MGMT 690 - Pset 3

Spring 2024

Instructions:

- This pset is due on Sunday, April 21 at 11:59pm.
- Completed psets should be submitted to Gradescope.
- **Exercises** are for your own review only. They do not need to be submitted and will not be graded.
- **Complete all problems 1–3.**

1 Exercises

Excercise 1

Lemma 1. *Consider* $f : \mathbb{R}^2 \to \mathbb{R}$ *by*

$$f(x) = |x_1| + 2|x_2|.$$

Then, $\partial f(1,0) = \{(1,y) : |y| \leq 2\}$. *Thus,* $(1,2) \in \partial f(1,0)$ *and* $f((1,0) - t(1,2)) > f(1,0)$ *for all* $t > 0$

Proof. We can write

$$f(x) = \max\{x_1 + 2x_2, -x_1 + 2x_2, x_1 - 2x_2, -x_1 - 2x_2\}.$$

Then, by Problem 1,

$$\partial(f(1,0)) = \text{conv}\left(\left\{\begin{pmatrix}1\\2\end{pmatrix}, \begin{pmatrix}1\\-2\end{pmatrix}\right\}\right) = \{(1,y) : |(|\,y) \leq 2\}.$$

Next,

$$f((1,0) - t(1,2)) = \begin{cases} 1+t & \text{if } 0 \leq t \leq 1 \\ 3t-1 & \text{else} \end{cases}$$

This is $> f(1,0)$ for all $t > 0$. ■

Problems

Problem 1

Lemma 2. *Let $f_i : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable for $i = 1, \ldots, n$. Let $F(x) \coloneqq \max_i f_i(x)$. Then, for any $x \in \mathbb{R}^d$,*

$$\partial F(x) = \operatorname{conv}(\{\nabla f_i(x) : f_i(x) = F(x)\}).$$

Proof. Fix $\bar{x} \in \mathbb{R}^d$. For convenience, let $\mathcal{I} \coloneqq \{i \in [n] : f_i(\bar{x}) = F(\bar{x})\}$. Also define

$$\mathcal{S} \coloneqq \operatorname{conv}(\{\nabla f_i(\bar{x}) : i \in \mathcal{I}\}).$$

Our goal is to show that $\partial F(\bar{x}) = \mathcal{S}$.

First, suppose $g \in \mathcal{S}$. By definition, there exists λ_i such that $\sum_{i \in \mathcal{I}} \lambda_i = 1$, $\lambda_i \geq 0$ for all $i \in \mathcal{I}$.

Now, consider the convex function

$$L(x) \coloneqq \sum_{i \in \mathcal{I}} \lambda_i f_i(x).$$

This function satisfies: $\nabla L(\bar{x}) = g$, $L(\bar{x}) = F(\bar{x})$ and $L(x) \leq F(x)$ everywhere. To see that the last statement is true:

$$L(x) = \sum_{i \in \mathcal{I}} \lambda_i f_i(x) \leq \max_{i \in \mathcal{I}} f_i(x) \leq \max_i f_i(x) = F(x).$$

We deduce that for all $x \in \mathbb{R}^d$,

$$F(x) \geq L(x) \geq L(\bar{x}) + \langle \nabla L(\bar{x}), x - \bar{x} \rangle = F(\bar{x}) + \langle g, x - \bar{x} \rangle .$$

By definition, this means that $g \in \partial F(\bar{x})$.

Now, suppose $g \notin \mathcal{S}$ and assume for the sake of contradiction that $g \in \partial F(\bar{x})$. As $\mathcal{S}$ is compact, there exists a $v \in \mathbb{R}^d$ with $\|v\| = 1$ so that

$$v^\intercal g > \max_{i \in \mathcal{I}} \langle v, \nabla f_i(\bar{x}) \rangle .$$

Let $\delta \coloneqq v^\intercal g - \max_{i \in \mathcal{I}} \langle v, \nabla f_i(\bar{x}) \rangle$.

