The University of Texas at Austin
Department of Electrical and Computer Engineering

EE381K: Large Scale Optimization — Fall 2015

Problem Set Eight Solutions

Constantine Caramanis

Written Problems

1. *Show that sub gradients have the following properties:*[1]

(a) $\partial(\alpha f(x)) = \alpha \partial f(x)$.
Solution For $\alpha = 0$, $\partial(\alpha f(x)) = \partial(0) = \{0\} = 0 \cdot \partial f(x)$. For $\alpha \neq 0$, we have

$$\begin{aligned} & x^* \in \alpha \partial f(x) \\ & \quad \Leftrightarrow \; x^*/\alpha \in \partial f(x) \\ & \quad \Leftrightarrow \; f(y) \geq f(x) + \langle x^*/\alpha, y - x \rangle, \; \forall y \in \mathbf{dom} f \\ & \quad \Leftrightarrow \; \alpha f(y) \geq \alpha f(x) + \langle x^*, y - x \rangle, \; \forall y \in \mathbf{dom} f \\ & \quad \Leftrightarrow \; x^* \in \partial(\alpha f(x)). \end{aligned}$$

(b) $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
Solution
Proof of $\partial(f_1 + f_2) \supseteq \partial f_1 + \partial f_2$

$$\begin{aligned} & x_1^* \in \partial f_1(x), \; x_2^* \in \partial f_2(x) \\ & \quad \Rightarrow f_1(y) \geq f_1(x) + \langle x_1^*, y - x \rangle, \; \forall y \in \mathbf{dom} f_1, \\ & \qquad f_2(y) \geq f_2(x) + \langle x_2^*, y - x \rangle, \; \forall y \in \mathbf{dom} f_2 \\ & \quad \Rightarrow f_1(y) + f_2(y) \geq f_1(x) + f_2(x) + \langle x_1^* + x_2^*, y - x \rangle, \; \forall y \in \mathbf{dom}(f_1 + f_2) \\ & \quad \Rightarrow x_1^* + x_2^* \in \partial(f_1(x) + f_2(x)) \end{aligned}$$

Proof of $\partial(f_1 + f_2) \subseteq \partial f_1 + \partial f_2$
Suppose $x^* \in \partial(f_1(x) + f_2(x))$, i.e., $f_1(z) + f_2(z) - \langle x^*, z \rangle \geq f_1(x) + f_2(x) - \langle x^*, x \rangle$ for every $z \in \mathbf{dom} f_1 \cap \mathbf{dom} f_2$. We will show that there exist vectors $x_1^* \in \partial f_1(x)$ and $x_2^* \in \partial f_2(x)$ such that $x^* = x_1^* + x_2^*$. To this end, we need an assumption $\mathbf{relint}(\mathbf{dom} f_1) \cap \mathbf{relint}(\mathbf{dom} f_2) \neq \emptyset$.

Consider a constrained optimization problem

$$\begin{aligned} \underset{z_1, z_2}{\text{minimize}} \quad & f_1(z_1) + f_2(z_2) - \langle x^*, z_1 \rangle \\ \text{subject to} \quad & z_1 = z_2, \; z_1 \in \mathbf{dom} f_1, \; z_2 \in \mathbf{dom} f_2 \end{aligned}$$

[1] We don't need convexity to define sub gradients, but we did use it in the definitions we gave in class. Therefore, in the problems below, you can assume that all functions are convex, coefficients nonnegative, etc.

