is negative.
- For the line search, you must first multiply the step size t by β until $x + t\Delta x_{\rm nt}$ is feasible (*i.e.*, strictly positive). If you don't, when you evaluate f you'll be taking the logarithm of negative numbers, and you'll never recover.

(b) *LP solver with strictly feasible starting point.* Using the centering code from part (1), implement a barrier method to solve the standard form LP

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b, \quad x \succeq 0, \end{array}$$

with variable $x \in \mathbf{R}^n$, given a strictly feasible starting point x_0. Your LP solver should take as argument A, b, c, and x_0, and return $x^\star$.

You can terminate your barrier method when the duality gap, as measured by n/t, is smaller than 10^{-3}. (If you make the tolerance much smaller, you might run into some numerical trouble.) Check your LP solver against the solution found by `cvx`, for several problem instances.

The comments in part (1) on how to generate random data hold here too.

Experiment with the parameter μ to see the effect on the number of Newton steps per centering step, and the total number of Newton steps required to solve the problem.

Plot the progress of the algorithm, for a problem instance with $n = 500$ and $m = 100$, showing duality gap (on a log scale) on the vertical axis, versus the cumulative total number of Newton steps (on a linear scale) on the horizontal axis.

Your algorithm should return a $2 \times k$ matrix `history`, (where k is the total number of centering steps), whose first row contains the number of Newton steps required for each centering step, and whose second row shows the duality gap at the end of each centering step. In order to get a plot that looks like the ones in the book (*e.g.*, figure 11.4, page 572), you should use the following code:

```
[xx, yy] = stairs(cumsum(history(1,:)),history(2,:));
semilogy(xx,yy);
```

(c) *LP solver.* Using the code from part (2), implement a general standard form LP solver, that takes arguments A, b, c, determines (strict) feasibility, and returns an optimal point if the problem is (strictly) feasible.

You will need to implement a phase I method, that determines whether the problem is strictly feasible, and if so, finds a strictly feasible point, which can then be fed to the code from part (2). In fact, you can use the code from part (2) to implement the phase I method.

To find a strictly feasible initial point x_0, we solve the phase I problem

$$\begin{array}{ll} \text{minimize} & t \\ \text{subject to} & Ax = b \\ & x \succeq (1-t)\mathbf{1}, \quad t \geq 0, \end{array}$$

with variables x and t. If we can find a feasible (x, t), with $t < 1$, then x is strictly feasible for the original problem. The converse is also true, so the original LP is strictly feasible if and only if $t^\star < 1$, where $t^\star$ is the optimal value of the phase I problem.

We can initialize x and t for the phase I problem with any x^0 satisfying $Ax^0 = b$, and $t^0 = 2 - \min_i x_i^0$. (Here we can assume that $\min x_i^0 \leq 0$; otherwise x^0 is already a strictly feasible point, and we are done.) You can use a change of variable $z = x + (t-1)\mathbf{1}$ to transform the phase I problem into the form in part (2).

Check your LP solver against `cvx` on several numerical examples, including both feasible and infeasible instances.

10 Mathematical background

10.1 *Some famous inequalities.* The Cauchy-Schwarz inequality states that

$$|a^T b| \leq \|a\|_2 \|b\|_2$$

for all vectors a, $b \in \mathbf{R}^n$ (see page 633 of the textbook).

(a) Prove the Cauchy-Schwarz inequality.
Hint. A simple proof is as follows. With a and b fixed, consider the function $g(t) = \|a + tb\|_2^2$ of the scalar variable t. This function is nonnegative for all t. Find an expression for $\inf_t g(t)$ (the minimum value of g), and show that the Cauchy-Schwarz inequality follows from the fact that $\inf_t g(t) \geq 0$.

(b) The 1-norm of a vector x is defined as $\|x\|_1 = \sum_{k=1}^n |x_k|$. Use the Cauchy-Schwarz inequality to show that

$$\|x\|_1 \leq \sqrt{n}\|x\|_2$$

for all x.

(c) The *harmonic mean* of a positive vector $x \in \mathbf{R}^n_{++}$ is defined as

$$\left(\frac{1}{n}\sum_{k=1}^n \frac{1}{x_k}\right)^{-1}.$$

Use the Cauchy-Schwarz inequality to show that the arithmetic mean $(\sum_k x_k)/n$ of a positive n-vector is greater than or equal to its harmonic mean.

10.2 *Schur complements.* Consider a matrix $X = X^T \in \mathbf{R}^{n\times n}$ partitioned as

$$X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix},$$

where $A \in \mathbf{R}^{k\times k}$. If $\det A \neq 0$, the matrix $S = C - B^T A^{-1} B$ is called the *Schur complement* of A in X. Schur complements arise in many situations and appear in many important formulas and theorems. For example, we have $\det X = \det A \det S$. (You don't have to prove this.)

(a) The Schur complement arises when you minimize a quadratic form over some of the variables. Let $f(u, v) = (u, v)^T X (u, v)$, where $u \in \mathbf{R}^k$. Let $g(v)$ be the minimum value of f over u, *i.e.*, $g(v) = \inf_u f(u, v)$. Of course $g(v)$ can be $-\infty$.
Show that if $A \succ 0$, we have $g(v) = v^T S v$.

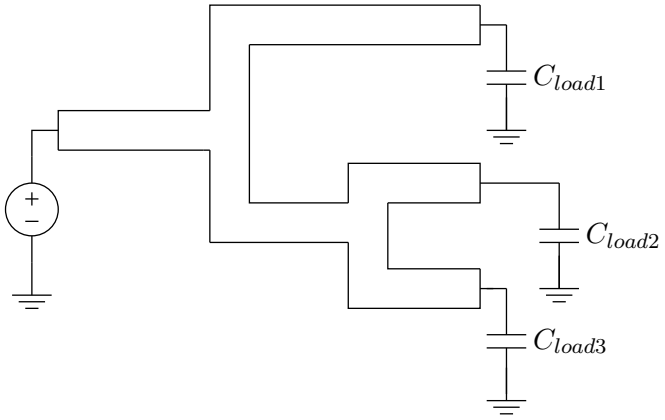
(b) The Schur complement arises in several characterizations of positive definiteness or semidefiniteness of a block matrix. As examples we have the following three theorems:

- $X \succ 0$ if and only if $A \succ 0$ and $S \succ 0$.
- If $A \succ 0$, then $X \succeq 0$ if and only if $S \succeq 0$.
- $X \succeq 0$ if and only if $A \succeq 0$, $B^T(I - AA^\dagger) = 0$ and $C - B^T A^\dagger B \succeq 0$, where $A^\dagger$ is the pseudo-inverse of A. ($C - B^T A^\dagger B$ serves as a generalization of the Schur complement in the case where A is positive semidefinite but singular.)

Prove *one* of these theorems. (You can choose which one.)

11 Circuit design

11.1 *Interconnect sizing.* In this problem we will size the interconnecting wires of the simple circuit shown below, with one voltage source driving three different capacitive loads C_{load1}, C_{load2}, and C_{load3}.



We divide the wires into 6 segments of fixed length l_i; our variables will be the widths w_i of the segments. (The height of the wires is related to the particular IC technology process, and is fixed.) The total area used by the wires is, of course,

$$A = \sum_i w_i l_i.$$

We'll take the lengths to be one, for simplicity. The wire widths must be between a minimum and maximum allowable value:

$$W_{\min} \leq w_i \leq W_{\max}.$$

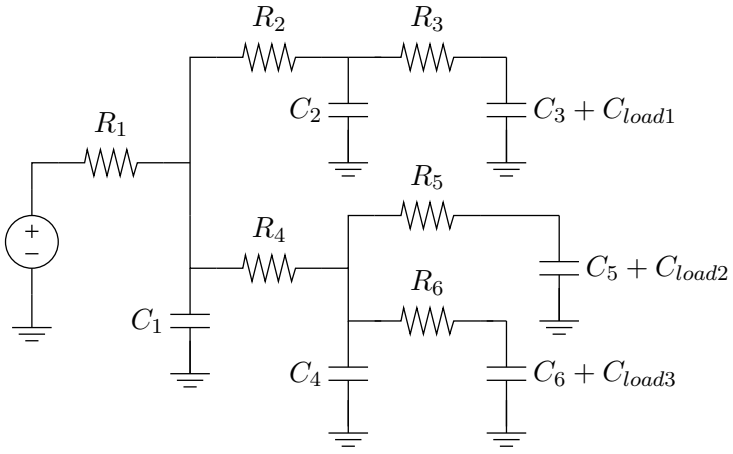
For our specific problem, we'll take $W_{\min} = 0.1$ and $W_{\max} = 10$.

Each of the wire segments will be modeled by a simple simple RC circuit, with the resistance inversely proportional to the width of the wire and the capacitance proportional to the width. (A far better model uses an extra constant term in the capacitance, but this complicates the equations.) The capacitance and resistance of the ith segment is thus

$$C_i = k_0 w_i, \qquad R_i = \rho / w_i,$$

where k_0 and ρ are positive constants, which we take to be one for simplicity. We also have $C_{load1} = 1.5$, $C_{load2} = 1$, and $C_{load3} = 5$.

Using the RC model for the wire segments yields the circuit shown below.



We will use the Elmore delay to model the delay from the source to each of the loads. The Elmore delay to loads 1, 2, and 3 are given by

$$\begin{aligned}
T_1 &= (C_3 + C_{load1})(R_1 + R_2 + R_3) + C_2(R_1 + R_2) + \\
&\quad +(C_1 + C_4 + C_5 + C_6 + C_{load2} + C_{load3})R_1 \\
T_2 &= (C_5 + C_{load2})(R_1 + R_4 + R_5) + C_4(R_1 + R_4) + \\
&\quad +(C_6 + C_{load3})(R_1 + R_4) + (C_1 + C_2 + C_3 + C_{load1})R_1 \\
T_3 &= (C_6 + C_{load3})(R_1 + R_4 + R_6) + C_4(R_1 + R_4) + \\
&\quad +(C_1 + C_2 + C_3 + C_{load1})R_1 + (C_5 + C_{load2})(R_1 + R_4).
\end{aligned}$$

Our main interest is in the maximum of these delays,

$$T = \max\{T_1, T_2, T_3\}.$$

(a) Explain how to find the optimal trade-off curve between area A and delay T.

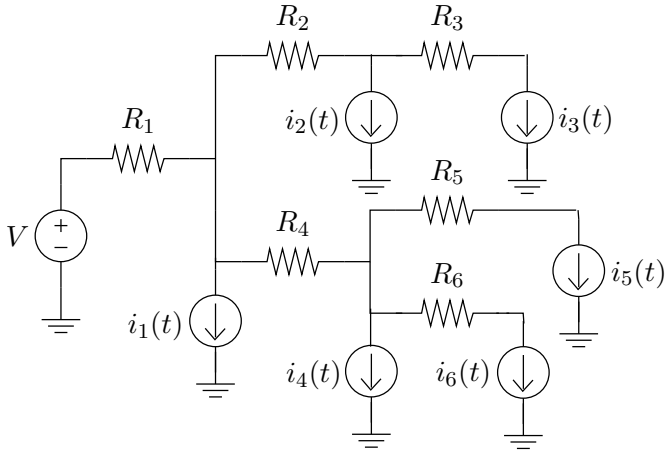
(b) *Optimal area-delay sizing.* For the specific problem parameters given, plot the area-delay trade-off curve, together with the individual Elmore delays. Comment on the results you obtain.

(c) *The simple method.* Plot the area-delay trade-off obtained when you assign all wire widths to be the same width (which varies between $W_{\min}$ and $W_{\max}$). Compare this curve to the optimal one, obtained in part (b). How much better does the optimal method do than the simple method? *Note:* for a large circuit, say with 1000 wires to size, the difference is *far larger*.

For this problem you can use the CVX in GP mode. We've also made available the function `elm_del_example.m`, which evaluates the three delays, given the widths of the wires.

11.2 *Optimal sizing of power and ground trees.* We consider a system or VLSI device with many subsystems or subcircuits, each of which needs one or more power supply voltages. In this problem we consider the case where the power supply network has a tree topology with the power supply (or external pin connection) at the root. Each node of the tree is connected to some subcircuit that draws power.

We model the power supply as a constant voltage source with value V. The m subcircuits are modeled as current sources that draw currents $i_1(t), \ldots, i_m(t)$ from the node (to ground) (see the figure below).



The subcircuit current draws have two components:

$$i_k(t) = i_k^{\mathrm{dc}} + i_k^{\mathrm{ac}}(t)$$

where i_k^{dc} is the DC current draw (which is a positive constant), and $i_k^{\mathrm{ac}}(t)$ is the AC draw (which has zero average value). We characterize the AC current draw by its RMS value, defined as

$$\mathrm{RMS}(i_k^{\mathrm{ac}}) = \left(\lim_{T\to\infty} \frac{1}{T} \int_0^T i_k^{\mathrm{ac}}(t)^2 \, dt \right)^{1/2}.$$

For each subcircuit we are given maximum values for the DC and RMS AC currents draws, *i.e.*, constants I_k^{dc} and I_k^{ac} such that

$$0 \leq i_k^{\mathrm{dc}} \leq I_k^{\mathrm{dc}}, \qquad \mathrm{RMS}(i_k^{\mathrm{ac}}) \leq I_k^{\mathrm{ac}}. \tag{28}$$

The n wires that form the distribution network are modeled as resistors R_k (which, presumably, have small value). (Since the circuit has a tree topology, we can use the following labeling convention: node k and the current source $i_k(t)$ are immediately following resistor R_k.) The resistance of the wires is given by

$$R_i = \alpha l_i / w_i,$$

where α is a constant and l_i are the lengths of the wires, which are known and fixed. The variables in the problem are the width of the wires, $w_1, \ldots, w_n$. Obviously by making the wires very wide, the resistances become very low, and we have a nearly ideal power network. The purpose of this problem is to optimally select wire widths, to minimize area while meeting certain specfications. Note that in this problem we ignore dynamics, *i.e.*, we do not model the capacitance or inductance of the wires.

As a result of the current draws and the nonzero resistance of the wires, the voltage at node k (which supplies subcircuit k) has a DC value less than the supply voltage, and also an AC voltage (which is called power supply ripple or noise). By superposition these two effects can be analyzed separately.

- The DC voltage drop $V - v_k^{\rm dc}$ at node k is equal to the sum of the voltage drops across wires on the (unique) path from node k to the root. It can be expressed as

$$V - v_k^{\rm dc} = \sum_{j=1}^{m} i_j^{\rm dc} \sum_{i \in \mathcal{N}(j,k)} R_i, \tag{29}$$

  where $\mathcal{N}(j,k)$ consists of the indices of the branches upstream from nodes j and k, *i.e.*, $i \in \mathcal{N}(j,k)$ if and only if R_i is in the path from node j to the root and in the path from node k to the root.

- The power supply noise at a node can be found as follows. The AC voltage at node k is equal to

$$v_k^{\rm ac}(t) = -\sum_{j=1}^{m} i_j^{\rm ac}(t) \sum_{i \in \mathcal{N}(j,k)} R_i.$$

  We assume the AC current draws are independent, so the RMS value of $v_k^{\rm ac}(t)$ is given by the squareroot of the sum of the squares of the RMS value of the ripple due to each other node, *i.e.*,

$$\mathrm{RMS}(v_k^{\rm ac}) = \left(\sum_{j=1}^{m} \left(\mathrm{RMS}(i_j^{\rm ac}) \sum_{i \in \mathcal{N}(j,k)} R_i \right)^2 \right)^{1/2}. \tag{30}$$

The problem is to choose wire widths w_i that minimize the total wire area $\sum_{i=k}^{n} w_k l_k$ subject to the following specifications:

- maximum allowable DC voltage drop at each node:

$$V - v_k^{\rm dc} \leq V_{\rm max}^{\rm dc}, \quad k = 1, \ldots, m, \tag{31}$$

  where $V - v_k^{\rm dc}$ is given by (29), and $V_{\rm max}^{\rm dc}$ is a given constant.

- maximum allowable power supply noise at each node:

$$\mathrm{RMS}(v_k^{\rm ac}) \leq V_{\rm max}^{\rm ac}, \quad k = 1, \ldots, m, \tag{32}$$

  where $\mathrm{RMS}(v_k^{\rm ac})$ is given by (30), and $V_{\rm max}^{\rm ac}$ is a given constant.

- upper and lower bounds on wire widths:

$$w_{\rm min} \leq w_i \leq w_{\rm max}, \quad i = 1, \ldots, n, \tag{33}$$

  where $w_{\rm min}$ and $w_{\rm max}$ are given constants.

- maximum allowable DC current density in a wire:

$$\left(\sum_{j \in \mathcal{M}(k)} i_j^{\rm dc} \right) \Big/ w_k \leq \rho_{\rm max}, \quad k = 1, \ldots, n, \tag{34}$$

  where $\mathcal{M}(k)$ is the set of all indices of nodes downstream from resistor k, *i.e.*, $j \in \mathcal{M}(k)$ if and only if R_k is in the path from node j to the root, and $\rho_{\rm max}$ is a given constant.

