Convex Optimization
Homework 5

Spring 1401
Due date: 16th of Ordibehesht

---

1. $\ell_{1.5}$ 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:

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

   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.

   (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).

2. *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),

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

   Another method minimizes instead the total variation,

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

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 $(i,j) \notin \mathcal{K}$.)

3. *Estimation with unknown sensor nonlinearity.* We consider the measurement setup

$$y_i = f\left(a_i^T x + b_i + v_i\right), \quad i = 1, \ldots, m,$$

where $x \in \mathbf{R}^n$ is the vector to be estimated, $y_i \in \mathbf{R}$ are the measurements, $a_i \in \mathbf{R}^n$, $b_i \in \mathbf{R}$ are known, and v_i are IID noises with log-concave probability density. The function $f : \mathbf{R} \to \mathbf{R}$, which represents a measurement nonlinearity, is not known. However, it is known that $f'(t) \in [l, u]$ for all t, where $0 < l < u$ are given. Explain how to use convex optimization to find a maximum likelihood estimate of x, as well as the function f. (This is an infinite-dimensional ML estimation problem, but you can be informal in your approach and explanation.)

Estimating a vector with unknown measurement nonlinearity. We want to estimate a vector $x \in \mathbf{R}^n$, given some measurements

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

Here $a_i \in \mathbf{R}^n$ are known, v_i are IID $\mathcal{N}\left(0, \sigma^2\right)$ 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.$*$, which includes a matrix $A \in \mathbf{R}^{m \times n}$, with rows $a_1^T, \ldots, a_m^T$. Give $\hat{x}_{\text{ml}}$, the maximum likelihood estimate of x. Plot your estimated function $\hat{\phi}_{\text{ml}}$. (You can do this by plotting $(\hat{z}_{\text{ml}})_i$ versus y_i, with y_i on the vertical axis and $(\hat{z}_{\text{ml}})_i$ 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.)

4. *Maximum likelihood estimation of a log-concave distribution.* We have a random variable X which takes values in $\{1, \ldots, n\}$. It has a distribution $p \in \mathbf{R}^n$, with $\text{prob}(X = i) = p_i$. However, we do not know p, and would like to determine it based on N independent samples of X. In those N samples, let m_i denote the number of samples for which $X = i$, so $\sum_i m_i = N$. The likelihood function is then

$$l(p) = \prod_{i=1}^{n} p_i^{m_i}.$$

We know that the distribution p is log-concave. Recall a discrete function $f : \mathbf{Z} \to \mathbf{R}$ is called concave if $f(i) \geq (1/2)(f(i-1) + f(i+1))$. For functions f defined on $\{1, \ldots, n\}$ we require this constraint to hold at $i = 2, \ldots, n-1$. The function p is called log-concave if $\log p$ is concave. Given $m_1, \ldots, m_n$, we would like to find the log-concave distribution p of maximum likelihood.

(a) Formulate this problem as a convex optimization problem.

(b) We have $n = 13$ and observe

$$m = (1, 5, 6, 15, 18, 20, 22, 11, 22, 8, 9, 4, 2).$$

Carry out your method from part (a) on this data. Plot m_i/N (the empirical distribution) and your estimate of p.

5. *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,

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

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.*. 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

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

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

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

6. *Minimax linear fitting.* Consider a linear measurement model $y = Ax + v$, where $x \in \mathbf{R}^n$ is a vector of parameters to be estimated, $y \in \mathbf{R}^m$ is a vector of measurements, $v \in \mathbf{R}^m$ is a set of measurement errors, and $A \in \mathbf{R}^{m \times n}$ with rank n, with $m \geq n$. We know y and A, but we don't know v; our goal is to estimate x. We make only one assumption about the measurement error v : $\|v\|_\infty \leq \epsilon$.
We will estimate x using a linear estimator $\hat{x} = By$; we must choose the estimation matrix $B \in \mathbf{R}^{n \times m}$. The estimation error is $e = \hat{x} - x$. We will choose B to minimize the maximum possible value of $\|e\|_\infty$, where the maximum is over all values of x and all values of v satisfying $\|v\|_\infty \leq \epsilon$.

(a) Show how to find B via convex optimization.

(b) Numerical example. Solve the problem instance given in minimax_fit_data.m. Display the $\hat{x}$ you obtain and report $\|\hat{x} - x^{\text{true}}\|_\infty$. Here x^{true} is the value of x used to generate the measurement y; it is given in the data file.

Good Luck!