MGMT 690 - Pset 4

Spring 2024

Instructions:

- This pset is due on **Thursday, May 2** 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.**

Problems

1. Let $f, g : \mathbb{R}^n \to \mathbb{R}$. The infimal convolution of f and g, denoted $f \Box g$, is a function $(f \Box g) : \mathbb{R}^n \to [-\infty, \infty]$ defined as

$$(f \Box g)(x) \coloneqq \inf_{z \in \mathbb{R}^n} f(z) + g(y - z).$$

Lemma 1. *If $f, g : \mathbb{R}^n \to \mathbb{R}$, then*

$$(f \Box g)^*(y) = f^*(y) + g^*(y) \qquad \forall y \in \mathbb{R}^n.$$

Proof. By definition,

$$\begin{aligned} (f \Box g)^*(y) &= \sup_{x \in \mathbb{R}^n} \langle y, x \rangle - (f \Box g)(x) \\ &= \sup_{x \in \mathbb{R}^n, z \in \mathbb{R}^n} \langle y, (x - z) + z \rangle - f(z) - g(x - z) \\ &= \sup_{w \in \mathbb{R}^n, z \in \mathbb{R}^n} \langle y, z \rangle - f(z) + \langle y, w \rangle - g(w) \\ &= f^*(y) + g^*(y). \end{aligned}$$

■

2. This question derives a mirror descent setup on the simplex with the ℓ_1 norm.

Let

$$\Delta_n \coloneqq \left\{ x \in \mathbb{R}^n : \begin{array}{l} x \geq 0 \\ 1^\intercal x \leq 1 \end{array} \right\}.$$

Define $\omega : \Delta_n \to \mathbb{R}$ by

$$\omega(x) = \sum_{i=1}^{n} x_i \log(x_i)$$

with the convention that $0 \log(0) = 0$. For convenience, define

$$\Delta_n^o \coloneqq \left\{ x \in \mathbb{R}^n : \begin{array}{c} x > 0 \\ 1^\intercal x \le 1 \end{array} \right\}.$$

Lemma 2. *ω is closed and convex and differentiable on* $\operatorname{dom}(\partial\omega) = \Delta_n^o$.

Proof. Note that $x \log x$ is continuous on $(0, \infty)$ and that $\lim_{x \to 0} x \log x = 0$. Furthermore, for $x > 0$,

$$\frac{d^2}{dx^2} x \log x = \frac{1}{x} > 0.$$

Thus, $x \mapsto x \log x$ is a convex function on $[0, \infty)$. Thus, $\sum_i x_i \log(x_i)$ is a real-valued convex function on $\mathbb{R}^n_+$. We deduce that ω is closed and convex on Δ_n.

We have that $\operatorname{dom}(\partial\omega) = \Delta_n^o$. On $\operatorname{dom}(\partial\omega)$, the gradient of $\omega(x)$ is given by

$$\nabla\omega(x) = \begin{pmatrix} 1 + \log(x_1) \\ \vdots \\ 1 + \log(x_n) \end{pmatrix}. \qquad \blacksquare$$

Lemma 3. *ω is 1-strongly convex on Δ_n.*

There are a few different proofs depending on what you may know from outside this course.

Proof 1. This proof uses only what we learned in this course.

Define

$$g(x) \coloneqq \omega(x) - \frac{1}{2} \|x\|_1^2 = \sum_i x_i \log(x_i) - \frac{1}{2} \left(\sum_i x_i \right)^2$$

on $\mathbb{R}^n_+$. Our goal is to check that g is convex on Δ_n. As g is continuous up to its boundary, it suffices to check that g is convex on Δ_n^o.