- maximum allowable total DC power dissipation in supply network:

$$\sum_{k=1}^{n} R_k \left(\sum_{j \in \mathcal{M}(k)} i_j^{\mathrm{dc}} \right)^2 \leq P_{\max}, \tag{35}$$

  where $P_{\max}$ is a given constant.

These specifications must be satisfied for all possible $i_k(t)$ that satisfy (28).

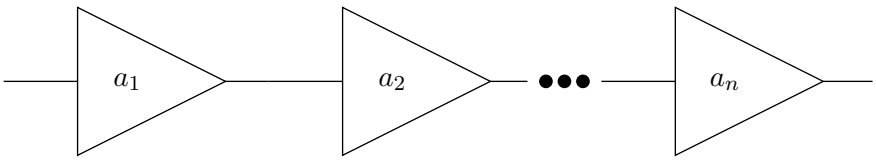
Formulate this as a convex optimization problem in the standard form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, p \\ & Ax = b. \end{array}$$

You may introduce new variables, or use a change of variables, but you must say very clearly

- what the optimization variable x is, and how it corresponds to the problem variables w (*i.e.*, is x equal to w, does it include auxiliary variables, ...?)
- what the objective f_0 and the constraint functions f_i are, and how they relate to the objectives and specifications of the problem description
- why the objective and constraint functions are convex
- what A and b are (if applicable).

11.3 *Optimal amplifier gains.* We consider a system of n amplifiers connected (for simplicity) in a chain, as shown below. The variables that we will optimize over are the gains $a_1, \ldots, a_n > 0$ of the amplifiers. The first specification is that the overall gain of the system, *i.e.*, the product $a_1 \cdots a_n$, is equal to A^{tot}, which is given.



We are concerned about two effects: noise generated by the amplifiers, and amplifier overload. These effects are modeled as follows.

We first describe how the noise depends on the amplifier gains. Let N_i denote the noise level (RMS, or root-mean-square) at the output of the ith amplifier. These are given recursively as

$$N_0 = 0, \quad N_i = a_i \left(N_{i-1}^2 + \alpha_i^2 \right)^{1/2}, \quad i = 1, \ldots, n$$

where $\alpha_i > 0$ (which is given) is the ('input-referred') RMS noise level of the ith amplifier. The *output noise level* N_{out} of the system is given by $N_{\mathrm{out}} = N_n$, *i.e.*, the noise level of the last amplifier. Evidently N_{out} depends on the gains $a_1, \ldots, a_n$.

Now we describe the amplifier overload limits. S_i will denote the signal level at the output of the ith amplifier. These signal levels are related by

$$S_0 = S_{\text{in}}, \qquad S_i = a_i S_{i-1}, \quad i = 1, \ldots, n,$$

where $S_{\text{in}} > 0$ is the *input signal level*. Each amplifier has a maximum allowable output level $M_i > 0$ (which is given). (If this level is exceeded the amplifier will distort the signal.) Thus we have the constraints $S_i \leq M_i$, for $i = 1, \ldots, n$. (We can ignore the noise in the overload condition, since the signal levels are much larger than the noise levels.)

The *maximum output signal level* S_{max} is defined as the maximum value of S_n, over all input signal levels S_{in} that respect the the overload constraints $S_i \leq M_i$. Of course $S_{\text{max}} \leq M_n$, but it can be smaller, depending on the gains $a_1, \ldots, a_n$.

The *dynamic range* D of the system is defined as $D = S_{\text{max}}/N_{\text{out}}$. Evidently it is a (rather complicated) function of the amplifier gains $a_1, \ldots, a_n$.

The goal is to choose the gains a_i to maximize the dynamic range D, subject to the constraint $\prod_i a_i = A^{\text{tot}}$, and upper bounds on the amplifier gains, $a_i \leq A_i^{\text{max}}$ (which are given).

Explain how to solve this problem as a convex (or quasiconvex) optimization problem. If you introduce new variables, or transform the variables, explain. Clearly give the objective and inequality constraint functions, explaining why they are convex if it is not obvious. If your problem involves equality constraints, give them explicitly.

Carry out your method on the specific instance with $n = 4$, and data

$$\begin{aligned} A^{\text{tot}} &= 10000, \\ \alpha &= (10^{-5}, 10^{-2}, 10^{-2}, 10^{-2}), \\ M &= (0.1, 5, 10, 10), \\ A^{\text{max}} &= (40, 40, 40, 20). \end{aligned}$$

Give the optimal gains, and the optimal dynamic range.

12 Signal processing and communications

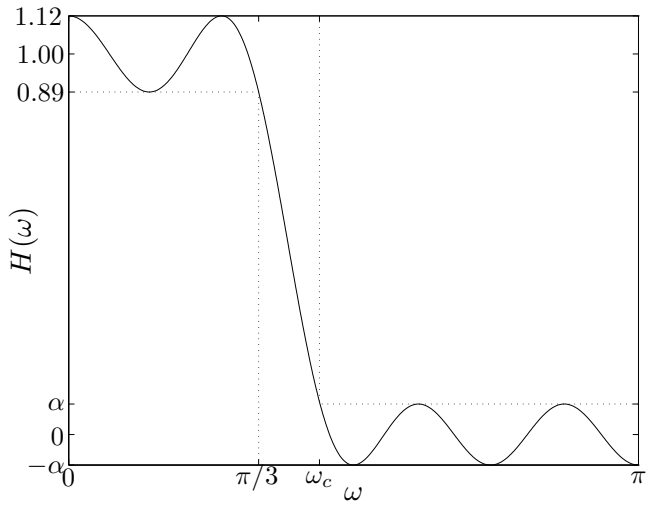
12.1 *FIR low-pass filter design.* Consider the (symmetric, linear phase) finite impulse response (FIR) filter described by its frequency response

$$H(\omega) = a_0 + \sum_{k=1}^{N} a_k \cos k\omega,$$

where $\omega \in [0, \pi]$ is the frequency. The design variables in our problems are the real coefficients $a = (a_0, \ldots, a_N) \in \mathbf{R}^{N+1}$, where N is called the order or length of the FIR filter. In this problem we will explore the design of a low-pass filter, with specifications:

- For $0 \leq \omega \leq \pi/3$, $0.89 \leq H(\omega) \leq 1.12$, *i.e.*, the filter has about $\pm$1dB ripple in the 'passband' $[0, \pi/3]$.
- For $\omega_c \leq \omega \leq \pi$, $|H(\omega)| \leq \alpha$. In other words, the filter achieves an attenuation given by α in the 'stopband' $[\omega_c, \pi]$. Here ω_c is called the filter 'cutoff frequency'.

(It is called a low-pass filter since low frequencies are allowed to pass, but frequencies above the cutoff frequency are attentuated.) These specifications are depicted graphically in the figure below.



For parts (a)–(c), explain how to formulate the given problem as a convex or quasiconvex optimization problem.

(a) *Maximum stopband attenuation.* We fix ω_c and N, and wish to maximize the stopband attenuation, *i.e.*, minimize α.

(b) *Minimum transition band.* We fix N and α, and want to minimize ω_c, *i.e.*, we set the stopband attenuation and filter length, and wish to minimize the 'transition' band (between $\pi/3$ and ω_c).

(c) *Shortest length filter.* We fix ω_c and α, and wish to find the smallest N that can meet the specifications, *i.e.*, we seek the shortest length FIR filter that can meet the specifications.

(d) Plot the optimal tradeoff curve of attenuation (α) versus cutoff frequency (ω_c) for $N = 7$.

For this subproblem, you may sample the constraints in frequency, which means the following. Choose $K \gg N$ (perhaps $K \approx 10N$), and set $\omega_k = k\pi/K$, $k = 0, \ldots, K$. Then replace the specifications with

- For k with $0 \leq \omega_k \leq \pi/3$, $0.89 \leq H(\omega_k) \leq 1.12$.
- For k with $\omega_c \leq \omega_k \leq \pi$, $|H(\omega_k)| \leq \alpha$.

12.2 *SINR maximization.* Solve the following instance of problem 4.20: We have $n = 5$ transmitters, grouped into two groups: $\{1, 2\}$ and $\{3, 4, 5\}$. The maximum power for each transmitter is 4, the total power limit for the first group is 6, and the total power limit for the second group is 9. The noise σ is equal to 0.5 and the limit on total received power is 6 for each receiver. Finally, the path gain matrix is given by

$$G = \begin{bmatrix} 1.0 & 0.1 & 0.2 & 0.1 & 0.0 \\ 0.1 & 1.0 & 0.1 & 0.1 & 0.0 \\ 0.2 & 0.1 & 2.0 & 0.2 & 0.2 \\ 0.1 & 0.1 & 0.2 & 1.0 & 0.1 \\ 0.0 & 0.0 & 0.2 & 0.1 & 1.0 \end{bmatrix}.$$

Find the optimal transmitter powers $p_1, \ldots, p_5$ that maximize the minimum SINR ratio over all receivers. Also report the maximum SINR value.

12.3 *Power control for sum rate maximization in interference channel.* We consider the optimization problem

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} \log\left(1 + \frac{p_i}{\sum_{j\neq i} A_{ij}p_j + v_i}\right) \\ \text{subject to} & \displaystyle\sum_{i=1}^{n} p_i = 1 \\ & p_i \geq 0, \quad i = 1, \ldots, n \end{array}$$

with variables $p \in \mathbf{R}^n$. The problem data are the matrix $A \in \mathbf{R}^{n\times n}$ and the vector $v \in \mathbf{R}^n$. We assume A and v are componentwise nonnegative ($A_{ij} \geq 0$ and $v_i \geq 0$), and that the diagonal elements of A are equal to one. If the off-diagonal elements of A are zero ($A = I$), the problem has a simple solution, given by the waterfilling method. We are interested in the case where the off-diagonal elements are nonzero.

We can give the following interpretation of the problem, which is not needed below. The variables in the problem are the transmission powers in a communications system. We limit the total power to one (for simplicity; we could have used any other number). The ith term in the objective is the Shannon capacity of the ith channel; the fraction in the argument of the log is the signal to interference plus noise ratio.

We can express the problem as

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{i=1}^{n} \log\left(\frac{\sum_{j=1}^{n} B_{ij}p_j}{\sum_{j=1}^{n} B_{ij}p_j - p_i}\right) \\ \text{subject to} & \displaystyle\sum_{i=1}^{n} p_i = 1 \\ & p_i \geq 0, \quad i = 1, \ldots, n, \end{array}$$

where $B \in \mathbf{R}^{n\times n}$ is defined as $B = A + v\mathbf{1}^T$, *i.e.*, $B_{ij} = A_{ij} + v_i$, $i, j = 1, \ldots, n$. Suppose B is nonsingular and

$$B^{-1} = I - C$$

with $C_{ij} \geq 0$. Express the problem above as a convex optimization problem. *Hint.* Use $y = Bp$ as variables.

12.4 *Radio-relay station placement and power allocation.* Radio relay stations are to be located at positions $x_1, \ldots, x_n \in \mathbf{R}^2$, and transmit at power $p_1, \ldots, p_n \geq 0$. In this problem we will consider the problem of simultaneously deciding on good locations *and* operating powers for the relay stations.

The received signal power S_{ij} at relay station i from relay station j is proportional to the transmit power and inversely proportional to the distance, *i.e.*,

$$S_{ij} = \frac{\alpha p_j}{\|x_i - x_j\|^2},$$

where $\alpha > 0$ is a known constant.

Relay station j must transmit a signal to relay station i at the rate (or bandwidth) $R_{ij} \geq 0$ bits per second; $R_{ij} = 0$ means that relay station j does not need to transmit any message (directly) to relay station i. The matrix of bit rates R_{ij} is given. Although it doesn't affect the problem, R would likely be sparse, *i.e.*, each relay station needs to communicate with only a few others.

To guarantee accurate reception of the signal from relay station j to i, we must have

$$S_{ij} \geq \beta R_{ij},$$

where $\beta > 0$ is a known constant. (In other words, the minimum allowable received signal power is proportional to the signal bit rate or bandwidth.)

The relay station positions $x_{r+1}, \ldots, x_n$ are fixed, *i.e.*, problem parameters. The problem variables are $x_1, \ldots, x_r$ and $p_1, \ldots, p_n$. The goal is to choose the variables to minimize the total transmit power, *i.e.*, $p_1 + \cdots + p_n$.

Explain how to solve this problem as a convex or quasiconvex optimization problem. If you introduce new variables, or transform the variables, explain. Clearly give the objective and inequality constraint functions, explaining why they are convex. If your problem involves equality constraints, express them using an affine function.

12.5 *Power allocation with coherent combining receivers.* In this problem we consider a variation on the power allocation problem described on pages 4-13 and 4-14 of the notes. In that problem we have m transmitters, each of which transmits (broadcasts) to n receivers, so the total number of receivers is mn. In this problem we have the converse: multiple transmitters send a signal to each receiver.

More specifically we have m receivers labeled $1, \ldots, m$, and mn transmitters labeled (j, k), $j = 1, \ldots, m$, $k = 1, \ldots, n$. The transmitters $(i, 1), \ldots, (i, n)$ all transmit the same message to the receiver i, for $i = 1, \ldots, m$.

Transmitter (j, k) operates at power p_{jk}, which must satisfy $0 \leq p_{jk} \leq P_{\max}$, where $P_{\max}$ is a given maximum allowable transmitter power.

The path gain from transmitter (j,k) to receiver i is $A_{ijk} > 0$ (which are given and known). Thus the power received at receiver i from transmitter (j,k) is given by $A_{ijk}p_{jk}$.

For $i \neq j$, the received power $A_{ijk}p_{jk}$ represents an interference signal. The total interference-plus-noise power at receiver i is given by

$$I_i = \sum_{j \neq i,\ k=1,\ldots,n} A_{ijk}p_{jk} + \sigma$$

where $\sigma > 0$ is the known, given (self) noise power of the receivers. Note that the *powers* of the interference and noise signals add to give the total interference-plus-noise power.

The receivers use *coherent detection and combining* of the desired message signals, which means the effective received signal power at receiver i is given by

$$S_i = \left(\sum_{k=1,\ldots,n} (A_{iik}p_{ik})^{1/2} \right)^2 .$$

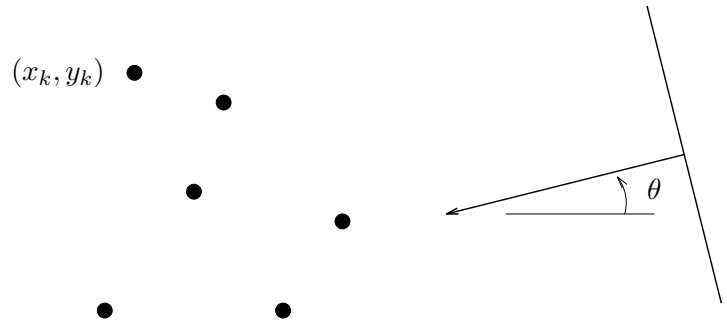
(Thus, the *amplitudes* of the desired signals add to give the effective signal amplitude.)

The total signal to interference-plus-noise ratio (SINR) for receiver i is given by $\gamma_i = S_i/I_i$.

The problem is to choose transmitter powers p_{jk} that maximize the minimum SINR $\min_i \gamma_i$, subject to the power limits.

Explain in detail how to solve this problem using convex or quasiconvex optimization. If you transform the problem by using a different set of variables, explain completely. Identify the objective function, and all constraint functions, indicating if they are convex or quasiconvex, etc.

12.6 *Antenna array weight design.* We consider an array of n omnidirectional antennas in a plane, at positions (x_k, y_k), $k = 1, \ldots, n$.



A unit plane wave with frequency ω is incident from an angle θ. This incident wave induces in the kth antenna element a (complex) signal $\exp(i(x_k \cos\theta + y_k \sin\theta - \omega t))$, where $i = \sqrt{-1}$. (For simplicity we assume that the spatial units are normalized so that the wave number is one, *i.e.*, the wavelength is $\lambda = 2\pi$.) This signal is demodulated, *i.e.*, multiplied by $e^{i\omega t}$, to obtain the baseband signal (complex number) $\exp(i(x_k \cos\theta + y_k \sin\theta))$. The baseband signals of the n antennas are combined linearly to form the output of the antenna array

$$\begin{aligned} G(\theta) &= \sum_{k=1}^{n} w_k \, e^{i(x_k \cos\theta + y_k \sin\theta)} \\ &= (w_{\mathrm{re},k} \cos\gamma_k(\theta) - w_{\mathrm{im},k} \sin\gamma_k(\theta)) + i\,(w_{\mathrm{re},k} \sin\gamma_k(\theta) + w_{\mathrm{im},k} \cos\gamma_k(\theta)) , \end{aligned}$$

if we define $\gamma_k(\theta) = x_k \cos\theta + y_k \sin\theta$. The complex weights in the linear combination,

$$w_k = w_{\mathrm{im},k} + i w_{\mathrm{re},k}, \quad k = 1, \ldots, n,$$

are called the *antenna array coefficients* or *shading coefficients*, and will be the design variables in the problem. For a given set of weights, the combined output $G(\theta)$ is a function of the angle of arrival θ of the plane wave. The design problem is to select weights w_i that achieve a desired directional pattern $G(\theta)$.

