Question 4.1 (Leave-one-out solutions in M-estimation): Consider the M-estimation (empirical risk minimization) problem:

\[
\min_b L_n(b) := \frac{1}{n} \sum_{i=1}^{n} \ell(y_i - x_i^T b)
\]

where \(\ell\) is twice-continuously differentiable, convex, and symmetric. In this problem, we will develop a method to efficiently approximate the minimizers of the leave-one-out \(\ell_2\) (ridge)-regularized objective in cross validation by solving a sequence of quadratic problems; even more, our method will involve only performing rank-one updates to a matrix inverse.\(^1\) Let \(\hat{\beta} = \arg\min_b \{L_n(b) + \frac{\lambda}{2} \|b\|^2\}\) be the minimizer of the empirical risk (the M-estimator), and for \(k = 1, 2, \ldots, n\), define the leave-one-out objective

\[
L_{-k}(b) := \frac{1}{n} \sum_{i \neq k} \ell(y_i - x_i^T b).
\]

Note that both \(L_n\) and \(L_{-k}\) are convex functions (bowl-shaped).

Consider finding minimizers of \(L_{-k}(b)\), equivalently, finding the perturbation to \(\hat{\beta}\):

\[
\Delta_k := \arg\min_{\Delta} \left\{ L_{-k}(\hat{\beta} + \Delta) + \frac{\lambda}{2} \|\Delta\|^2 \right\}.
\]

In this problem, you will show that this is very well-approximated by a simpler quadratic minimizer (at least under appropriate conditions). By Taylor’s theorem, for \(\Delta \in \mathbb{R}^d\) we have

\[
L_{-k}(\hat{\beta} + \Delta) = L_{-k}(\hat{\beta}) + \nabla L_{-k}(\hat{\beta})^T \Delta + \frac{1}{2} \Delta^T \nabla^2 L_{-k}(\hat{\beta}) \Delta
\]

for some \(\beta_\Delta\) between \(\hat{\beta}\) and \(\hat{\beta} + \Delta\). Define the gradients and Hessian matrices

\[
g_k := \nabla L_{-k}(\hat{\beta}) + \lambda \hat{\beta}, \quad H = \nabla^2 L_n(\hat{\beta}) + \lambda I, \quad H_k = \nabla^2 L_{-k}(\hat{\beta}) + \lambda I,
\]

so that if we define the empirical errors \(\tilde{\varepsilon}_i = y_i - x_i^T \hat{\beta}\), these have the explicit forms

\[
g_k = -\frac{1}{n} \sum_{i \neq k} \ell'(\tilde{\varepsilon}_i)x_i + \lambda \hat{\beta} = \frac{1}{n} \ell'(\tilde{\varepsilon}_k)x_k,
\]

\[
H = \frac{1}{n} \sum_{i=1}^{n} \ell''(\tilde{\varepsilon}_i)x_i x_i^T + \lambda I, \quad H_k = H - \frac{1}{n} \ell''(\tilde{\varepsilon}_i)x_i x_i^T.
\]

(To simplify \(g_k\), we used that \(\nabla L_n(\hat{\beta}) + \lambda \hat{\beta} = 0\).) Then

\[
L_{-k}(\hat{\beta} + \Delta) + \frac{\lambda}{2} \|\Delta\|^2 \approx L_{-k}(\hat{\beta}) + g_k^T \Delta + \frac{1}{2} \Delta^T H_k \Delta.
\]

We define \(\hat{\Delta}_k\) to be the minimizer of this quadratic approximation to \(L_{-k} + \frac{\lambda}{2} \|\cdot\|^2\) around \(\hat{\beta}\),

\[
\hat{\Delta}_k = \arg\min_{\Delta} \left\{ L_{-k}(\hat{\beta}) + g_k^T \Delta + \frac{1}{2} \Delta H_k \Delta \right\} = -H_k^{-1} g_k.
\]

\(^1\)In Question 4.2 (which is extra credit), we work through an argument making approximations here rigorous and showing that, in fact, we lose very little by our computationally efficient approach.
(a) Assume you have computed $H^{-1}$. Show how to compute $H_k^{-1}g_k$ efficiently, that is, without recomputing the full inverse of $H_k$. (It is sufficient to simply give the formula.)