Now, let $x_\alpha \coloneqq \bar{x} + \alpha v$. As the f_i are continuous, it holds that $F(x_\alpha) = \max_{i \in \mathcal{I}} f_i(x_\alpha)$ for all small enough $\alpha > 0$. Now, as $g \in \partial F(\bar{x})$, we have that

$$\begin{aligned}
F(x_\alpha) &\geq F(\bar{x}) + \langle g, x_\alpha - \bar{x} \rangle \\
&= F(\bar{x}) + \alpha \langle g, v \rangle \\
&\geq \max_{i \in \mathcal{I}} \left(F(\bar{x}) + \alpha \langle v, \nabla f_i(\bar{x}) \rangle \right) + \alpha\delta \\
&\geq F(x_\alpha) - o(\alpha) + \alpha\delta \\
&> F(x_\alpha).
\end{aligned}$$

The last two inequalities both hold for all $\alpha > 0$ small enough. This is a contradiction. ∎

Problem 2

Let $\gamma > 1$ and consider the following function $f : \mathbb{R}^2 \to \mathbb{R}$

$$f(x) = \begin{cases} \sqrt{x_1^2 + \gamma x_2^2} & \text{if } |x_2| \le x_1 \\ \frac{x_1 + \gamma |x_2|}{\sqrt{1+\gamma}} & \text{else} \end{cases}$$

This function is convex and $\sqrt{\gamma}$-Lipschitz (you do not need to prove this).

Consider the subgradient method with *exact* line-search initialized at $x^{(0)} = (\gamma, 1)$, i.e., for $t \ge 1$, let $g \in \partial f(x^{(t-1)})$ and set

$$x^{(t)} = \operatorname*{arg\,min}_{x \in x^{(t-1)} - \mathbb{R}_+ g} f(x)$$

We will show that this method behaves poorly. We will need the following lemma.

Lemma 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$.*

Proof. As $f : \mathbb{R}^n \to \mathbb{R}$ is convex, the set $\partial f(x)$ is nonempty. Let $g \in \partial f(x)$. Let $u \in \mathbb{R}^n$ and consider the one dimensional function

$$t \mapsto f(x + tu).$$

Now, for all $t > 0$

$$\langle \nabla f(x), u \rangle = \lim_{t \to 0} \frac{f(x+tu) - f(x)}{t} \ge \lim_{t \to 0} \frac{t \langle g, u \rangle}{t} = \langle g, u \rangle .$$

Here, we have used that $f(x + tu) \ge f(x) + t \langle g, u \rangle$. We deduce that

$$\langle \nabla f(x), u \rangle \ge \langle g, u \rangle .$$

As u was arbitrary, we conclude that $g = \nabla f(x)$. ■

For convenience, define

$$\alpha := \frac{\gamma - 1}{\gamma + 1}.$$

Proposition 1. *For all $t \ge 0$, it holds that*

$$x^{(t)} = \left(\gamma \alpha^t, (-\alpha)^t\right) .$$

In particular, for all $t \ge 0$, it holds that $f(x^{(t)}) \ge 0$, despite $\inf_x f(x) = -\infty$.

Proof. By definition, $x^{(0)} = (\gamma, 1) = (\gamma\alpha^0, (-\alpha)^0)$. Thus, the claim holds for $t = 0$.

Now, consider $t > 0$. For convenience, let $y = x^{(t-1)}$. By induction, we have that

$$y = (\gamma\alpha^{t-1}, (-\alpha)^{t-1}).$$

As $\gamma > 1$, we have that $|y_2| < y_1$ so that f is differentiable at y with gradient

$$\nabla f(y) = \frac{1}{\sqrt{y_1^2 + \gamma y_2^2}} \begin{pmatrix} y_1 \\ \gamma y_2 \end{pmatrix}.$$