We now describe a basic weight design problem. We require unit gain in a target direction θ^{tar}, *i.e.*, $G(\theta^{\mathrm{tar}}) = 1$. We want $|G(\theta)|$ small for $|\theta - \theta^{\mathrm{tar}}| \geq \Delta$, where 2Δ is our beamwidth. To do this, we can minimize

$$\max_{|\theta-\theta^{\mathrm{tar}}|\geq\Delta} |G(\theta)|,$$

where the maximimum is over all $\theta \in [-\pi, \pi]$ with $|\theta - \theta^{\mathrm{tar}}| \geq \Delta$. This number is called the sidelobe level for the array; our goal is to minimize the sidelobe level. If we acheve a small sidelobe level, then the array is relatively insensitive to signals arriving from directions more than Δ away from the target direction. This results in the optimization problem

$$\begin{array}{ll} \text{minimize} & \max_{|\theta|\geq\Delta} |G(\theta)| \\ \text{subject to} & G(0) = 1, \end{array}$$

with $w \in \mathbf{C}^n$ as variables.

The objective function can be approximated by discretizing the angle of arrival with (say) N values (say, uniformly spaced) $\theta_1, \ldots, \theta_N$ over the interval $[-\pi, \pi]$, and replacing the objective with

$$\max\{|G(\theta_k)| \mid |\theta_k - \theta^{\mathrm{tar}}| \geq \Delta\}$$

(a) Formulate the antenna array weight design problem as an SOCP.

(b) Solve an instance using CVX, with $n = 40$, $\theta^{\mathrm{tar}} = 15^\circ$, $\Delta = 15^\circ$, $N = 400$, and antenna positions generated using

```
>> rand('state',0);
>> n = 40;
>> x = 30 * rand(n,1);
>> y = 30 * rand(n,1);
```

Compute the optimal weights and make a plot of $|G(\theta)|$ (on a logarithmic scale) versus θ. *Hint.* CVX can directly handle complex variables, and recognizes the modulus `abs(x)` of a complex number as a convex function of its real and imaginary parts, so you do not need to explicitly form the SOCP from part (a).

12.7 *Power allocation problem with analytic solution.* Consider a system of n transmitters and n receivers. The ith transmitter transmits with power x_i, $i = 1, \ldots, n$. The vector x will be the variable in this problem. The path gain from each transmitter j to each receiver i will be denoted A_{ij} and is assumed to be known (obviously, $A_{ij} \geq 0$, so the matrix A is elementwise nonnegative, and $A_{ii} > 0$). The signal received by each receiver i consists of three parts: the desired signal, arriving from transmitter i with power $A_{ii}x_i$, the interfering signal, arriving from the other receivers with

power $\sum_{j\neq i} A_{ij}x_j$, and noise β_i (which are positive and known). We are interested in allocating the powers x_i in such a way that the signal to noise plus interference ratio at each of the receivers exceeds a level α. (Thus α is the minimum acceptable SNIR for the receivers; a typical value might be around $\alpha = 3$, *i.e.*, around 10dB). In other words, we want to find $x \succeq 0$ such that for $i = 1, \ldots, n$

$$A_{ii}x_i \geq \alpha \left(\sum_{j\neq i} A_{ij}x_j + \beta_i \right).$$

Equivalently, the vector x has to satisfy

$$x \succeq 0, \quad Bx \succeq \alpha\beta \tag{36}$$

where $B \in \mathbf{R}^{n\times n}$ is defined as

$$B_{ii} = A_{ii}, \quad B_{ij} = -\alpha A_{ij}, \; j \neq i.$$

(a) Show that (36) is feasible if and only if B is invertible and $z = B^{-1}\mathbf{1} \succeq 0$ ($\mathbf{1}$ is the vector with all components 1). Show how to construct a feasible power allocation x from z.

(b) Show how to find the largest possible SNIR, *i.e.*, how to maximize α subject to the existence of a feasible power allocation.

To solve this problem you may need the following:

Hint. Let $T \in \mathbf{R}^{n\times n}$ be a matrix with nonnegative elements, and $s \in \mathbf{R}$. Then the following are equivalent:

(a) $s > \rho(T)$, where $\rho(T) = \max_i |\lambda_i(T)|$ is the spectral radius of T.

(b) $sI - T$ is nonsingular and the matrix $(sI - T)^{-1}$ has nonnegative elements.

(c) there exists an $x \succeq 0$ with $(sI - T)x \succ 0$.

(For such s, the matrix $sI - T$ is called a *nonsingular M-matrix.*)

Remark. This problem gives an analytic solution to a very special form of transmitter power allocation problem. Specifically, there are exactly as many transmitters as receivers, and no power limits on the transmitters. One consequence is that the receiver noises β_i play no role at all in the solution — just crank up all the transmitters to overpower the noises!

12.8 *Optimizing rates and time slot fractions.* We consider a wireless system that uses time-domain multiple access (TDMA) to support n communication flows. The flows have (nonnegative) rates $r_1, \ldots, r_n$, given in bits/sec. To support a rate r_i on flow i requires transmitter power

$$p = a_i(e^{br} - 1),$$

where b is a (known) positive constant, and a_i are (known) positive constants related to the noise power and gain of receiver i.

TDMA works like this. Time is divided up into periods of some fixed duration T (seconds). Each of these T-long periods is divided into n time-slots, with durations $t_1, \ldots, t_n$, that must satisfy $t_1 + \cdots + t_n = T$, $t_i \geq 0$. In time-slot i, communications flow i is transmitted at an instantaneous

rate $r = Tr_i/t_i$, so that over each T-long period, Tr_i bits from flow i are transmitted. The power required during time-slot i is $a_i(e^{bTr_i/t_i} - 1)$, so the average transmitter power over each T-long period is

$$P = (1/T)\sum_{i=1}^{n} a_i t_i (e^{bTr_i/t_i} - 1).$$

When t_i is zero, we take $P = \infty$ if $r_i > 0$, and $P = 0$ if $r_i = 0$. (The latter corresponds to the case when there is zero flow, and also, zero time allocated to the flow.)

The problem is to find rates $r \in \mathbf{R}^n$ and time-slot durations $t \in \mathbf{R}^n$ that maximize the log utility function

$$U(r) = \sum_{i=1}^{n} \log r_i,$$

subject to $P \leq P^{\max}$. (This utility function is often used to ensure 'fairness'; each communication flow gets at least some positive rate.) The problem data are a_i, b, T and $P^{\max}$; the variables are t_i and r_i.

(a) Formulate this problem as a convex optimization problem. Feel free to introduce new variables, if needed, or to change variables. Be sure to justify convexity of the objective or constraint functions in your formulation.

(b) Give the optimality conditions for your formulation. Of course we prefer simpler optimality conditions to complex ones. *Note:* We do not expect you to *solve* the optimality conditions; you can give them as a set of equations (and possibly inequalities).

Hint. With a log utility function, we cannot have $r_i = 0$, and therefore we cannot have $t_i = 0$; therefore the constraints $r_i \geq 0$ and $t_i \geq 0$ cannot be active or tight. This will allow you to simplify the optimality conditions.

12.9 *Optimal jamming power allocation.* A set of n jammers transmit with (nonnegative) powers $p_1, \ldots, p_n$, which are to be chosen subject to the constraints

$$p \succeq 0, \qquad Fp \preceq g.$$

The jammers produce interference power at m receivers, given by

$$d_i = \sum_{j=1}^{n} G_{ij} p_j, \quad i = 1, \ldots, m,$$

where G_{ij} is the (nonnegative) channel gain from jammer j to receiver i.

Receiver i has capacity (in bits/s) given by

$$C_i = \alpha \log(1 + \beta_i/(\sigma_i^2 + d_i)), \quad i = 1, \ldots, m,$$

where α, β_i, and σ_i are positive constants. (Here β_i is proportional to the signal power at receiver i and σ_i^2 is the receiver i self-noise, but you won't need to know this to solve the problem.)

Explain how to choose p to *minimize* the sum channel capacity, $C = C_1 + \cdots + C_m$, using convex optimization. (This corresponds to the most effective jamming, given the power constraints.) The problem data are F, g, G, α, β_i, σ_i.

If you change variables, or transform your problem in any way that is not obvious (for example, you form a relaxation), you must explain fully how your method works, and why it gives the solution. If your method relies on any convex functions that we have not encountered before, you must show that the functions are convex.

12.10 *2D filter design.* A symmetric convolution kernel with support $\{-(N-1), \ldots, N-1\}^2$ is characterized by N^2 coefficients

$$h_{kl}, \quad k, l = 1, \ldots, N.$$

These coefficients will be our variables. The corresponding 2D frequency response (Fourier transform) $H : \mathbf{R}^2 \to \mathbf{R}$ is given by

$$H(\omega_1, \omega_2) = \sum_{k,l=1,\ldots,N} h_{kl} \cos((k-1)\omega_1) \cos((l-1)\omega_2),$$

where ω_1 and ω_2 are the frequency variables. Evidently we only need to specify H over the region $[0, \pi]^2$, although it is often plotted over the region $[-\pi, \pi]^2$. (It won't matter in this problem, but we should mention that the coefficients h_{kl} above are not exactly the same as the impulse response coefficients of the filter.)

We will design a 2D filter (*i.e.*, find the coefficients h_{kl}) to satisfy $H(0,0) = 1$ and to minimize the maximum response R in the rejection region $\Omega_{\text{rej}} \subset [0, \pi]^2$,

$$R = \sup_{(\omega_1, \omega_2) \in \Omega_{\text{rej}}} |H(\omega_1, \omega_2)|.$$

(a) Explain why this 2D filter design problem is convex.

(b) Find the optimal filter for the specific case with $N = 5$ and

$$\Omega_{\text{rej}} = \{(\omega_1, \omega_2) \in [0, \pi]^2 \mid \omega_1^2 + \omega_2^2 \geq W^2\},$$

with $W = \pi/4$.

You can approximate R by sampling on a grid of frequency values. Define

$$\omega^{(p)} = \pi(p-1)/M, \quad p = 1, \ldots, M.$$

(You can use $M = 25$.) We then replace the exact expression for R above with

$$\hat{R} = \max\{|H(\omega^{(p)}, \omega^{(q)})| \mid p, q = 1, \ldots, M, \ (\omega^{(p)}, \omega^{(q)}) \in \Omega_{\text{rej}}\}.$$

Give the optimal value of $\hat{R}$. Plot the optimal frequency response using `plot_2D_filt(h)`, available on the course web site, where `h` is the matrix containing the coefficients h_{kl}.

13 Finance

13.1 *Transaction cost.* Consider a market for some asset or commodity, which we assume is infinitely divisible, *i.e.*, can be bought or sold in quantities of shares that are real numbers (as opposed to integers). The *order book* at some time consists of a set of offers to sell or buy the asset, at a given price, up to a given quantity of shares. The N offers to sell the asset have positive prices per share $p_1^{\text{sell}}, \ldots, p_N^{\text{sell}}$, sorted in increasing order, in positive share quantities $q_1^{\text{sell}}, \ldots, q_N^{\text{sell}}$. The M offers to buy the asset have positive prices $p_1^{\text{buy}}, \ldots, p_N^{\text{buy}}$, sorted in decreasing order, and positive quantities $q_1^{\text{buy}}, \ldots, q_M^{\text{buy}}$. The price p_1^{sell} is called the (current) *ask price* for the asset; p_1^{buy} is the *bid price* for the asset. The ask price is larger than the bid price; the difference is called the *spread*. The average of the ask and bid prices is called the *mid-price*, denoted p^{mid}.

Now suppose that you want to purchase $q > 0$ shares of the asset, where $q \leq q_1^{\text{sell}} + \cdots + q_N^{\text{sell}}$, *i.e.*, your purchase quantity does not exceed the total amount of the asset currently offered for sale. Your purchase proceeds as follows. Suppose that

$$q_1^{\text{sell}} + \cdots + q_k^{\text{sell}} < q \leq q_{k+1}^{\text{sell}}.$$

Then you pay an amount

$$A = p_1^{\text{sell}} q_1^{\text{sell}} + \cdots + p_k^{\text{sell}} q_k^{\text{sell}} + p_{k+1}^{\text{sell}}(q - q_1^{\text{sell}} - \cdots - q_k^{\text{sell}}).$$

Roughly speaking, you work you way through the offers in the order book, from the least (ask) price, and working your way up the order book until you fill the order. We define the *transaction cost* as

$$T(q) = A - p^{\text{mid}} q.$$

This is the difference between what you pay, and what you would have paid had you been able to purchase the shares at the mid-price. It is always positive.

We handle the case of selling the asset in a similar way. Here we take $q < 0$ to mean that we sell $-q$ shares of the asset. Here you sell shares at the bid price, up the quantity q^{buy} (or $-q$, whichever is smaller); if needed, you sell shares at the price p_2^{buy}, and so on, until all $-q$ shares are sold. Here we assume that $-q \leq q_1^{\text{buy}} + \cdots + q_M^{\text{buy}}$, *i.e.*, you are not selling more shares than the total quantity of offers to buy. Let A denote the amount you receive from the sale. Here we define the transaction cost as

$$T(q) = p^{\text{mid}} - A,$$

the difference between the amount you would have received had you sold the shares at the mid-price, and the amount you received. It is always positive. We set $T(0) = 0$.

(a) Show that T is a convex piecewise linear function.

(b) Show that $T(q) \geq (s/2)|q|$, where s is the spread. When would we have $T(q) = (s/2)|q|$ for all q (in the range between the total shares offered to purchase or sell)?

(c) Give an interpretation of the conjugate function $T^*(y) = \sup_q(yq - T(q))$. *Hint.* Suppose you can purchase or sell the asset in another market, at the price p^{other}.

13.2 *Risk-return trade-off in portfolio optimization.* We consider the portfolio risk-return trade-off problem of page 185, with the following data:

$$\bar{p} = \begin{bmatrix} 0.12 \\ 0.10 \\ 0.07 \\ 0.03 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 0.0064 & 0.0008 & -0.0011 & 0 \\ 0.0008 & 0.0025 & 0 & 0 \\ -0.0011 & 0 & 0.0004 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

(a) Solve the quadratic program

$$\begin{array}{ll} \text{minimize} & -\bar{p}^T x + \mu x^T \Sigma x \\ \text{subject to} & \mathbf{1}^T x = 1, \quad x \succeq 0 \end{array}$$

for a large number of positive values of μ (for example, 100 values logarithmically spaced between 1 and 10^7). Plot the optimal values of the expected return $\bar{p}^T x$ versus the standard deviation $(x^T \Sigma x)^{1/2}$. Also make an area plot of the optimal portfolios x versus the standard deviation (as in figure 4.12).

(b) Assume the price change vector p is a Gaussian random variable, with mean $\bar{p}$ and covariance Σ. Formulate the problem

$$\begin{array}{ll} \text{maximize} & \bar{p}^T x \\ \text{subject to} & \mathbf{prob}(p^T x \leq 0) \leq \eta \\ & \mathbf{1}^T x = 1, \quad x \succeq 0, \end{array}$$

as a convex optimization problem, where $\eta < 1/2$ is a parameter. In this problem we maximize the expected return subject to a constraint on the probability of a negative return. Solve the problem for a large number of values of η between 10^{-4} and 10^{-1}, and plot the optimal values of $\bar{p}^T x$ versus η. Also make an area plot of the optimal portfolios x versus η.

Hint: The Matlab functions `erfc` and `erfcinv` can be used to evaluate $\Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} e^{-t^2/2} dt$ and its inverse:

$$\Phi(u) = \frac{1}{2}\mathbf{erfc}(-u/\sqrt{2}).$$

Since you will have to solve this problem for a large number of values of η, you may find the command `cvx_quiet(true)` helpful...

(c) *Monte Carlo simulation.* Let x be the optimal portfolio found in part (b), with $\eta = 0.05$. This portfolio maximizes the expected return, subject to the probability of a loss being no more than 5%. Generate 10000 samples of p, and plot a histogram of the returns. Find the empirical mean of the return samples, and calculate the percentage of samples for which a loss occurs.

Hint: You can generate samples of the price change vector using

```
p=pbar+sqrtm(Sigma)*randn(4,1);
```

13.3 *Simple portfolio optimization.* We consider a portfolio optimization problem as described on pages 155 and 185–186 of *Convex Optimization*, with data that can be found in the file `simple_portfolio_data.m`.

(a) Find minimum-risk portfolios with the same expected return as the uniform portfolio ($x = (1/n)\mathbf{1}$), with risk measured by portfolio return variance, and the following portfolio constraints (in addition to $\mathbf{1}^T x = 1$):

- No (additional) constraints.
- Long-only: $x \succeq 0$.
- Limit on total short position: $\mathbf{1}^T(x_-) \leq 0.5$, where $(x_-)_i = \max\{-x_i, 0\}$.

Compare the optimal risk in these portfolios with each other and the uniform portfolio.

