Yen-Huan Li · Follow
Last edited by Yen-Huan Li on Apr 20, 2025
Contributed by

CSIE5410 Optimization Algorithms: HW1

Yen-Huan Li @ CSIE, National Taiwan University

> There are a total of 115 points. However, your grades will be capped at 100.
>
> If you use any result that is not included in the lecture slides, please ensure that the TAs do not need to spend time verifying its correctness. Otherwise, you will lose points.

1. Find the book *Lectures on Convex Optimization* by Yurii Nesterov, which can be freely downloaded on Springer's website using NTU's VPN service. **For each of the following two optimization problems, find a *first-order optimization algorithm* in the book that can solve it, specify the parameters in the algorithm, and state the algorithm's optimiztion error guarantee.**

   You will get partial points if the performance of the algorithm you choose can be dominated by other algorithms in the book.

   1. (10 points) Let $(x_1, y_1), \dots (x_n, y_n) \in \mathbb{R}^d \times \{\pm 1\}$. Suppose that $n < d$. Consider the optimization problem

      $$w_\star \in \operatorname*{argmin}_{w \in \mathbb{R}^d} f(w), \quad f(w) := \frac{1}{n} \sum_{i=1}^n \log \left(1 + \mathrm{e}^{-y_i \langle x_i, w \rangle} \right).$$

   2. (10 points) Let $\lambda > 0$. Let f be defined as above. Consider the slightly modified optimization problem

      $$w_{\star, \mu} \in \operatorname*{argmin}_{w \in \mathbb{R}^d} f(w) + \frac{\lambda}{2} \| w \|_2^2 .$$

2. Let $\varphi : \mathbb{R}^d \to]-\infty, \infty]$ be a function twice differentiable on its domain.

   1. (10 points) We say that the function φ is *logarithmically homogeneous* if

      $$\varphi(\gamma x) = \varphi(x) - \log \gamma, \quad \forall x \in \mathbb{R}^d, \gamma > 0.$$

      Show that if the function φ is logarithmically homogeneous, then the following equalities hold for all $x \in \operatorname{dom} \varphi$:

      $$\begin{aligned} &\langle \nabla \varphi(x), x \rangle = -1, \\ &\nabla \varphi(x) = -\nabla^2 \varphi(x) x, \\ &\langle x, \nabla^2 \varphi(x) x \rangle = 1. \end{aligned}$$

   2. (10 points) We say the function φ is strictly convex if

      $$\nabla^2 \varphi(x) > 0, \quad \forall x \in \operatorname{dom} \varphi.$$

      Show that if the function φ is strictly convex and logarithmically homogeneous, then it holds that

      $$\nabla^2 \varphi(x) \geq \nabla \varphi(x) (\nabla \varphi(x))^\top, \quad \forall x \in \operatorname{dom} \varphi,$$

      where $(\nabla \varphi(x))^\top$ denotes the transpose of $\nabla \varphi(x)$.

   3. (10 points) **Show that the following three statements are equivalent:**

      - The function $\mathrm{e}^{-\varphi}$ is concave.

      - The function φ satisfies

        $$\varphi(y) \geq \varphi(x) - \log \left(1 - \langle \nabla \varphi(x), y - x \rangle \right), \quad \forall x \in \operatorname{dom} \varphi.$$

      - The function φ satisfies

        $$\nabla^2 \varphi(x) \geq \nabla \varphi(x) (\nabla \varphi(x))^\top, \quad \forall x, y \in \operatorname{dom} \varphi.$$

3. Alice wants to make money through long-term investment in the stock market. Suppose there are d stocks. On the t-th day, Alice distributes her total wealth to the d stocks following the ratios specified in a weight vector $x_t \in \Delta$, where Δ denotes the probability simplex

   $$\left\{ x = (x[1], \dots, x[d]) \in \mathbb{R}_+^d : \sum_{i=1}^d x[i] = 1 \right\},$$

   where $x[i]$ denotes the i-th entry of the vector x; then, Alice observes the price relative–the ratio of the closing price to the opening price–of each stock. Let us list the price relatives in a vector $a_t \in \mathbb{R}_+^d$; the entry $a_t[i]$ denotes the price relative of the i-th stock on the t-th day.

