EECS 127/227AT Optimization Models in Engineering UC Berkeley Spring 2023

Homework 8

This homework is due at 11 PM on March 16, 2023.

Submission Format: Your homework submission should consist of a single PDF file that contains all of your answers (any handwritten answers should be scanned).

1. **Convergence of Gradient Descent for Ridge Regression**

   Let $A \in \mathbb{R}^{m \times n}$, $\vec{y} \in \mathbb{R}^m$, and $\lambda > 0$. Consider a slight variation of the ridge regression problem where the least squares loss is normalized by the number of data points:

   $$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\}. \tag{1}$$

   In this problem, we will examine the behavior of gradient descent (GD) on this problem, and in particular the interplay between the learning rate $\eta > 0$ and regularization parameter $\lambda > 0$ in determining convergence.

   (a) Show that the unique solution to the problem in Equation (1) is

   $$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{2}$$

   (b) Show that the GD update

   $$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda \vec{x}_t\right) \tag{3}$$

   can be rearranged into the form

   $$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{4}$$

   Use this to show that

   $$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star). \tag{5}$$

   for every positive integer t.

   (c) We now discuss the insight that the SVD can give us regarding the convergence of GD. Let $A = U\Sigma V^\top$ be a full SVD of A. Let $\vec{z}_t = V^\top \vec{x}_t$ and $\vec{z}_\lambda^\star = V^\top \vec{x}_\lambda^\star$. Show that

   $$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star), \tag{6}$$

   and, moreover, show that for each $i \in \{1, \ldots, n\}$, we have

   $$(\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{7}$$

   where $\sigma_i\{A\}$ is the i^{th} largest singular value of A. This shows that the *rate of convergence* of $\vec{z}_t$ to $\vec{z}_\lambda^\star$ along the i^{th} component is influenced by the interaction between $\sigma_i\{A\}$, λ, and η, but critically no other singular values. Thus, one considers the V basis to be the "natural" basis for thinking about GD for ridge regression.

(d) Show that $\lim_{t\to\infty} \vec{z}_t = \vec{z}^\star_\lambda$ for all initializations $\vec{x}_0 = V\vec{z}_0$ if and only if

$$\max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{8}$$

Use this to show that GD converges for all initializations $\vec{x}_0$ if and only if

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{9}$$

where $\sigma_{\max}\{A\} = \sigma_1\{A\}$ is the largest singular value of A.

(e) **(OPTIONAL)** One way we can derive an "optimal" learning rate $\eta^\star$ is to minimize the largest rate of convergence:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\dots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|. \tag{10}$$

One important property of $\eta^\star$ is that it makes the rates of convergence $\left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right|$ associated with the largest and smallest singular values of A equal. Use this property to show that

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2 + \sigma_{\min}\{A\}^2 + 2\lambda m} \tag{11}$$

where $\sigma_{\min}\{A\} = \sigma_n\{A\}$ is the n^{th} largest singular value of A.

NOTE: There are several useful notions of optimal learning rate; this is just one of them.

(f) The attached notebook, `gd_convergence.ipynb`, will examine the computational aspects of GD on ridge regression. Implement the GD and stochastic gradient descent (SGD) functions at the top of the notebook, which are marked with `TODO`s.

(g) Click through the notebook and run the sections $n = 1$, $n = 2$, and $n \gg 2$. Change the values of λ and η and re-run the cells a few times. Write down your observations about how the convergence of GD works under different values of λ and η.

(h) In the sections $n = 1$, $n = 2$, and $n \gg 2$, change the calls to GD to instead call SGD. Write down your observations about how the convergence of SGD works under different values of λ and η. Compare the behavior of GD and SGD.

(i) You might have noticed that if we think of convergence in the "last iterate" sense, i.e., $\lim_{T\to\infty} \vec{x}_T = \vec{x}^\star_\lambda$, then *SGD rarely converges*. This is because even if we reach the global optimum, the gradient estimate used by SGD is in general nonzero, and so the iterates end up bouncing around near the optimum. Another different, weaker, notion of convergence under which one might show that SGD actually does converge is convergence "in time average", i.e., $\lim_{T\to\infty} \overline{\vec{x}}_T = \vec{x}^\star_\lambda$ where $\overline{\vec{x}}_T \doteq \frac{1}{T}\sum_{t=1}^{T} \vec{x}_t$. Visualize this by adding the argument `time_avg=True` to each plotting function; the plot will now visualize the sequence of $\overline{\vec{x}}_t$. Re-run the notebook. Write down your observations, especially regarding the stability of SGD and convergence in the last-iterate sense versus the time-average sense.

2. Homework Process

With whom did you work on this homework? List the names and SIDs of your group members.

NOTE: If you didn't work with anyone, you can put "none" as your answer.