CSIE5410 Optimization Algorithms: HW0

Yen-Huan Li @ CSIE, National Taiwan University

> There are a total of 12 problems, each worth 10 points. However, your grades will be capped at 100.

We say a set $\mathcal{X} \subseteq \mathbb{R}^d$ is *convex* if

$$x, y \in \mathcal{X} \rightarrow \alpha x + (1-\alpha) y \in \mathcal{X}, \quad \forall \alpha \in [0,1].$$

We say a differentiable function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ is *convex* on a set $\mathcal{X} \subseteq \mathbb{R}^d$ if

$$\varphi(y) \geq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle,$$

for any $x, y \in \mathcal{X}^d$ such that $\varphi(x) < +\infty$ and $\nabla \varphi(x)$ exists, where $\nabla \varphi(x)$ denotes the gradient of φ at x and $\langle \cdot, \cdot \rangle$ denotes the dot product. We say an optimization problem of the form

$$\min_{x \in \mathcal{X}} \varphi(x) \quad \text{or} \quad \operatorname*{argmin}_{x \in \mathcal{X}} \varphi(x)$$

is *convex* if both the *objective function* φ and the *constraint set* $\mathcal{X}$ are convex. The first characterizes the minimum value of φ on $\mathcal{X}$, while the second characterizes the set of minimizers of φ on $\mathcal{X}$.

1. Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be the *data points*. We seek a parameter $w^\natural$ such that y_i is close to $\langle x_i, w^\natural \rangle$ for all $1 \leq i \leq n$. The *least squares approach* yields:

   $$w^\natural \in \operatorname*{argmin}_{w \in \mathbb{R}^d} g(w), \quad g(w) := \frac{1}{2n} \sum_{i=1}^n (y_i - \langle x_i, w \rangle)^2,$$

   where the factor $1/(2n)$ is introduced for convenience. Suppose that $n > d$ and that the matrix $X \in \mathbb{R}^{n \times d}$, whose i-th row is given by the transpose of x_i, has full column rank.

   1. (10 points) Show that the optimization problem defining $w^\natural$ is convex.
   2. (10 points) Consider solving the optimization problem by Newton's method, which is given by:

   - Let $w_1 \in \mathbb{R}^d$.
   - For each $t \in \mathbb{N}$, compute

     $$w_{t+1} = w_t - \left(\nabla^2 g(w_t) \right)^{-1} \nabla g(w_t),$$

     where $\nabla^2 g(w)$ denotes the Hessian matrix of g at w.
     Show that if we choose $w_1 = 0$, the all-zeros vector, then Newton's method solves the optimization problem in one iteration (i.e., $w_2 = w^\natural$).
     Hint. You can freely cite the results in the Wikipedia page on the Moore-Penrose inverse.

2. Let $x_1, \ldots, x_n \in \mathbb{R}^p$. Let $y_1, \ldots, y_n$ be independent $\{0,1\}$-valued random variables satisfying

   $$\mathsf{P}(y_i = 1) = 1 - \mathsf{P}(y_i = 0) = \frac{1}{1 + \mathrm{e}^{-\langle x_i, \theta^\natural \rangle}},$$

   for some unknown parameter $\theta^\natural \in \mathbb{R}^p$.

   1. (10 points) Given $(x_1, y_1), \ldots, (x_n, y_n)$, show that the *maximum-likelihood estimator* for $\theta^\natural$ is given by

      $$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} L(\theta), \quad L(\theta) := \frac{1}{n} \sum_{i=1}^n \log \left(1 + \mathrm{e}^{-2(y_i - 1/2)\langle x_i, \theta \rangle} \right).$$