As $g(x)$ is twice differentiable on Δ_n^o, it suffices to show that for all $x \in \Delta_n^o$, that

$$\nabla^2 g(x) \succeq 0.$$

Let $x \in \Delta_n^o$ and $y \in \mathbb{R}^n$. We compute

$$\begin{aligned}
\left\langle y, \nabla^2 g(x) y \right\rangle &= \sum_i \frac{y_i^2}{x_i} - \left(\sum_i y_i \right)^2 \\
&\geq \sum_i \frac{y_i^2}{x_i} - \left(2 - \sum_i x_i \right) \left(\sum_i y_i \right)^2 \\
&= \sum_i \frac{y_i^2}{x_i} - 2 \sum_{i=1}^n y_i \sum_{j=1}^n y_j + \sum_i x_i \left(\sum_i y_i \right)^2 \\
&= \sum_i x_i \left(\frac{y_i^2}{x_i^2} - 2 \frac{y_i}{x_i} \sum_{j=1}^n y_j + \left(\sum_{j=1}^n y_j \right)^2 \right) \\
&= \sum_i x_i \left(\frac{y_i}{x_i} - \sum_{j=1}^n y_j \right)^2 \\
&\geq 0. \qquad \blacksquare
\end{aligned}$$

Proof 2. This proof uses what we learned in the course and the Sherman–Morrison formula.

Define

$$g(x) \coloneqq \omega(x) - \frac{1}{2} \|x\|_1^2 = \sum_i x_i \log(x_i) - \frac{1}{2} \left(\sum_i x_i \right)^2$$

on $\mathbb{R}_+^n$. Our goal is to check that g is convex on Δ_n. As g is continuous up to its boundary, it suffices to check that g is convex on

$$\Omega \coloneqq \left\{ x \in \mathbb{R}^n : \begin{array}{c} x > 0 \\ 1^\intercal x < 1 \end{array} \right\}.$$

As $g(x)$ is twice differentiable on Ω, it suffices to show that for all $x \in \Omega$, that

$$\nabla^2 g(x) \succeq 0.$$

Let $x \in \Omega$. Let

$$\alpha = \frac{1}{1 - 1^\intercal x},$$

which exists by the assumption $x \in \Omega$. We will write the Hessian explicitly and recognize the Sherman–Morrison formula:

$$\begin{aligned}\nabla^2 g(x) &= \mathrm{Diag}(x)^{-1} - \mathbf{1}\mathbf{1}^\intercal \\ &= \mathrm{Diag}(x)^{-1} - \frac{\alpha \mathbf{1}\mathbf{1}^\intercal}{1+\alpha 1^\intercal x} \\ &= \mathrm{Diag}(x)^{-1} - \frac{\alpha\,\mathrm{Diag}(x)^{-1} x x^\intercal \,\mathrm{Diag}(x)^{-1}}{1+\alpha x^\intercal \,\mathrm{Diag}(x)^{-1} x} \\ &= (\mathrm{Diag}(x) + \alpha x x^\intercal)^{-1} \\ &\succ 0. \end{aligned}$$

■

Proof 3. This proof uses the fact that a twice-differentiable function f is 1-strongly convex in a norm $\|\cdot\|$ if and only if $\langle y, \nabla^2 f(x) y\rangle \geq \|y\|^2$ for all $x \in \mathrm{dom}(f)$ and $y \in \mathbb{R}^n$.

Our goal is to check that ω is 1-strongly convex on Δ_n. As ω is continuous up to its boundary, it suffices to check that ω is 1-strongly convex on Δ_n^o. Note that ω is twice-differentiable on Δ_n^o, thus it suffices to check that for all $x \in \Delta_n^o$ and $y \in \mathbb{R}^n$, that $\langle y, \nabla^2 \omega(x) y\rangle \geq \|y\|^2$. We compute:

$$\begin{aligned}\langle y, \nabla^2 \omega(x) y\rangle &= \langle y, \mathrm{Diag}(x)^{-1} y\rangle \\ &\geq \sum_i \frac{y_i^2}{x_i} \sum_i x_i \\ &\geq \left(\sum_i y_i\right)^2. \end{aligned}$$

Here, the last line follows by Cauchy-Schwarz. ■

Lemma 4. *Let $\hat{x} \in (\Delta_n)_{++}$, $g \in \mathbb{R}^n$, and $\eta > 0$. Define*

$$\theta = \min\left(-\log\left(\sum_i \exp\left(1 + \log(\hat{x}_i) - \eta g_i\right)\right), -1\right), \qquad \textit{and}$$
$$\tilde{x} = \left(\exp(1 + \log(\hat{x}_i) - \eta g_i + \theta)\right)_i.$$

Then, $\tilde{x}$ is the unique minimizer of

$$\min_{x \in \Delta_n} \left\{\langle \eta g - \nabla\omega(\hat{x}), x\rangle + \omega(x)\right\}.$$

Proof. For convenience, set $\hat{g} = \eta g - \nabla\omega(\hat{x})$. Let $\theta \in \mathbb{R}$ to be fixed momentarily and define $\tilde{x} \in \mathbb{R}^n_{++}$ by

$$\tilde{x}_i = \exp(-\hat{g}_i)\cdot\exp(\theta).$$

Note that $\sum_i \tilde{x}_i = \exp(\theta)\cdot\sum_i \exp(-\hat{g}_i)$.

