# EE364M EXERCISE SET 10

Due date: March 20, 11:59pm

In this exercise set, we will explore a few consequences of the model-based minimization methods we have developed. This problem set is *completely optional*, but it provides (what I think) are some interesting additional developments to the lecture's contents, including showing how we relate stationarity to gradient mappings, even in problems beyond convexity.

**Question 10.1** (Models in convex optimization)**:** Consider the minimization problem

$$\begin{array}{ll} \text{minimize} & f(x) := h(x) + r(x) \\ \text{subject to} & x \in \Omega \end{array} \tag{10.1}$$

where $\Omega \subset \mathbb{R}^n$, $h$ has $L$-Lipschitz gradient on $\Omega$, and $r$ is convex.

(a) Show that the *prox-linear* model

$$f_x(y) := h(x) + \langle \nabla h(x), y - x \rangle + r(x)$$

is quadratically accurate for $f$, that is, $|f_x(y) - f(y)| \le \frac{L}{2} \|x - y\|_2^2$, and $f_x(y) \le f(y)$ for all $y$.

(b) Let $x^\star \in \Omega$ solve problem (10.1). Show that for an appropriate stepsize $\alpha > 0$, the iteration

$$x_{k+1} := \operatorname*{argmin}_{x \in \Omega} \left\{ f_{x_k}(x) + \frac{1}{2\alpha} \|x - x_k\|_2^2 \right\}$$

satisfies

$$f(x_{k+1}) - f(x^\star) \le \frac{L \|x_1 - x^\star\|_2^2}{2k}.$$

(c) Show how to solve the one-step update in part (b) for $r(x) = \lambda \|x\|_1$, where $\Omega \in \mathbb{R}^n$. (This method is called *iterative shrinkage and thresholding*, or *ISTA*.)

(d) Let $\Omega$ be the the collection of positive semidefinite matrices, and for $X \in \Omega$ let $r(X) = \lambda \operatorname{tr} X$ be the trace of $X$. Show how to solve the one-step update in part (b).

**Question 10.2** (A variational principle)**:** Ekeland's variational principle relates near optimality of points to being (nearly) stationary in ways that Question 10.3 explores.

**Theorem 10.2.1** (Ekeland [2])**.** *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be proper (so that $\inf_x f(x) > -\infty$) and closed, and let $x_0$ satisfy $f(x_0) - \inf_x f(x) \le \epsilon$ for some $0 \le \epsilon < \infty$. Then for any $\delta > 0$ there exists a point $\widehat{x}$ satisfying*

*i.* $\|\widehat{x} - x_0\|_2 \le \frac{\epsilon}{\delta}$

*ii.* $f(\widehat{x}) \le f(x_0)$

*iii.* *$\widehat{x}$ uniquely minimizes $f(x) + \delta \|x - \widehat{x}\|_2$ over $x \in \mathbb{R}^n$.*

(a) Define the sublevel-like set

$$S := \{x \in \mathbb{R}^n \mid f(x) + \delta \|x - x_0\|_2 \le f(x_0)\}.$$

Argue that $S$ is non-empty and compact and hence a minimizer of $\widehat{x} \in S$ of $f$ on $S$ exists.

(b) Show that
$$\|\widehat{x} - x_0\|_2 \le \frac{\epsilon}{\delta} \quad \text{and} \quad f(\widehat{x}) \le f(x_0).$$

(c) Finalize the proof of the theorem to demonstrate that $\widehat{x}$ as above satisfies its conclusions.

(d) Give an interpretation of Theorem 10.2.1. (There is no uniquely correct answer for this.)

