Midterm for EE227C (Spring 2018): Convex Optimization and Approximation

Instructor: Moritz Hardt

Email: hardt+ee227c@berkeley.edu

Graduate Instructor: Max Simchowitz

Email: msimchow+ee227c@berkeley.edu

March 20, 2018

Instructions:

(A) The midterm is due on Gradescope at 11:59pm Thursday, March 22nd.

(B) Collaboration is forbidden.

(C) Please ask all questions via *private* messages on Piazza. Do not email me and do not post publicly.

(D) I will be maintaining a list of errata on Piazza. Please consult this before asking a question about a typo that's already been addressed.

(E) The midterm must be LaTeXed. Points will be deducted otherwise.

(F) You may use anything in the course notes or one of the course texts as given. In particular, you can cite theorems and equations from coast notes and materials, but please point me to the lecture number and the statement you are citing. You may not Google anything or use the web.

Problem 1: Quasi-Pseudo-Mega-Star-Convexity

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable, but not necessarily convex function.

Definition 0.1 (Correlation)**.** A vector g is said to be $(\alpha, \beta, \epsilon)$-correlated with $x^* \in \mathbb{R}^n$ at x if

$$\langle g, x - x^* \rangle \geqslant \alpha \|x - x^*\|_2^2 + \beta \|g\|_2^2 - \epsilon \tag{1}$$

(A) Show that if f is 2α-strongly convex, $\frac{1}{2\beta}$-smooth, then there exists a unique $x^* \in \mathbb{R}^n$ such that, for all $x \in \mathbb{R}^n$, $\nabla f(x)$ is $(\alpha, \beta, 0)$-correlated with x^*. What is x^*?

(B) Suppose that f is possibly non-convex, but has a unique global minimizer x^* such that the gradients $\nabla f(x)$ are $(\alpha, \beta, \epsilon)$-correlated with x^*. Fix an $x_0 \in \mathbb{R}^n$, and consider the update rule $x_t \leftarrow x_{t-1} - \eta \nabla f(x_{t-1})$. Prove that, for any $\eta \leqslant \min\{2\beta, \frac{1}{2\alpha}\}$,

$$\|x_{t+1} - x^*\|_2^2 \leqslant (1 - 2\alpha\eta)\|x_t - x^*\|_2^2 + 2\eta\epsilon \tag{2}$$

Conclude that

$$\|x_t - x^*\|_2^2 \leqslant (1 - 2\alpha\eta)\|x_0 - x^*\|_2^2 + \epsilon/\alpha \tag{3}$$

(C) Consider the function $f(x) = g(x/\|x_2\|) \cdot \frac{1}{2}\|x\|_2^2$, where g is a differentiable possibly, non-convex function, with such that, for all $x \in \mathcal{S}^{n-1}$, $0 < A \leqslant g(x) \leqslant B$ and $\|\nabla g(x)\|_2 \leqslant M$.

(C.1) Suppose $x \neq 0$. Compute $\nabla f(x)$.

(C.2) Prove that for an explicit choice of step-size η, the algorithm $x_s = x_{s-1} - \eta \nabla f(x)$ satisfies the following convergence bound (you may be loose by constants):

$$\|x_s - x_*\|_2^2 \leqslant \left(1 - \frac{AB}{2(B + M/2)^2})\right)^s \|x_0 - x_*\|_2^2 \tag{4}$$

(C.3) Give a counter-example to show that f need not be convex.

Problem 2: The P Method

Throughout, we fix functions f and h. We suppose that f is convex and $M > 0$ smooth, and that h is convex. Let $F(x) = f(x) + h(x)$. For $L > 0$, define

$$Q_L(x, y) := f(y) + \langle x - y, \nabla f(y) \rangle + \frac{L}{2}\|x - y\|_2^2 + h(x) \tag{5}$$

$$p_L(y) := \arg\min\{Q_L(x, y) : x \in \mathbb{R}^n\} \tag{6}$$

(A) Show that $p_\alpha(y)$ is well defined; i.e., a unique point.

(B) Show that

$$p_L(y) = \mathsf{P}_{1/L,h}(y - \frac{1}{L}\nabla f(Y)) \tag{7}$$

$$\text{where } \mathsf{P}_{\alpha,h}(y) := \arg\min\{h(x) + \frac{1}{2\alpha}\|x - y\|_2^2 : x \in \mathbb{R}^d\} \tag{8}$$

(C) Show that if z and y satisfy $z = p_L(y)$, then

$$\nabla f(y) + L(z - y) \in -\partial h(z) \tag{9}$$

(D) Suppose that, for a given value of y, $F(p_L(y)) \leqslant Q(p_L(y), y)$. Show that

$$F(x) - F(p_L(y)) \geqslant \frac{L}{2}\|p_L(y) - y\|_2^2 + L\langle y - x, p_L(y) - y\rangle \tag{10}$$

(E) Prove that if f is $M \leqslant L$ smooth, then $F(x) \leqslant Q(x, y)$ for all y. Conclude that $F(p_L(y)) \leqslant Q(p_L(y), y) \leqslant Q(y, y) = F(y)$.

(F) State and prove an $O(1/t)$ convergence rate for the iterations $x_t \leftarrow p_L(x_{t-1})$, when $M \leqslant L$, when minimizing $F(\cdot)$. You may want to look at the notes I posted on Piazza for inspiration.

(G) Reasoning by analogy to accelerated gradient descent for smooth functions, modify the update rule in the previous subproblem to obtain an accelerated rate. You do not need to prove the rate, just give the algorithm.

(H) When might the update rule $x_t \leftarrow p_L(x_{t-1})$ be preferable to gradient descent?