Optimization Methods in ML Spring 2022/23 - HW 1

Lecturer: Dan Garber

Due date: 2.6.2022

Guidelines:

- you may submit in pairs
- you may consult with your fellow classmates ("high-level" discussions) - but you may not copy answers!
- programming is allowed in whatever environment you prefer

Question 1. *Prove that given a real-valued function f differentiable over a convex and closed set $\mathcal{K} \subseteq \mathbb{R}^d$, f is convex on $\mathcal{K}$ if and only if f satisfies the gradient inequality over $\mathcal{K}$. Hint: recall the definition of the directional derivative of a function.*

Question 2. *Let f be twice differentiable over $\mathcal{K} \subseteq \mathbb{R}^d$ closed and convex. Prove the following 2nd-order sufficient conditions:*

1. *If $\forall \mathbf{x} \in \mathcal{K} : \nabla^2 f(\mathbf{x}) \succeq 0$ then $f(\mathbf{x})$ is convex over $\mathcal{K}$*
2. *If $\forall \mathbf{x} \in \mathcal{K} : \nabla^2 f(\mathbf{x}) \preceq \beta \mathbf{I}$ then $f(\mathbf{x})$ is β-smooth over $\mathcal{K}$ (smoothness ineq. holds)*
3. *If $\forall \mathbf{x} \in \mathcal{K} : \nabla^2 f(\mathbf{x}) \succeq \alpha \mathbf{I}$ then $f(\mathbf{x})$ is α-strongly convex over $\mathcal{K}$ (stronger gradient-ineq. holds)*

Question 3. *Let $f(\mathbf{x}) := \max_{1 \leq i \leq n} g_i(\mathbf{x})$ such that each $g_i : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex and differentiable. Prove that for any $\mathbf{x} \in \mathbb{R}^d$, $\nabla g_{i^*}(\mathbf{x})$, where $i^* \in \arg\max_{1 \leq j \leq n} g_j(\mathbf{x})$, is a subgradient of f at $\mathbf{x}$.*

Question 4. *In class we have established the convergence of the subgradient descent method with a fixed step-size $\eta = \frac{D}{G\sqrt{T}}$, where T is a pre-fixed number of iterations. However, in practice many times a decaying step-size is preferred which does not require to pre-fix the number of iterations. Prove that our convergence theorem for subgradient descent still holds (potentially with a slightly worse universal constant) in case we choose the step-size on iteration t to be $\eta_t = \frac{D}{G\sqrt{t}}$.*

Question 5 (Beyond the black-box first-order model). *Consider the following **composite** optimization problem*

$$\min_{\mathbf{x}\in\mathcal{K}}\{f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})\},$$

*where $\mathcal{K} \subset \mathbb{R}^d$ is convex and compact, $g : \mathbb{R}^d \to \mathbb{R}$ is convex and β-smooth, and $h : \mathbb{R}^d \to \mathbb{R}$ is convex but **nonsmooth**.*

For instance, a famous problem that matches this model is LASSO Regression:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

Consider the following modified projected gradient method for composite optimization, which applies the following updates:

$$\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{K}} \left\|\mathbf{x} - \left(\mathbf{x}_t - \frac{1}{\beta}\nabla g(\mathbf{x}_t)\right)\right\|_2^2 + \frac{2}{\beta}h(\mathbf{x})$$

Answer the following questions:

- *Prove this method converges with rate $O(\beta D^2/t)$.*
- *Prove that if $h(\mathbf{x})$ is α-strongly convex (though still not smooth) this method converges with rate $O\left(\exp\left(-\Theta\left(\frac{\alpha}{\beta}t\right)\right)\right)$.*
- *Contrast the above results with the lower-bounds we know for nonsmooth minimization with first-order methods. Explain why they do not contradict each other.*

Question 6 (Conditional Gradient method for nonconvex optimization). *Consider the optimization problem:*

$$\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x}),$$

*where $\mathcal{K} \subset \mathbb{R}^d$ is convex and compact, and $f : \mathbb{R}^d \to \mathbb{R}$ is β smooth over $\mathcal{K}$ but **not convex**.*

A stationary point $\mathbf{x}_0 \in \mathcal{K}$ for the above problem can be equivalently defined as such that

$$\forall \mathbf{y} \in \mathcal{K} : \quad (\mathbf{y} - \mathbf{x}_0)^\top \nabla f(\mathbf{x}_0) \geq 0.$$

That is, there are no feasible descent directions from $\mathbf{x}_0$.

