Math 126 Problem 20: Subsampled Regression and Leverage Scores

Spring 2027

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Monday, April 19, 2027 (Pacific time).**

- You are encouraged to discuss ideas and work with your classmates. However, you *must acknowledge* your collaborators at the top of each solution on which you collaborated with others and you *must write* your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown.
- I recommend that you write your solutions in a Python notebook using Google Colab.
- You should submit your solutions as a *single PDF* via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by *not* grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.
Original notebook: Google Colab.

Subsampled Regression and Leverage Scores

Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a feature matrix with full column rank ($n \geq d$), and let $\mathbf{b} \in \mathbb{R}^n$ be the target vector. Instead of multiplying $\mathbf{A}$ by a dense Johnson-Lindenstrauss matrix to compress it, another strategy in randomized numerical linear algebra is to directly subsample the rows of $\mathbf{A}$ and $\mathbf{b}$.

However, uniform sampling is risky: if a single row contains critical information (e.g., a rare but important feature), uniform sampling might drop it. To avoid this, we sample rows proportional to their *leverage scores*.

The leverage score ℓ_i of the i-th row $\mathbf{a}_i^\top$ is defined as the i-th diagonal element of the "hat matrix" $\mathbf{H} = \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top$. Mathematically, $\ell_i = \mathbf{a}_i^\top(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{a}_i$.

Part A: The Sum of Leverage Scores

In order to use leverage scores as a probability distribution for sampling, we need to know their total sum. Prove that the sum of all leverage scores is exactly equal to d, the dimension of the feature space.

Hint: Use the cyclic property of the trace.

Part B: Bounds on Leverage Scores

Show that $\mathbf{H}$ is idempotent (i.e., $\mathbf{H}^2 = \mathbf{H}$). Then, use this property to prove that the leverage scores are bounded such that $0 \leq \ell_i \leq 1$ for all $i \in \{1, \ldots, n\}$.

Part C: Empirical Comparison

In practice, we sample rows from $\mathbf{A}$ and $\mathbf{b}$ to form a smaller sketched system $\mathbf{SAx} \approx \mathbf{Sb}$, where $\mathbf{S} \in \mathbb{R}^{m \times n}$ is a sampling matrix.

If we sample row i with probability p_i, we must scale the sampled row by $\frac{1}{\sqrt{mp_i}}$ to ensure our sketched matrix $\mathbf{SA}$ is an unbiased estimator of $\mathbf{A}^\top \mathbf{A}$ (i.e., $\mathbb{E}[(\mathbf{SA})^\top (\mathbf{SA})] = \mathbf{A}^\top \mathbf{A}$).

For uniform sampling, $p_i = \frac{1}{n}$. For leverage score sampling, $p_i = \frac{\ell_i}{d}$.

On a real linear regression problem (from e.g., `scikit-learn`), compare the error from the uniform sampled solution and the leverage score sampled solution. In particular, for 10 values of m in the range d to $10d$, compute the error of the approximate solution produced by each method. For an estimate $\widehat{\mathbf{x}}$, plot the relative error

$$\frac{\|\mathbf{A}\widehat{\mathbf{x}} - \mathbf{b}\|_2^2}{\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2}$$

by m.