(b) Plot the optimal risk-return trade-off curves for the long-only portfolio, and for total short-position limited to 0.5, in the same figure. Follow the style of figure 4.12 (top), with horizontal axis showing standard deviation of portfolio return, and vertical axis showing mean return.

13.4 *Bounding portfolio risk with incomplete covariance information.* Consider the following instance of the problem described in §4.6, on p171–173 of *Convex Optimization*. We suppose that Σ_{ii}, which are the squares of the price volatilities of the assets, are known. For the off-diagonal entries of Σ, all we know is the sign (or, in some cases, nothing at all). For example, we might be given that $\Sigma_{12} \geq 0$, $\Sigma_{23} \leq 0$, etc. This means that we do not know the correlation between p_1 and p_2, but we do know that they are nonnegatively correlated (*i.e.*, the prices of assets 1 and 2 tend to rise or fall together).

Compute σ_{wc}, the worst-case variance of the portfolio return, for the specific case

$$x = \begin{bmatrix} 0.1 \\ 0.2 \\ -0.05 \\ 0.1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0.2 & + & + & \pm \\ + & 0.1 & - & - \\ + & - & 0.3 & + \\ \pm & - & + & 0.1 \end{bmatrix},$$

where a "+" entry means that the element is nonnegative, a "−" means the entry is nonpositive, and "±" means we don't know anything about the entry. (The negative value in x represents a *short position*: you sold stocks that you didn't have, but must produce at the end of the investment period.) In addition to σ_{wc}, give the covariance matrix Σ_{wc} associated with the maximum risk. Compare the worst-case risk with the risk obtained when Σ is diagonal.

13.5 *Log-optimal investment strategy.* In this problem you will solve a specific instance of the log-optimal investment problem described in exercise 4.60, with $n = 5$ assets and $m = 10$ possible outcomes in each period. The problem data are defined in `log_opt_invest.m`, with the rows of the matrix `P` giving the asset return vectors p_j^T. The outcomes are equiprobable, *i.e.*, we have $\pi_j = 1/m$. Each column of the matrix `P` gives the return of the associated asset in the different posible outcomes. You can examine the columns to get an idea of the types of assets. For example, the last asset gives a fixed and certain return of 1%; the first asset is a very risky one, with occasional large return, and (more often) substantial loss.

Find the log-optimal investment strategy $x^\star$, and its associated long term growth rate $R_{\mathrm{lt}}^\star$. Compare this to the long term growth rate obtained with a uniform allocation strategy, *i.e.*, $x = (1/n)\mathbf{1}$, and also with a pure investment in each asset.

For the optimal investment strategy, and also the uniform investment strategy, plot 10 sample trajectories of the accumulated wealth, *i.e.*, $W(T) = W(0)\prod_{t=1}^{T}\lambda(t)$, for $T = 0, \ldots, 200$, with initial wealth $W(0) = 1$.

To save you the trouble of figuring out how to simulate the wealth trajectories or plot them nicely, we've included the simulation and plotting code in `log_opt_invest.m`; you just have to add the code needed to find $x^\star$.

Hint: The current version of CVX handles the logarithm via an iterative method, which can be slow and unreliable. You're better off using `geo_mean()`, which is directly handled by CVX, to solve the problem.

13.6 *Optimality conditions and dual for log-optimal investment problem.*

(a) Show that the optimality conditions for the log-optimal investment problem described in exercise 4.60 can be expressed as: $\mathbf{1}^T x = 1$, $x \succeq 0$, and for each i,

$$x_i > 0 \;\Rightarrow\; \sum_{j=1}^{m} \pi_j \frac{p_{ij}}{p_j^T x} = 1, \quad x_i = 0 \;\Rightarrow\; \sum_{j=1}^{m} \pi_j \frac{p_{ij}}{p_j^T x} \leq 1.$$

We can interpret this as follows. $p_{ij}/p_j^T x$ is a random variable, which gives the ratio of the investment gain with asset i only, to the investment gain with our mixed portfolio x. The optimality condition is that, for each asset we invest in, the expected value of this ratio is one, and for each asset we do not invest in, the expected value cannot exceed one. Very roughly speaking, this means our portfolio does as well as any of the assets that we choose to invest in, and cannot do worse than any assets that we do not invest in.

Hint. You can start from the simple criterion given in §4.2.3 or the KKT conditions.

(b) In this part we will derive the dual of the log-optimal investment problem. We start by writing the problem as,

$$\begin{array}{ll} \text{minimize} & -\sum_{j=1}^{m} \pi_j \log y_j \\ \text{subject to} & y = P^T x, \quad x \succeq 0, \quad \mathbf{1}^T x = 1. \end{array}$$

Here, P has columns $p_1, \ldots, p_m$, and we have the introduced new variables $y_1, \ldots, y_m$, with the implicit constraint $y \succ 0$. We will associate dual variables ν, λ and ν_0 with the constraints $y = P^T x$, $x \succeq 0$, and $\mathbf{1}^T x = 1$, respectively. Defining $\tilde{\nu}_j = \nu_j/\nu_0$ for $j = 1, \ldots, m$, show that the dual problem can be written as

$$\begin{array}{ll} \text{maximize} & \sum_{j=1}^{m} \pi_j \log(\tilde{\nu}_j/\pi_j) \\ \text{subject to} & P\tilde{\nu} \preceq \mathbf{1}, \end{array}$$

with variable $\tilde{\nu}$. The objective here is the (negative) Kullback-Leibler divergence between the given distribution π and the dual variable $\tilde{\nu}$.

13.7 *Arbitrage and theorems of alternatives.* Consider an event (for example, a sports game, political elections, the evolution of the stockmarket over a certain period) with m possible outcomes. Suppose that n wagers on the outcome are possible. If we bet an amount x_j on wager j, and the outcome of the event is i ($i = 1, \ldots, m$), then our return will be equal to $r_{ij} x_j$. The return $r_{ij} x_j$ is the net gain: we pay x_j initially, and receive $(1 + r_{ij})x_j$ if the outcome of the event is i. We allow the bets x_j to be positive, negative, or zero. The interpretation of a negative bet is as follows. If $x_j < 0$, then initially we *receive* an amount of money $|x_j|$, with an obligation to *pay* $(1 + r_{ij})|x_j|$ if outcome i occurs. In that case, we lose $r_{ij}|x_j|$, *i.e.*, our net is gain $r_{ij} x_j$ (a negative number).

We call the matrix $R \in \mathbf{R}^{m \times n}$ with elements r_{ij} the *return matrix*. A *betting strategy* is a vector $x \in \mathbf{R}^n$, with as components x_j the amounts we bet on each wager. If we use a betting strategy x, our total return in the event of outcome i is equal to $\sum_{j=1}^{n} r_{ij} x_j$, *i.e.*, the ith component of the vector Rx.

Country	Odds
Holland	3.5
Italy	5.0
Spain	5.5
France	6.5
Germany	7.0
England	10.0
Belgium	14.0
Sweden	16.0

Country	Odds
Czech Republic	17.0
Romania	18.0
Yugoslavia	20.0
Portugal	20.0
Norway	20.0
Denmark	33.0
Turkey	50.0
Slovenia	80.0

Table 1: Odds for the 2000 European soccer championships.

(a) *The arbitrage theorem.* Suppose you are given a return matrix R. Prove the following theorem: there is a betting strategy $x \in \mathbf{R}^n$ for which

$$Rx \succ 0$$

if and only if there exists no vector $p \in \mathbf{R}^m$ that satisfies

$$R^T p = 0, \qquad p \succeq 0, \qquad p \neq 0.$$

We can interpret this theorem as follows. If $Rx \succ 0$, then the betting strategy x guarantees a positive return for all possible outcomes, *i.e.*, it is a sure-win betting scheme. In economics, we say there is an *arbitrage opportunity*.

If we normalize the vector p in the second condition, so that $\mathbf{1}^T p = 1$, we can interpret it as a probability vector on the outcomes. The condition $R^T p = 0$ means that

$$\mathbf{E}\, Rx = p^T Rx = 0$$

for all x, *i.e.*, the expected return is zero for all betting strategies. In economics, p is called a risk neutral probability.

We can therefore rephrase the arbitrage theorem as follows: There is no sure-win betting strategy (or arbitrage opportunity) if and only if there is a probability vector on the outcomes that makes all bets fair (*i.e.*, the expected gain is zero).

(b) *Betting.* In a simple application, we have exactly as many wagers as there are outcomes ($n = m$). Wager i is to bet that the outcome will be i. The returns are usually expressed as *odds*. For example, suppose that a bookmaker accepts bets on the result of the 2000 European soccer championships. If the odds against Belgium winning are 14 to one, and we bet \$100 on Belgium, then we win \$1400 if they win the tournament, and we lose \$100 otherwise.

In general, if we have m possible outcomes, and the odds against outcome i are λ_i to one, then the return matrix $R \in \mathbf{R}^{m \times m}$ is given by

$$\begin{array}{ll} r_{ij} = \lambda_i & \text{if } j = i \\ r_{ij} = -1 & \text{otherwise.} \end{array}$$

Show that there is no sure-win betting scheme (or arbitrage opportunity) if

$$\sum_{i=1}^{m} \frac{1}{1+\lambda_i} = 1.$$

In fact, you can verify that if this equality is not satisfied, then the betting strategy

$$x_i = \frac{1/(1+\lambda_i)}{1 - \sum_{i=1}^{m} 1/(1+\lambda_i)}$$

always results in a profit.

The common situation in real life is

$$\sum_{i=1}^{m} \frac{1}{1+\lambda_i} > 1,$$

because the bookmakers take a cut on all bets.

13.8 *Log-optimal investment.* We consider an instance of the log-optimal investment problem described in exercise 4.60 of *Convex Optimization*. In this exercise, however, we allow x, the allocation vector, to have negative components. Investing a negative amount $x_i W(t)$ in an asset is called *shorting* the asset. It means you borrow the asset, sell it for $|x_i W(t)|$, and have an obligation to purchase it back later and return it to the lender.

(a) Let R be the $n \times m$-matrix with columns r_j:

$$R = \begin{bmatrix} r_1 & r_2 & \cdots & r_m \end{bmatrix}.$$

We assume that the elements r_{ij} of R are all positive, which implies that the log-optimal investment problem is feasible. Show the following property: if there exists a $v \in \mathbf{R}^n$ with

$$\mathbf{1}^T v = 0, \qquad R^T v \succeq 0, \qquad R^T v \neq 0 \tag{37}$$

then the log-optimal investment problem is unbounded (assuming that the probabilities p_j are all positive).

(b) Derive a Lagrange dual of the log-optimal investment problem (or an equivalent problem of your choice). Use the Lagrange dual to show that the condition in part a is also necessary for unboundedness. In other words, the log-optimal investment problem is bounded if and only if there does not exist a v satisfying (37).

(c) Consider the following small example. We have four scenarios and three investment options. The return vectors for the four scenarios are

$$r_1 = \begin{bmatrix} 2 \\ 1.3 \\ 1 \end{bmatrix}, \qquad r_2 = \begin{bmatrix} 2 \\ 0.5 \\ 1 \end{bmatrix}, \qquad r_3 = \begin{bmatrix} 0.5 \\ 1.3 \\ 1 \end{bmatrix}, \qquad r_4 = \begin{bmatrix} 0.5 \\ 0.5 \\ 1 \end{bmatrix}.$$

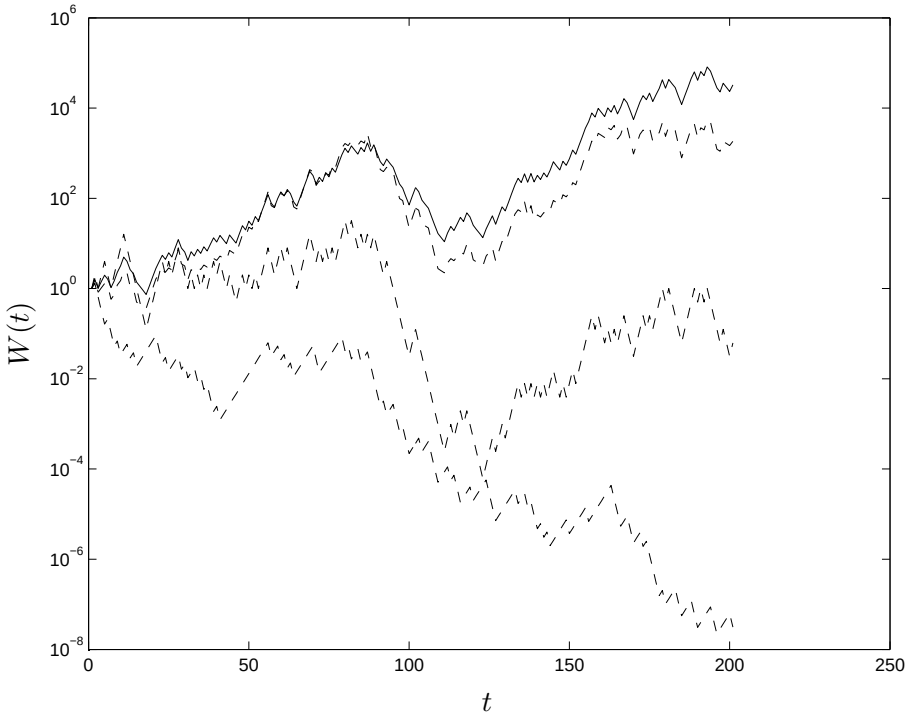
The probabilities of the three scenarios are

$$p_1 = 1/3, \qquad p_2 = 1/6, \qquad p_3 = 1/3, \qquad p_4 = 1/6.$$

The interpretation is as follows. We can invest in two stocks. The first stock doubles in value in each period with a probability 1/2, or decreases by 50% with a probability 1/2. The second stock either increases by 30% with a probability 2/3, or decreases by 50% with a probability 1/3. The fluctuations in the two stocks are independent, so we have four scenarios: both stocks go up (probability 2/6), stock 1 goes up and stock 2 goes down (probability 1/6), stock 1 goes down and stock 2 goes up (probability 1/3), both stocks go down (probability 1/6). The fractions of our capital we invest in stocks 1 and 2 are denoted by x_1 and x_2, respectively. The rest of our capital, $x_3 = 1 - x_1 - x_2$ is not invested.

What is the expected growth rate of the log-optimal strategy x? Compare with the strategies $(x_1, x_2, x_3) = (1, 0, 0)$, $(x_1, x_2, x_3) = (0, 1, 0)$ and $(x_1, x_2, x_3) = (1/2, 1/2, 0)$. (Obviously the expected growth rate for $(x_1, x_2, x_3) = (0, 0, 1)$ is zero.)

Remark. The figure below shows a simulation that compares three investment strategies over 200 periods. The solid line shows the log-optimal investment strategy. The dashed lines show the growth for strategies $x = (1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$.



13.9 *Maximizing house profit in a gamble and imputed probabilities.* A set of n participants bet on which one of m outcomes, labeled $1, \ldots, m$, will occur. Participant i offers to purchase up to $q_i > 0$ gambling contracts, at price $p_i > 0$, that the true outcome will be in the set $S_i \subset \{1, \ldots, m\}$. The house then sells her x_i contracts, with $0 \leq x_i \leq q_i$. If the true outcome j is in S_i, then participant i receives \$1 per contract, *i.e.*, x_i. Otherwise, she loses, and receives nothing. The house collects a total of $x_1 p_1 + \cdots + x_n p_n$, and pays out an amount that depends on the outcome j,

$$\sum_{j \in S_i} x_i.$$

The difference is the house profit.

(a) *Optimal house strategy.* How should the house decide on x so that its worst-case profit (over the possible outcomes) is maximized? (The house determines x after examining all the participant offers.)

(b) *Imputed probabilities.* Suppose $x^\star$ maximizes the worst-case house profit. Show that there exists a probability distribution π on the possible outcomes (*i.e.*, $\pi \in \mathbf{R}^m_+$, $\mathbf{1}^T\pi = 1$) for which $x^\star$ also maximizes the expected house profit. Explain how to find π.

Hint. Formulate the problem in part (a) as an LP; you can construct π from optimal dual variables for this LP.

Remark. Given π, the 'fair' price for offer i is $p_i^{\text{fair}} = \sum_{j\in S_i} \pi_j$. All offers with $p_i > p_i^{\text{fair}}$ will be completely filled (*i.e.*, $x_i = q_i$); all offers with $p_i < p_i^{\text{fair}}$ will be rejected (*i.e.*, $x_i = 0$).

Remark. This exercise shows how the probabilities of outcomes (*e.g.*, elections) can be guessed from the offers of a set of gamblers.

(c) *Numerical example.* Carry out your method on the simple example below with $n = 5$ participants, $m = 5$ possible outcomes, and participant offers

Participant i	p_i	q_i	S_i
1	0.50	10	{1,2}
2	0.60	5	{4}
3	0.60	5	{1,4,5}
4	0.60	20	{2,5}
5	0.20	10	{3}

Compare the optimal worst-case house profit with the worst-case house profit, if all offers were accepted (*i.e.*, $x_i = q_i$). Find the imputed probabilities.