where $z_1 = z_2 = x$ solves the problem. The Lagrangian is given by

$$\begin{aligned} L(z_1, z_2; \lambda) &= f_1(z_1) + f_2(z_2) - \langle x^*, z_1 \rangle + \lambda^\top (z_1 - z_2) \\ &= f_1(z_1) - \langle x^* - \lambda, z_1 \rangle + f_2(z_2) - \langle \lambda, z_2 \rangle. \end{aligned}$$

on $z_1 \in \mathbf{dom} f_1$ and $z_2 \in \mathbf{dom} f_2$. Let λ^* be the dual optimum. It follows from the assumption that the strong duality holds, and thus we have

$$\begin{aligned} L(z_1, x; \lambda^*) \geq L(x, x; \lambda^*), \quad \forall z_1 \in \mathbf{dom} f_1, \\ L(x, z_2; \lambda^*) \geq L(x, x; \lambda^*), \quad \forall z_2 \in \mathbf{dom} f_2, \end{aligned}$$

which means that

$$\begin{aligned} f_1(z_1) - \langle x^* - \lambda^*, z_1 \rangle \geq f_1(x) - \langle x^* - \lambda^*, x \rangle, \quad \forall z_1 \in \mathbf{dom} f_1, \\ f_2(z_2) - \langle \lambda^*, z_2 \rangle \geq f_2(x) - \langle \lambda^*, x \rangle, \quad \forall z_2 \in \mathbf{dom} f_2. \end{aligned}$$

Therefore, $x^* - \lambda^* \in \partial f_1(x)$, and $\lambda^* \in \partial f_2(x)$. This completes the proof.

(c) *If $g(x) = f(Ax + b)$, then $\partial g(x) = A^\top \partial f(Ax + b)$.*

Solution

Proof of $\partial g(x) \supseteq A^\top \partial f(Ax + b)$

$$\begin{aligned} & x^* \in \partial f(Ax + b) \\ & \Rightarrow \ f(y) \geq f(Ax + b) + \langle x^*, y - Ax - b \rangle, \ \forall y \in \mathbf{dom} f \\ & \Rightarrow \ f(Az + b) \geq f(Ax + b) + \langle x^*, Az + b - Ax - b \rangle \\ & \qquad\qquad = f(Ax + b) + \langle x^*, A(z - x) \rangle \\ & \qquad\qquad = f(Ax + b) + \langle A^\top x^*, z - x \rangle, \ \forall z \in \mathbf{dom} g = \{x : Ax + b \in \mathbf{dom} f\} \\ & \Rightarrow \ A^\top x^* \in \partial g(x) \end{aligned}$$

Proof of $\partial g(x) \subseteq A^\top \partial f(Ax + b)$

Suppose $x^* \in \partial g(x)$, i.e., $f(Az + b) - \langle x^*, z \rangle \geq f(Ax + b) - \langle x^*, x \rangle$ for every $z \in \mathbf{dom} g$. We will show that there exists a vector $\lambda^* \in \partial f(Ax + b)$ such that $x^* = A^\top \lambda^*$. To this end, we need an assumption $(\text{Range}(A) + b) \cap \mathbf{relint}(\mathbf{dom} f) \neq \emptyset$.

Consider a constrained optimization problem

$$\begin{array}{cl} \underset{y \in \mathbf{dom} f,\ z}{\text{minimize}} & f(y) - \langle x^*, z \rangle \\ \text{subject to} & y = Az + b \end{array}$$

where $(Ax + b, x)$ are the optimum for (y, z). The Lagrangian is given by

$$\begin{aligned} L(y, z; \lambda) &= f(y) - \langle x^*, z \rangle + \lambda^\top (Az + b - y) \\ &= f(y) - \langle \lambda, y \rangle + \langle A^\top \lambda - x^*, z \rangle + \langle \lambda, b \rangle. \end{aligned}$$

on $y \in \mathbf{dom} f$. Let λ^* be the dual optimum. Due to the assumption, the strong duality holds, and thus $y = Ax + b$ and $z = x$ also minimize $L(y, z; \lambda^*)$. As $L(y, z; \lambda)$ is smooth over z, we have

$$\nabla_z L = 0 \quad \Rightarrow \quad x^* = A^\top \lambda^* \tag{1}$$

We also have $L(y, x; \lambda^*) \geq L(Ax + b, x; \lambda^*)$ for every $y \in \mathbf{dom} f$. Applying (1) and subtracting $\langle \lambda, b \rangle$ from the both sides, we get

$$f(y) - \langle \lambda^*, y \rangle \geq f(Ax + b) - \langle \lambda^*, Ax + b \rangle, \quad \forall y \in \mathbf{dom} f$$

Therefore, $\lambda^* \in \partial f(Ax + b)$. This completes the proof.