(b) Let $\hat{y}_{-k} = x_k^T(\hat{\beta} + \Delta_k)$ be the (approximate) leave-one-out prediction. Show that $\hat{y}_{-k} = \hat{y}_k - x_k^T H_k^{-1} g_k$ and that $\hat{\eps}_{-k} = y_k - \hat{y}_{-k} = \hat{\eps}_k + x_k^T H_k^{-1} g_k$. Using these identities, show how to compute the vector $[\hat{y}_{-k}]_{n=1}^n$ in time $O(nd^2)$, assuming that you have $\hat{\beta}$ already.

(c) Implement the (approximate) leave-one-out cross validation procedure above to choose the level $\lambda$ of ridge regularization when solving the M-estimation problem

$$
\min_b L_n(b) + \frac{\lambda}{2} \| b \|^2_2,
$$

where $\ell(t) = \log(1 + e^t) + \log(1 + e^{-t})$. Use your procedure to evaluate the (approximate) leave-one-out (LOO) validation error for the randomly generated data in the file `generate-outlier-data.*`. Note that the actual error you should be plotting is

$$
\text{LOO} = \frac{1}{n} \sum_{i=1}^n \ell(\hat{\eps}_{-i}).
$$

Plot your LOO error against $\lambda$ for 25 logarithmically spaced values of $\lambda$ from .01 to 10 (your horizontal axis should be on the log scale as well). Include your code in your solution.

(d) Choose the $\hat{\lambda}$ minimizing the LOO error, and let $\hat{\beta}$ be the minimizer of the $\ell_2$-regularized objective (4.1) on the training data. Evaluate its median absolute prediction error on the held-out test set generated by `generate.data` (call this $\text{ERR}_{\text{test}}$). For 25 values of $\lambda$ logarithmically spaced between 2 and 50, let

$$
\hat{\beta}_\lambda^\ell = \arg\min_b \left\{ \frac{1}{2n} \| Xb - y \|^2 + \frac{\lambda}{2} \| b \|^2_2 \right\}
$$

be the $\ell_2$-regularized least-squares solution, and let $\text{ERR}_{\text{test}}^\ell(\lambda)$ be its median absolute prediction error on the held-out test set. Plot $\text{ERR}_{\text{test}}^\ell(\lambda) - \text{ERR}_{\text{test}}$ against $\lambda$. What do you observe?

(e) Repeat the same experiment as in part (d), except that you should vary the number of outliers (see the method `generate.data`) in $\{0, 5, 10, 15, 20, 25\}$. Plot the same results as above. What do you observe?

---

2This file provides a method for minimizing the robust loss with ridge regularization, which you should use to find your initial estimate of $\hat{\beta}$. To run it, you will need to install CVXPY or CVXR, which solve convex optimization problems.
Question 4.2 (Extra Credit): We revisit the setting in Question 4.1, but in this variant, we shall develop a perturbation guarantee to show that our heuristic of using the second-order approximation is actually quite accurate. Roughly, you will show that if the Hessian $H = \nabla^2 L_n(\hat{\beta})$ is positive definite (a relatively easy condition to satisfy), then

$$\max_{k \leq n} \|\Delta_k - \hat{\Delta}_k\| = O\left(\frac{1}{n^2}\right)$$

for each $k \in [n]$. That is, the difference between the “true” leave-one-out minimizer and the approximation is of much smaller order than any statistical/sampling error.

Unfortunately, to do this fully rigorously requires nontrivial setup and a bit of analysis. We review the tools and assumptions here. For simplicity through this derivation, we assume the first and second derivatives satisfy the boundedness conditions

$$\sup_{t \in \mathbb{R}} |\ell'(t)| \leq 1, \quad \sup_{t \in \mathbb{R}} |\ell''(t)| \leq 1, \quad |\ell''(t) - \ell''(s)| \leq |t - s|, \quad \text{all } t, s \in \mathbb{R}.$$

We also assume that the covariate vectors $x_i \in \mathbb{R}^d$ satisfy $\|x_i\|_2 \leq D_x$ for all $i$. Let $X = [x_1 \cdots x_n]^T \in \mathbb{R}^{n \times d}$ be the usual design matrix.