Now, there are two cases. First, suppose $\exp(-1)\sum_i \exp(-\hat{g}_i) \le 1$. Then, we can set $\theta = -1$ and have $\tilde{x} \in (\Delta_n)_{++}$. Note that

$$\begin{aligned}(\hat{g} + \nabla\omega(\tilde{x}))_i &= \hat{g}_i + 1 + \log(\tilde{x}_i) \\ &= 1 + \theta = 0.\end{aligned}$$

We see that $\tilde{x}$ is optimal.

In the second case, $\exp(-1)\sum_i \exp(-\hat{g}_i) > 1$. Set θ so that $\sum_i \tilde{x}_i = 1$. This is achieved by setting $\theta = -\log(\sum_i \exp(-\hat{g}_i)) < -1$. Now, we have $\tilde{x} \in (\Delta_n)_{++}$ and it remains to check that

$$\begin{aligned}(\hat{g} + \nabla\omega(\tilde{x}))_i &= \hat{g}_i + 1 + \log(\tilde{x}_i) \\ &= \theta + 1.\end{aligned}$$

As ω is convex, we deduce that $\tilde{x}$ is optimal. ■

3. This problem improves the Frank–Wolfe convergence rate by assuming that the domain is strongly convex and the objective is strongly convex.

   Fix an arbitrary norm on $\mathbb{R}^n$. We say that a set $\Omega \subseteq \mathbb{R}^n$ is μ-strongly convex if for all $x, y \in \Omega$, $\gamma \in [0,1]$

$$\mathbb{B}((1-\gamma)x + \gamma y, \gamma(1-\gamma)\frac{\mu}{2}\left\|x-y\right\|^2) \subseteq \Omega.$$

   Here, $\mathbb{B}(x_0, r) = \{x \in \mathbb{R}^n : \|x_0 - x\| \le r\}$.

   Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L_f-smooth μ_f-strongly convex function w.r.t. $\|\cdot\|$. Let $\Omega \subseteq \mathbb{R}^n$ be a compact convex set with diameter D. Assume that Ω is μ_Ω-strongly convex.

   Now, consider the following algorithm

Algorithm 1 Frank–Wolfe for strongly convex sets and objectives

Given: $x_0 \in \Omega$

- For $t = 0, \dots,$
  - Set $y_t \in \arg\min_{y\in\Omega} \langle \nabla f(x_t), y_t\rangle$
  - Set $x_{t+1} = (1-\eta_t)x_t + \eta_t y_t$ where

$$\eta_t = \begin{cases} 1 & \text{if } t = 0 \\ 1 & \text{if } L_f \le \frac{\mu_\Omega}{2}\left\|\nabla f(x_t)\right\|_* \\ \frac{\mu_\Omega \|\nabla f(x_t)\|_*}{2L} & \text{else} \end{cases}.$$

Let $\delta_t \coloneqq f(x_t) - f^\star$.

Lemma 5. *It holds that $\delta_1 \le \frac{LD^2}{2}$.*

Proof. We compute

$$\begin{aligned}
\delta_1 &:= f(x_1) - f(x^\star) \\
&= f(y_0) - f(x^\star) \\
&\leq \langle \nabla f(x_0), y_0 - x_0 \rangle + \frac{L}{2} \|x_0 - y_0\|^2 - \langle \nabla f(x_0), x^\star - x_0 \rangle \\
&\leq \frac{LD^2}{2}.
\end{aligned}$$

Here, the second line follows by L-smoothness and convexity, and the last line follows by the optimality of y_0. ■

Lemma 6. *For all $t \geq 1$, it holds that*

$$\langle \nabla f(x_t), x_t - y_t \rangle \geq \frac{\mu_\Omega}{2} \|x_t - y_t\|^2 \|\nabla f(x_t)\|_* .$$

Proof. Let $t \geq 1$.