(d) *If* $f(x) = \max_{1 \leq i \leq m} f_i(x)$, *then*

$$\partial f(x) = \text{conv} \bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}.$$

Solution $\underline{\partial f(x) \supseteq \text{conv} \bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}}$
Suppose $x^* = \sum_{i=1}^m \lambda_i x_i^*$ where $x_i^* \in \partial f_i(x)$, $\sum_{i=1}^m \lambda_i = 1$, and $\lambda_i = 0$ for $f_i(x) \neq f(x)$. This means that $x^* \in \text{conv} \bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}$. Then it follows that

$$\begin{aligned} f(y) &= \sum_{i=1}^m \lambda_i f(y) = \sum_{i=1}^m \lambda_i f_i(y) \\ &\geq \sum_{i=1}^m \lambda_i \left(f_i(x) + \langle x_i^*, y - x \rangle \right) \\ &= \max_{1 \leq i \leq m} f_i(x) + \langle x^*, y - x \rangle, \quad y \in \mathbf{dom} f. \end{aligned}$$

$\underline{\partial f(x) \subseteq \text{conv} \bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}}$
Suppose $x^* \in \partial f(x)$, i.e., $f(z) - \langle x^*, z \rangle \geq f(x) - \langle x^*, x \rangle$ for every $z \in \mathbf{dom} f$. Consider a constrained optimization problem

$$\begin{aligned} \underset{y,z}{\text{minimize}} \quad & y - \langle x^*, z \rangle \\ \text{subject to} \quad & y \geq f_i(z),\ i = 1, \ldots, m \end{aligned}$$

where $(f(x), x)$ is the optimum for (y, z). The Lagrangian is given by

$$\begin{aligned} L(y, z; \lambda) &= y - \langle x^*, z \rangle + \sum_{i=1}^m \lambda_i (f_i(z) - y) \\ &= y \left(1 - \sum_{i=1}^m \lambda_i \right) + \sum_{i=1}^m \lambda_i f_i(z) - \langle x^*, z \rangle \end{aligned}$$

where $\lambda_i \geq 0$ for $i = 1, \ldots, m$. Let λ^* be the dual optimum. It follows from the strong duality that $y = f(x)$ and $z = x$ also minimize $L(y, z; \lambda^*)$. As $L(y, z; \lambda)$ is smooth over y and λ, we have

$$\begin{aligned} \nabla_y L = 0 \quad & \Rightarrow \quad \sum_{i=1}^m \lambda_i = 1, \\ \lambda_i^* (f_i(x) - f(x)) = 0 \quad & \Rightarrow \quad \lambda_i^* = 0 \text{ or } f(x) = f_i(x) \end{aligned}$$

We also have $L(f(x), z; \lambda^*) \geq L(f(x), x; \lambda^*)$ for every $z \in \mathbf{dom} f$. Then it follows that

$$\sum_{i=1}^{m} \lambda_i^* f_i(z) - \langle x^*, z \rangle \geq \sum_{i=1}^{m} \lambda_i^* f_i(x) - \langle x^*, x \rangle, \quad \forall z \in \mathbf{dom} f$$

Therefore, $x^* \in \partial \left(\sum_{i=1}^m \lambda_i^* f_i(x) \right) = \lambda_1^* \partial f_1(x) + \ldots + \lambda_m^* \partial f_m(x) \subseteq \operatorname{conv} \bigcup_i \{\partial f_i(x),\ f_i(x) = f(x)\}$. This completes the proof.