We can parameterize the ray beginning at y in the direction of $-\nabla f(y)$ as

$$x_\beta = \begin{pmatrix} \gamma\alpha^{t-1}(1-\beta) \\ (-\alpha)^{t-1}(1-\gamma\beta) \end{pmatrix}$$

for $\beta \in \mathbb{R}_+$. Note that when $\beta = 1 - \alpha$

$$x_\beta = \begin{pmatrix} \gamma\alpha^t \\ (-\alpha)^{t-1}(1-\gamma(1-\alpha)) \end{pmatrix} = \begin{pmatrix} \gamma\alpha^t \\ (-\alpha)^t \end{pmatrix}.$$

Thus, it remains to show that

$$\beta \mapsto \sqrt{(\gamma\alpha^{t-1}(1-\beta))^2 + \gamma((-\alpha)^{t-1}(1-\gamma\beta))^2}$$

is minimized at $\beta = 1 - \alpha$. It suffices to check that term inside the radical achieves its minimum at $\beta = 1 - \alpha$. The derivative in β of the term inside the radical evaluated at $\beta = 1 - \alpha$ is

$$\begin{aligned} &-2(\gamma\alpha^{t-1})^2(1-\beta) - 2\gamma^2\alpha^{2(t-1)}(1-\gamma\beta) \\ &\quad = -2\gamma^2\alpha^{2t-1} + 2\gamma^2\alpha^{2t-1} \\ &\quad = 0. \end{aligned}$$

■

Problem 3

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a L-Lipschitz convex function with minimizer $x^\star$ and minimum value f^*. Suppose that f satisfies the following growth condition parameterized by $\delta > 0, \alpha > 0$:

$$f(x) - f^\star \leq \delta \quad \implies \quad f(x) - f^* \geq \alpha \left\| x - x^\star \right\|^2.$$

Suppose we are given $x_0 \in \mathbb{R}^n$ with $\|x_0 - x^\star\| \leq R$.

Consider the following *restarted* subgradient method.

Algorithm 1 Restarted subgradient method

Given: $L, R, \alpha, \delta, x_0$

- For each $k = 0, \ldots$
  - Run the subgradient method with constant stepsizes (see Corollary 7 in the notes) with initial iterate x_k for

$$T_k = \begin{cases} \left\lfloor \left(\frac{LR}{\delta}\right)^2 \right\rfloor & \text{if } k = 0 \\ \left\lfloor \frac{2L^2}{\alpha\delta 2^k} \right\rfloor & \text{else} \end{cases}$$

    iterations. Let x_{k+1} to be the output of the subgradient method.

By setting $T_0 = \left\lfloor \left(\frac{LR}{\delta}\right)^2 \right\rfloor$, we can ensure the following property:

Lemma 4. *It holds that $f(x_1) - f^\star \leq \delta$.*

Proof. By Corollary 7 of the notes,

$$f(x_1) - f^\star \leq \frac{LR}{\sqrt{T_0 + 1}} = \frac{LR}{\sqrt{L^2 R^2/\delta^2}} = \delta. \qquad \blacksquare$$

For $k \geq 1$, define $\delta_k = \frac{2}{2^k}\delta \leq \delta$. By setting $T_k = \left\lfloor \frac{4L^2}{\alpha\delta_k} \right\rfloor$ for $k \geq 1$, we can ensure the following property:

Lemma 5. *It holds that $f(x_k) - f^\star \leq \delta_k$.*

Proof. By the previous lemma, we have that $f(x_1) - f^\star \leq \delta_1$.

