SI251 - Convex Optimization homework 2

Deadline: 2022-11-16 23:59:59

1. You can use Word, Latex or handwriting to complete this assignment. If you want to submit a handwritten version, scan it clearly.
2. The **report** has to be submitted as a PDF file to Gradescope, other formats are not accepted.
3. You have to write your assignment in English, otherwise you will get a 50% penalty of your score.
4. The submitted file name is **student_id+your_student_name.pdf**.
5. Late policy: You have 4 free late days for the quarter and may use up to 2 late days per assignment with no penalty. Once you have exhausted your free late days, we will deduct a late penalty of 25% per additional late day. Note: The timeout period is recorded in days, even if you delay for 1 minute, it will still be counted as a 1 late day.
6. You are required to follow ShanghaiTech's academic honesty policies. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious sanctions.

Any plagiarism will get Zero point.

1. (20 pts) **Subgradient Methods**. Consider the following problem

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m, \end{array} \tag{1}$$

where $f, g_1, g_2, \ldots, g_m : \mathbb{E} \to \mathbb{R}$ are real-valued convex functions. Please prove the following problems.

(1) Let x^* be an optimal solution of (1), and assume that there exists $\bar{x}$, that satisfies $g_i(\bar{x}) < 0$ for all $g_i(x)$. Then there exist $\lambda_1, \ldots, \lambda_m \geq 0$ for which

$$\mathbf{0} \in \partial f\left(\mathbf{x}^*\right) + \sum_{i=1}^{m} \lambda_i \partial g_i\left(\mathbf{x}^*\right) \tag{2}$$

$$\lambda_i g_i\left(\mathbf{x}^*\right) = 0, \quad i = 1, 2, \ldots, m. \tag{3}$$

(2) If $\mathrm{x}^* \in \mathbb{E}$ satisfies conditions (2) and (3) for some $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$, then it is an optimal solution of problem (1).

2. (30 pts) **L-smooth functions**. Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable. Please prove that the following relations holds for all $x, y \in \mathbb{R}$ if f with an L-smooth conditions,

$$[1] \Rightarrow [2] \Rightarrow [3] \Rightarrow [4]$$

[1] $\langle \nabla f(x) - \nabla f(y), x - y \rangle \leq L\|x - y\|^2$,

[2] $f(y) \leq f(x) + \nabla f(x)^T (y - x) + \dfrac{L}{2}\|y - x\|^2$,

[3] $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \dfrac{1}{2L}\|\nabla f(y) - \nabla(x)\|^2$,

[4] $f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha) f(y) - \dfrac{\alpha(1 - \alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \quad \alpha \in [0, 1]$.

3. **Mirror Descent Methods**.

(1) (10 pts) Let φ be proper convex and differentiable. Suppose $\mathbf{y} = \nabla\varphi(\mathbf{x})$, from conjugate subgradient theorem, show that

$$\varphi(\mathbf{x}) + \varphi^*(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle,$$

where $\varphi^*(\mathbf{y})$ is Fenchel conjugate of φ.

Hint: The conjugate subgradient theorem states that if f is closed proper convex, then the following statements are equivalent for a pair of vectors $(\mathbf{x}, \mathbf{y})$: (i) $\langle \mathbf{x}, \mathbf{y} \rangle = f(\mathbf{x}) + f^*(\mathbf{y})$; (ii) $\mathbf{y} \in \partial f(\mathbf{x})$; (iii) $\mathbf{x} \in \partial f^*(\mathbf{y})$, where $f^*(\mathbf{y})$ is Fenchel conjugate of f.

(2) (10 pts) Using Bregman divergence, show that mirror descent has an alternative form, which reads

$$\mathbf{x}^{t+1} = \nabla\varphi^*(\nabla\varphi(\mathbf{x}^t) - \eta_t \mathbf{g}^t),$$

where $\varphi^*(\mathbf{x})$ is Fenchel conjugate of φ, $\mathbf{g}^t \in \partial f(\mathbf{x}^t)$ and $\eta_t > 0$ is the stepsize. (For simplicity, assume the constraints set $\mathcal{C} = \mathbb{R}^n$.)