2. *Compute the sub gradient of the $\|\cdot\|_{2,1}$ norm on matrices: For M a matrix with columns M_i, this is defined as:*

$$\|M\|_{2,1} = \sum_i \|M_i\|_2.$$

Solution Applying the property in Problem 1(d), we obtain

$$\begin{aligned}
\partial \|M_i\|_2 &= \partial \left(\sup_{\|x\| \leq 1} x^\top M_i \right) \\
&= \operatorname{conv} \left\{ \partial \left(x^\top M_i \right) : x^\top M_i = \|M_i\|_2,\ \|x\| \leq 1 \right\} \\
&= \operatorname{conv} \left\{ x e_i^\top : x^\top M_i = \|M_i\|_2,\ \|x\| \leq 1 \right\} \\
&= \left\{ x e_1^\top : x = \frac{M_i}{\|M_i\|} \text{ if } M_i \neq 0,\ \|x\|_2 \leq 1 \text{ if } M_i = 0 \right\}
\end{aligned}$$

Then it follows that

$$\begin{aligned}
\partial \|M\|_{2,1} &= \partial \left(\sum_{i=1}^{m} \|M_i\|_2 \right) \\
&= \left\{ M_1^* e_1^\top : M_1^* = \frac{M_1}{\|M_1\|} \text{ if } M_1 \neq 0,\ \|M_1^*\|_2 \leq 1 \text{ if } M_1 = 0 \right\} \\
&\quad + \ldots + \left\{ M_n^* e_n^\top : M_n^* = \frac{M_n}{\|M_n\|} \text{ if } M_n \neq 0,\ \|M_n^*\|_2 \leq 1 \text{ if } M_n = 0 \right\} \\
&= \{ M^* : M_i^* = M_i / \|M_i\|_2 \text{ if } M_i \neq 0, \|M_i^*\|_2 \leq 1 \text{ if } M_i = 0 \}
\end{aligned}$$

3. *(more tricky) Suppose $A_0, A_1, \ldots, A_m$ are symmetric matrices. Consider the function*

$$f(x) = \lambda_{\max}(A(x)),$$

where

$$A(x) = A_0 + x_1 A_1 + \cdots + x_m A_m.$$

Compute the sub gradient of $f(x)$. Hint: use the fact that

$$f(x) = \sup_{\|y\|_2 = 1} y^\top A(x) y,$$

and the last property you proved from the first problem.

Solution $y^\top A(x)y$ for a fixed y is an affine function of x, so we have $\partial(y^\top A(x)y) = (y^\top A_1 y, \dots, y^\top A_m y)^\top$. Using the property in Problem 1(d), we obtain

$$\partial f(x) = \operatorname{conv}\left\{ \begin{bmatrix} y^\top A_1 y \\ \vdots \\ y^\top A_m y \end{bmatrix} : A(x)y = \lambda_{\max}(A(x))y, \|y\| = 1 \right\}.$$

4. *The indicator function of a set $\mathcal{X}$ is defined as:*

$$I_{\mathcal{X}}(x) = \begin{cases} 0 & x \in \mathcal{X} \\ +\infty & \textit{otherwise.} \end{cases}$$

(a) *Show that the sub differential of $I_{\mathcal{X}}$ is the normal cone to $\mathcal{X}$ at the point x.*

Solution The subdifferential of $f(x) = I_{\mathcal{X}}(x)$ for $x \in \mathcal{X}$ is the set of x^* such that

$$\begin{aligned} \langle x^*, x\rangle &= f(x) + f^*(x^*) \\ &= I_{\mathcal{X}}(x) + \sup_z\{\langle x^*, z\rangle - I_{\mathcal{X}}(z)\} \\ &= \sup_{z\in\mathcal{X}}\langle x^*, z\rangle \end{aligned}$$

This is identical to the set of x^* such that for any $\hat{x} \in \mathcal{X}$

$$\langle x^*, \hat{x} - x\rangle \le \sup_{z\in\mathcal{X}}\langle x^*, z - x\rangle = \sup_{z\in\mathcal{X}}\langle x^*, z\rangle - \langle x^*, x\rangle = 0,$$

which is the normal cone to $\mathcal{X}$ at x.