Naturally, a point $\mathbf{x}_0$ will be called ϵ-stationary if it holds that

$$\min_{y\in\mathcal{K}} (\mathbf{y} - \mathbf{x}_0)^\top \nabla f(\mathbf{x}_0) \geq -\epsilon.$$

Consider now the conditional gradient method with line-search for the above problem, which applies the following steps:

$$\begin{aligned}
\mathbf{v}_t &\leftarrow \arg\min_{\mathbf{v}\in\mathcal{K}} \mathbf{v}^\top \nabla f(\mathbf{x}_t) \\
\eta_t &\leftarrow \arg\min_{\eta\in[0,1]} f((1-\eta)\mathbf{x}_t + \eta\mathbf{v}_t) \\
\mathbf{x}_{t+1} &\leftarrow (1-\eta_t)\mathbf{x}_t + \eta_t\mathbf{v}_t,
\end{aligned}$$

where $\mathbf{x}_1$ *is some arbitrary point in* $\mathcal{K}$.

Prove that the sequence $\{\mathbf{x}_t\}_{t\geq 1}$ *produced by the above updates satisfies that for all* $t \geq 1$, *at least one of the iterates* $\mathbf{x}_1, \ldots, \mathbf{x}_t$ *is a* $O(\frac{\beta D^2}{\sqrt{t}})$*-stationary point, where* D *is the Euclidean diameter of* $\mathcal{K}$.

Question 7 (strong convexity and the Polyak-Lojasiewicz property). *We have seen that when minimzing a function* $f : \mathbb{R}^d \to \mathbb{R}$ *which is* α*-strongly convex and* β*-smooth over* $\mathbb{R}^d$ *(without constraints), the gradient descent method converges with rate* $\exp(-\Theta(\alpha/\beta)t)$. *We say a differentiable function has the* ***Polyak-Lojasiewicz*** *(PL) property with constant* α *if for any* x *it holds that* $\|\nabla f(x)\|^2 \geq \frac{\alpha}{2}(f(x) - f^*)$. *Answer the following questions:*

1. *Prove that if* f *is* α*-strongly convex it also has the PL property with constant* α.

2. *Prove that the least-squares function* $f(x) = \frac{1}{2}\|Ax - b\|^2$, *with* A *that has linearly-independent rows but is not full rank, is on one hand not strongly convex, but does satisfy the PL property. Give an expression for the PL constant.*

3. *Prove that the gradient descent method (with step size* $1/\beta$*) converges with rate* $\exp(-\Theta(\alpha/\beta)t)$ *for* $f(\cdot)$ *which is* β*-smooth and is PL with parameter* α.

Question 8 (programming question). *You are requested to empirically compare the performances of the (sub)gradient method for non-smooth optimization (with decaying step-sizes* $\frac{D}{G\sqrt{t}}$*), gradient descent for smooth convex optimization, and the accelerated gradient method on the linear regression optimization task:*

$$\min_{\mathbf{x}\in\mathbb{R}^d} \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2.$$

Generate the data as follows: take $\mathbf{A}$ *to be a random matrix (random as you choose) with fixed values of* $\sigma_{\max}(\mathbf{A}), \sigma_{\min}(\mathbf{A})$ *(of your choosing). Choose a solution* $\mathbf{x}^*$ *and set* $\mathbf{b} = \mathbf{A}\mathbf{x}^* + \xi$, *where* ξ *is a random noise of low magnitude. Compare the convergence rate of the algorithms (i.e,. function value vs. number of iterations). Experiment both in the case in which* $\mathbf{A}^\top\mathbf{A}$ *is not positive definite and in the case in which it is (then the problem is strongly convex). You may set the parameters* (D, G, β, α) *based directly on the data* $(\mathbf{A}, \mathbf{b})$ *(though this is not likely in real-life). Since data is random, plot the average of several i.i.d. experiments. Briefly discuss your observations of the experiment and contrast with the theory we have developed. Submit:*

- *code for experiments (zip file)*
- *documentation - how did you generate the data and how did you set the parameters for the algorithms. Conclusions from experiments.*
- *plot of the requested graphs*