      Google and find the definition of maximum-likelihood estimator if you are unfamiliar with this concept.
   2. (10 points) Show that the optimization problem defining the maximum-likelihood estimator is convex.
   3. (10 points) Let $X \in \mathbb{R}^{n \times p}$ whose i-th row is given by the transpose of x_i. Let y be an n-dimensional random vector whose i-th entry is y_i. Show that

      $$\nabla L(\theta^\natural) = -\frac{1}{n} X^\top (y - \mathsf{E}[y]),$$

      where $\mathsf{E}[y]$ denotes the expectation of y.
   4. (10 points) Let X be as defined in the previous problem. Show that

      $$\nabla^2 L(\theta^\natural) = \frac{1}{n} X^\top D X,$$

      where $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose (i,i)-th entry is given by the variance of y_i and $X^\top$ denotes the transpose of the matrix X.
   5. (10 points) For any matrices A and B, we write $A \leq B$ if the matrix $B - A$ is positive semi-definite. Show that

      $$0 \leq \nabla^2 L(\theta) \leq \frac{\lambda_{\max}(X^\top X)}{4n} I, \quad \forall \theta \in \mathbb{R}^p,$$

      where 0 denotes the all-zeros matrix, $\lambda_{\max}(X^\top X)$ the largest eigenvalue of the matrix $X^\top X$, and I the identity matrix.

3. The *growth-optimal portfolio* in mathematical finance is formulated as

   $$x_\star \in \operatorname*{argmin}_{x \in \Delta_d} f(x), \quad f(x) := -\sum_{i=1}^n w_i \log \langle a_i, x \rangle, \tag{P}$$

   where Δ_d denotes the probability simplex in $\mathbb{R}^d$, i.e.,

   $$\Delta_d := \left\{ x = (x[1], \ldots, x[d]) \in \mathbb{R}^d \,\middle|\, x[i] \geq 0, \sum_{i=1}^d x[i] = 1 \right\},$$

   $\log$ denotes the natural log, and w_i are given non-negative numbers satisfying $\sum_i w_i = 1$ and entrywise-nonnegative vectors $a_i \in \mathbb{R}^d$ such that $a_i \neq 0$. Without loss of generality, let us assume that $a_i \in [0,1]^d$ for all i.

   For convenience, let us write the optimization problem (P) equivalently as:

   $$\begin{aligned} & u_\star \in \operatorname*{argmin}_{u \in \mathcal{U}} \tilde{f}(u), && \text{(P')} \\ & \tilde{f}(u) := -\sum_{i=1}^n w_i \log \left\langle a_i, \left(u[1], \ldots, u[d-1], 1 - \sum_{i=1}^{d-1} u[i] \right) \right\rangle, \\ & \mathcal{U} := \left\{ (u[1], \ldots, u[d-1]) \in \mathbb{R}^{d-1} \,\middle|\, u[i] \geq 0 \ \forall i, \sum_{i=1}^{d-1} u[i] \leq 1 \right\}. \end{aligned}$$

   Find the book draft "Techniques in Optimization and Sampling" at https://github.com/YinTat/optimizationbook/. Consider solving the optimization problem ~~(P)~~ (P') using Algorithm 5 in the book draft, which we will refer to as the *ellipsoid method*.

   1. (10 points) Show that the optimization problem ~~(P)~~ (P') is convex.
   2. (10 points) Show how the mapping $g(\cdot)$ in the ellipsoid method can be chosen for solving (P') .
   3. (10 points) In the proof of Lemma 3.3 in the book draft, it is claimed that *"we can do a transformation so that $A^{(k)} = I$, $x^{(k)} = 0$ and $v(x^{(k)}) = e_1$."* ~~Please specify the associated affine transformation.~~ Please find an affine transformation such that "$A^{(k)} = I$, $x^{(k)} = 0$"; then, argue that it suffices to consider the case where "$v(x^{(k)}) = e_1$". Moreover, ~~this~~ the quoted sentence contains a typo. Please correct the typo.

   Note that then, Lemma 3.3 implies that the ellipsoid method identifies a set of volume ε that contains the minimizer $x_\star$ in $O(\log(1/\varepsilon))$ iterations.

   4. (10 points) Argue that the convergence guarantee offered by Lemma 3.3 may not be meaningful.
   5. (10 points) Show that the proof of Theorem 3.7 does not apply to the optimization problem ~~(P)~~ (P'); hence, Theorem 3.9 that provides a more meaningful convergence guarantee does not apply. Please clearly point out which sentences in the proof of Theorem 3.7 do not hold and explain the reason.