Let $\tilde{x} = (1-\alpha)x_t + \alpha y_t + \alpha(1-\alpha)\frac{\mu}{2} \|x_t - y_t\|^2 z \in \Omega$ where $\alpha \in [0, 1)$ and $z \in \mathbb{R}^n$ with $\|z\| \leq 1$ will be chosen momentarily.

Then, by optimality of y_t, we have that

$$\begin{aligned}
\langle \nabla f(x_t), y_t \rangle \leq \langle \nabla f(x_t), \tilde{x} \rangle \\
&= (1-\alpha) \langle \nabla f(x_t), x_t \rangle + \alpha \langle \nabla f(x_t), y_t \rangle \\
&\quad + \alpha(1-\alpha)\frac{\mu}{2} \|x_t - y_t\|^2 \langle \nabla f(x_t), z \rangle .
\end{aligned}$$

Subtracting $\alpha \langle \nabla f(x_t), y_t \rangle$ and dividing by $(1-\alpha) > 0$ gives

$$\langle \nabla f(x_t), y_t \rangle \leq \langle \nabla f(x_t), x_t \rangle + \alpha \frac{\mu}{2} \|x_t - y_t\|^2 \langle \nabla f(x_t), z \rangle .$$

We may now take the infimum of the right hand side over z with $\|z\| \leq 1$ and $\alpha \in [0, 1)$ to get:

$$\langle \nabla f(x_t), x_t - y_t \rangle \geq \frac{\mu}{2} \|x_t - y_t\|^2 \|\nabla f(x_t)\|_* .$$ ■

Lemma 7. *For all $t \geq 1$, it holds that*

$$\langle \nabla f(x_t), x_t - y_t \rangle \geq \frac{1}{2}\delta_t + \frac{\mu}{4} \|x_t - y_t\|^2 \|\nabla f(x_t)\|_* .$$

Proof. Recall that

$$\langle \nabla f(x_t), x_t - y_t \rangle \geq f(x_t) - f^\star = \delta_t.$$

The lemma follows by taking the average of this inequality with the previous lemma. ■

Lemma 8. *For all $t \geq 1$,*

$$\delta_{t+1} \leq \max\left(\frac{1}{2}, \left(1 - \frac{\mu \|\nabla f(x_t)\|_*}{4L}\right)\right) \cdot \delta_t.$$

Proof. Let $t \geq 1$. It holds that

$$\begin{aligned}
\delta_{t+1} &\leq \delta_t + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_t - x_{t+1}\|^2 \\
&= \delta_t - \eta_t \langle \nabla f(x_t), x_t - y_t \rangle + \frac{L\eta_t^2}{2} \|x_t - y_t\|^2 \\
&\leq \left(1 - \frac{\eta_t}{2}\right) \delta_t + \frac{\|x_t - y_t\|^2}{2} \left(L\eta_t^2 - \frac{\eta_t \mu}{2} \|\nabla f(x_t)\|_*\right).
\end{aligned}$$

If $L \leq \frac{\mu}{2} \|\nabla f(x_t)\|_*$, then by definition, $\eta_t = 1$ so that

$$\delta_{t+1} \leq \frac{\delta_t}{2} + (\text{something nonpositive}).$$

On the other hand, if $L > \frac{\mu}{2} \|\nabla f(x_t)\|_*$, then by definition, $\eta_t = \frac{\mu\|\nabla f(x_t)\|_*}{2L}$ so that

$$\delta_{t+1} \leq \left(1 - \frac{\mu \|\nabla f(x_t)\|_*}{4L}\right) \delta_t.$$

■

Lemma 9. *Let $0 < \epsilon \ll 1$. Then $\delta_T \leq \epsilon$ for*

$$T = O\left(\frac{L}{\mu_\Omega \sqrt{\mu_f} \sqrt{\epsilon}}\right).$$

Proof. Let $\epsilon > 0$ and consider the sequence

$$\delta_0, \delta_1, \ldots$$

By our previous lemmas, we know that $\delta_1 \leq \frac{LD^2}{2}$ and that the δ_t are nonincreasing. Let T be the smallest index so that $\delta_T \leq \epsilon$. For each index $i \in [1, T-2]$, we will place index i into bin $\mathcal{B}_k$ where

$$\frac{LD^2}{2^{k+1}} < \delta_i \leq \frac{LD^2}{2^k}.$$

The bins are indexed by $k \in \left[1, \left\lfloor \log_2\left(\frac{LD^2}{2\epsilon}\right)\right\rfloor\right]$.

