EECS 127/227AT Optimization Models in Engineering UC Berkeley Spring 2024

Homework 8

This homework is due at 11 PM on March 15, 2024.

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. Quadratic inequalities

Consider the set S defined by the following inequalities:

$$(x_1 \geq -x_2 + 1 \text{ and } x_1 \leq 0) \text{ or } (x_1 \leq -x_2 + 1 \text{ and } x_1 \geq 0). \tag{1}$$

To be more precise,

$$S_1 = \{\vec{x} \in \mathbb{R}^2 \mid x_1 \geq -x_2 + 1, x_1 \leq 0\} \tag{2}$$

$$S_2 = \{\vec{x} \in \mathbb{R}^2 \mid x_1 \leq -x_2 + 1, x_1 \geq 0\} \tag{3}$$

$$S = S_1 \cup S_2. \tag{4}$$

(a) Draw the set S. Is it convex?

(b) Show that the set S, can be described as a single quadratic inequality of the form $q(\vec{x}) = \vec{x}^\top A\vec{x} + 2\vec{b}^\top \vec{x} + c \leq 0$, for matrix $A = A^\top \in \mathbb{R}^{2\times 2}$, $\vec{b} \in \mathbb{R}^2$ and $c \in \mathbb{R}$ i.e S can be written as $S = \{\vec{x} \in \mathbb{R}^2 \mid q(\vec{x}) \leq 0\}$). Find $A, \vec{b}, c$.
Hint: Can you combine the constraints to make one quadratic constraint?

(c) Recall the definition of the convex hull of a set $A \subseteq \mathbb{R}^n$ is the set of all convex combinations of points in A, i.e.,

$$\text{conv}(A) = \left\{ \sum_{i=1}^{k} \theta_i \vec{x}_i \;\middle|\; k \in \mathbb{N}, \theta_1, \ldots, \theta_k \geq 0, \sum_{i=1}^{k} \theta_i = 1, \vec{x}_1, \ldots, \vec{x}_k \in A \right\}. \tag{5}$$

What is the convex hull of S?

(d) We will now consider some convex optimization problems over S_1 that illustrate the role of the constraints in the optimization problem. For each of the following optimization problems find the optimal point, $\vec{x}^*$. Describe the constraints that are active in attaining the optimal value. *Hint: Suppose that there exists a point $\vec{x}$ such that $\nabla f(\vec{x}) = 0$. From the first order characterization of a convex function $\vec{x}$ would be an optimum value for f subject to no constraints. If $\vec{x}$ is not in the constraint set S_1, then the optimum point must be on the boundary of the set, i.e. it satisfies at least one of the constraints defining S_1 with equality.*

i. Minimize $f(\vec{x}) = (x_1 + 1)^2 + (x_2 - 3)^2$ subject to $\vec{x} \in S_1$.

ii. Minimize $f(\vec{x}) = (x_1 + 2)^2 + (x_2 - 2)^2$ subject to $\vec{x} \in S_1$.

iii. Minimize $f(\vec{x}) = x_1^2 + x_2^2$ subject to $\vec{x} \in S_1$.

2. Direction of Steepest Ascent

For a differentiable function $f : \mathbb{R}^n \mapsto \mathbb{R}$ we want to show that the gradient $\nabla f(\vec{x})$ is the direction of steepest ascent at the point $\vec{x}$.

(a) Let us define the rate of change of the function $f(\vec{x})$ at the point $\vec{x}$ along an arbitrary unit vector $\vec{u}$ as:

$$D_{\vec{u}} f(\vec{x}) = \lim_{h \to 0} \frac{f(\vec{x} + h\vec{u}) - f(\vec{x})}{h}. \tag{6}$$

We call this the directional derivative. Show that the directional derivative can be equivalently expressed as $D_{\vec{u}} f(\vec{x}) = \vec{u}^\top [\nabla f(\vec{x})]$.