4. **Natural gradient descent**. Natural gradient descent is an optimization method motivated by the difference between the parameter space (e.g., parameters of a neural network) and the distribution space of the output (e.g., categorical distribution given by a classifier), and performs well for many applications as an alternative to stochastic gradient descent. Now consider a simple classification problem as below.

$$\min_{\theta} \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{L}(p(\cdot|x;\theta), y_{true})\right] \tag{4}$$

where x and y_{true} are the feature and label of the sample respectively, ρ_{data} is the distribution of the samples, $\mathcal{L}$ is some loss function, e.g., cross-entropy loss, and $p(\cdot|x;\theta)$ is the classifier that we want to obtain, which outputs the probabilities that the sample x belongs to class ys.

For some reason, we cannot optimize the classification problem above directly. In such case, we will often find a surrogate function to approximate the original problem in each iteration.

However, The surrogate function is just an approximation to the original problem, and there will be a large bias when the updated classifier is far from itself before. Hence, we need to restrict the update within a trust region.

That is what natural gradient descent does. Mathematically, the formulation of it is

$$\begin{aligned} \min_{\theta} \quad & \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{S}_{\theta_{old}}(p(\cdot|x;\theta), y_{true})\right] \\ \text{subject to} \quad & \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \leq \delta \end{aligned} \tag{5}$$

where θ_{old} is the parameters of the classifier $f(x;\theta)$ before the update, $\mathcal{S}_{\theta_{old}}$ is the surrogate function of $\mathcal{L}$ at the point θ_{old}, and the averaged KL divergence

$$\mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \leq \delta \tag{6}$$

is to restrict the update to the trust region of θ_{old}.

(Hint: $\mathrm{kl}(p,q) = \sum_i p_i \log \frac{p_i}{q_i}$)

The story ends.

(1) (10 pts) Please prove that, when $\theta = \theta_{old}$, $\nabla_\theta \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] = 0$, which is known as score function in probability and statistics.

(2) (20 pts) To perform natural gradient descent, we should do an approximation as below on the objective and constraint further.

$$\begin{aligned} \min_{\theta} \quad & g^T\theta \\ \text{subject to} \quad & \frac{1}{2}(\theta-\theta_{old})^T H(\theta-\theta_{old}) \leq \delta \end{aligned} \tag{7}$$

where $g = \nabla_\theta \mathbb{E}_{x,y_{true}\sim\rho_{data}} \left[\mathcal{S}_{\theta_{old}}(p(\cdot|x;\theta), y_{true})\right], H = \nabla^2_\theta \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right]$. As we have proved above, there is no need to consider the derivative of the constraint.

Now that please prove that the update formula of natural gradient descent is

$$\theta = \theta_{old} - \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g \tag{8}$$

(3) (Bonus 20 pts) In practice, $p(\cdot|x;\theta)$ is usually a complicated model such as neural network. Hence the hessian matrix H of the constraint is hard to compute.

However, the hessian matrix H here is also known as Fisher Information Matrix in the probability and statics, which holds an amazing equality that we can infer H directly just computing the gradient of the log likelihood.

Below is that equality. Please prove it.

$$\begin{aligned} &\nabla_\theta^2 \mathbb{E}_{x\sim\rho_{data}} \left[\mathrm{kl}\left(p(\cdot|x;\theta_{old}), p(\cdot|x;\theta)\right)\right] \\ &= \mathbb{E}_{x\sim\rho_{data}} \left[\mathbb{E}_{y\sim p(\cdot|x;\theta_{old})} \left[\nabla_\theta \log\left(p(y|x;\theta)\right)^T \nabla_\theta \log\left(p(y|x;\theta)\right)\right]\right] \end{aligned} \tag{9}$$