13.10 *Optimal investment to fund an expense stream.* An organization (such as a municipality) knows its operating expenses over the next T periods, denoted $E_1, \ldots, E_T$. (Normally these are positive; but we can have negative E_t, which corresponds to income.) These expenses will be funded by a combination of investment income, from a mixture of bonds purchased at $t = 0$, and a cash account.

The bonds generate investment income, denoted $I_1, \ldots, I_T$. The cash balance is denoted $B_0, \ldots, B_T$, where $B_0 \geq 0$ is the amount of the initial deposit into the cash account. We can have $B_t < 0$ for $t = 1, \ldots, T$, which represents borrowing.

After paying for the expenses using investment income and cash, in period t, we are left with $B_t - E_t + I_t$ in cash. If this amount is positive, it earns interest at the rate $r_+ > 0$; if it is negative, we must pay interest at rate r_-, where $r_- \geq r_+$. Thus the expenses, investment income, and cash balances are linked as follows:

$$B_{t+1} = \begin{cases} (1+r_+)(B_t - E_t + I_t) & B_t - E_t + I_t \geq 0 \\ (1+r_-)(B_t - E_t + I_t) & B_t - E_t + I_t < 0, \end{cases}$$

for $t = 1, \ldots, T-1$. We take $B_1 = (1+r_+)B_0$, and we require that $B_T - E_T + I_T = 0$, which means the final cash balance, plus income, exactly covers the final expense.

The initial investment will be a mixture of bonds, labeled $1, \ldots, n$. Bond i has a price $P_i > 0$, a coupon payment $C_i > 0$, and a maturity $M_i \in \{1, \ldots, T\}$. Bond i generates an income stream

given by

$$a_t^{(i)} = \begin{cases} C_i & t < M_i \\ C_i + 1 & t = M_i \\ 0 & t > M_i, \end{cases}$$

for $t = 1, \ldots, T$. If x_i is the number of units of bond i purchased (at $t = 0$), the total investment cash flow is

$$I_t = x_1 a_t^{(1)} + \cdots + x_n a_t^{(n)}, \quad t = 1, \ldots, T.$$

We will require $x_i \geq 0$. (The x_i can be fractional; they do not need to be integers.)

The total initial investment required to purchase the bonds, and fund the initial cash balance at $t = 0$, is $x_1 P_1 + \cdots + x_n P_n + B_0$.

(a) Explain how to choose x and B_0 to minimize the total initial investment required to fund the expense stream.

(b) Solve the problem instance given in `opt_funding_data.m`. Give optimal values of x and B_0. Give the optimal total initial investment, and compare it to the initial investment required if no bonds were purchased (which would mean that all the expenses were funded from the cash account). Plot the cash balance (versus period) with optimal bond investment, and with no bond investment.

13.11 *Planning production with uncertain demand.* You must order (nonnegative) amounts $r_1, \ldots, r_m$ of raw materials, which are needed to manufacture (nonnegative) quantities $q_1, \ldots, q_n$ of n different products. To manufacture one unit of product j requires at least A_{ij} units of raw material i, so we must have $r \succeq Aq$. (We will assume that A_{ij} are nonnegative.) The per-unit cost of the raw materials is given by $c \in \mathbf{R}_+^m$, so the total raw material cost is $c^T r$.

The (nonnegative) demand for product j is denoted d_j; the number of units of product j sold is $s_j = \min\{q_j, d_j\}$. (When $q_j > d_j$, $q_j - d_j$ is the amount of product j produced, but not sold; when $d_j > q_j$, $d_j - q_j$ is the amount of unmet demand.) The revenue from selling the products is $p^T s$, where $p \in \mathbf{R}_+^n$ is the vector of product prices. The profit is $p^T s - c^T r$. (Both d and q are real vectors; their entries need not be integers.)

You are given A, c, and p. The product demand, however, is not known. Instead, a set of K possible demand vectors, $d^{(1)}, \ldots, d^{(K)}$, with associated probabilities $\pi_1, \ldots, \pi_K$, is given. (These satisfy $\mathbf{1}^T \pi = 1$, $\pi \succeq 0$.)

You will explore two different optimization problems that arise in choosing r and q (the variables).

I. Choose r and q ahead of time. You must choose r and q, knowing only the data listed above. (In other words, you must order the raw materials, and commit to producing the chosen quantities of products, before you know the product demand.) The objective is to maximize the expected profit.

II. Choose r ahead of time, and q after d is known. You must choose r, knowing only the data listed above. Some time after you have chosen r, the demand will become known to you. This means that you will find out which of the K demand vectors is the true demand. Once you

know this, you must choose the quantities to be manufactured. (In other words, you must order the raw materials before the product demand is known; but you can choose the mix of products to manufacture after you have learned the true product demand.) The objective is to maximize the expected profit.

(a) Explain how to formulate each of these problems as a convex optimization problem. Clearly state what the variables are in the problem, what the constraints are, and describe the roles of any auxiliary variables or constraints you introduce.

(b) Carry out the methods from part (a) on the problem instance with numerical data given in `planning_data.m`. This file will define `A`, `D`, `K`, `c`, `m`, `n`, `p` and `pi`. The K columns of D are the possible demand vectors. For both of the problems described above, give the optimal value of r, and the expected profit.

13.12 *Gini coefficient of inequality.* Let $x_1, \ldots, x_n$ be a set of nonnegative numbers with positive sum, which typically represent the wealth or income of n individuals in some group. The *Lorentz curve* is a plot of the fraction f_i of total wealth held by the i poorest individuals,

$$f_i = (1/\mathbf{1}^T x) \sum_{j=1}^{i} x_{(j)}, \quad i = 0, \ldots, n,$$

versus i/n, where $x_{(j)}$ denotes the jth smallest of the numbers $\{x_1, \ldots, x_n\}$, and we take $f_0 = 0$. The Lorentz curve starts at $(0, 0)$ and ends at $(1, 1)$. Interpreted as a continuous curve (as, say, $n \to \infty$) the Lorentz curve is convex and increasing, and lies on or below the straight line joining the endpoints. The curve coincides with this straight line, *i.e.*, $f_i = (i/n)$, if and only if the wealth is distributed equally, *i.e.*, the x_i are all equal.

The *Gini coefficient* is defined as twice the area between the straight line corresponding to uniform wealth distribution and the Lorentz curve:

$$G(x) = (2/n) \sum_{i=1}^{n} ((i/n) - f_i).$$

The Gini coefficient is used as a measure of wealth or income inequality: It ranges between 0 (for equal distribution of wealth) and $1 - 1/n$ (when one individual holds all wealth).

(a) Show that G is a quasiconvex function on $x \in \mathbf{R}^n_+ \setminus \{0\}$.

(b) *Gini coefficient and marriage.* Suppose that individuals i and j get married ($i \neq j$) and therefore pool wealth. This means that x_i and x_j are both replaced with $(x_i + x_j)/2$. What can you say about the change in Gini coefficient caused by this marriage?

13.13 *Internal rate of return for cash streams with a single initial investment.* We use the notation of example 3.34 in the textbook. Let $x \in \mathbf{R}^{n+1}$ be a cash flow over n periods, with x indexed from 0 to n, where the index denotes period number. We assume that $x_0 < 0$, $x_j \geq 0$ for $j = 1, \ldots, n$, and $x_0 + \cdots + x_n > 0$. This means that there is an initial positive investment; thereafter, only payments are made, with the total of the payments exceeding the initial investment. (In the more general setting of example 3.34, we allow additional investments to be made after the initial investment.)

(a) Show that $\mathrm{IRR}(x)$ is quasilinear in this case.

(b) *Blending initial investment only streams.* Use the result in part (a) to show the following. Let $x^{(i)} \in \mathbf{R}^{n+1}$, $i = 1, \ldots, k$, be a set of k cash flows over n periods, each of which satisfies the conditions above. Let $w \in \mathbf{R}^k_+$, with $\mathbf{1}^T w = 1$, and consider the blended cash flow given by $x = w_1 x^{(1)} + \cdots + w_k x^{(k)}$. (We can think of this as investing a fraction w_i in cash flow i.) Show that $\mathrm{IRR}(x) \leq \max_i \mathrm{IRR}(x^{(i)})$. Thus, blending a set of cash flows (with initial investment only) will not improve the IRR over the best individual IRR of the cash flows.

13.14 *Efficient solution of basic portfolio optimization problem.* This problem concerns the simplest possible portfolio optimization problem:

$$\begin{array}{ll} \text{maximize} & \mu^T w - (\lambda/2) w^T \Sigma w \\ \text{subject to} & \mathbf{1}^T w = 1, \end{array}$$

with variable $w \in \mathbf{R}^n$ (the normalized portfolio, with negative entries meaning short positions), and data μ (mean return), $\Sigma \in \mathbf{S}^n_{++}$ (return covariance), and $\lambda > 0$ (the risk aversion parameter). The return covariance has the factor form $\Sigma = FQF^T + D$, where $F \in \mathbf{R}^{n \times k}$ (with rank K) is the *factor loading matrix*, $Q \in \mathbf{S}^k_{++}$ is the factor covariance matrix, and D is a diagonal matrix with positive entries, called the *idiosyncratic risk* (since it describes the risk of each asset that is independent of the factors). This form for Σ is referred to as a 'k-factor risk model'. Some typical dimensions are $n = 2500$ (assets) and $k = 30$ (factors).

(a) What is the flop count for computing the optimal portfolio, if the low-rank plus diagonal structure of Σ is *not* exploited? You can assume that $\lambda = 1$ (which can be arranged by absorbing it into Σ).

(b) Explain how to compute the optimal portfolio more efficiently, and give the flop count for your method. You can assume that $k \ll n$. You do not have to give the best method; any method that has linear complexity in n is fine. You can assume that $\lambda = 1$.

Hints. You may want to introduce a new variable $y = F^T w$ (which is called the vector of factor exposures). You may want to work with the matrix

$$G = \begin{bmatrix} \mathbf{1} & F \\ 0 & -I \end{bmatrix} \in \mathbf{R}^{(n+k) \times (1+k)},$$

treating it as dense, ignoring the (little) exploitable structure in it.

(c) Carry out your method from part (b) on some randomly generated data with dimensions $n = 2500$, $k = 30$. For comparison (and as a check on your method), compute the optimal portfolio using the method of part (a) as well. Give the (approximate) CPU time for each method, using `tic` and `toc`. *Hints.* After you generate D and Q randomly, you might want to add a positive multiple of the identity to each, to avoid any issues related to poor conditioning. Also, to be able to invert a block diagonal matrix efficiently, you'll need to recast it as sparse.

(d) *Risk return trade-off curve.* Now suppose we want to compute the optimal portfolio for M values of the risk aversion parameter λ. Explain how to do this efficiently, and give the complexity in terms of M, n, and k. Compare to the complexity of using the method of part (b) M times. *Hint.* Show that the optimal portfolio is an affine function of $1/\lambda$.

14 Mechanical engineering

14.1 *Optimal design of a tensile structure.* A tensile structure is modeled as a set of n masses in $\mathbf{R}^2$, some of which are fixed, connected by a set of N springs. The masses are in equilibrium, with spring forces, connection forces for the fixed masses, and gravity balanced. (This equilibrium occurs when the position of the masses minimizes the total energy, defined below.)

We let $(x_i, y_i) \in \mathbf{R}^2$ denote the position of mass i, and $m_i > 0$ its mass value. The first p masses are fixed, which means that $x_i = x_i^{\text{fixed}}$ and $y_i = y_i^{\text{fixed}}$, for $i = 1, \ldots, p$. The gravitational potential energy of mass i is gm_iy_i, where $g \approx 9.8$ is the gravitational acceleration.

Suppose spring j connects masses r and s. Its elastic potential energy is

$$(1/2)k_j \left((x_r - x_s)^2 + (y_r - y_s)^2 \right),$$

where $k_j \geq 0$ is the stiffness of spring j.

To describe the topology, *i.e.*, which springs connect which masses, we will use the incidence matrx $A \in \mathbf{R}^{n \times N}$, defined as

$$A_{ij} = \begin{cases} 1 & \text{head of spring } j \text{ connects to mass } i \\ -1 & \text{tail of spring } j \text{ connects to mass } i \\ 0 & \text{otherwise.} \end{cases}$$

Here we arbitrarily choose a head and tail for each spring, but in fact the springs are completely symmetric, and the choice can be reversed without any effect. (Hopefully you will discover why it is convenient to use the incidence matrix A to specify the topology of the system.)

The total energy is the sum of the gravitational energies, over all the masses, plus the sum of the elastic energies, over all springs. The equilibrium positions of the masses is the point that minimizes the total energy, subject to the constraints that the first p positions are fixed. (In the equilibrium positions, the total force on each mass is zero.) We let $E_{\min}$ denote the total energy of the system, in its equilibrium position. (We assume the energy is bounded below; this occurs if and only if each mass is connected, through some set of springs with positive stiffness, to a fixed mass.)

The total energy $E_{\min}$ is a measure of the stiffness of the structure, with larger $E_{\min}$ corresponding to stiffer. (We can think of $E_{\min} = -\infty$ as an infinitely unstiff structure; in this case, at least one mass is not even supported against gravity.)

(a) Suppose we know the fixed positions $x_1^{\text{fixed}}, \ldots, x_p^{\text{fixed}}$, $y_1^{\text{fixed}}, \ldots, y_p^{\text{fixed}}$, the mass values $m_1, \ldots, m_n$, the spring topology A, and the constant g. You are to choose nonnegative $k_1, \ldots, k_N$, subject to a budget constraint $\mathbf{1}^T k = k_1 + \cdots + k_N = k^{\text{tot}}$, where k^{tot} is given. Your goal is to maximize $E_{\min}$.

Explain how to do this using convex optimization.

(b) Carry out your method for the problem data given in `tens_struct_data.m`. This file defines all the needed data, and also plots the equilibrium configuration when the stiffness is evenly distributed across the springs (*i.e.*, $k = (k^{\text{tot}}/N)\mathbf{1}$).

Report the optimal value of $E_{\min}$. Plot the optimized equilibrium configuration, and compare it to the equilibrium configuration with evenly distributed stiffness. (The code for doing this is in the file `tens_struct_data.m`, but commented out.)

14.2 *Equilibrium position of a system of springs.* We consider a collection of n masses in $\mathbf{R}^2$, with locations $(x_1, y_1), \ldots, (x_n, y_n)$, and masses $m_1, \ldots, m_n$. (In other words, the vector $x \in \mathbf{R}^n$ gives the x-coordinates, and $y \in \mathbf{R}^n$ gives the y-coordinates, of the points.) The masses m_i are, of course, positive.

For $i = 1, \ldots, n-1$, mass i is connected to mass $i+1$ by a spring. The potential energy in the ith spring is a function of the (Euclidean) distance $d_i = \|(x_i, y_i) - (x_{i+1}, y_{i+1})\|_2$ between the ith and $(i+1)$st masses, given by

$$E_i = \begin{cases} 0 & d_i < l_i \\ (k_i/2)(d_i - l_i)^2 & d_i \geq l_i \end{cases}$$

where $l_i \geq 0$ is the rest length, and $k_i > 0$ is the stiffness, of the ith spring. The gravitational potential energy of the ith mass is gm_iy_i, where g is a positive constant. The total potential energy of the system is therefore

$$E = \sum_{i=1}^{n-1} E_i + gm^T y.$$

The locations of the first and last mass are fixed. The equilibrium location of the other masses is the one that minimizes E.

(a) Show how to find the equilibrium positions of the masses $2, \ldots, n-1$ using convex optimization. Be sure to justify convexity of any functions that arise in your formulation (if it is not obvious). The problem data are m_i, k_i, l_i, g, x_1, y_1, x_n, and y_n.

(b) Carry out your method to find the equilibrium positions for a problem with $n = 10$, $m_i = 1$, $k_i = 10$, $l_i = 1$, $x_1 = y_1 = 0$, $x_n = y_n = 10$, with g varying from $g = 0$ (no gravity) to $g = 10$ (say). Verify that the results look reasonable. Plot the equilibrium configuration for several values of g.

14.3 *Elastic truss design.* In this problem we consider a truss structure with m bars connecting a set of nodes. Various external forces are applied at each node, which cause a (small) displacement in the node positions. $f \in \mathbf{R}^n$ will denote the vector of (components of) external forces, and $d \in \mathbf{R}^n$ will denote the vector of corresponding node displacements. (By 'corresponding' we mean if f_i is, say, the z-coordinate of the external force applied at node k, then d_i is the z-coordinate of the displacement of node k.) The vector f is called a *loading* or *load*.

The structure is linearly elastic, *i.e.*, we have a linear relation $f = Kd$ between the vector of external forces f and the node displacements d. The matrix $K = K^T \succ 0$ is called the *stiffness matrix* of the truss. Roughly speaking, the 'larger' K is (*i.e.*, the stiffer the truss) the smaller the node displacement will be for a given loading.