Now, by induction, suppose that $f(x_k) - f^\star \leq \delta_k$. By Corollary 7 of the notes and the growth property, we have that

$$\begin{aligned} f(x_{k+1}) - f^\star &\leq \frac{L\left\|x_k - x^\star\right\|}{\sqrt{T_k + 1}} \\ &\leq \frac{L\sqrt{(f(x_k) - f^\star)/\alpha}}{\sqrt{T_k + 1}} \\ &\leq \frac{L\sqrt{\delta_k/\alpha}}{\sqrt{T_k + 1}} \\ &\leq \frac{\delta_k}{2} \\ &= \delta_{k+1}. \end{aligned} \qquad \blacksquare$$

We conclude that:

Proposition 2. *The restarted subgradient method with constant stepsizes and horizons $T_0 = \left\lfloor \left(\frac{LR}{\delta}\right)^2 \right\rfloor$ and $T_k = \left\lfloor \frac{4L^2}{\alpha\delta_k} \right\rfloor$ for all $k \geq 1$ achieves a gap $f(x) - f^\star \leq \epsilon$*

after at most

$$O\left(\frac{L^2R^2}{\delta^2} + \frac{L^2}{\alpha\epsilon}\right)$$

total (inner) iterations. Thus for $\epsilon \ll \frac{\delta^2}{\alpha R^2}$*, this convergence rate is* $O\left(\frac{L^2}{\alpha\epsilon}\right)$.

Proof. Set

$$k = \left\lceil \log_2\left(\frac{2\delta}{\epsilon}\right)\right\rceil.$$

Note that with this choice of k, it holds that $\delta_k \le \epsilon$ and $2^{k-1} < \frac{2\delta}{\epsilon}$.

It remains to compute

$$T_1 + T_2 + \cdots + T_k \le \frac{L^2}{\alpha\delta}\left(2^1 + 2^2 + \cdots + 2^k\right) \le \frac{L^2 2^{k+1}}{\alpha\delta} = \frac{8L^2}{\alpha\epsilon}.$$

■

Problem 4

Consider a minimization problem of the form

$$\min_{x\in\Omega} F(x)$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is an arbitrary function and $\Omega \subseteq \mathbb{R}^n$ is an arbitrary set. We say that

$$\texttt{prox} : \mathbb{R}^n \to \Omega$$

is a *smooth prox-oracle* for this problem if `prox` satisfies the following property: Given $y \in \mathbb{R}^n$, define $g(y) \coloneqq L(y - \texttt{prox}(y))$. Then, for all $z \in \Omega$, it holds that

$$F(\texttt{prox}(y)) \le F(z) + \langle g(y), y - z\rangle - \frac{\|g(y)\|^2}{2L}. \quad (1)$$

We will replace the gradient step in accelerated gradient descent with the prox oracle:

Algorithm 2 Accelerated gradient descent for smoothly proxable problems

Given $x_0 \in \mathbb{R}^d$, $F : \mathbb{R}^n \to \mathbb{R}$ and $\texttt{prox} : \mathbb{R}^n \to \Omega$

- Set $y_0 = x_0$ and $\lambda_{-1} = 1$
- For $t = 0, \ldots$

$$\lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2}$$
$$\gamma_t = \frac{\lambda_{t-1} - 1}{\lambda_t}$$
$$x_{t+1} = \texttt{prox}(y_t) = y_t - \frac{1}{L} g(y_t)$$
$$y_{t+1} = x_{t+1} + \gamma_t (x_{t+1} - x_t)$$

Theorem 1. *Suppose $F : \mathbb{R}^n \to \mathbb{R}$ and $\Omega \subseteq \mathbb{R}^n$ and suppose $\texttt{prox} : \mathbb{R}^n \to \Omega$ is a smooth prox-oracle for $\min_{x \in \Omega} F(x)$. Furthermore, suppose F has minimizer $x^\star$ with minimum value $F^\star$. Then, it holds that*

$$F(x_T) - F^\star = O\left(\frac{L \left\| x_0 - x^\star \right\|^2 + F(x_0) - F^\star}{T^2} \right).$$

Proof. Let $\delta_t = F(x_t) - F^\star$, $g_t = g(y_t)$, and $\Delta_t = y_t - x_t$.

