Solutions to Assignment 3 of CPSC 368/516 (Spring'23)

February 21, 2023

1 Problem P1

Problem 1.1. *Consider the generalized negative entropy function $f(x) = \sum_{i=1}^n x_i \log x_i - x_i$ over $\mathbb{R}^n_{>0}$.*

1. *Write the gradient and Hessian of f.*
2. *Prove f is strictly convex.*
3. *Prove that f is not strongly convex with respect to the ℓ_2-norm.*
4. *Write the Bregman divergence D_f. Is $D_f(x, y) = D_f(y, x)$ for all $x, y \in \mathbb{R}^n_{>0}$?*
5. *Prove that f is 1-strongly convex with respect to ℓ_1-norm when restricted to points in the subdomain $\{x \in \mathbb{R}^n_{>0} : \sum_{i=1}^n x_i = 1\}$.*

Part 1. For any $i \in [n]$, the partial derivative $\frac{\partial f(x)}{\partial x_i}$ is $\log x_i$. Thus, the gradient is

$$\nabla f(x) = [\log x_1, \log x_2, \ldots, \log x_n]^\top .$$

For any $i, j \in [n]$, the partial derivative of $f(x)$ with respect to x_i and x_j is

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \begin{cases} x_i^{-1} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the Hessian is the diagonal matrix

$$\nabla^2 f(x) = \operatorname{diag}(x_1^{-1}, x_2^{-1}, \ldots, x_n^{-1}). \tag{1}$$

Part 2. Equation (1) shows that the eigenvalues of $\nabla^2 f(x)$ are $x_1^{-1}, x_2^{-1}, \ldots, x_n^{-1}$, which are strictly positive for any $x \in \mathbb{R}^n_{>0}$. Because the eigenvalues of the Hessian are strictly positive for all points in $\mathbb{R}^n_{>0}$ (the domain of $f(x)$), it follows that $f(x)$ is strictly convex.

Part 3.

Proposition 1.2. *Suppose $K \subseteq \mathbb{R}^n$ is convex and open. If $f \colon K \to \mathbb{R}$ is twice differentiable and λ-strongly convex with respect to the ℓ_2-norm, then for all $x \in K$, $\nabla^2 f(x) \succeq \lambda I$.*

Proof. Suppose $f \colon K \to \mathbb{R}$ is twice differentiable and λ-strongly convex, for some $\lambda > 0$. For any $x \in K$ and any unit vector $v \in \mathbb{R}^n$, since K is open, there is a $\tau > 0$ such that $x + tv \in K$ for all $0 < t \leq \tau$. Fix any $0 < t \leq \tau$. Because f is λ-strongly convex over K and both $x, x + tv \in K$, we have that

$$f(x + tv) \geq f(x) + \langle \nabla f(x), tv \rangle + \frac{\lambda}{2} \|tv\|^2 ,$$
$$f(x) \geq f(x + tv) + \langle \nabla f(x + tv), -tv \rangle + \frac{\lambda}{2} \|tv\|^2 .$$

Adding the above inequalities, we get

$$\langle \nabla f(x+tv) - \nabla f(x), tv \rangle \geq \lambda \|tv\|^2 = \lambda t^2. \qquad \text{(Using that } v \text{ is a unit vector)} \quad (2)$$

Observe that $\langle \nabla f(x), v \rangle = Df(x)[v]$ and $\langle \nabla f(x+tv), v \rangle = Df(x+tv)[v]$. Substituting this in Equation (2), we get

$$t \cdot (Df(x+tv)[v] - Df(x)[v]) \geq \lambda t^2 \overset{(t>0)}{\Longrightarrow} \frac{1}{t}(Df(x+tv)[v] - Df(x)[v]) \geq \lambda.$$

By letting $t \to 0$ in the above equation, we get

$$D^2 f(x)[v,v] \coloneqq \lim_{t \to 0} \frac{1}{t}(Df(x+tv)[v] - Df(x)[v]) \geq \lambda.$$