We assume that the geometry (unloaded bar lengths and node positions) of the truss is fixed; we are to design the cross-sectional areas of the bars. These cross-sectional areas will be the design variables x_i, $i = 1, \ldots, m$. The stiffness matrix K is a linear function of x:

$$K(x) = x_1K_1 + \cdots + x_mK_m,$$

where $K_i = K_i^T \succeq 0$ depend on the truss geometry. You can assume these matrices are given or known. The total weight W_{tot} of the truss also depends on the bar cross-sectional areas:

$$W_{\text{tot}}(x) = w_1x_1 + \cdots + w_mx_m,$$

where $w_i > 0$ are known, given constants (density of the material times the length of bar i). Roughly speaking, the truss becomes stiffer, but also heavier, when we increase x_i; there is a tradeoff between stiffness and weight.

Our goal is to design the stiffest truss, subject to bounds on the bar cross-sectional areas and total truss weight:

$$l \leq x_i \leq u, \;\; i = 1, \ldots, m, \qquad W_{\rm tot}(x) \leq W,$$

where l, u, and W are given. You may assume that $K(x) \succ 0$ for all feasible vectors x. To obtain a specific optimization problem, we must say how we will measure the stiffness, and what model of the loads we will use.

(a) There are several ways to form a scalar measure of how stiff a truss is, for a given load f. In this problem we will use the *elastic stored energy*

$$\mathcal{E}(x, f) = \frac{1}{2} f^T K(x)^{-1} f$$

to measure the stiffness. Maximizing stiffness corresponds to minimizing $\mathcal{E}(x, f)$.

Show that $\mathcal{E}(x, f)$ is a convex function of x on $\{x \mid K(x) \succ 0\}$.
Hint. Use Schur complements to prove that the epigraph is a convex set.

(b) We can consider several different scenarios that reflect our knowledge about the possible loadings f that can occur. The simplest is that f is a single, fixed, known loading. In more sophisticated formulations, the loading f might be a random vector with known distribution, or known only to lie in some set $\mathcal{F}$, etc.

Show that each of the following four problems is a convex optimization problem, with x as variable.

- *Design for a fixed known loading.* The vector f is known and fixed. The design problem is

$$\begin{array}{ll} \text{minimize} & \mathcal{E}(x, f) \\ \text{subject to} & l \leq x_i \leq u, \quad i = 1, \ldots, m \\ & W_{\rm tot}(x) \leq W. \end{array}$$

- *Design for multiple loadings.* The vector f can take any of N known values $f^{(i)}$, $i = 1, \ldots, N$, and we are interested in the worst-case scenario. The design problem is

$$\begin{array}{ll} \text{minimize} & \max_{i=1,\ldots,N} \mathcal{E}(x, f^{(i)}) \\ \text{subject to} & l \leq x_i \leq u, \quad i = 1, \ldots, m \\ & W_{\rm tot}(x) \leq W. \end{array}$$

- *Design for worst-case, unknown but bounded load.* Here we assume the vector f can take arbitrary values in a ball $B = \{f \mid \|f\| \leq \alpha\}$, for a given value of α. We are interested in minimizing the worst-case stored energy, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \sup_{\|f\| \leq \alpha} \mathcal{E}(x, f^{(i)}) \\ \text{subject to} & l \leq x_i \leq u, \quad i = 1, \ldots, m \\ & W_{\rm tot}(x) \leq W. \end{array}$$

- *Design for a random load with known statistics.* We can also use a stochastic model of the uncertainty in the load, and model the vector f as a random variable with known mean and covariance:

$$\mathbf{E}\, f = f^{(0)}, \quad \mathbf{E}(f - f^{(0)})(f - f^{(0)})^T = \Sigma.$$

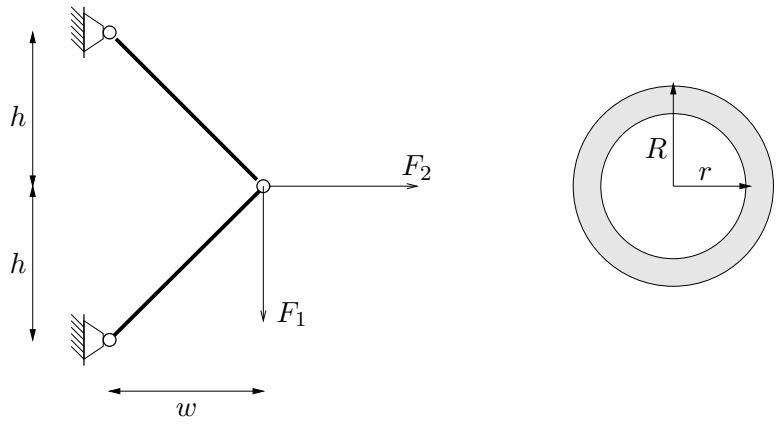
In this case we would be interested in minimizing the expected stored energy, *i.e.*,

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\, \mathcal{E}(x, f^{(i)}) \\ \text{subject to} & l \leq x_i \leq u, \quad i = 1, \ldots, m \\ & W_{\text{tot}}(x) \leq W. \end{array}$$

Hint. If v is a random vector with zero mean and covariance Σ, then $\mathbf{E}\, v^T A v = \mathbf{E}\, \mathbf{tr}\, A v v^T = \mathbf{tr}\, A\, \mathbf{E}\, v v^T = \mathbf{tr}\, A\Sigma$.

(c) Formulate the four problems in (b) as semidefinite programming problems.

14.4 *A structural optimization problem.* [BSS93] The figure shows a two-bar truss with height $2h$ and width w. The two bars are cylindrical tubes with inner radius r and outer radius R. We are interested in determining the values of r, R, w, and h that minimize the weight of the truss subject to a number of constraints. The structure should be strong enough for two loading scenarios. In the first scenario a vertical force F_1 is applied to the node; in the second scenario the force is horizontal with magnitude F_2.



The weight of the truss is proportional to the total volume of the bars, which is given by

$$2\pi(R^2 - r^2)\sqrt{w^2 + h^2}$$

This is the cost function in the design problem.

The first constraint is that the truss should be strong enough to carry the load F_1, *i.e.*, the stress caused by the external force F_1 must not exceed a given maximum value. To formulate this constraint, we first determine the forces in each bar when the structure is subjected to the vertical load F_1. From the force equilibrium and the geometry of the problem we can determine that the magnitudes of the forces in two bars are equal and given by

$$\frac{\sqrt{w^2 + h^2}}{2h} F_1.$$

The maximum force in each bar is equal to the cross-sectional area times the maximum allowable stress σ (which is a given constant). This gives us the first constraint:

$$\frac{\sqrt{w^2+h^2}}{2h}F_1 \leq \sigma\pi(R^2-r^2).$$

The second constraint is that the truss should be strong enough to carry the load F_2. When F_2 is applied, the magnitudes of the forces in two bars are again equal and given by

$$\frac{\sqrt{w^2+h^2}}{2w}F_2,$$

which gives us the second constraint:

$$\frac{\sqrt{w^2+h^2}}{2w}F_2 \leq \sigma\pi(R^2-r^2).$$

We also impose limits $w_{\min} \leq w \leq w_{\max}$ and $h_{\min} \leq h \leq h_{\max}$ on the width and the height of the structure, and limits $1.1r \leq R \leq R_{\max}$ on the outer radius.

In summary, we obtain the following problem:

$$\begin{array}{ll}
\text{minimize} & 2\pi(R^2-r^2)\sqrt{w^2+h^2} \\
\text{subject to} & \dfrac{\sqrt{w^2+h^2}}{2h}F_1 \leq \sigma\pi(R^2-r^2) \\
& \dfrac{\sqrt{w^2+h^2}}{2w}F_2 \leq \sigma\pi(R^2-r^2) \\
& w_{\min} \leq w \leq w_{\max} \\
& h_{\min} \leq h \leq h_{\max} \\
& 1.1r \leq R \leq R_{\max} \\
& R > 0, \quad r > 0, \quad w > 0, \quad h > 0.
\end{array}$$

The variables are R, r, w, h.

Formulate this as a geometric programming problem.

14.5 *Optimizing the inertia matrix of a 2D mass distribution.* An object has density $\rho(z)$ at the point $z = (x, y) \in \mathbf{R}^2$, over some region $\mathcal{R} \subset \mathbf{R}^2$. Its mass $m \in \mathbf{R}$ and center of gravity $c \in \mathbf{R}^2$ are given by

$$m = \int_{\mathcal{R}} \rho(z)\, dxdy, \qquad c = \frac{1}{m}\int_{\mathcal{R}} \rho(z) z\, dxdy,$$

and its inertia matrix $M \in \mathbf{R}^{2\times 2}$ is

$$M = \int_{\mathcal{R}} \rho(z)(z-c)(z-c)^T\, dxdy.$$

(You do not need to know the mechanics interpretation of M to solve this problem, but here it is, for those interested. Suppose we rotate the mass distribution around a line passing through the

center of gravity in the direction $q \in \mathbf{R}^2$ that lies in the plane where the mass distribution is, at angular rate ω. Then the total kinetic energy is $(\omega^2/2)q^T M q$.)

The goal is to choose the density ρ, subject to $0 \leq \rho(z) \leq \rho^{\max}$ for all $z \in \mathcal{R}$, and a fixed total mass $m = m^{\text{given}}$, in order to maximize $\lambda_{\min}(M)$.

To solve this problem numerically, we will discretize $\mathcal{R}$ into N pixels each of area a, with pixel i having constant density ρ_i and location (say, of its center) $z_i \in \mathbf{R}^2$. We will assume that the integrands above don't vary too much over the pixels, and from now on use instead the expressions

$$m = a\sum_{i=1}^{N} \rho_i, \quad c = \frac{a}{m}\sum_{i=1}^{N} \rho_i z_i, \quad M = a\sum_{i=1}^{N} \rho_i (z_i - c)(z_i - c)^T.$$

The problem below refers to these discretized expressions.

(a) Explain how to solve the problem using convex (or quasiconvex) optimization.

(b) Carry out your method on the problem instance with data in `inertia_dens_data.m`. This file includes code that plots a density. Give the optimal inertia matrix and its eigenvalues, and plot the optimal density.

15 Graphs and networks

15.1 A *hypergraph* with nodes $1, \ldots, m$ is a set of nonempty subsets of $\{1, 2, \ldots, m\}$, called *edges*. An ordinary graph is a special case in which the edges contain no more than two nodes.

We consider a hypergraph with m nodes and assume coordinate vectors $x_j \in \mathbf{R}^p$, $j = 1, \ldots, m$, are associated with the nodes. Some nodes are fixed and their coordinate vectors x_j are given. The other nodes are free, and their coordinate vectors will be the optimization variables in the problem. The objective is to place the free nodes in such a way that some measure of the physical size of the nets is small.

As an example application, we can think of the nodes as modules in an integrated circuit, placed at positions $x_j \in \mathbf{R}^2$. Every edge is an interconnect network that carries a signal from one module to one or more other modules.

To define a measure of the size of a net, we store the vectors x_j as columns of a matrix $X \in \mathbf{R}^{p\times m}$. For each edge S in the hypergraph, we use X_S to denote the $p \times |S|$ submatrix of X with the columns associated with the nodes of S. We define

$$f_S(X) = \inf_y \|X_S - y\mathbf{1}^T\|. \tag{38}$$

as the *size* of the edge S, where $\|\cdot\|$ is a matrix norm, and $\mathbf{1}$ is a vector of ones of length $|S|$.

(a) Show that the optimization problem

$$\text{minimize} \quad \textstyle\sum_{\text{edges } S} f_S(X)$$

is convex in the free node coordinates x_j.

(b) The size $f_S(X)$ of a net S obviously depends on the norm used in the definition (38). We consider five norms.

- *Frobenius norm:*
$$\|X_s - y\mathbf{1}^T\|_F = \left(\sum_{j\in S}\sum_{i=1}^{p}(x_{ij} - y_i)^2\right)^{1/2}.$$
- *Maximum Euclidean column norm:*
$$\|X_S - y\mathbf{1}^T\|_{2,1} = \max_{j\in S}\left(\sum_{i=1}^{p}(x_{ij} - y_i)^2\right)^{1/2}.$$
- *Maximum column sum norm:*
$$\|X_S - y\mathbf{1}^T\|_{1,1} = \max_{j\in S}\sum_{i=1}^{p}|x_{ij} - y_i|.$$
- *Sum of absolute values norm:*
$$\|X_s - y\mathbf{1}^T\|_{\text{sav}} = \sum_{j\in S}\sum_{i=1}^{p}|x_{ij} - y_i|$$

- *Sum-row-max norm:*

$$\|X_s - y\mathbf{1}^T\|_{\text{srm}} = \sum_{i=1}^{p} \max_{j\in S} |x_{ij} - y_i|$$

For which of these norms does f_S have the following interpretations?

(i) $f_S(X)$ is the radius of the smallest Euclidean ball that contains the nodes of S.

(ii) $f_S(X)$ is (proportional to) the perimeter of the smallest rectangle that contains the nodes of S:

$$f_S(X) = \frac{1}{4}\sum_{i=1}^{p}(\max_{j\in S} x_{ij} - \min_{j\in S} x_{ij}).$$

(iii) $f_S(X)$ is the squareroot of the sum of the squares of the Euclidean distances to the mean of the coordinates of the nodes in S:

$$f_S(X) = \left(\sum_{j\in S} \|x_j - \bar{x}\|_2^2\right)^{1/2} \quad \text{where} \quad \bar{x}_i = \frac{1}{|S|}\sum_{k\in S} x_{ik}, \quad i = 1, \ldots, p.$$

(iv) $f_S(X)$ is the sum of the ℓ_1-distances to the (coordinate-wise) median of the coordinates of the nodes in S:

$$f_S(X) = \sum_{j\in S} \|x_j - \hat{x}\|_1 \quad \text{where} \quad \hat{x}_i = \text{median}(\{x_{ik} \mid k \in S\}), \quad i = 1, \ldots, p.$$

15.2 Let $W \in \mathbf{S}^n$ be a symmetric matrix with nonnegative elements w_{ij} and zero diagonal. We can interpret W as the representation of a weighted undirected graph with n nodes. If $w_{ij} = w_{ji} > 0$, there is an edge between nodes i and j, with weight w_{ij}. If $w_{ij} = w_{ji} = 0$ then nodes i and j are not connected. The *Laplacian* of the weighted graph is defined as

$$L(W) = -W + \mathbf{diag}(W\mathbf{1}).$$

This is a symmetric matrix with elements

$$L_{ij}(W) = \begin{cases} \sum_{k=1}^n w_{ik} & i = j \\ -w_{ij} & i \neq j. \end{cases}$$

The Laplacian has the useful property that

$$y^T L(W) y = \sum_{i\leq j} w_{ij}(y_i - y_j)^2$$

for all vectors $y \in \mathbf{R}^n$.

(a) Show that the function $f : \mathbf{S}^n \to \mathbf{R}$,

$$f(W) = \inf_{\mathbf{1}^T x = 0} n\lambda_{\max}\left(L(W) + \mathbf{diag}(x)\right),$$

is convex.

(b) Give a simple argument why $f(W)$ is an upper bound on the optimal value of the combinatorial optimization problem

$$\begin{array}{ll} \text{maximize} & y^T L(W) y \\ \text{subject to} & y_i \in \{-1, 1\}, \quad i = 1, \ldots, n. \end{array}$$

This problem is known as the *max-cut* problem, for the following reason. Every vector y with components ± 1 can be interpreted as a partition of the nodes of the graph in a set $S = \{i \mid y_i = 1\}$ and a set $T = \{i \mid y_i = -1\}$. Such a partition is called a *cut* of the graph. The objective function in the max-cut problem is

$$y^T L(W) y = \sum_{i \leq j} w_{ij} (y_i - y_j)^2.$$

If y is ± 1-vector corresponding to a partition in sets S and T, then $y^T L(W) y$ equals four times the sum of the weights of the edges that join a point in S to a point in T. This is called the weight of the cut defined by y. The solution of the max-cut problem is the cut with the maximum weight.

(c) The function f defined in part 1 can be evaluated, for a given W, by solving the optimization problem

$$\begin{array}{ll} \text{minimize} & n\lambda_{\max}(L(W) + \mathbf{diag}(x)) \\ \text{subject to} & \mathbf{1}^T x = 0, \end{array}$$

with variable $x \in \mathbf{R}^n$. Express this problem as an SDP.

(d) Derive an alternative expression for $f(W)$, by taking the dual of the SDP in part 3. Show that the dual SDP is equivalent to the following problem:

$$\begin{array}{ll} \text{maximize} & \sum_{i \leq j} w_{ij} \|p_i - p_j\|_2^2 \\ \text{subject} & \|p_i\|_2 = 1, \quad i = 1, \ldots, n, \end{array}$$

with variables $p_i \in \mathbf{R}^n$, $i = 1, \ldots, n$. In this problem we place n points p_i on the unit sphere in $\mathbf{R}^n$ in such a way that the weighted sum of their squared pair-wise distances is maximized.