HINT: Use Taylor approximation of the function around the point $\vec{x}$ and evaluate it at the point $\vec{x} + h\vec{u}$.

(b) Show that

$$\frac{\nabla f(\vec{x})}{\|\nabla f(\vec{x})\|_2} = \underset{\|\vec{u}\|_2 = 1}{\operatorname{argmax}} \vec{u}^\top [\nabla f(\vec{x})]. \tag{7}$$

3. 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:

$$\min_{\vec{x} \in \mathbb{R}^n} f_\lambda(\vec{x}) \qquad \text{where} \qquad f_\lambda(\vec{x}) \doteq \frac{1}{2}\left\{\frac{1}{m}\|A\vec{x} - \vec{y}\|_2^2 + \lambda \|\vec{x}\|_2^2\right\}. \tag{8}$$

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 (8) is

$$\vec{x}_\lambda^\star = \left(A^\top A + \lambda m I\right)^{-1} A^\top \vec{y}. \tag{9}$$

(b) Show that the GD update

$$\vec{x}_{t+1} = \vec{x}_t - \eta\left(\frac{1}{m}A^\top(A\vec{x}_t - \vec{y}) + \lambda \vec{x}_t\right) \tag{10}$$

can be rearranged into the form

$$\vec{x}_{t+1} - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)(\vec{x}_t - \vec{x}_\lambda^\star). \tag{11}$$

Use this to show that

$$\vec{x}_t - \vec{x}_\lambda^\star = \left(I - \eta\left(\frac{A^\top A}{m} + \lambda I\right)\right)^t (\vec{x}_0 - \vec{x}_\lambda^\star). \tag{12}$$

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

$$\vec{z}_t - \vec{z}_\lambda^\star = \left(I - \eta\left(\frac{\Sigma^\top \Sigma}{m} + \lambda I\right)\right)^t (\vec{z}_0 - \vec{z}_\lambda^\star), \tag{13}$$

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

$$(\vec{z}_t)_i - (\vec{z}_\lambda^\star)_i = \left(1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right)^t ((\vec{z}_0)_i - (\vec{z}_\lambda^\star)_i) \tag{14}$$

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 $\lim_{t\to\infty} \vec{z}_t = \vec{z}_\lambda^\star$ for all initializations $\vec{x}_0 = V\vec{z}_0$ if and only if

$$\max_{i \in \{1,\ldots,n\}} \left|1 - \eta\left(\frac{\sigma_i\{A\}^2}{m} + \lambda\right)\right| < 1. \tag{15}$$

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

$$\eta \in \left(0, \frac{2m}{\sigma_{\max}\{A\}^2 + \lambda m}\right) \tag{16}$$

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:

$$\eta^\star \in \operatorname*{argmin}_{\eta\in\mathbb{R}} \max_{i\in\{1,\dots,n\}} \left|1-\eta\left(\frac{\sigma_i\{A\}^2}{m}+\lambda\right)\right|. \tag{17}$$

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

$$\eta^\star = \frac{2m}{\sigma_{\max}\{A\}^2+\sigma_{\min}\{A\}^2+2\lambda m} \tag{18}$$

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.

4. Dual of the dual of a linear program

Consider a standard linear program P:

$$\min_{\vec{x}} \quad \vec{c}^\top \vec{x} \tag{19}$$

$$\text{s.t.} \quad A\vec{x} = \vec{b} \tag{20}$$

$$\vec{x} \geq \vec{0}. \tag{21}$$

where $\vec{x}, \vec{c} \in \mathbb{R}^n, \vec{b} \in \mathbb{R}^m, A \in \mathbb{R}^{m \times n}$. $\vec{x} \geq \vec{0}$ means $x_i \geq 0$ for all $i = 1, \ldots, n$.

(a) Formulate the Lagrangian of the problem P, and write the dual problem.

Note: The dual problem should not have the variable $\vec{x}$.

(b) Express the dual problem as an equivalent minimization problem. Find the dual of this minimization problem, i.e., the dual of the dual. Compare it to the original linear program formulation.