Because $D^2 f(x)[v,v] = v^\top \nabla^2 f(x) v$, we have that $v^\top \nabla^2 f(x) v \geq \lambda$. Since $v \in \mathbb{R}^n$ is an arbitrary unit vector and x an arbitrary point in K, it follows that $\nabla^2 f(x) \succeq \lambda I$ for any $x \in K$. $\square$

Toward a contradiction suppose that f is λ-strongly convex with respect to the ℓ_2-norm for some $\lambda > 0$. Since the domain of f ($\mathbb{R}^n_{>0}$) is a convex and open set and f is twice differentiable, Proposition 1.2 implies that if f λ-strongly convex then for all $x \in \mathbb{R}^n_{>0}$, $\nabla^2 f(x) \succeq \lambda I$. Consider the point $x_\lambda = [2/\lambda, 2/\lambda, \ldots, 2/\lambda]^\top \in \mathbb{R}^n_{>0}$. From Equation (1), we know the all eigenvalues of $\nabla^2 f(x_\lambda)$ are equal to $\lambda/2$, and hence, $\nabla^2 f(x_\lambda) \preceq 0.5\lambda I$. We have a contradiction, and so, f cannot be λ-strongly convex with respect to the ℓ_2-norm for any $\lambda > 0$.

Part 4. The Bregman divergence $D_f(x,y)$ at some points $x, y \in \mathbb{R}^n_{>0}$ is

$$\begin{aligned} D_f(x,y) &\coloneqq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= \sum_{i=1}^n y_i \log y_i - y_i - \left(\sum_{i=1}^n x_i \log x_i - x_i \right) + \sum_{i=1}^n (x_i - y_i) \log x_i \\ &= \sum_{i=1}^n y_i \log \left(\frac{y_i}{e x_i} \right) + \sum_{i=1}^n x_i. \end{aligned} \quad (3)$$

Next, we will prove that there exist points $x, y \in \mathbb{R}^n_{>0}$ such that $D_f(x,y) \neq D_f(y,x)$: Consider any point $x \in \mathbb{R}^n_{>0}$ and let $y = ex$. Note that y is also a point in $\mathbb{R}^n_{>0}$. From Equation (3), we have that

$$\begin{aligned} D_f(x,y) &= \sum_{i=1}^n e x_i \log(1) + \sum_{i=1}^n x_i = \sum_{i=1}^n x_i, \\ D_f(y,x) &= \sum_{i=1}^n x_i \log \left(\frac{x_i}{e^2 x_i} \right) + \sum_{i=1}^n e x_i = (e-2) \cdot \sum_{i=1}^n x_i. \end{aligned}$$

Because $x_i > 0$ for all $i \in [n]$ and $e - 2 \in (0,1)$, we have that $D_f(x,y) > D_f(y,x) > 0$.

Part 5. We will use Pinsker's inequality in our proof.

Fact 1.3 (**Pinsker's inequality**). *For any $x, y \in \mathbb{R}^n_{\geq 0}$, if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$, then*

$$\sum_{i=1}^n y_i \log \frac{y_i}{x_i} \geq \frac{1}{2} \left(\sum_{i=1}^n |y_i - x_i| \right)^2.$$

Let $K \coloneqq \{x \in \mathbb{R}^n_{>0} : \sum_{i=1}^n x_i = 1\}$. Fix any points $x, y \in K$. We need to prove that the following inequality holds

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2} \|y - x\|_1^2.$$

Equivalently, we have to prove that

$$\sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i \geq \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n x_i + \sum_{i=1}^n (y_i - x_i) \cdot \log x_i + \frac{1}{2}\left(\sum_{i=1}^n |y_i - x_i|\right)^2$$

$$\iff \qquad \sum_{i=1}^n y_i \log y_i \geq \sum_{i=1}^n x_i \log x_i + \sum_{i=1}^n (y_i - x_i) \cdot \log x_i + \frac{1}{2}\left(\sum_{i=1}^n |y_i - x_i|\right)^2$$

(Using that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$)

$$\iff \qquad \sum_{i=1}^n y_i \log \frac{y_i}{x_i} \geq \frac{1}{2}\left(\sum_{i=1}^n |y_i - x_i|\right)^2. \tag{4}$$

Since for all $x, y \in K$, $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i = 1$, Fact 1.3 implies that Inequality (4) holds for all $x, y \in K$.