15.3 *Utility versus latency trade-off in a network.* We consider a network with m edges, labeled $1, \ldots, m$, and n flows, labeled $1, \ldots, n$. Each flow has an associated nonnegative flow rate f_j; each edge or link has an associated positive capacity c_i. Each flow passes over a fixed set of links (its route); the total traffic t_i on link i is the sum of the flow rates over all flows that pass through link i. The flow routes are described by a routing matrix $R \in \mathbf{R}^{m \times n}$, defined as

$$R_{ij} = \begin{cases} 1 & \text{flow } j \text{ passes through link } i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the vector of link traffic, $t \in \mathbf{R}^m$, is given by $t = Rf$. The link capacity constraint can be expressed as $Rf \preceq c$. The (logarithmic) network utility is defined as $U(f) = \sum_{j=1}^n \log f_j$.

The (average queuing) delay on link i is given by

$$d_i = \frac{1}{c_i - t_i}$$

(multiplied by a constant, that doesn't matter to us). We take $d_i = \infty$ for $t_i = c_i$. The delay or latency for flow j, denoted l_j, is the sum of the link delays over all links that flow j passes through. We define the maximum flow latency as

$$L = \max\{l_1, \ldots, l_n\}.$$

We are given R and c; we are to choose f.

(a) How would you find the flow rates that maximize the utility U, ignoring flow latency? (In particular, we allow $L = \infty$.) We'll refer to this maximum achievable utility as $U^{\max}$.

(b) How would you find the flow rates that minimize the maximum flow latency L, ignoring utility? (In particular, we allow $U = -\infty$.) We'll refer to this minimum achievable latency as $L^{\min}$.

(c) Explain how to find the optimal trade-off between utility U (which we want to maximize) and latency L (which we want to minimize).

(d) Find $U^{\max}$, $L^{\min}$, and plot the optimal trade-off of utility versus latency for the network with data given in `net_util_data.m`, showing $L^{\min}$ and $U^{\max}$ on the same plot. Your plot should cover the range from $L = 1.1L^{\min}$ to $L = 11L^{\min}$. Plot U vertically, on a linear scale, and L horizontally, using a log scale.

Note. For parts (a), (b), and (c), your answer can involve solving one or more convex optimization problems. But if there is a simpler solution, you should say so.

16 Energy and power

16.1 *Power flow optimization with 'N − 1' reliability constraint.* We model a network of power lines as a graph with n nodes and m edges. The power flow along line j is denoted p_j, which can be positive, which means power flows along the line in the direction of the edge, or negative, which means power flows along the line in the direction opposite the edge. (In other words, edge orientation is only used to determine the direction in which power flow is considered positive.) Each edge can support power flow in either direction, up to a given maximum capacity $P_j^{\max}$, *i.e.*, we have $|p_j| \leq P_j^{\max}$.

Generators are attached to the first k nodes. Generator i provides power g_i to the network. These must satisfy $0 \leq g_i \leq G_i^{\max}$, where $G_i^{\max}$ is a given maximum power available from generator i. The power generation costs are $c_i > 0$, which are given; the total cost of power generation is $c^T g$.

Electrical loads are connected to the nodes $k+1, \ldots, n$. We let $d_i \geq 0$ denote the demand at node $k+i$, for $i = 1, \ldots, n-k$. We will consider these loads as given. In this simple model we will neglect all power losses on lines or at nodes. Therefore, power must balance at each node: the total power flowing into the node must equal the sum of the power flowing out of the node. This power balance constraint can be expressed as

$$Ap = \begin{bmatrix} -g \\ d \end{bmatrix},$$

where $A \in \mathbf{R}^{n \times m}$ is the node-incidence matrix of the graph, defined by

$$A_{ij} = \begin{cases} +1 & \text{edge } j \text{ enters node } i, \\ -1 & \text{edge } j \text{ leaves node } i, \\ 0 & \text{otherwise.} \end{cases}$$

In the basic power flow optimization problem, we choose the generator powers g and the line flow powers p to minimize the total power generation cost, subject to the constraints listed above. The (given) problem data are the incidence matrix A, line capacities $P^{\max}$, demands d, maximum generator powers $G^{\max}$, and generator costs c.

In this problem we will add a basic (and widely used) reliability constraint, commonly called an '$N-1$ constraint'. (N is not a parameter in the problem; '$N-1$' just means 'all-but-one'.) This states that the system can still operate even if any one power line goes out, by re-routing the line powers. The case when line j goes out is called 'failure contingency j'; this corresponds to replacing $P_j^{\max}$ with 0. The requirement is that there must exist a contingency power flow vector $p^{(j)}$ that satisfies all the constraints above, with $p_j^{(j)} = 0$, using the same given generator powers. (This corresponds to the idea that power flows can be re-routed quickly, but generator power can only be changed more slowly.) The '$N-1$ reliability constraint' requires that for each line, there is a contingency power flow vector. The '$N-1$ reliability constraint' is (implicitly) a constraint on the generator powers.

The questions below concern the specific instance of this problem with data given in `rel_pwr_flow_data.m`. (Executing this file will also generate a figure showing the network you are optimizating.) Especially for part (b) below, you must explain exactly how you set up the problem as a convex optimization problem.

(a) *Nominal optimization.* Find the optimal generator and line power flows for this problem instance (without the $N-1$ reliability constraint). Report the optimal cost and generator powers. (You do not have to give the power line flows.)

(b) *Nominal optimization with $N-1$ reliability constraint.* Minimize the nominal cost, but you must choose generator powers that meet the $N-1$ reliability requirement as well. Report the optimal cost and generator powers. (You do not have to give the nominal power line flows, or any of the contingency flows.)

16.2 *Optimal generator dispatch.* In the *generator dispatch problem*, we schedule the electrical output power of a set of generators over some time interval, to minimize the total cost of generation while exactly meeting the (assumed known) electrical demand. One challenge in this problem is that the generators have dynamic constraints, which couple their output powers over time. For example, every generator has a maximum rate at which its power can be increased or decreased.

We label the generators $i = 1, \ldots, n$, and the time periods $t = 1, \ldots, T$. We let $p_{i,t}$ denote the (nonnegative) power output of generator i at time interval t. The (positive) electrical demand in period t is d_t. The total generated power in each period must equal the demand:

$$\sum_{i=1}^{n} p_{i,t} = d_t, \quad t = 1, \ldots, T.$$

Each generator has a minimum and maximum allowed output power:

$$P_i^{\min} \leq p_{i,t} \leq P_i^{\max}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T.$$

The cost of operating generator i at power output u is $\phi_i(u)$, where ϕ_i is an increasing strictly convex function. (Assuming the cost is mostly fuel cost, convexity of ϕ_i says that the thermal efficiency of the generator decreases as its output power increases.) We will assume these cost functions are quadratic: $\phi_i(u) = \alpha_i u + \beta_i u^2$, with α_i and β_i positive.

Each generator has a maximum ramp-rate, which limits the amount its power output can change over one time period:

$$|p_{i,t+1} - p_{i,t}| \leq R_i, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T-1.$$

In addition, changing the power output of generator i from u_t to u_{t+1} incurs an additional cost $\psi_i(u_{t+1} - u_t)$, where ψ_i is a convex function. (This cost can be a real one, due to increased fuel use during a change of power, or a fictitious one that accounts for the increased maintenance cost or decreased lifetime caused by frequent or large changes in power output.) We will use the power change cost functions $\psi_i(v) = \gamma_i|v|$, where γ_i are positive.

Power plants with large capacity (*i.e.*, $P_i^{\max}$) are typically more efficient (*i.e.*, have smaller α_i, β_i), but have smaller ramp-rate limits, and higher costs associated with changing power levels. Small gas-turbine plants ('peakers') are less efficient, have less capacity, but their power levels can be rapidly changed.

The total cost of operating the generators is

$$C = \sum_{i=1}^{n} \sum_{t=1}^{T} \phi_i(p_{i,t}) + \sum_{i=1}^{n} \sum_{t=1}^{T-1} \psi_i(p_{i,t+1} - p_{i,t}).$$

Choosing the generator output schedules to minimize C, while respecting the constraints described above, is a convex optimization problem. The problem data are d_t (the demands), the generator power limits $P_i^{\min}$ and $P_i^{\max}$, the ramp-rate limits R_i, and the cost function parameters α_i, β_i, and γ_i. We will assume that problem is feasible, and that $p_{i,t}^\star$ are the (unique) optimal output powers.

(a) *Price decomposition.* Show that there are power prices $Q_1, \ldots, Q_T$ for which the following holds: For each i, $p_{i,t}^\star$ solves the optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{t=1}^{T} \left(\phi_i(p_{i,t}) - Q_t p_{i,t}\right) + \sum_{t=1}^{T-1} \psi_i(p_{i,t+1} - p_{i,t}) \\ \text{subject to} & P_i^{\min} \leq p_{i,t} \leq P_i^{\max}, \quad t = 1, \ldots, T \\ & |p_{i,t+1} - p_{i,t}| \leq R_i, \quad t = 1, \ldots, T-1. \end{array}$$

The objective here is the portion of the objective for generator i, minus the revenue generated by the sale of power at the prices Q_t. Note that this problem involves *only* generator i; it can be solved independently of the other generators (once the prices are known). How would you find the prices Q_t?

You do not have to give a full formal proof; but you must explain your argument fully. You are welcome to use results from the text book.

(b) Solve the generator dispatch problem with the data given in `gen_dispatch_data.m`, which gives (fake, but not unreasonable) demand data for 2 days, at 15 minute intervals. This file includes code to plot the demand, optimal generator powers, and prices. (You must replace these variables with their correct values.) Comment on anything you see in your solution that might at first seem odd. Using the prices found, solve the problems in part (a) for the generators separately, to be sure they give the optimal powers (up to some small numerical errors).

Remark. While beyond the scope of this course, we mention that there are very simple price update mechanisms that adjust the prices in such a way that when the generators independently schedule themselves using the prices (as described above), we end up with the total power generated in each period matching the demand, *i.e.*, the optimal solution of the whole (coupled) problem. This gives a decentralized method for generator dispatch.

16.3 *Optimizing a portfolio of energy sources.* We have n different energy sources, such as coal-fired plants, several wind farms, and solar farms. Our job is to size each of these, *i.e.*, to choose its capacity. We will denote by c_i the capacity of plant i; these must satisfy $c_i^{\min} \leq c_i \leq c_i^{\max}$, where $c_i^{\min}$ and $c_i^{\max}$ are given minimum and maximum values.

Each generation source has a cost to build and operate (including fuel, maintenance, government subsidies and taxes) over some time period. We lump these costs together, and assume that the cost is proportional to c_i, with (given) coefficient b_i. Thus, the total cost to build and operate the energy sources is $b^T c$ (in, say, \$/hour).

Each generation source is characterized by an availability a_i, which is a random variable with values in $[0, 1]$. If source i has capacity c_i, then the power available from the plant is $c_i a_i$; the total power available from the portfolio of energy sources is $c^T a$, which is a random variable. A coal fired plant has $a_i = 1$ almost always, with $a_i < 1$ when one of its units is down for maintenance. A wind farm, in contrast, is characterized by strong fluctations in availability with $a_i = 1$ meaning a strong wind

is blowing, and $a_i = 0$ meaning no wind is blowing. A solar farm has $a_i = 1$ only during peak sun hours, with no cloud cover; at other times (such as night) we have $a_i = 0$.

Energy demand $d \in \mathbf{R}_+$ is also modeled as a random variable. The components of a (the availabilities) and d (the demand) are *not* independent. Whenever the total power available falls short of the demand, the additional needed power is generated by (expensive) peaking power plants at a fixed positive price p. The average cost of energy produced by the peakers is

$$\mathbf{E}\, p(d - c^T a)_+,$$

where $x_+ = \max\{0, x\}$. This average cost has the same units as the cost $b^T c$ to build and operate the plants.

The objective is to choose c to minimize the overall cost

$$C = b^T c + \mathbf{E}\, p(d - c^T a)_+.$$

Sample average approximation. To solve this problem, we will minimize a cost function based on a sample average of peaker cost,

$$C^{\mathrm{sa}} = b^T c + \frac{1}{N} \sum_{j=1}^{N} p(d^{(j)} - c^T a^{(j)})_+$$

where $(a^{(j)}, d^{(j)})$, $j = 1, \ldots, N$, are (given) samples from the joint distribution of a and d. (These might be obtained from historical data, weather and demand forecasting, and so on.)

Validation. After finding an optimal value of c, based on the set of samples, you should double check or validate your choice of c by evaluating the overall cost on another set of (validation) samples, $(\tilde{a}^{(j)}, \tilde{d}^{(j)})$, $j = 1, \ldots, N^{\mathrm{val}}$,

$$C^{\mathrm{val}} = b^T c + \frac{1}{N^{\mathrm{val}}} \sum_{j=1}^{N^{\mathrm{val}}} p(\tilde{d}^{(j)} - c^T \tilde{a}^{(j)})_+.$$

(These could be another set of historical data, held back for validation purposes.) If $C^{\mathrm{sa}} \approx C^{\mathrm{val}}$, our confidence that each of them is approximately the optimal value of C is increased.

Finally we get to the problem. Get the data in `energy_portfolio_data.m`, which includes the required problem data, and the samples, which are given as a $1 \times N$ row vector `d` for the scalars $d^{(j)}$, and an $n \times N$ matrix `A` for $a^{(j)}$. A second set of samples is given for validation, with the names `d_val` and `A_val`.

Carry out the optimization described above. Give the optimal cost obtained, C^{sa}, and compare to the cost evaluated using the validation data set, C^{val}.

Compare your solution with the following naive ('certainty-equivalent') approach: Replace a and d with their (sample) means, and then solve the resulting optimization problem. Give the optimal cost obtained, C^{ce} (using the average values of a and d). Is this a lower bound on the optimal value of the original problem? Now evaluate the cost for these capacities on the validation set, $C^{\mathrm{ce,val}}$. Make a brief statement.

16.4 *Optimizing processor speed.* A set of n tasks is to be completed by n processors. The variables to be chosen are the processor speeds $s_1, \ldots, s_n$, which must lie between a given minimum value $s_{\min}$ and a maximum value $s_{\max}$. The computational load of task i is α_i, so the time required to complete task i is $\tau_i = \alpha_i/s_i$.

The power consumed by processor i is given by $p_i = f(s_i)$, where $f : \mathbf{R} \to \mathbf{R}$ is positive, increasing, and convex. Therefore, the total energy consumed is

$$E = \sum_{i=1}^{n} \frac{\alpha_i}{s_i} f(s_i).$$

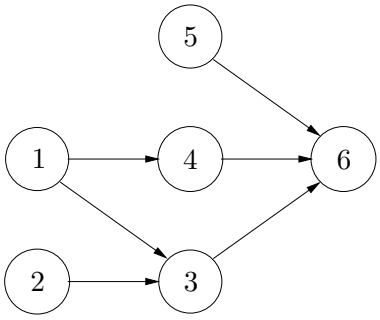
(Here we ignore the energy used to transfer data between processors, and assume the processors are powered down when they are not active.)

There is a set of *precedence constraints* for the tasks, which is a set of m ordered pairs $\mathcal{P} \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$. If $(i, j) \in \mathcal{P}$, then task j cannot start until task i finishes. (This would be the case, for example, if task j requires data that is computed in task i.) When $(i, j) \in \mathcal{P}$, we refer to task i as a *precedent* of task j, since it must precede task j. We assume that the precedence constraints define a directed acyclic graph (DAG), with an edge from i to j if $(i, j) \in \mathcal{P}$.

If a task has no precedents, then it starts at time $t = 0$. Otherwise, each task starts as soon as all of its precedents have finished. We let T denote the time for all tasks to be completed.

To be sure the precedence constraints are clear, we consider the very small example shown below, with $n = 6$ tasks and $m = 6$ precedence constraints.

$$\mathcal{P} = \{(1,4),\ (1,3),\ (2,3),\ (3,6),\ (4,6),\ (5,6)\}.$$



In this example, tasks 1, 2, and 5 start at time $t = 0$ (since they have no precedents). Task 1 finishes at $t = \tau_1$, task 2 finishes at $t = \tau_2$, and task 5 finishes at $t = \tau_5$. Task 3 has tasks 1 and 2 as precedents, so it starts at time $t = \max\{\tau_1, \tau_2\}$, and ends τ_3 seconds later, at $t = \max\{\tau_1, \tau_2\} + \tau_3$. Task 4 completes at time $t = \tau_1 + \tau_4$. Task 6 starts when tasks 3, 4, and 5 have finished, at time $t = \max\{\max\{\tau_1, \tau_2\} + \tau_3, \tau_1 + \tau_4, \tau_5\}$. It finishes τ_6 seconds later. In this example, task 6 is the last task to be completed, so we have

$$T = \max\{\max\{\tau_1, \tau_2\} + \tau_3, \tau_1 + \tau_4, \tau_5\} + \tau_6.$$

(a) Formulate the problem of choosing processor speeds (between the given limits) to minimize completion time T, subject to an energy limit $E \leq E_{\max}$, as a convex optimization problem.

