Étude Problem

This is one of the étude problems for this winter. Please turn it in by uploading to Gradescope by 5pm on Thursday, with a regraded version by 5pm on Friday. We will not be lenient with upload times: because of the nature of the regrading, we will be posting solutions as close as possible to 5:01pm on Thursdays, so turning the étude in late will result in no credit. Please upload a placeholder to Gradescope well before the due date.

You may use any books, notes, or computer programs, but you may not discuss the étude with anyone—including online—until the solutions are posted. The only exception is that you can ask us for clarification, via the course staff email address. We’ve tried pretty hard to make these questions unambiguous and clear, so we’re unlikely to say much.

Please submit your étude via Gradescope.

We will deduct points from long, needlessly complex solutions, even if they are correct. Our solutions are not long, so if you find that your solution to a problem goes on and on for many pages, you should try to figure out a simpler one. We expect neat, legible études from everyone, including those enrolled Cr/N.

When a problem involves computation you must give all of the following: a clear discussion and justification of exactly what you did, the source code that produces the result, and the final numerical results or plots.

Files containing problem data can be found in the following location:

http://www.stanford.edu/~jduchi/teaching/364a/data/

Please respect the honor code. Although we allow you to work on homework assignments in small groups, you cannot discuss the études with anyone.

Some études are (quite) straightforward. Others, not so much.

Be sure you are using the most recent version of CVX, CVXPY, CVXR, or Convex.jl.

Some problems involve applications. But you do not need to know anything about the problem area to solve the problem; the problem statement contains everything you need.
5. Robustness in a linear modeling problem. In this problem, you investigate the choice of losses in a problem of fitting a linear predictor to given data. We assume the generative model

\[ y = X\theta_{\text{gen}} + w, \]

where \( X \in \mathbb{R}^{m \times n} \) is an \( m \times n \) data matrix, \( y \in \mathbb{R}^m \) are targets, \( w \) is (unobserved) noise, and \( \theta_{\text{gen}} \in \mathbb{R}^n \) is a vector we attempt to find given the pair \((X, y)\). We consider a setting in which the data may be corrupted—either adversarially or because of mis-measurement—yet we still wish to estimate \( \theta_{\text{gen}} \) by minimizing a convex loss. We investigate a few possibilities here.

We consider three losses applying to triples \((\theta, x, y)\) \( \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \), each giving different robustness properties: the squared error

\[ \ell_{\text{sq}}(\theta, x, y) = \frac{1}{2}(x^T \theta - y)^2, \]

the absolute error

\[ \ell_{\text{abs}}(\theta, x, y) = |x^T \theta - y|, \]

and the normalized error

\[ \ell_{\text{norm}}(\theta, x, y) = \frac{1}{\max\{|x\|_2, 1\}}|x^T \theta - y|. \]

Each is convex in \( \theta \). (Note that for this problem, \( x \) and \( y \) are problem data, not variables.)

For these three losses, you will estimate \( \theta_{\text{gen}} \) by solving

\[
\text{minimize} \sum_{i=1}^m \ell(\theta, x_i, y_i)
\]

in the variable \( \theta \), where \( X = [x_1 \cdots x_m]^T \) has rows \( x_i^T \) and \( y = [y_1 \cdots y_m]^T \), for different choices of data matrix \( X \), target vector \( y \), and loss \( \ell \).

The data for this problem is available in robust_linear_models_data.*. There are two data matrices \( X \) and two target vectors \( y \) in the file, \( X_{\text{std}}, X_{\text{outliers}}; y_{\text{std}}, y_{\text{outliers}} \). The pair \( X_{\text{std}}, y_{\text{std}} \) corresponds to data generated via the well-specified model (1), with \( w \) a mean-zero vector. The matrix \( X_{\text{outliers}} \) has its first 10 rows corrupted by large noise, and similarly, the vector \( y_{\text{outliers}} \) has its first 10 entries corrupted.

(a) For the squared loss \( \ell_{\text{sq}} \), solve problem (2) with the following four pairs of data: \((X_{\text{std}}, y_{\text{std}}), (X_{\text{std}}, y_{\text{outliers}}), (X_{\text{outliers}}, y_{\text{std}}), \) and \((X_{\text{outliers}}, y_{\text{outliers}})\). Give the error \( \|\theta^* - \theta_{\text{gen}}\|_2 \), where \( \theta^* \) denotes the solution to problem (2), for each of the data pairs.

(b) Repeat part (a), but use the absolute loss \( \ell_{\text{abs}} \) instead of the squared loss.

(c) Repeat part (a), but use the normalized absolute loss \( \ell_{\text{norm}} \) instead of the squared loss.

(d) In a sentence or two, explain why you might expect the results you see.

Include your code in your solutions.