By definition of the $\texttt{prox}$ oracle, we have that for all $t \geq 0$ and all $z \in \Omega$,

$$F(x_{t+1}) \leq F(z) + \langle g_t, y_t - z \rangle - \frac{\left\| g_t \right\|^2}{2L}$$

Taking $z = x_t$ gives

$$\delta_{t+1} - \delta_t \leq \langle g_t, \Delta_t \rangle - \frac{1}{2L} \left\| g_t \right\|^2.$$

Taking $z = x^\star$ gives

$$\delta_{t+1} \leq \langle g_t, \Delta_t + x_t - x^\star \rangle - \frac{1}{2L} \left\| g_t \right\|^2$$

Now, let us take the first inequality weighted by $(\lambda_t - 1)$ and add it to the second inequality to get

$$\lambda_t \delta_{t+1} - (\lambda_t - 1)\delta_t \leq \langle g_t, \lambda_t \Delta_t + (x_t - x^\star) \rangle - \frac{\lambda_t}{2L} \left\| g_t \right\|^2.$$

We will complete the square on the right hand side to write it as

$$\begin{aligned}
&\langle g_t, \lambda_t \Delta_t + (x_t - x^\star)\rangle - \frac{\lambda_t}{2L}\|g_t\|^2 \\
&\quad = \frac{L}{2\lambda_t}\left(2\left\langle \frac{\lambda_t g_t}{L}, \lambda_t\Delta_t + (x_t - x^\star)\right\rangle - \left\|\frac{\lambda_t g_t}{L}\right\|^2\right) \\
&\quad = \frac{L}{2\lambda_t}\left(\|\lambda_t\Delta_t + (x_t - x^\star)\|^2 - \left\|\lambda_t\Delta_t + (x_t - x^\star) - \frac{\lambda_t g_t}{L}\right\|^2\right).
\end{aligned}$$

By our choice of λ_t and γ_t, we have that

$$\begin{aligned}
\lambda_t\Delta_t + (x_t - x^\star) - \frac{\lambda_t g_t}{L} &= \lambda_t(y_t - x_t) + x_t - x^\star - \lambda_t(y_t - x_{t+1}) \\
&= (1-\lambda_t)x_t - x^\star + \lambda_t x_{t+1} \\
&= (\lambda_t - 1)(x_{t+1} - x_t) + (x_{t+1} - x^\star) \\
&= \frac{\lambda_t - 1}{\gamma_{t+1}}\Delta_{t+1} + (x_{t+1} - x^\star) \\
&= \lambda_{t+1}\Delta_{t+1} + (x_{t+1} - x^\star).
\end{aligned}$$

Thus,

$$\lambda_t^2\delta_{t+1} - (\lambda_t - 1)\lambda_t\delta_t \le \frac{L}{2}\left(\|\lambda_t\Delta_t + (x_t - x^\star)\|^2 - \|\lambda_{t+1}\Delta_{t+1} + (x_{t+1} - x^\star)\|^2\right).$$

Note that $(\lambda_t - 1)\lambda_t = \lambda_{t-1}^2$. Finally, we may telescope this inequality to get

$$\lambda_{T-1}^2\delta_T \le \frac{L}{2}\|x_0 - x^\star\|^2 + \delta_0. \qquad \blacksquare$$

Lemma 6. *Suppose $F : \mathbb{R}^n \to \mathbb{R}$ is an L-smooth convex function and $\Omega \subseteq \mathbb{R}^n$ is nonempty, closed, and convex. Define*

$$\texttt{prox}(y) \coloneqq \arg\min_{x\in\Omega}\left\{F(y) + \langle\nabla F(y), x - y\rangle + \frac{L}{2}\|x - y\|^2\right\}.$$

This map is well-defined, is equal to

$$\texttt{prox}(y) = \Pi_\Omega\left(y - \frac{1}{L}\nabla F(y)\right),$$

and is a smooth prox-oracle for $\min_{x\in\Omega} F(x)$.