The data in this problem are $\mathcal{P}$, $s_{\min}$, $s_{\max}$, $\alpha_1, \ldots, \alpha_n$, $E_{\max}$, and the function f. The variables are $s_1, \ldots, s_n$.

Feel free to change variables or to introduce new variables. Be sure to explain clearly why your formulation of the problem is convex, and why it is equivalent to the problem statement above.

Important:

- Your formulation must be convex for any function f that is positive, increasing, and convex. You cannot make any further assumptions about f.
- This problem refers to the general case, not the small example described above.

(b) Consider the specific instance with data given in `proc_speed_data.m`, and processor power

$$f(s) = 1 + s + s^2 + s^3.$$

The precedence constraints are given by an $m \times 2$ matrix `prec`, where m is the number of precedence constraints, with each row giving one precedence constraint (the first column gives the precedents).

Plot the optimal trade-off curve of energy E versus time T, over a range of T that extends from its minimum to its maximum possible value. (These occur when all processors operate at $s_{\max}$ and $s_{\min}$, respectively, since T is monotone nonincreasing in s.) On the same plot, show the energy-time trade-off obtained when all processors operate at the same speed $\bar{s}$, which is varied from $s_{\min}$ to $s_{\max}$.

Note: In this part of the problem there is no limit $E^{\max}$ on E as in part (a); you are to find the optimal trade-off of E versus T.

16.5 *Minimum energy processor speed scheduling.* A single processor can adjust its speed in each of T time periods, labeled $1, \ldots, T$. Its speed in period t will be denoted s_t, $t = 1, \ldots, T$. The speeds must lie between given (positive) minimum and maximum values, $S^{\min}$ and $S^{\max}$, respectively, and must satisfy a slew-rate limit, $|s_{t+1} - s_t| \leq R$, $t = 1, \ldots, T-1$. (That is, R is the maximum allowed period-to-period change in speed.) The energy consumed by the processor in period t is given by $\phi(s_t)$, where $\phi : \mathbf{R} \to \mathbf{R}$ is increasing and convex. The total energy consumed over all the periods is $E = \sum_{t=1}^{T} \phi(s_t)$.

The processor must handle n jobs, labeled $1, \ldots, n$. Each job has an availability time $A_i \in \{1, \ldots, T\}$, and a deadline $D_i \in \{1, \ldots, T\}$, with $D_i \geq A_i$. The processor cannot start work on job i until period $t = A_i$, and must complete the job by the end of period D_i. Job i involves a (nonnegative) total work W_i. You can assume that in each time period, there is at least one job available, *i.e.*, for each t, there is at least one i with $A_i \leq t$ and $D_i \geq t$.

In period t, the processor allocates its effort across the n jobs as θ_t, where $\mathbf{1}^T \theta_t = 1$, $\theta_t \succeq 0$. Here θ_{ti} (the ith component of θ_t) gives the fraction of the processor effort devoted to job i in period t. Respecting the availability and deadline constraints requires that $\theta_{ti} = 0$ for $t < A_i$ or $t > D_i$. To complete the jobs we must have

$$\sum_{t=A_i}^{D_i} \theta_{ti} s_t \geq W_i, \quad i = 1, \ldots, n.$$

(a) Formulate the problem of choosing the speeds $s_1, \ldots, s_T$, and the allocations $\theta_1, \ldots, \theta_T$, in order to minimize the total energy E, as a convex optimization problem. The problem data are $S^{\min}$, $S^{\max}$, R, ϕ, and the job data, A_i, D_i, W_i, $i = 1, \ldots, n$. Be sure to justify any change of variables, or introduction of new variables, that you use in your formulation.

(b) Carry out your method on the problem instance described in `proc_sched_data.m`, with quadratic energy function $\phi(s_t) = \alpha + \beta s_t + \gamma s_t^2$. (The parameters α, β, and γ are given in the data file.) Executing this file will also give a plot showing the availability times and deadlines for the jobs.

Give the energy obtained by your speed profile and allocations. Plot these using the command `bar((s*ones(1,n)).*theta,1,'stacked')`, where s is the $T \times 1$ vector of speeds, and θ is the $T \times n$ matrix of allocations with components θ_{ti}. This will show, at each time period, how much effective speed is allocated to each job. The top of the plot will show the speed s_t. (You don't need to turn in a color version of this plot; B&W is fine.)

16.6 *AC power flow analysis via convex optimization.* This problem concerns an AC (alternating current) power system consisting of m transmission lines that connect n nodes. We describe the topology by the node-edge incidence matrix $A \in \mathbf{R}^{n \times m}$, where

$$A_{ij} = \begin{cases} +1 & \text{line } j \text{ leaves node } i \\ -1 & \text{line } j \text{ enters node } i \\ 0 & \text{otherwise.} \end{cases}$$

The power flow on line j is p_j (with positive meaning in the direction of the line as defined in A, negative meaning power flow in the opposite direction).

Node i has voltage phase angle ϕ_i, and external power input s_i. (If a generator is attached to node i we have $s_i > 0$; if a load is attached we have $s_i < 0$; if the node has neither, $s_i = 0$.) Neglecting power losses in the lines, and assuming power is conserved at each node, we have $Ap = s$. (We must have $\mathbf{1}^T s = 0$, which means that the total power pumped into the network by generators balances the total power pulled out by the loads.)

The line power flows are a nonlinear function of the difference of the phase angles at the nodes they connect to:

$$p_j = \kappa_j \sin(\phi_k - \phi_l),$$

where line j goes from node k to node l. Here κ_j is a known positive constant (related to the inductance of the line). We can write this in matrix form as $p = \mathbf{diag}(\kappa) \sin(A^T \phi)$, where sin is applied elementwise.

The *DC power flow equations* are

$$Ap = s, \qquad p = \mathbf{diag}(\kappa) \sin(A^T \phi).$$

In the power analysis problem, we are given s, and want to find p and ϕ that satisfy these equations. We are interested in solutions with voltage phase angle differences that are smaller than $\pm 90^\circ$. (Under normal conditions, real power lines are never operated with voltage phase angle differences more than $\pm 20^\circ$ or so.)

You will show that the DC power flow equations can be solved by solving the convex optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=j}^{m} \psi_j(p_j) \\ \text{subject to} & Ap = s, \end{array}$$

with variable s, where

$$\psi_j(u) = \kappa_j \int_0^u \sin^{-1}(v/\kappa_j)\, dv = u \sin^{-1}(u/\kappa_j) + \kappa_j(\sqrt{1-(u/\kappa_j)^2} - 1),$$

with domain $\mathbf{dom}\, \psi_j = (-\kappa_j, \kappa_j)$. (The second expression will be useless in this problem.)

(a) Show that the problem above is convex.

(b) Suppose the problem above has solution $p^\star$, with optimal dual variable $\nu^\star$ associated with the equality constraint $Ap = s$. Show that $p^\star$, $\phi = \nu^\star$ solves the DC power flow equation. *Hint.* Write out the optimality conditions for the problem above.

17 Miscellaneous applications

17.1 *Earth mover's distance.* In this exercise we explore a general method for constructing a distance between two probability distributions on a finite set, called the *earth mover's distance*, *Wasserstein metric*, or *Dubroshkin metric*. Let x and y be two probability distributions on $\{1, \ldots, n\}$, *i.e.*, $\mathbf{1}^T x = \mathbf{1}^T y = 1$, $x \succeq 0$, $y \succeq 0$. We imagine that x_i is the amount of earth stored at location i; our goal is to move the earth between locations to obtain the distribution given by y. Let C_{ij} be the cost of moving one unit of earth from location j to location i. We assume that $C_{ii} = 0$, and $C_{ij} = C_{ji} > 0$ for $i \neq j$. (We allow $C_{ij} = \infty$, which means that earth cannot be moved directly from node j to node i.) Let $S_{ij} \geq 0$ denote the amount of earth moved from location j to location i. The total cost is $\sum_{i,j=1}^n S_{ij} C_{ij} = \mathbf{tr}\, C^T S$. The shipment matrix S must satisfy the balance equations,

$$\sum_{j=1}^n S_{ij} = y_i, \quad i = 1, \ldots, n, \qquad \sum_{i=1}^n S_{ij} = x_j, \quad j = 1, \ldots, n,$$

which we can write compactly as $S\mathbf{1} = y$, $S^T\mathbf{1} = x$. (The first equation states that the total amount shipped into location i equals y_i; the second equation states that the total shipped out from location j is x_j.) The earth mover's distance between x and y is given by the minimal cost of earth moving required to transform x to y, *i.e.*, the optimal value of the problem

$$\begin{array}{ll} \text{minimize} & \mathbf{tr}\, C^T S \\ \text{subect to} & S_{ij} \geq 0, \quad i, j = 1, \ldots, n \\ & S\mathbf{1} = y, \quad S^T\mathbf{1} = x, \end{array}$$

with variables $S \in \mathbf{R}^{n \times n}$. We can also give a probabality interpretation: We seek the joint distribution that minimizes the expected value of C, with given marginals x and y.

The earth mover's distance is used to compare, for example, 2D images, with C_{ij} equal to the distance between pixels i and j. If x and y represent two photographs of the same scene, from slightly different viewpoints and with an offset in camera position (say), $d(x, y)$ will be small, but the distance between x and y measured by most common norms (*e.g.*, $\|x - y\|_1$) will be large.

(a) Show that d is a metric, *i.e.*, $d(x, y) = d(y, x)$, $d(x, x) = 0$, $d(x, y) > 0$ for $x \neq y$, and that the triangle inequality holds: $d(x, z) = d(x, y) + d(y, z)$ (where z is another distribution).

(b) Show that $d(x, y)$ is the optimal value of the problem

$$\begin{array}{ll} \text{maximize} & \nu^T x + \mu^T y \\ \text{subject to} & \nu_i + \mu_j \leq C_{ij}, \quad i, j = 1, \ldots, n, \end{array}$$

with variables $\nu,\ \mu \in \mathbf{R}^n$.

(c) Now consider the special case with $C_{i,i+1} = 1$, $i = 1, \ldots, n-1$, $C_{ii} = 0$, and $C_{ij} = \infty$ otherwise. Express d in terms of the cumulative distributions of x and y, $f_i = \sum_{j=1}^i x_j$, $g_i = \sum_{j=1}^i y_j$. (You do not need to give a fully formal argument here; an informal derivation is fine.)

17.2 *Radiation treatment planning.* In radiation treatment, radiation is delivered to a patient, with the goal of killing or damaging the cells in a tumor, while carrying out minimal damage to other tissue. The radiation is delivered in beams, each of which has a known pattern; the level of each beam can

be adjusted. (In most cases multiple beams are delivered at the same time, in one 'shot', with the treatment organized as a sequence of 'shots'.) We let b_j denote the level of beam j, for $j = 1, \ldots, n$. These must satisfy $0 \leq b_j \leq B^{\max}$, where $B^{\max}$ is the maximum possible beam level. The exposure area is divided into m voxels, labeled $i = 1, \ldots, m$. The dose d_i delivered to voxel i is linear in the beam levels, *i.e.*, $d_i = \sum_{j=1}^{n} A_{ij} b_j$. Here $A \in \mathbf{R}_+^{m \times n}$ is a (known) matrix that characterizes the beam patterns. We now describe a simple radiation treatment planning problem.

A (known) subset of the voxels, $\mathcal{T} \subset \{1, \ldots, m\}$, corresponds to the tumor or target region. We require that a minimum radiation dose D^{target} be administered to each tumor voxel, *i.e.*, $d_i \geq D^{\text{target}}$ for $i \in \mathcal{T}$. For all other voxels, we would like to have $d_i \leq D^{\text{other}}$, where D^{other} is a desired maximum dose for non-target voxels. This is generally not feasible, so instead we settle for minimizing the penalty

$$E = \sum_{i \notin \mathcal{T}} ((d_i - D^{\text{other}})_+)^2,$$

where $(\cdot)_+$ denotes the nonnegative part. We can interpret E as the sum of the squares of the nontarget excess doses.

(a) Show that the treatment planning problem is convex. The optimization variable is $b \in \mathbf{R}^n$; the problem data are $B^{\max}$, A, $\mathcal{T}$, D^{target}, and D^{other}.

(b) Solve the problem instance with data given in the file `treatment_planning_data.m`. Here we have split the matrix A into `Atarget`, which contains the rows corresponding to the target voxels, and `Aother`, which contains the rows corresponding to other voxels. Give the optimal value. Plot the dose histogram for the target voxels, and also for the other voxels. Make a brief comment on what you see. *Remark.* The beam pattern matrix in this problem instance is randomly generated, but similar results would be obtained with realistic data.

17.3 *Flux balance analysis in systems biology.* Flux balance analysis is based on a very simple model of the reactions going on in a cell, keeping track only of the gross rate of consumption and production of various chemical species within the cell. Based on the known stoichiometry of the reactions, and known upper bounds on some of the reaction rates, we can compute bounds on the other reaction rates, or cell growth, for example.

We focus on m metabolites in a cell, labeled $M_1, \ldots, M_m$. There are n reactions going on, labeled $R_1, \ldots, R_n$, with nonnegative reaction rates $v_1, \ldots, v_n$. Each reaction has a (known) stoichiometry, which tells us the rate of consumption and production of the metabolites per unit of reaction rate. The stoichiometry data is given by the *stoichiometry matrix* $S \in \mathbf{R}^{m \times n}$, defined as follows: S_{ij} is the rate of production of M_i due to unit reaction rate $v_j = 1$. Here we consider consumption of a metabolite as negative production; so $S_{ij} = -2$, for example, means that reaction R_j causes metabolite M_i to be consumed at a rate $2v_j$.

As an example, suppose reaction R_1 has the form $M_1 \to M_2 + 2M_3$. The consumption rate of M_1, due to this reaction, is v_1; the production rate of M_2 is v_1; and the production rate of M_3 is $2v_1$. (The reaction R_1 has no effect on metabolites $M_4, \ldots, M_m$.) This corresponds to a first column of S of the form $(-1, 1, 2, 0, \ldots, 0)$.

Reactions are also used to model flow of metabolites into and out of the cell. For example, suppose that reaction R_2 corresponds to the flow of metabolite M_1 into the cell, with v_2 giving the flow rate. This corresponds to a second column of S of the form $(1, 0, \ldots, 0)$.

The last reaction, R_n, corresponds to biomass creation, or cell growth, so the reaction rate v_n is the cell growth rate. The last column of S gives the amounts of metabolites used or created per unit of cell growth rate.

Since our reactions include metabolites entering or leaving the cell, as well as those converted to biomass within the cell, we have conservation of the metabolites, which can be expressed as $Sv = 0$. In addition, we are given upper limits on *some* of the reaction rates, which we express as $v \preceq v^{\max}$, where we set $v_j^{\max} = \infty$ if no upper limit on reaction rate j is known. The goal is to find the maximum possible cell growth rate (*i.e.*, largest possible value of v_n) consistent with the constraints

$$Sv = 0, \qquad v \succeq 0, \qquad v \preceq v^{\max}.$$

The questions below pertain to the data found in `fba_data.m`.

(a) Find the maximum possible cell growth rate $G^\star$, as well as optimal Lagrange multipliers for the reaction rate limits. How sensitive is the maximum growth rate to the various reaction rate limits?

(b) *Essential genes and synthetic lethals.* For simplicity, we'll assume that each reaction is controlled by an associated gene, *i.e.*, gene G_i controls reaction R_i. Knocking out a set of genes associated with some reactions has the effect of setting the reaction rates (or equivalently, the associated $v^{\max}$ entries) to zero, which of course reduces the maximum possible growth rate. If the maximum growth rate becomes small enough or zero, it is reasonable to guess that knocking out the set of genes will kill the cell. An *essential gene* is one that when knocked out reduces the maximum growth rate below a given threshold $G^{\min}$. (Note that G_n is always an essential gene.) A *synthetic lethal* is a pair of non-essential genes that when knocked out reduces the maximum growth rate below the threshold. Find all essential genes and synthetic lethals for the given problem instance, using the threshold $G^{\min} = 0.2G^\star$.

References

[And79] T. Ando. Concavity of certain maps on positive definite matrices and applications to Hadamard products. *Linear Algebra and Its Applications*, 26:203–241, 1979.

[BB65] E. F. Beckenbach and R. Bellman. *Inequalities*. Springer, second edition, 1965.

[BN78] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. John Wiley & Sons, 1978.

[BSL08] A. Beck, P. Stoica, and J. Li. Exact and approximate solutions of source localization problems. *IEEE Transactions on Signal Processing*, 56:1770–1778, 2008.

[BSS93] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming. Theory and Algorithms.* John Wiley & Sons, second edition, 1993.

[HLP52] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition, 1952.

[Mar05] P. Maréchal. On a functional operation generating convex functions, part 1: duality. *Journal of Optimization Theory and Applications*, 126(1):175–189, 2005.

[MO79] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, 1979.

[Roc70] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[RV73] A. W. Roberts and D. E. Varberg. *Convex Functions*. Academic Press, 1973.

[vT84] J. van Tiel. *Convex Analysis. An Introductory Text*. John Wiley & Sons, 1984.