1. *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

$$y = X\theta_{\text{gen}} + w, \tag{1}$$

where $X \in \mathbf{R}^{m\times n}$ is an $m \times n$ data matrix, $y \in \mathbf{R}^m$ are targets, w is (unobserved) noise, and $\theta_{\text{gen}} \in \mathbf{R}^n$ 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 θ_{gen} by minimizing a convex loss. We investigate a few possibilities here.

We consider three losses applying to triples $(\theta, x, y) \in \mathbf{R}^n \times \mathbf{R}^n \times \mathbf{R}$, 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 θ. (Note that for this problem, x and y are problem *data*, not variables.)

For these three losses, you will estimate θ_{gen} by solving

$$\text{minimize} \quad \sum_{i=1}^{m} \ell(\theta, x_i, y_i) \tag{2}$$

in the variable θ, 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 ℓ.

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_{std}, X_{outliers}, y_{std}, and y_{outliers}. The pair X_{std}, y_{std} corresponds to data generated via the well-specified model (1), with w a mean-zero vector. The matrix X_{outliers} has its first 10 rows corrupted by large noise, and similarly, the vector y_{outliers} has its first 10 entries corrupted.

(a) For the squared loss ℓ_{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 $\|\theta^\star - \theta_{\text{gen}}\|_2$, where $\theta^\star$ denotes the solution to problem (2), for each of the data pairs.

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

(c) Repeat part (a), but use the normalized absolute loss ℓ_{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.