2 Problem P2

Problem 2.1. *Consider the following subset of* $\mathbb{R}^n$

$$P := \{x \in \mathbb{R}^n : \ |\langle a_i, x\rangle| \leq 1 \quad \text{for } i = 1, 2, \ldots, m\},$$

where $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ *are vectors. Let* d *denote the dimension of the linear subspace of* $\mathbb{R}^n$ *spanned by* $a_1, a_2, \ldots, a_m$*. We define a function*

$$F(x) := -\sum_{i=1}^m \log(1 - \langle a_i, x\rangle^2)$$

for all $x \in \mathbb{R}^n$ *where the above formula makes sense and set* $F(x) = +\infty$ *otherwise.*

1. *Prove that the set* P *is bounded*[1] *if and only if* $d = n$*.*
2. *Compute the gradient* $g(x)$ *and the Hessian* $H(x)$ *of* F*.*
3. *Prove that* F *is a convex function. What is the domain of* F *(the set of points where* F *is finite)?*
4. *What is the global minimum of* F*? (Assume* $d = n$*.)*
5. *For any* x *in the domain of* F *define* $\mathcal{E}_x := \{h \in \mathbb{R}^n : h^\top H(x) h \leq 1\}$*. Prove that* $\mathcal{E}_x$ *is a convex set and that* $\mathcal{E}_x \subseteq P$*.*

2.1 Part 1

Note that because $a_1, \ldots, a_m$ are vectors in $\mathbb{R}^n$ they cannot span a linear subspace of dimension higher than n, i.e., $d \leq n$. We consider two cases: $d < n$ and $d = n$.

Case A ($d < n$)**:** In this case, $a_1, a_2, \ldots, a_m$ do not span $\mathbb{R}^n$, and hence, there exists a vector $v \in \mathbb{R}^n$ that is linearly independent of $a_1, a_2, \ldots, a_m$. Without loss of generality assume that v is a unit vector. Toward a contradiction suppose that P is bounded. Then there must exist a $r > 0$ such that for every $x \in P$, $\|x\|_2 \leq r$. Consider the vector $(r+1)v$. We claim that $(r+1)v \in P$. This can be verified as follows: Since v is linearly independent of $a_1, a_2, \ldots, a_m$, for every $i \in [m]$, we have that $|\langle a_i, (r+1)v\rangle| = 0 < 1$, and hence, $(r+1)v \in P$. But we have a contradiction because $\|(r+1)v\|_2^2 = (r+1)^2 > r^2$. Thus, in this case, P is unbounded.

Case B ($d = n$)**:** Let $A \in \mathbb{R}^{n \times m}$ be the $n \times m$ matrix whose i-th column is a_i for all $i \in [m]$. Because $a_1, a_2, \ldots, a_m$ span a subspace of dimension n and $m \geq n$, it follows that A is a full-rank matrix. This implies

[1] A set $K \subseteq \mathbb{R}^n$ is called bounded if there exists an $r > 0$ such that $\|x\| \leq r$ for every $x \in K$.

that $AA^\top$ is PD. Thus, the smallest eigenvalue, say $\lambda_{\min}$, of $AA^\top$ is strictly positive. Let $\lambda_{\min} > 0$ be the smallest eigenvalue of $AA^\top$. Consider any vector $x \in P$. Then we have that

$$\begin{aligned}
n &\geq \sum_{i=1}^{m} \langle a_i, x\rangle^2 && \text{(Using that for all } x \in P \text{ and } i \in [m],\ |\langle a_i, x\rangle| \leq 1) \\
&= \left\|A^\top x\right\|_2^2 \\
&= x^\top A A^\top x \\
&\geq \lambda_{\min} \left\|x\right\|_2^2. && (\lambda_{\min} \text{ is the smallest eigenvalue of } AA^\top.)
\end{aligned}$$