Now, let $k \in \left[1, \left\lfloor \log_2\left(\frac{LD^2}{2\epsilon}\right)\right\rfloor\right]$. We will upper bound the number of indices in $\mathcal{B}_k$. For concreteness, suppose $\mathcal{B}_k = [\ell, r]$. We say an index $t \in [\ell, r]$ is "blue" if $\delta_{t+1} \leq \delta_t/2$. Otherwise, it is "red." We will count $\mathcal{B}_k$ in three parts: blue indices, the singleton $\{r\}$, and the red indices in $[\ell, r-1]$.

There is at most one blue index in $\mathcal{B}_k$. Indeed, if $t \in \mathcal{B}_k$ is blue, then

$$\delta_{t+1} \leq \frac{1}{2}\delta_t \leq \frac{LD^2}{2^{k+1}}.$$

For all red indices $t \in [\ell, r-1]$, we have that

$$\delta_t - \delta_{t+1} \geq \frac{\mu_\Omega \left\|\nabla f(x_t)\right\|_*}{4L}\delta_t.$$

By the μ_f-strong convexity of f, we may bound

$$\delta_t \leq \frac{1}{2\mu_f}\left\|\nabla f(x_t)\right\|_*^2.$$

In particular, every red $t \in [\ell, r-1]$ satisfies

$$\delta_t - \delta_{t+1} \geq \frac{\mu_\Omega \sqrt{\mu_f}}{\sqrt{2}L}\delta_t^{3/2} \geq \frac{\mu_\Omega \sqrt{\mu_f}}{2\sqrt{2}L}\left(\frac{LD^2}{2^{k+1}}\right)^{3/2}.$$

The last inequality follows by $\delta_t > \frac{LD^2}{2^{k+1}}$.
We now sum up these decreases $\delta_t - \delta_{t+1}$ over the red indices $t \in [\ell, r-1]$. We have that

$$\begin{aligned}
\left|\{t \in [\ell, r-1] : \text{red}\}\right| \frac{\mu_\Omega \sqrt{\mu_f}}{2\sqrt{2}L}\left(\frac{LD^2}{2^{k+1}}\right)^{3/2} &\leq \sum_{\substack{t \in [\ell, r-1] \\ \text{red}}} (\delta_t - \delta_{t+1}) \\
&\leq \sum_{t \in [\ell, r-1]} (\delta_t - \delta_{t+1}) \\
&= \delta_\ell - \delta_r \\
&\leq \frac{LD^2}{2^{k+1}}.
\end{aligned}$$

Combining these bounds gives

$$|\mathcal{B}_k| \leq 2 + \left(\frac{2^{k+1}}{LD^2}\right)^{1/2} \frac{2\sqrt{2}L}{\mu_\Omega \sqrt{\mu_f}}.$$

Finally, we count the total number of indices as

$$\begin{aligned}
T &\leq \sum_{k=1}^{\left\lfloor \log_2\left(\frac{LD^2}{2\epsilon}\right)\right\rfloor} \left(2 + \left(\frac{2^{k+1}}{LD^2}\right)^{1/2} \frac{2\sqrt{2}L}{\mu_\Omega \sqrt{\mu_f}}\right) \\
&= O\left(\log_2\left(\frac{LD^2}{2\epsilon}\right)\right) + \frac{2\sqrt{2}L}{\mu_\Omega \sqrt{\mu_f}\sqrt{LD^2}} \sum_{k=1}^{\left\lfloor \log_2\left(\frac{LD^2}{2\epsilon}\right)\right\rfloor} 2^{(k+1)/2} \\
&= O\left(\log_2\left(\frac{LD^2}{2\epsilon}\right)\right) + O\left(\frac{L}{\mu_\Omega \sqrt{\mu_f}\sqrt{\epsilon}}\right).
\end{aligned}$$

For all $\epsilon > 0$ small enough, this bound is dominated by the term on the right. ■