   Suppose that the price relatives $a_1, \dots, a_T$ are independent and identically distributed random vectors following a common probability distribution P known to Alice. Consider the following strategy for Alice:

   $$x_t \in \operatorname*{argmin}_{x \in \Delta} f(x); \quad f(x) := \mathsf{E} \left[-\log \langle a_t, x \rangle \right], \quad \forall t \in \mathbb{N}.$$

   Obviously, this strategy suggests maintaining a *constant rebalanced portfolio* over time.

   Assume that the probability distribution of a_t ensures that the function f is strictly convex. Also, for simplicity, assume that the random vector a_t has a finite support throughout this problem series.

   1. (5 points) Suppose that Alice has 1 unit of wealth before the first day. **Show that after T days, her wealth is given by**

      $$W_T^\star := \langle a_1, x_1 \rangle \cdots \langle a_T, x_T \rangle .$$

   2. (10 points) **Show that Alice's strategy competes with any constant rebalanced portfolio in the sense that**

      $$\lim_{T \to \infty} \mathsf{E} \left[\frac{W_T(x)}{W_T^\star} \right] \leq 1, \quad \forall x \in \Delta,$$

      where the expectation is with respect to the price realtives. Here, $W_T(x)$ denotes the total wealth yielded by the constant rebalanced portfolio specified by $x \in \Delta$ after T days, i.e.,

      $$W_T(x) = \langle a_1, x \rangle \cdots \langle a_T, x \rangle .$$

   3. (10 points) Consider the optimization problem that defines x_t. Supppose that $a_1 \in [\alpha, 1]^d$ with probability 1, for some $0 < \alpha < 1$. **Find the paper "Logarithmic regret algorithms for online convex optimization" by Elad Hazan et al., and show how it provides an algorithm for computing x_t using only the first-order oracle of the function f.** Please write down the algorithm, its parameters (e.g., step size), and its optimization error bounds explicitly.

4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function that may be *non-convex*. Suppose that ∇f is L-Lipschitz with respect to some norm $\| \cdot \|$ on $\mathbb{R}^d$, which is not necessarily the ℓ_2-norm, and $L > 0$; that is,

   $$\| \nabla f(y) - \nabla f(x) \|_* \leq L \| y - x \|, \quad \forall x, y \in \mathbb{R}^d,$$

   where $\| \cdot \|_*$ denotes the dual norm, i.e.,

   $$\| u \|_* := \max_{x \in \mathbb{R}^d, \| x \| \leq 1} \langle u, x \rangle .$$

   We say that a point x is ε-stationary for some $\varepsilon > 0$ if $\| \nabla f(x) \|_* \leq \varepsilon$. Here and for the remainder of this problem, the inner product is defined with respect to the ℓ_2-norm and is not necessarily consistent with the general norm. Note that a general norm, such as the ℓ_1- or ℓ_∞-norm, may not correspond to any inner product.

   1. (10 points) **Show that**

      $$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \| y - x \|^2, \quad \forall x, y \in \mathbb{R}^d .$$

   2. (10 points) Consider the algorithm below:

      - Let $x_1 \in \mathbb{R}^d$.
      - For every $t \in \mathbb{N}$, compute

        $$x_{t+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2} \| x - x_t \|^2 .$$

      Notice that this algorithm can be viewed as a generalization of gradient descent under general norms. **Show that**

      $$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} \| \nabla f(x_t) \|_*^2, \quad \forall t \in \mathbb{N}.$$

   3. (10 points) **Show that the algorithm satisfies**

      $$\min_{1 \leq \tau \leq t} \| \nabla f(x_\tau) \|_*^2 \leq \frac{2L \left[f(x_1) - f(x_{t+1}) \right]}{t}, \quad \forall t \in \mathbb{N}.$$

   4. (10 points) Consider the following algorithm:

      - Let $x_1 \in \mathbb{R}^d$.
      - For every $t \in \mathbb{N}$, compute

        $$x_{t+1} = x_t - \frac{\| \nabla f(x_t) \|_1}{L} \operatorname{sign}(\nabla f(x_t)),$$

        where we define

        $$\operatorname{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0; \\ -1 & \text{otherwise}, \end{cases}$$

        for any $x \in \mathbb{R}$ and

        $$\operatorname{sign}(v) = (\operatorname{sign}(v[1]), \dots, \operatorname{sign}(v[d])),$$

        for any $v = (v[1], \dots, v[d]) \in \mathbb{R}^d$.
        Interpret the algorithm using the results above.

Yen-Huan Li · Follow
Associate Professor at the Computer Science Dept., National Taiwan University. Machine learning, convex opt., high-dim. stat., quantum information.