1. **Convex functions**

(a) If f_i are convex functions, show that $f(x) := \sup_i f_i(x)$ is also convex.

(b) Show that the largest eigenvalue of a symmetric matrix is a convex function of the matrix (i.e. $\lambda_{\max}(M)$ is a convex function of M). Is the same true for the eigenvalue of largest magnitude ?

(c) Consider a weighted graph with edge weight vector w. Fix two nodes a and b. The *weighted shortest path* from a to b is the path whose sum of edge weights is the minimum, among all paths with one endpoint at a and another at b. Let $f(w)$ be the weight of this path. Show that f is a concave function of w.

Xueying Lu
xl4888

(a). Pf. $\forall\ 0 \le \lambda \le 1,\ \forall\ x, y.$

$$f(\lambda x + (1-\lambda)y) = \sup_i f_i(\lambda x + (1-\lambda)y)$$
$$\le \sup_i \left(\lambda f_i(x) + (1-\lambda) f_i(y)\right).$$
$$\le \lambda \sup_i f_i(x) + (1-\lambda) \sup_i f_i(y) = \lambda f(x) + (1-\lambda) f(y)$$ □

(b) Pf. $\forall\ 0 \le \mu \le 1.\ \forall\ M, N \in \mathbb{R}^{n\times n}$ symmetric.

$$\lambda_{\max}(M) I - M \succeq 0. \quad \lambda_{\max}(N) I - N \succeq 0.$$

$$\Rightarrow \mu \lambda_{\max}(M) I + (1-\mu)\lambda_{\max}(N) I - \mu M - (1-\mu) N \succeq 0.$$

$$\Rightarrow \mu \lambda_{\max}(M) + (1-\mu)\lambda_{\max}(N) \ge \lambda_{\max}(\mu M + (1-\mu) N).$$

This is also true for the largest magnitude of eigenvalue.

Let $f(M) := \max\{|\lambda_{\max}(M)|, |\lambda_{\min}(M)|\}$.

then $f(M) I \pm M \succeq 0.$ $f(N) I \pm N \succeq 0.$

$\forall\ 0 \le \mu \le 1.$ $\mu f(M) I + (1-\mu) f(N) I \pm (\mu M + (1-\mu) N) \succeq 0.$

$$\Rightarrow \mu f(M) + (1-\mu) f(N) \ge \lambda_{\max}(\mu M + (1-\mu) N).$$

and $\mu f(M) + (1-\mu) f(N) \ge -\lambda_{\min}(\mu M + (1-\mu) N).$

$$\Rightarrow \mu f(M) + (1-\mu) f(N) \ge f(\mu M + (1-\mu) N).$$ □

(c). Pf. Let l denote a path and $f_l(w)$ denote the weight of that path. Then $\exists$ a vector I_l with elements of 0 and 1 such that $f_l(w) = I_l^T w.$ $\Rightarrow f_l(w)$ is linear

As proved in (a), $\sup_l -f_l(w)$ is convex.

$\Rightarrow f(w) = -\left(\sup_l -f_l(w)\right)$ is concave. □

2. **Projection**.

(a) Suppose that $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed and bounded convex set. Let $\mathbf{y} \in \mathbb{R}^2$ be any point. The projection of $\mathbf{y}$ on $\mathcal{X}$ is defined by

$$\Pi_{\mathcal{X}}(\mathbf{y}) = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{y} - \mathbf{x}\|_2^2.$$

Show that the solution to the optimization problem is unique.

Pf. Suppose there are two distinct projections x_1, x_2, $x_1 \neq x_2$.

then $\frac{1}{2}\|y - x_1 + y - x_2\|^2 = \|y - x_1\|^2 + \|y - x_2\|^2 - \frac{\|x_2 - x_1\|^2}{2}$.

$\Rightarrow \frac{1}{2}\|y - x_1 + y - x_2\|^2 < \|y - x_1\|^2 + \|y - x_2\|^2 = 2\|y - x_1\|^2$.

i.e. $\|y - \frac{x_1 + x_2}{2}\|^2 < \|y - x_1\|^2$.

Since $\mathcal{X}$ is convex, $\frac{x_1 + x_2}{2} \in \mathcal{X}$. This contradicts that x_1 is a projection. □

(b) Show that for $\mathcal{X}$ as above, if $\mathbf{y} \notin \mathcal{X}$, there exists a hyperplane with $\mathcal{X}$ on one side, and $\mathbf{y}$ strictly on the other side. That is, show that there is a vector $\mathbf{s}$ and a scalar b, such $\langle \mathbf{s}, \mathbf{x} \rangle \leq b$ for all $\mathbf{x} \in \mathcal{X}$, and $\langle \mathbf{s}, \mathbf{y} \rangle > b$.

