Coursework (4) for *Introductory Lectures on Optimization*

Zhao Hongjue
3190104515

Nov. 17, 2022

Excercise 1. Prove the following results. For proximal point method, if f is closed and convex and optimal value f^* is finite and attained at $\boldsymbol{x}_*$. We have

$$f(\boldsymbol{x}_k) - f^* \le \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2}{2\sum_{i=0}^{k-1} t_i}, \quad \text{for } k \ge 1.$$

Proof of Excercise 1:

Lemma 1. $\boldsymbol{u} = \operatorname{prox}_h(\boldsymbol{x})$ is equivalent to the following

1. $\boldsymbol{x} - \boldsymbol{u} \in \partial h(\boldsymbol{u})$,
2. $h(\boldsymbol{z}) \ge h(\boldsymbol{u}) + (\boldsymbol{x} - \boldsymbol{u})^\top (\boldsymbol{z} - \boldsymbol{u})$ for all $\boldsymbol{z}$.

The updating rule of proximal point method can be concluded as

$$\boldsymbol{x}' = \operatorname{prox}_{tf}(\boldsymbol{x}).$$

According to Lemma. 1, we can obtain

$$f(\boldsymbol{z}) \ge f(\boldsymbol{x}') + \frac{1}{t}(\boldsymbol{x} - \boldsymbol{x}')^\top (\boldsymbol{z} - \boldsymbol{x}')$$

Let $\boldsymbol{z} = \boldsymbol{x}$, we can obtain

$$f(\boldsymbol{x}') \le f(\boldsymbol{x}) - \frac{1}{t}\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2.$$

Therefore, this algorithm is a descent method. Let $\boldsymbol{z} = \boldsymbol{x}_*$, we can get

$$\begin{aligned}
f(\boldsymbol{x}') - f(\boldsymbol{x}_*) &\le \frac{1}{t}(\boldsymbol{x} - \boldsymbol{x}')^\top (\boldsymbol{x}' - \boldsymbol{x}_*) \\
&= \frac{1}{t}[(\boldsymbol{x} - \boldsymbol{x}_*) - (\boldsymbol{x}' - \boldsymbol{x}_*)]^\top (\boldsymbol{x}' - \boldsymbol{x}_*) \\
&= \frac{1}{t}\left[(\boldsymbol{x} - \boldsymbol{x}_*)^\top (\boldsymbol{x}' - \boldsymbol{x}_*) - \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2^2\right] \\
&\le \frac{1}{t}\left[\|\boldsymbol{x} - \boldsymbol{x}_*\|_2 \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2 - \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2^2\right] \\
&\le \frac{1}{t}\left[\frac{\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2^2}{2} - \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2^2\right] \\
&= \frac{1}{2t}\left(\|\boldsymbol{x} - \boldsymbol{x}_*\|_2^2 - \|\boldsymbol{x}' - \boldsymbol{x}_*\|_2^2\right)
\end{aligned}$$

Let $t = t_i$, $\boldsymbol{x} = \boldsymbol{x}_i$, $\boldsymbol{x}' = \boldsymbol{x}_{i+1}$. For $i = 0, \ldots, k-1$

$$t_i(f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_*)) \leq \frac{1}{2}\left(\|\boldsymbol{x}_i - \boldsymbol{x}_*\|_2^2 + \|\boldsymbol{x}_{i+1} - \boldsymbol{x}_*\|_2^2\right).$$

Adding inequalities from $i = 0$ to $k-1$ gives

$$\left(\sum_{i=0}^{k-1} t_i\right)(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_*)) \leq \sum_{i=0}^{k-1} t_i(f(\boldsymbol{x}_{i+1}) - f(\boldsymbol{x}_*)) \leq \frac{1}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2.$$

Thus we have $f(\boldsymbol{x}_k) - f^* \leq (\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2^2)/(2\sum_{i=0}^{k-1} t_i)$ for $k \geq 1$. $\square$

Excercise 2. Derive the the dual problem of hard margin SVM.

Solution of Excercise 2: First we derive the primal optimization problem of hard margin SVM. For hard margin SVM, it assumes the existence of a hyperplane that perfectly separates the training sample into two populations of positively and negatively labeled points.