5. Minimizing a Sum of Logarithms

Consider the following problem, which arises in estimation of transition probabilities of a discrete-time Markov chain:

$$p^\star = \max_{\vec{x}\in\mathbb{R}^n} \quad \sum_{i=1}^n \alpha_i \log(x_i) \tag{22}$$

$$\text{s.t.} \quad \vec{x} \geq 0, \quad \vec{1}^\top \vec{x} = c, \tag{23}$$

where $c > 0$ and $\alpha_i > 0$, $i = 1, \ldots, n$. (Recall that if $\vec{x}$ is a vector then by "$\vec{x} \geq 0$" we mean "$x_i \geq 0$ for each i.") We will determine in closed-form a minimizer, and show that the optimal objective value of this problem is

$$p^\star = \alpha \log(c/\alpha) + \sum_{i=1}^n \alpha_i \log(\alpha_i), \tag{24}$$

where $\alpha \doteq \sum_{i=1}^n \alpha_i$. We will show this in a series of steps.

(a) First, express the problem as a minimization problem which has optimal value $p^\star_{\min}$.

(b) In optimization, we often "relax" problems of the form $p^\star_{\min} = \min_{\vec{x}\in\mathcal{X}} f_0(\vec{x})$, i.e., replacing the constraint set $\mathcal{X}$ with a larger constraint set $\mathcal{X}_r$, and instead solving $p^\star_r = \min_{\vec{x}\in\mathcal{X}_r} f_0(\vec{x})$, then showing a connection between $p^\star_{\min}$ and $p^\star_r$. In this problem, a particular relaxation we will use is to replace the equality constraint $\vec{1}^\top \vec{x} = c$ with an inequality constraint $\vec{1}^\top \vec{x} \leq c$.

Show that the relaxed problem has the same optimal value as the original problem, i.e., $p^\star_r = p^\star_{\min}$, and the two problems have the same solutions.

HINT: First argue that $p^\star_r \leq p^\star_{\min}$. Then, suppose for the sake of contradiction that $p^\star_r < p^\star_{\min}$. Let $\vec{x}^r$ be a solution to the relaxed minimization problem which has objective value $p^\star_r$. Consider the vector $\vec{x}$ given by

$$\vec{x} \doteq \begin{bmatrix} c - 1^\top \vec{x}^r + x^r_1 \\ x^r_2 \\ \vdots \\ x^r_n \end{bmatrix}. \tag{25}$$

Show that $\vec{x}$ is feasible for the original problem and has objective value $< p^\star_r$. Argue that this implies $p^\star_{\min} < p^\star_r$ and derive a contradiction. Finally, argue that any solution to the relaxed problem is a solution to the original problem, and vice-versa — you might need to use a construction similar to $\vec{x}$.

(c) After relaxing the equality constraint to an inequality constraint, form the Lagrangian $\mathcal{L}(\vec{x}, \vec{\lambda}, \mu)$ for the relaxed minimization problem, where λ_i is the dual variable corresponding to the inequality $x_i \geq 0$, and μ is the dual variable corresponding to the inequality constraint $\vec{1}^\top \vec{x} \leq c$.

(d) Now derive the dual function $g(\vec{\lambda}, \mu)$ for the relaxed minimization problem, and solve the dual problem $d^\star_r = \max_{\substack{\vec{\lambda}\geq\vec{0} \\ \mu\geq 0}} g(\vec{\lambda}, \mu)$. What are the optimal dual variables $\vec{\lambda}^\star, \mu^\star$?

(e) Show that strong duality holds for the relaxed problem, so $p^\star_r = d^\star_r$.

(f) From the $\vec{\lambda}^\star, \mu^\star$ obtained in the previous part, how do we obtain the optimal primal variable $\vec{x}^\star$? What is the optimal objective function value $p^\star_r$? Finally, what is $p^\star$?

6. 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.