Proof. For the first point, note that

$$\begin{aligned}
\texttt{prox}(y) &\coloneqq \arg\min_{x\in\Omega}\left\{F(y) + \langle\nabla F(y), x - y\rangle + \frac{L}{2}\|x - y\|^2\right\} \\
&= \arg\min_{x\in\Omega}\left\{\frac{L}{2}\left\|x - y + \frac{\nabla F(x)}{L}\right\|^2 + \text{constant}\right\} \\
&= \arg\min_{x\in\Omega}\left\|x - y + \frac{\nabla F(x)}{L}\right\|.
\end{aligned}$$

As Ω is closed and convex, this arg min exists and is unique and is, by definition, the projection of $y - \nabla F(x)/L$ onto Ω.

We need to check that for all $y, z \in \Omega$ that

$$F(\texttt{prox}(y)) \leq F(z) + \langle g(y), y - z\rangle - \frac{\|g(y)\|^2}{2L},$$

where $g(y) \coloneqq L(y - \texttt{prox}(y))$. For convenience, let $y' \coloneqq \texttt{prox}(y)$.

We compute:

$$\begin{aligned}
F(z) &\geq F(y) + \langle \nabla F(y), z - y\rangle \\
&= F(y) + \langle \nabla F(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 - \frac{L}{2}\|z - y\|^2 \\
&\geq F(y) + \langle \nabla F(y), y' - y\rangle + \frac{L}{2}\|y' - y\|^2 + \frac{L}{2}\|y' - z\|^2 - \frac{L}{2}\|z - y\|^2 \\
&\geq F(y') + \frac{L}{2}\|y' - y + y - z\|^2 - \frac{L}{2}\|z - y\|^2 \\
&= F(y') + \frac{1}{2L}\|g(y)\|^2 - \langle g(y), y - z\rangle .
\end{aligned}$$

Here, the second inequality follows from the fact that $F(y) + \langle \nabla F(y), z - y\rangle + \frac{L}{2}\|z - y\|^2$ is L-strongly convex in z with minimizer y'. ■

Lemma 7. *Suppose $f_1, \ldots, f_k : \mathbb{R}^n \to \mathbb{R}$ are L-smooth convex functions and define*

$$F(x) \coloneqq \max_{i\in[k]} f_i(x)$$
$$\texttt{prox}(y) \coloneqq \arg\min_{x\in\mathbb{R}^n} \max_{i\in[k]} \left\{ f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{L}{2}\|x - y\|^2 \right\}.$$

This map is well-defined and is a smooth prox-oracle for $\min_{x\in\Omega} F(x)$.

Proof. The objective function in the definition of $\texttt{prox}$ is L-strongly convex, thus the arg min exists and is unique. This proves that $\texttt{prox}$ is well-defined.

Now, note that the objective function in the definition of $\texttt{prox}$ is L-strongly convex. Now, let $y, z \in \mathbb{R}^n$ be arbitrary and for convenience, let $y' = \texttt{prox}(y)$. Then,

$$\begin{aligned}
F(z) &\geq \max_{i\in[k]} \{ f_i(y) + \langle \nabla f(y), z - y\rangle \} \\
&= \max_{i\in[k]} \left\{ f_i(y) + \langle \nabla f(y), z - y\rangle + \frac{L}{2}\|z - y\|^2 \right\} - \frac{L}{2}\|z - y\|^2 \\
&\geq \max_{i\in[k]} \left\{ f_i(y) + \langle \nabla f(y), y' - y\rangle + \frac{L}{2}\|y' - y\|^2 \right\} + \frac{L}{2}\|y' - z\|^2 - \frac{L}{2}\|z - y\|^2 \\
&\geq F(y') + \frac{L}{2}\|y' - y + y - z\|^2 - \frac{L}{2}\|z - y\|^2 \\
&= F(y') + \frac{1}{2L}\|g(y)\|^2 - \langle g(y), y - z\rangle .
\end{aligned}$$

Here, the second inequality follows from the fact that the first half of the second line is L-strongly convex in z with minimizer y'. ■