(b) *Now consider the constrained optimization problem:*

$$\begin{aligned} \text{min} : &\quad f(x) \\ \text{s.t.} : &\quad x \in \mathcal{X}, \end{aligned}$$

for f and $\mathcal{X}$ convex. This can be rewritten as the equivalent unconstrained problem:

$$\text{min} : \ f(x) + I_{\mathcal{X}}(x).$$

This is again a convex function. Write down conditions for a point x^* to be optimal to the unconstrained problem.

Solution The optimality condition is written as

$$0 \in \partial(f(x) + I_{\mathcal{X}}(x)) = \partial f(x) + \partial I_{\mathcal{X}}(x) = \partial f(x) + \mathcal{N}_{\mathcal{X}}(x)$$

where the first equality holds as shown in Problem 1(b), and the second equality holds as shown in the previous problem. This is equivalent to the optimality condition for the constrained optimization problem.

5. *(Monotonicity). Show that the subdifferential of a convex function $f(\cdot)$ is an example of what is known as a* monotone operator. *That is, show that*

$$\langle u - v, x - y\rangle \geq 0, \quad \forall u \in \partial f(x),\ v \in \partial f(y).$$

Solution For any $u \in \partial f(x)$ and $v \in \partial f(y)$, we have

$$\langle u, x\rangle = f(x) + f^*(u) = f(x) + \sup_z\{\langle u, z\rangle - f(z)\} \geq f(x) + \langle u, y\rangle - f(y)$$
$$\langle v, y\rangle = f(y) + f^*(v) = f(y) + \sup_z\{\langle v, z\rangle - f(z)\} \geq f(y) + \langle v, x\rangle - f(x)$$

Getting sum of the two inequalities, we have

$$\langle u, x\rangle + \langle v, y\rangle \geq f(x) + \langle u, y\rangle - f(y) + f(y) + \langle v, x\rangle - f(x)$$

Rearranging the inequality, we obtain the monotonicity.

6. *Consider the ℓ_1-regularized regression problem*

$$\min_x \quad \frac{1}{2}\|y - Ax\|_2^2 \ + \ \lambda\|x\|_1 \tag{2}$$

Where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. Show that a point $\bar{x}$ is an optimum of this problem if and only if there exists a $z \in \mathbb{R}^n$ such that both the following hold:
(a) $A^\top(y - A\bar{x}) + \lambda z = 0$
(b) For every $i \in [n]$, $z_i = sign(\bar{x}_i)$ if $\bar{x}_i \neq 0$, and $|z_i| \leq 1$ if $\bar{x}_i = 0$.

Solution

Since it is an unconstrained optimization, $\bar{x}$ is an optimal solution if and only if

$$0 \in \partial\left(\frac{1}{2}\|y - A\bar{x}\|_2^2 + \lambda\|\bar{x}\|_1\right) = A^\top(y - A\bar{x}) + \lambda\partial\|\bar{x}\|_1.$$

This is equivalent to the existence of z such that

$$0 = A^\top(y - A\bar{x}) + \lambda z, \quad z \in \partial\|\bar{x}\|_1.$$

As the condition $z \in \partial\|\bar{x}\|_1$ is equivalent to (b), this completes the proof.

7. *In class we saw that the convergence of sub gradient descent is given by*

$$f_{k,best} - f^* \ \leq \ \frac{R^2 + G^2\sum_{i\leq k} h_i^2}{\sum_{i\leq k} h_i}$$

where h_i were the step sizes. We also learnt that, for a fixed *k, the lowest this bound could be is $\frac{RG}{\sqrt{k+1}}$; this is achieved by choosing every $h_i = \frac{R/G}{\sqrt{k+1}}$. Thus, each k needs a different sequence to achieve the best lower bound.*

Suggest a single *sequence of h_i's that makes the above bound decay as $O(\frac{\log k}{\sqrt{k}})$, for* every *$k$. Prove your result.*

Solution

Choosing $h_i = \frac{1}{\sqrt{i}}$ will result in the bound decaying as $O(\frac{\log k}{\sqrt{k}})$.

