IE 539 Convex Optimization Assignment 3

Fall 2024

Out: 30th October 2024

Due: 10th November 2024 at 11:59pm

Instructions

- Submit a PDF document with your solutions through the assignment portal on KLMS by the due date. Please ensure that your name and student ID are on the front page.
- Late assignments will be subject to a penalty. Special consideration should be applied for in this case.
- It is **required** that you typeset your solutions in LaTeX. Handwritten solutions will not be accepted.
- Spend some time ensuring your arguments are **coherent** and your solutions **clearly** communicate your ideas.

Question:	1	2	3	4	5	6	Total
Points:	20	10	20	15	15	20	100

1. Consider the binary classification problem with n data points $\{(x_i, y_i) : i = 1, \ldots, n\}$ where $x_i \in \mathbb{R}^d$ are features and $y_i \in \{-1, 1\}$ are labels. From these, we want to learn a separating hyperplane to classify new points $x \in \mathbb{R}^d$ as $+1$ or -1. Specifically, we want to find a hyperplane $w^\top x = b$ so that if $w^\top x \geq b$, then we classify x as $+1$, and if $w^\top x < b$, then we label x as -1.

   (a) (5 points) Given $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ that gives rise to a hyperplane, we define the "penalty" of (w, b) as the number of misclassifications among the training data set $\{(x_i, y_i) : i = 1, \ldots, n\}$. Explain that the penalty of (w, b) can be expressed as

   $$\sum_{i=1}^{n} \mathbf{1}\left(y_i \neq \operatorname{sign}\left(w^\top x_i - b\right)\right).$$

   (b) (5 points) We can find a hyperplane minimizing the penalty by solving the following optimization problem.

   $$\min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \sum_{i=1}^{n} \mathbf{1}\left(y_i \neq \operatorname{sign}\left(w^\top x_i - b\right)\right). \tag{1} \quad \boxed{\texttt{svm-1}}$$

   Explain that

   $$\min_{(w,b) \in \mathbb{R}^d \times \mathbb{R}} \sum_{i=1}^{n} \max\left\{0,\ 1 - y_i\left(w^\top x_i - b\right)\right\} \tag{2} \quad \boxed{\texttt{svm-2}}$$

   is an upper bound on the value of (1).

   (c) (10 points) Prove that the loss function

   $$\frac{1}{n} \sum_{i=1}^{n} \max\left\{0,\ 1 - y_i\left(w^\top x_i - b\right)\right\}$$

   is convex with respect to (w, b).

2. (10 points) The perceptron algorithm takes as input n data points $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in \mathbb{R}^d$ are features and $y_i \in \{-1, 1\}$ are labels. As in the previous question, we want to determine a hyperplane $w^\top x = 0$ that classifies the data points. Prove that the loss function

   $$\frac{1}{n} \sum_{i=1}^{n} \max\left\{-y_i(w^\top x_i), 0\right\}$$

   is convex in w.

3. (20 points) Prove that for a positive definite matrix A,

   $$f(x) = \frac{1}{2} x^\top A x + b^\top x + c$$

   is smooth and strongly convex in the ℓ_2-norm. Write down the smoothness constant and the strong convexity constant.

4. (15 points) In this question we prove the convergence of the projected subgradient method for functions that are strongly convex and Lipschitz continuous. Let $f : C \to \mathbb{R}$ be a function that is α-strongly convex with respect to the ℓ_2 norm and L-Lipschitz continuous in the ℓ_2 norm over a convex domain C. Recall that the projected subgradient method proceeds as follows.

   - Choose $x_1 \in C$.
   - For $t = 1, 2, 3, \ldots, T - 1$:
     - Select any subgradient $g_t \in \partial f(x_t)$ and step size $\eta_t > 0$.
     - Compute $x_{t+1} = \operatorname{Proj}_C\{x_t - \eta_t g_t\}$.

   (a) Set $\eta_t = \frac{2}{\alpha(t+1)}$. Show that

   $$f\left(\sum_{t=1}^{T} \frac{2t}{T(T+1)} x_t\right) - f(x^*) \leq \frac{2L^2}{\alpha(T+1)}$$

   where $x^* \in \arg\min_{x \in C} f(x)$.

(b) Set $\eta_t = \frac{1}{\alpha t}$. Show that

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{L^2(1+\log T)}{2\alpha T}$$

where $x^* \in \arg\min_{x\in C} f(x)$.

5. (15 points) In this question we prove the convergence of stochastic gradient descent for functions that are strongly convex and Lipschitz continuous. Let $f : C \to \mathbb{R}$ be a function that is α-strongly convex with respect to the ℓ_2 norm and L-Lipschitz continuous in the ℓ_2 norm over a convex domain C. Recall that stochastic gradient descent proceeds as follows.

   - Choose $x_1 \in C$.
   - For $t = 1, 2, 3, \ldots, T-1$:
     - Obtain an unbiased estimator $\hat{g}_{x_t}$ of some $g \in \partial f(x_t)$.
     - Update $x_{t+1} = \mathrm{Proj}_C\left\{x_t - \eta_t \hat{g}_{x_t}\right\}$ for a step size $\eta_t > 0$.

   Set $\eta_t = \frac{1}{\alpha t}$. Assuming $\|\hat{g}_{x_t}\|_2 \leq L$ for all t, show that

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) \leq \frac{L^2(1+\log T)}{2\alpha T}$$

   where $x^* \in \arg\min_{x\in C} f(x)$.

6. (20 points) In this question, we consider a basic version of mini-batch SGD. At each point x taken by SGD, we sample unbiased estimators $\hat{g}_x^1, \ldots, \hat{g}_x^B$ of a subgradient $g_x \in \partial f(x)$ independently at random. Assume that

$$\|g_x\|_2 \leq L \quad \text{for all } g_x \in \partial f(x)$$

   and that

$$\mathbb{E}\left[\|\hat{g}_x^i - g_x\|_2^2 \mid x\right] \leq \sigma^2.$$

   Then mini-batch SGD uses

$$\hat{g}_x = \frac{1}{B}\left(\hat{g}_x^1 + \cdots + \hat{g}_x^B\right)$$

   as an unbiased estimator of g_x. Prove that mini-batch SGD for $\min_{x\in\mathbb{R}^d} f(x)$ with step size $\eta = 1/\sqrt{T}$ guarantees that

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)\right] - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\sqrt{T}} + \frac{1}{2\sqrt{T}}\left(L^2 + \frac{1}{B}\sigma^2\right)$$

   where x^* is an optimal solution to $\min_{x\in\mathbb{R}^d} f(x)$.