A few useful inequalities and notational simplifications follow. For symmetric matrices $A, B$, we say $A \succeq B$ if $A - B$ is positive semidefinite, that is, $\lambda_{\min}(A - B) \geq 0$. We also have Weyl’s inequality, that is, $|\lambda_i(A + B) - \lambda_i(A)| \leq \|B\|_{\text{op}}$ for any symmetric $A, B$, where $\|B\|_{\text{op}} = \sup_{\|v\| = 1} \|Bv\|_2$ is the usual operator norm. A useful convexity inequality is the following: if $f$ is a convex function, and for some $\lambda, c > 0$ its second derivative satisfies $\nabla^2 f(b) \succeq \lambda I$ for all $b \in \mathbb{R}^d$ satisfying $\|b - \beta\|_2 \leq c$, then

$$f(b) \geq f(\beta) + \nabla f(\beta)^T (b - \beta) + \frac{\lambda}{2} \min\{c, \|b - \beta\|_2\} \|b - \beta\|_2. \quad (4.2)$$

The rough proof outline is the following: first, we show that the Hessian $\nabla^2 L_{-k}$ is positive definite, so that $L_{-k}$ has reasonable growth away from $\hat{\beta}$. This implies that the minimizer $\Delta_k$ cannot be too large, as otherwise, $L_{-k}(\hat{\beta} + \Delta) > L_{-k}(\hat{\beta})$. Once we have this, then we can perform a more careful second-order Taylor approximation of $L_{-k}(\hat{\beta} + \Delta)$, solving directly.

(a) Show that

$$\nabla^2 L_n(\hat{\beta} + \Delta) \succeq \nabla^2 L_n(\hat{\beta}) - D_x \|\Delta\|_2 \frac{1}{n} X^T X.$$

Conclude that if that $H = \nabla^2 L_n(\hat{\beta}) \succeq 2\lambda I$, then whenever $\|\Delta\|_2 \leq \lambda/(2D_x \|n^{-1} X^T X\|_{\text{op}})$,

$$\nabla^2 L_n(\hat{\beta} + \Delta) \succeq \frac{3\lambda}{2} I.$$

(b) Show that if $H = \nabla^2 L_n(\hat{\beta}) \succeq 2\lambda I$, then

$$\nabla^2 L_{-k}(\hat{\beta} + \Delta) \succeq \lambda I$$

when $\|\Delta\|_2 \leq \lambda/(2D_x \|n^{-1} X^T X\|_{\text{op}})$ and $n$ is large enough that $\frac{2D_x^2}{n} \leq \lambda$, i.e., $n \geq \frac{2D_x^2}{\lambda}$.

\[3\] A similar argument is precisely what shows that robust M-estimators are indeed robust.
(c) As before assume $H = \nabla^2 L_n(\tilde{\beta}) \succeq 2\lambda I$. Assume that $n$ is large enough that
\[ n \geq \frac{4D_x^2 \|n^{-1}X^TX\|_{op}}{\lambda^2}. \]
Argue that the minimizer
\[ \Delta_k := \arg\min_{\Delta} L_{-k}(\tilde{\beta} + \Delta) \]
satisfies $\|\Delta_k\|_2 \leq \frac{2D_x}{\lambda n}$.

(d) Assume the conditions in parts (a)–(c). Show that under these,
\[ \|\hat{\Delta}_k - \Delta_k\|_2 = O\left(\frac{1}{n^2}\right). \]

In our answer, we obtain
\[ \|\hat{\Delta}_k - \Delta_k\|_2 \leq \frac{2D_x^5}{\lambda^2 - 2D_x^4/n} \cdot \frac{1}{n^2}. \]

Hint: Perform a Taylor approximation, recognizing that $0 = \nabla L_{-k}(\tilde{\beta} + \Delta_k) \approx g_k + H_k \Delta_k$.
The following bound on matrix inverses may be useful: if $A$ is positive definite and the error matrix $E$ satisfies $\|E\|_{op} \leq \lambda_{\text{min}}(A)$, then $\|(A + E)^{-1} - A^{-1}\|_{op} \leq \frac{\|E\|_{op}}{\lambda_{\text{min}}(A) - \|E\|_{op}}$.\(^4\)

\(^4\)This identity follows from the equality $(A + E)^{-1} = A^{-1} + \sum_{i=1}^{\infty} (-1)^i (A^{-1} E)^i A^{-1}$. 

4