On rearranging the inequality (using that $\lambda_{\min} > 0$), we get that

$$\left\|x\right\|_2 \leq \sqrt{\frac{n}{\lambda_{\min}}}.$$

Since x was an arbitrary point in P, it follows that P is bounded.

2.2 Part 2

Note that $F(x)$ is the function:

$$F(x) \coloneqq \begin{cases} \sum_{i=1}^{m} -\log(1 - \langle a_i, x\rangle^2) & \text{if for all } i \in [m], \quad \langle a_i, x\rangle^2 < 1, \\ \infty & \text{otherwise.} \end{cases} \tag{5}$$

We will compute the gradient and Hessian of F at any point x where F is finite, i.e., x satisfies that for all $i \in [m]$, $\langle a_i, x\rangle^2 < 1$. Using the chain rule, for all $i \in [n]$, we have that

$$\begin{aligned}
\frac{\partial F(x)}{\partial x_i} &= \frac{\partial}{\partial x_i} \sum_{k=1}^{m} - \log(1 - \langle a_k, x\rangle^2) \\
&= \sum_{k=1}^{m} \frac{2\langle a_k, x\rangle}{1 - \langle a_k, x\rangle^2} a_{ki}.
\end{aligned} \tag{6}$$

Thus, the gradient is

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_n}(x)\right]^\top = \sum_{k=1}^{m} \frac{2\langle a_k, x\rangle}{1 - \langle a_k, x\rangle^2} a_k.$$

For any $i, j \in [n]$, we can compute $\frac{\partial^2 F}{\partial x_i \partial x_j}(x)$ as follows

$$\begin{aligned}
\frac{\partial^2 F}{\partial x_i \partial x_j}(x) &\overset{(6)}{=} \frac{\partial}{\partial x_j} \sum_{k=1}^{m} \frac{2\langle a_k, x\rangle}{1 - \langle a_k, x\rangle^2} a_{ki} \\
&= \frac{\partial}{\partial x_j} \sum_{k=1}^{m} \left(\frac{1}{1 - \langle a_k, x\rangle} - \frac{1}{1 + \langle a_k, x\rangle}\right) a_{ki} \\
&= \sum_{k=1}^{m} \left(\frac{a_{kj}}{(1 - \langle a_k, x\rangle)^2} + \frac{a_{kj}}{(1 + \langle a_k, x\rangle)^2}\right) a_{ki} \\
&= \sum_{k=1}^{m} \left(\frac{1}{(1 - \langle a_k, x\rangle)^2} + \frac{1}{(1 + \langle a_k, x\rangle)^2}\right) a_{ki} a_{kj}.
\end{aligned}$$

Thus, the Hessian is

$$\nabla^2 F(x) = \sum_{k=1}^{m} \left(\frac{1}{(1 - \langle a_k, x\rangle)^2} + \frac{1}{(1 + \langle a_k, x\rangle)^2}\right) a_k a_k^\top. \tag{7}$$

2.3 Part 3

From Equation (5) it follows that F is finite at a point $x \in \mathbb{R}^n$ if and only if for all $i \in [m]$, $\langle a_i, x\rangle^2 < 1$; or equivalently if for all $i \in [m]$, $|\langle a_i, x\rangle| < 1$. Thus, the set of points where F is finite is

$$S \coloneqq \{x \in \mathbb{R}^n : \text{for all } i \in [m],\ |\langle a_i, x\rangle| < 1\} = \operatorname{Int}(P).$$

Because S is an interior of another set, S is open. Since P is defined by linear-inequality constraints, it follows that it is an intersection of half-spaces, and hence, is convex. Thus, S (which is the interior of P) is also convex. Because S is a convex and open set and $F(x)$ is twice differentiable on S, if we can prove that $\nabla^2 F(x)$ is PSD for all $x \in S$, then it follows that F is convex on S.