The sequence $\{h_i\}_i$ is nonnegative and monotonically decreasing, so we can bound the sums by

$$\int_1^{k+1} h(x)dx \leq \sum_{i=1}^{k} h_i \leq \int_0^k h(x)dx, \quad \int_1^{k+1} h^2(x)dx \leq \sum_{i=1}^{k} h_i^2 \leq \int_0^k h^2(x)dx$$

where $h(x) = 1/\sqrt{x}$. This implies that $\sum_{i=1}^k h_i$ and $\sum_{i=1}^k h_i^2$ scales as

$$\sum_{i=1}^{k} h_i = O\left(\int_0^k \frac{1}{\sqrt{x}}dx\right) = O(\sqrt{k}), \quad \sum_{i=1}^{k} h_i^2 = O\left(\int_0^k \frac{1}{x}dx\right) = O(\log k)$$

as k tends to infinity. Since the sum $\sum h_i^2$ diverges as k increases, the term R^2 will not affect the scaling behavior of the numerator. Then the bound decays as

$$\frac{R^2 + G^2 \sum_{i\leq k} h_i^2}{\sum_{i\leq k} h_i} = \frac{O(\log k)}{O(\sqrt{k})} = O\left(\frac{\log k}{\sqrt{k}}\right).$$

as k tends to infinity.

8. *Compute the Legendre-Fenchel Transform of the following:*

   (a) *Quadratic function*

   $$f(x) = \frac{1}{2}x^\top Qx,$$

   where Q is positive definite.

   Solution As Q is positive definite, the inverse Q^{-1} exists. Then we have

   $$\begin{aligned} f^*(z) &= \sup_x \left\{\langle z, x\rangle - \frac{1}{2}x^\top Qx\right\} \\ &= \sup_x \left\{-\frac{1}{2}(x - Q^{-1}z)^\top Q(x - Q^{-1}z) + \frac{1}{2}z^\top Q^{-1}z\right\} \\ &= \frac{1}{2}z^\top Q^{-1}z \end{aligned}$$

   The last equality follows from that the first term $-\frac{1}{2}(x - Q^{-1}z)^\top Q(x - Q^{-1}z)$ cannot be greater than zero because Q is positive definite.

   (b) *Negative logarithm.*

   $$f(x) = -\log x.$$

   Solution $f(x)$ is only defined on $x > 0$ as a smooth function, so the subgradient is just the derivative of $f(x)$.

   $$\partial f(x) = \{-x^{-1}\}, \quad \mathbf{dom} f = (0, \infty)$$

(c) *Norm.*

$$f(x) = \|x\|,$$

for some norm $\|\cdot\|$.

Solution The dual norm can be expressed as

$$\begin{aligned}
\|z\|_* &= \sup\{\langle z, x\rangle : \|x\| \le 1\} \\
&= \sup\{\langle z, x\rangle : \|x\| = 1\} \\
&= \sup_x \left\langle z, \frac{x}{\|x\|} \right\rangle,
\end{aligned}$$

where the second equality holds because if some x with $\|x\| < 1$ maximizes the inner product, we can increase x by a factor of $1/\|x\|$ to increase the inner product, resulting in a contradiction.

Then we have

$$\begin{aligned}
f^*(z) &= \sup_x \{\langle z, x\rangle - \|x\|\} \\
&= \sup_x \left\{ \|x\| \left(\left\langle z, \frac{x}{\|x\|} \right\rangle - 1 \right) \right\} \\
&= \begin{cases} 0, & \|z\|_* \le 1 \\ +\infty, & \text{otherwise} \end{cases}
\end{aligned}$$

The last equality holds based on the following argument. If $\|z\|_* \le 1$, for any point x, $\left\langle z, \frac{x}{\|x\|} \right\rangle - 1 \le 0$, so $\|x\| \left(\left\langle z, \frac{x}{\|x\|} \right\rangle - 1 \right)$ cannot be greater than zero. If $\|z\|^* > 1$, we can choose sufficiently large x where $\left\langle z, \frac{x}{\|x\|} \right\rangle - 1 > 0$.