Pf. Let $s = y - \Pi_{\mathcal{X}}(y)$.

$$\langle s, y \rangle = \langle y - \Pi_{\mathcal{X}}(y),\ y - \Pi_{\mathcal{X}}(y) + \Pi_{\mathcal{X}}(y) \rangle$$

$$= \|y - \Pi_{\mathcal{X}}(y)\|_2^2 > 0.$$

$\forall\ x \in \mathcal{X}$.

$\langle s, x \rangle = \langle y - \Pi_{\mathcal{X}}(y),\ x - \Pi_{\mathcal{X}}(y) \rangle \leq 0$. This is because

$\|y - \Pi_{\mathcal{X}}(y)\|_2^2 \leq \|y - (\Pi_{\mathcal{X}}(y) + \lambda(x - \Pi_{\mathcal{X}}(y))\|^2$ $\quad \forall\ 0 \leq \lambda \leq 1$.

$$= \|y - \Pi_{\mathcal{X}}(y)\|_2^2 + \lambda^2 \|x - \Pi_{\mathcal{X}}(y)\|^2$$

$$- 2\lambda \langle y - \Pi_{\mathcal{X}}(y),\ x - \Pi_{\mathcal{X}}(y) \rangle.$$

□.

(c) Projected gradient descent is used for constrained optimization problems. Instead of just taking gradient steps, we follow up each gradient step with a projection onto the feasible set. That is, the update is given by:

$$x^{(k+1)} = \text{Proj}_{\mathcal{X}}(x^{(k)} - t_k \nabla f(x^{(k)})).$$

Show that this is equivalent to the update:

$$x^{(k+1)} = \arg\min_{x \in \mathcal{X}} \left\{ \langle x, \nabla f(x^{(k)}) \rangle + \frac{1}{2t_k} \|x - x^{(k)}\|_2^2 \right\}.$$

Pf $\min f(x)$

s.t $x \in X \subseteq \mathbb{R}^d$. X is convex.

$$\text{Proj}_X(x^{(k)} - t_k \nabla f(x^{(k)})) =$$

$$\arg\min_{x \in X} \|x^{(k)} - t_k \nabla f(x^{(k)}) - x\|^2. \quad (*)$$

$$\|x^{(k)} - t_k \nabla f(x^{(k)}) - x\|^2$$

$$= \|x - x^{(k)}\|^2 + 2t_k \langle \nabla f(x^k), x - x^k \rangle + t_k^2 \|\nabla f(x^k)\|^2.$$

$$= 2t_k \left[\langle x - x^{(k)} \rangle + \frac{1}{2t_k} \|x - x^{(k)}\|^2 \right]$$

$$- 2t_k \langle \nabla f(x^k), x^k \rangle + t_k^2 \|\nabla f(x^k)\|^2.$$

$$\Rightarrow (*) = \arg\min_{x \in X} \left\{ \langle x - x^{(k)} \rangle + \frac{1}{2t_k} \|x - x^{(k)}\|^2 \right\}.$$

□

3. **Computing Projections**. For the given convex set $\mathcal{X}$, compute the projection of a point z.

(a) $\mathcal{X} = \mathbb{R}^n_+$.

(b) Euclidean ball: $\{x : \|x\|_2 \leq 1\}$.

(c) Positive semidefinite cone: $\mathbb{S}^n_+ = \{M \in \mathbb{S}^n : x^\top M x \geq 0, \ \forall x \in \mathbb{R}^n\}$. Assume for this problem that $z \in \mathbb{S}^n$.

(d) $\mathcal{X}$ is a rectangle defined by vectors L and U that satisfy $U_i \geq L_i$. Thus, $\mathcal{X} = \{x : L_i \leq x_i \leq U_i, \ i = 1, \ldots, n\}$.

Sol. (a). $\text{Proj}_{\mathbb{R}^n_+}(z)_i = \max\{z_i, 0\}$.

(b). $\text{Proj}_{B_1(0)}(z) = \frac{z}{\|z\|_2}$.