For the final problem, we require a few new definitions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is $\lambda$-*weakly convex* if for some $x_0 \in \mathbb{R}^n$, the function
$$f(x) + \frac{\lambda}{2} \|x - x_0\|_2^2$$
is convex in $x$. (Convince yourself that the choice of $x_0$ is immaterial here.) The Fréchet subdifferential $\partial^{\mathrm{F}} f(x)$ of such a function consists of those vectors $g$ for which
$$f(y) \ge f(x) + \langle g, y - x\rangle - O(\|y - x\|^2)$$
as $y \to x$, which is equivalent in the $\lambda$-weakly convex case (this is not completely trivial) to the set of vectors
$$g \in \partial_z \left\{ f(z) + \frac{\lambda}{2} \|z - x\|_2^2 \right\} \bigg|_{z=x},$$
that is, the regular subgradient of the convex function $z \mapsto f(z) + \frac{\lambda}{2} \|z - x\|_2^2$ evaluated at $z = x$. Note that in this case, if the function $f$ in Ekeland's variational principle (Theorem 10.2.1) is weakly convex, then the final part becomes equivalent to the claim that
$$0 \in \partial^{\mathrm{F}} f(\widehat{x}) + \delta \mathbb{B}_2^n.$$

**Question 10.3** (Gradient mappings and weakly convex minimization)**:**

(a) Give an example of a convex function $f : \mathbb{R} \to \mathbb{R}$ for which $|f'(x)| \ge 1$ except when $x = \operatorname{argmin}_x f(x)$, so that even in the convex case, we would not (generally) expect algorithms to find points with small subgradients.

(b) Let $f(x) = h(c(x))$, where $h$ is convex and $L_h$-Lipschitz and $c$ has $L_c$-Lipschitz derivative. Show that $f$ is $L_c L_h$-weakly convex, and even more, that
$$\partial^{\mathrm{F}} f(x) \supset \nabla c(x) \partial h(c(x)).$$

*Hint.* It is enough to show that $f + \frac{M}{2} \|\cdot\|_2^2$ is subdifferentiable.

(c) Recall that for a model $f_x$ of $f$ at the point $x$, we define the gradient mapping $\mathsf{G}_\alpha$ via
$$x_\alpha = \operatorname*{argmin}_{x \in X} \left\{ f_{x_0}(x) + \frac{1}{2\alpha} \|x - x_0\|_2^2 \right\} \quad \text{and} \quad \mathsf{G}_\alpha(x_0) := \frac{1}{\alpha}(x_0 - x_\alpha).$$

Let $f_x$ be convex and quadratically accurate, so that $|f_x(y) - f(y)| \le \frac{M}{2} \|x - y\|_2^2$ for all $y$. Use Ekeland's variational principle (Theorem 10.2.1) show that for any $x_0$ and $\alpha > 0$, there exists $\widehat{x}$ satisfying

i. Point proximity: $\|\widehat{x} - x_\alpha\|_2 \le \alpha \|\mathsf{G}_\alpha(x_0)\|_2$.

ii. Value proximity: $f(\widehat{x}) \le f(x_\alpha) + \frac{M\alpha^2+\alpha}{2} \left\| \mathsf{G}_\alpha(x_0) \right\|_2^2$.

iii. Near stationarity: $\operatorname{dist}(0, \partial^{\mathrm{F}} f(\widehat{x})) \le (2M\alpha + 1) \left\| \mathsf{G}_\alpha(x_0) \right\|_2$.

In short, once the gradient mapping $\mathsf{G}_\alpha(x_0)$ is small, there is a point $\widehat{x}$ near the updated point $x_\alpha$ that is nearly stationary for $f$.

*Hint.* Following Drusvyatskiy and Lewis [1], define the function $\varphi(y) := f(y) + \frac{M+\alpha^{-1}}{2} \left\| y - x_0 \right\|_2^2$. Argue that $\varphi(x_\alpha) - \varphi^\star \le M \left\| x_\alpha - x_0 \right\|_2^2$, where $\varphi^\star = \inf_y \varphi(y)$. Then apply Ekeland's variational principle with $\epsilon = M\alpha \left\| \mathsf{G}_\alpha(x_0) \right\|_2^2$, and observe that the $\widehat{x}$ it guarantees satisfies $0 \in \partial^{\mathrm{F}} \varphi(\widehat{x}) + \delta \mathbb{B}_2^n$, where $\partial^{\mathrm{F}} \varphi(x) = \partial^{\mathrm{F}} f(x) + \frac{M+\alpha^{-1}}{2}(x - x_0)$.

## References


[1] D. Drusvyatskiy and A. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

[2] I. Ekeland. On the variational principle. *Journal of Mathematical Analysis and Applications*, 47(2):324–353, 1974.