However, there are then infinitely many such separating hyperplanes. Consider the training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^m$ for $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^N$ and $y_i \in \mathcal{Y} = \{+1, -1\}$, we define the geometric margin of linear classifier $h : \boldsymbol{x} \mapsto \boldsymbol{w}^\top \boldsymbol{x} + b$ at $\boldsymbol{x}$ as follows:

$$\rho_h(\boldsymbol{x}) = \frac{\left|\boldsymbol{w}^\top \boldsymbol{x} + b\right|}{\|\boldsymbol{w}\|_2}.$$

Then we select hyperplane which can maximize the margin ρ:

$$\rho = \max_{\boldsymbol{w}, b} \min_{i \in [m]} \frac{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|_2}.$$

Observe that the last expression is invariant to multiplication $(\boldsymbol{w}, b)$ by a positive scalar. Thus we restrict ourselves to pairs $(\boldsymbol{w}, b)$ scaled such that $\min_{i \in [m]} y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1$:

$$\rho = \max_{\boldsymbol{w}, b} \frac{1}{\|\boldsymbol{w}\|_2}.$$

Since maximizing $1/\|\boldsymbol{w}\|_2$ is equivalent to minimizing $\frac{1}{2}\|\boldsymbol{w}\|_2$, the primal optimization problem can be concluded as

$$\begin{aligned} &\min_{\boldsymbol{w}, b} \ \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{s.t. } y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) \geq 1, \ i = 1, \ldots, m \end{aligned} \tag{1}$$

The Lagrangian of problem. 1 can then be defined for all $\boldsymbol{w} \in \mathbb{R}^N$, $b \in \mathbb{R}$, and $\boldsymbol{\alpha} \in \mathbb{R}_+^m$, by

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^m \alpha_i[y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) - 1]. \tag{2}$$

The KKT conditions are obtained by setting the gradient of the Lagrangian with respect to the primal variables w and b to zero and by writing the complementarity conditions:

$$\nabla_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i = 0 \quad \Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i \tag{3}$$

$$\nabla_b L = -\sum_{i=1}^m \alpha_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^m \alpha_i y_i = 0 \tag{4}$$

$$\forall i,\ \alpha_i[y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) - 1] = 0 \quad \Longrightarrow \quad \alpha_i = 0 \vee y_i(\boldsymbol{w}^\top \boldsymbol{x} + b) = 1 \tag{5}$$

To derive the dual form of the constrained optimization problem 1, we plug into the Lagrangian the definition of $\boldsymbol{w}$ in terms of the dual variables as expressed in (3) and apply the constraint (4). This yields

$$\begin{aligned} L &= \frac{1}{2}\left\|\sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i\right\|^2 - \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^\top \boldsymbol{x}_j) - \sum_{i=1}^{m} \alpha_i y_i b + \sum_{i=1}^{m} \alpha_i \\ &= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^\top \boldsymbol{x}_j). \end{aligned}$$

This leads to the following dual optimization problem for hard margin SVM:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i^\top \boldsymbol{x}_j) \\ &\text{s.t.} \alpha_i \geq 0 \wedge \sum_{i=1}^{m} \alpha_i y_i = 0,\ i = 1, \ldots, m. \end{aligned}$$

Obviously, the objective function is infinitely differentiable. Since the constraints are affine and convex, this dual problem is a convex optimization problem. According to the KKT conditions, we can obtain the solution of the primal optimization problem. □

Excercise 3. For KL divergence defined on the probability simplex, prove that the upper bound of $\triangle_\psi(x^*, x_1)$ is $\log n$, for $x_1 = [\frac{1}{n}, \ldots, \frac{1}{n}]$.

Proof of Excercise 3:

$$\begin{aligned} \Delta_\psi(\boldsymbol{x}^*, \boldsymbol{x}_1) &= \sum_{i=1}^{n} x_i^* \log \frac{x_i^*}{x_{1i}} \\ &= \sum_{i=1}^{n} x_i^* \log(n x_i^*) \\ &= \sum_{i=1}^{n} x_i^* (\log n + \log x_i^*) \\ &= \log n \underbrace{\sum_{i=1}^{n} x_i^*}_{=1} + \underbrace{\sum_{i=1}^{n} x_i^* \log x_i^*}_{\leq 0} \\ &\leq \log n. \end{aligned}$$

□