Towards this, consider any $x \in S$ and $u \in \mathbb{R}^n$, we have that

$$\begin{aligned}
u^\top \nabla^2 F(x) u &= \sum_{k=1}^{m} \left(\frac{1}{(1-\langle a_k, x\rangle)^2} + \frac{1}{(1+\langle a_k, x\rangle)^2} \right) u^\top a_k a_k^\top u \\
&= \sum_{k=1}^{m} \left(\frac{1}{(1-\langle a_k, x\rangle)^2} + \frac{1}{(1+\langle a_k, x\rangle)^2} \right) \langle a_k, u\rangle^2 \\
&\geq 0. \qquad \left(\text{Using that } \langle a_k, u\rangle^2 \geq 0 \text{ and that for all } x \in \mathbb{R},\ \tfrac{1}{(1-x)^2} + \tfrac{1}{(1+x)^2} > 0\right)
\end{aligned} \tag{8}$$

Equation (8) implies that for all $x \in S$, $\nabla^2 F(x)$ is PSD. Thus, it follows that the restriction of F to S is a convex function. This implies that for any $\lambda \in [0, 1]$ and $x, y \in S$, it holds that

$$\lambda F(x) + (1-\lambda)F(y) \geq F(\lambda x + (1-\lambda)y). \tag{9}$$

Next, we show that F is convex over all of $\mathbb{R}^n$. Consider any $\lambda \in [0, 1]$ and $x, y \in \mathbb{R}^n$. We analyze four cases:

Case A ($x \in S, y \in S$): From Equation (9) we have that $\lambda F(x) + (1-\lambda)F(y) \geq F(\lambda x + (1-\lambda)y)$.

Case B ($x \in S, y \notin S$): We have that $F(y) = \infty$. If $\lambda < 1$, then it holds that

$$\begin{aligned}
\lambda F(x) + (1-\lambda)F(y) &\overset{(y \notin S)}{=} \lambda F(x) + (1-\lambda)\infty \\
&\overset{(1-\lambda > 0)}{=} \infty \\
&\geq F(\lambda x + (1-\lambda)y).
\end{aligned}$$

If $\lambda = 1$, then $\lambda F(x) + (1-\lambda)F(y) = F(x) = F(\lambda x + (1-\lambda)y)$.

Case C ($x \notin S, y \in S$): This case is similar to Case B. By swapping x and y in the calculation in Case B, we get that for all $\lambda \in [0, 1]$: $\lambda F(x) + (1-\lambda)F(y) \geq F(\lambda x + (1-\lambda)y)$.

Case D ($x \notin S, y \notin S$): In this case, both $F(x)$ and $F(y)$ are infinite. Thus, we have that for all $\lambda \in [0, 1]$: $\lambda F(x) + (1-\lambda)F(y) = F(\lambda x + (1-\lambda)y)$.

Combining the results from all four cases, it follows that F is convex on all of $\mathbb{R}^n$.

2.4 Part 4

We can lower bound F as follows

$$\begin{aligned}
F(x) &\coloneqq \begin{cases} \sum_{i=1}^{m} -\log(1 - \langle a_i, x\rangle^2) & \text{if } x \in S, \\ \infty & \text{otherwise.} \end{cases} \\
&\geq \begin{cases} 0 & \text{if } x \in S, \\ \infty & \text{otherwise.} \end{cases} \\
&\qquad \left(\text{Using that for all } 0 < z \leq 1,\ -\log z \geq 0 \text{ and that } 0 < 1 - \langle a_i, x\rangle^2 \leq 1 \text{ for all } x \in S\right) \\
&\geq 0.
\end{aligned}$$

Then, since $F(0) = \sum_{i=1}^{m} -\log(1 - \langle a_i, 0\rangle^2) = 0$ and for all $x \in \mathbb{R}^n$, we get that 0 is a global minimum of F. (One can also show that 0 is the unique global minimizer.)

