Optimization Methods in ML Spring 2020 - HW 1

Lecturer: Dan Garber

Due date: 24.5.2020

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
- submit to me via email with title: "Optimization Methods in ML - HW 1" (expect my acknowledgment) (include names and IDs in email body)
- feel free to contact me with questions

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 $g_i : \mathbb{R}^d \to \mathbb{R}$ 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 (subgradient descent with noisy subgradients)**.** *You have seen that the projected subgradient descent method with step-size $\eta = \frac{R}{L\sqrt{t}}$ (recall R is the radius of a ball containing the feasible set $\mathcal{X}$ and L is an upper-bound on the Euclidean norm of subgradients) converges with rate $RL/\sqrt{t}$. Suppose now that we are given a **noisy** first order-oracle. That is, instead of observing a subgradient $g_s \in \partial f(x_s)$, we only observe some vector $\hat{g}_s$ such that $\|\hat{g}_s\| \leq L$ and $\|g_s - \hat{g}_s\| \leq 1/\sqrt{t}$.*

Prove that the noisy subgradient method with updates $x_{s+1} = \Pi_{\mathcal{X}}(x_s - \eta \hat{g}_s)$ also converges with rate $O(RL/\sqrt{t})$.

Question 5. *Recall that computing Euclidean projections onto convex sets is a central building-block of the methods we have studied so far. A feasible set which comes up in numerous applications is the unit simplex:* $\Delta_d := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{x} \geq \mathbf{0},\ \sum_{i=1}^d \mathbf{x}_i = 1\}$. *Prove that given a point* $\mathbf{y} \in \mathbb{R}^d$, *its projection onto the simplex is given by:* $\Pi_{\Delta_d}(\mathbf{y}) = \max\{\mathbf{y} - a \cdot \vec{\mathbf{1}}, \vec{\mathbf{0}}\}$, *where* max *operates entry-wise, and* a *is a real number satisfying:* $\sum_{i=1}^d \max\{\mathbf{y}_i - a, 0\} = 1$. *What is the best running-time you can think of for computing* $\Pi_{\Delta_d}(\mathbf{y})$*? Hint: recall* $\Pi_{\Delta_d}(\mathbf{y})$ *should be the optimal solution to a convex optimization problem and that global optimality can be verified via the gradient inequality.*

Question 6 (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:

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

Prove this method converges with rate $O(\beta R^2/t)$. *Contrast the above result with the lower-bounds we know for nonsmooth minmization with first-order methods. Explain why they do not contradict each other.*

Question 7 (strong convexity, quadratic growth and the Polyak-Lojasiewicz properties)**.** *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* $\exp(-\Theta(\alpha/\beta)t)$.

1. *We say a function* f *has the Quadratic Growth (QG) property with constant* α *if for any* x *it holds that* $dist(x, \mathcal{X}^*)^2 \leq \frac{2}{\alpha}(f(x) - f^*)$. *Here* $\mathcal{X}^*$ *denotes the set of optimal solutions,* $dist(x, \mathcal{X}^*)$ *is the distance of the point* x *from the set* $\mathcal{X}^*$, *i.e.,* $dist(x, \mathcal{X}^*) = \min_{y\in\mathcal{X}^*}\|x - y\|$, *and* f^* *is the optimal value.*

2. *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 QG property with constant* α.

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

3. *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.*

4. *Prove the the gradient descent method converges with rate $\exp(-\Theta(\alpha/\beta)t)$ for $f(\cdot)$ which is β-smooth and is QG with parameter α.*

5. *Prove the the gradient descent method 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 $\frac{R}{L\sqrt{t}}$), 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 (R, L, 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*