2.5 Part 5

$\mathcal{E}_\mathbf{x}$ is convex for all $\mathbf{x} \in \mathbf{S}$. To prove that $\mathcal{E}_x$ is convex for all $x \in S$, we need to show that for all $\lambda \in [0,1]$ and $u, v \in S$,

$$\lambda u + (1-\lambda)v \in \mathcal{E}_x \quad \iff \quad (\lambda u + (1-\lambda)v)^\top H(x)\,(\lambda u + (1-\lambda)v) \le 1.$$

From Part 3, recall that for any $x \in S$, $H(x)$ is PSD. Combining this with the fact that for any PSD matrix A the function $f(u) \coloneqq \sqrt{u^\top A u}$ satisfies the triangle inequality (Problem 2.13), we get that

$$\begin{aligned}
\sqrt{(\lambda u + (1-\lambda)v)^\top H(x)\,(\lambda u + (1-\lambda)v)} &\le \sqrt{(\lambda u)^\top H(x)\,(\lambda u)} + \sqrt{((1-\lambda)v)^\top H(x)\,((1-\lambda)v)} \\
&= |\lambda| \sqrt{u^\top H(x) u} + |1-\lambda| \sqrt{v^\top H(x) v}.
\end{aligned}$$

Squaring both sides, we get that

$$\begin{aligned}
(\lambda u + (1-\lambda)v)^\top H(x)\,(\lambda u + (1-\lambda)v) &\le \lambda^2 u^\top H(x) u + (1-\lambda)^2 v^\top H(x) v + 2\,|\lambda(1-\lambda)| \sqrt{u^\top H(x) u} \cdot \sqrt{v^\top H(x) v} \\
&\le \lambda^2 + (1-\lambda)^2 + 2\,|\lambda(1-\lambda)| && \text{(Using that } x, y \in \mathcal{E}_x\text{)} \\
&= 1. && \text{(Using that } \lambda \in [0,1]\text{)}
\end{aligned}$$

Thus, $\lambda u + (1-\lambda)v \in \mathcal{E}_x$.

$\mathcal{E}_\mathbf{x} \subseteq \mathbf{P}$ for all $\mathbf{x} \in \mathbf{S}$. Fix any $x \in S$ and any $h \in \mathcal{E}_x$. It suffices to show that $h \in P$, i.e., for all $i \in [m]$, $\langle h, a_i\rangle \le 1$. We will prove a stronger statement that $\sum_{i=1}^{m} 2\langle h, a_i\rangle^2 < 1$. First, note that since $h \in \mathcal{E}_x$ we have that

$$h^\top H(x) h \le 1. \tag{10}$$

We can show that $\sum_{i=1}^{m} 2\langle h, a_i\rangle^2 \le h^\top H(x) h$ as follows

$$\begin{aligned}
h^\top H(x) h &= h^\top \left(\sum_{i=1}^{m} \left(\frac{1}{(1 - \langle a_k, x\rangle)^2} + \frac{1}{(1 + \langle a_k, x\rangle)^2} \right) a_k a_k^\top \right) h \\
&= \sum_{i=1}^{m} \left(\frac{1}{(1 - \langle a_k, x\rangle)^2} + \frac{1}{(1 + \langle a_k, x\rangle)^2} \right) \langle a_k, h\rangle^2 \\
&\ge \sum_{i=1}^{m} 2\langle a_k, h\rangle^2
\end{aligned}$$

$$\text{(Using that for all } x \in S,\ |\langle a_k, x\rangle| < 1 \text{ and that for all } x \in [-1,1],\ \tfrac{1}{(1-x)^2} + \tfrac{1}{(1+x)^2} > 2) \tag{11}$$

We get the required result by chaining the Inequalities (10) and (11).