(c). Let $M = \sum_i \lambda_i v_i v_i^T$ be its eigen decomposition.

$$\text{Proj}_{S^n_+}(M) = \sum_{i=1}^n \max\{\lambda_i, 0\} v_i v_i^T.$$

(d). $\text{Proj}_X(z)_i = \begin{cases} U_i, & \text{if } z_i > U_i \\ z_i & \text{if } L_i \leq z_i \leq U_i \\ L_i & \text{if } z_i < L_i. \end{cases}$ $\forall 1 \leq i \leq n.$

4. **Computing More Projections**. Now for two more tricky ones.

(a) 1-norm ball: $\{x : \sum_i |x_i| \le 1\}$.

(b) Probability simplex: $\mathcal{X} = \{x : \sum_i x_i = 1,\ x_i \ge 0,\ i = 1, \dots, n\}$.

Sol (a). $\min \|x - z\|_2^2$

s.t. $\sum_i |x_i| \le 1$.

The Lagrangian is

$$\mathcal{L}(x, \lambda) = \sum (x_i - z_i)^2 + \lambda \left(\sum_i |x_i| - 1 \right), \quad \lambda \ge 0.$$

$$= \sum_i \left[(x_i - z_i)^2 + \lambda |x_i| \right] - \lambda.$$

The dual function is

$$g(\lambda) = \inf_x \mathcal{L}(x, \lambda)$$

This can be solved component wise to obtain

$$x_i^* = \operatorname{sgn}(z_i) \max\left(|z_i| - \tfrac{1}{2}\lambda, 0 \right).$$

$$\Rightarrow \sum_{i=1}^{n} \operatorname{sgn}(z_i) \max\left(|z_i| - \tfrac{1}{2}\lambda, 0 \right) - 1 = 0.$$

Since $h(\lambda) := \sum_{i=1}^{n} \operatorname{sgn}(z_i) \max\left(|z_i| - \tfrac{1}{2}\lambda, 0 \right) - 1$

is piece-wise linear,

Its derivative is.

$$\frac{dh(\lambda)}{d\lambda} = \sum_{i=1}^{n} -\tfrac{1}{2} \left\{ |z_i| - \tfrac{1}{2}\mu > 0 \right\}.$$

We can solve for λ. by Newton's iterations.

(b). $\min \|x - z\|_2^2$.

s.t. $\sum_i x_i = 1$

$x \geq 0$.

The Lagrangian is

$$\mathcal{L}(x, \lambda, \mu) = \sum (x_i - z_i)^2 - \lambda^T x + \mu(\sum x_i - 1). \quad \lambda \geq 0.$$

By the KKT condition.

$$\begin{cases} 2x - 2z - \lambda + \mu \mathbb{I} = 0. \\ x \geq 0. \quad \lambda \geq 0 \\ \lambda_i x_i = 0. \quad \forall i. \\ \sum x_i - 1 = 0. \end{cases}$$

$\Rightarrow$ If $x_i \neq 0$. $\lambda_i = 0$. then $x_i = z_i - \frac{1}{2}\mu$.

$\Rightarrow \sum_{i=1}^{n} \max(z_i - \frac{1}{2}\mu, 0) - 1 = 0$.

Since $h(\mu) := \sum_{i=1}^{n} \max(z_i - \frac{1}{2}\mu, 0) -$ is piece-wise linear.

Its derivative is.

$$\frac{dh(\mu)}{d\mu} = \sum_{i=1}^{n} -\frac{1}{2}\{z_i - \frac{1}{2}\mu > 0\}.$$

We can solve for μ by Newton's iterations.

5. Show that sub gradients have the following properties:[2] Note that the "$\subseteq$" inclusion can be much more tricky than the reverse inclusion to show. Do your best for this direction.

(a) $\partial(\alpha f(x)) = \alpha \partial f(x)$.
(b) $\partial(f_1 + f_2) = \partial f_1 + \partial f_2$.
(c) If $g(x) = f(Ax + b)$, then $\partial g(x) = A^\top \partial f(Ax + b)$.
(d) If $f(x) = \max_{1 \le i \le m} f_i(x)$, then

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

(a). Pf $\forall\ g \in \partial f(x)$. $f(y) - f(x) \geq \langle g, y - x \rangle$.

$\Rightarrow \alpha f(y) - \alpha f(x) \geq \langle \alpha g, y - x \rangle$. $\alpha \geq 0$.

$\Rightarrow \alpha g \in \partial(\alpha f(x))$. $\Rightarrow$. $\alpha \partial f(x) \subseteq \partial(\alpha f(x))$.

$\forall\ g \in \partial(\alpha f(x))$. $\alpha f(y) - \alpha f(x) \geq \langle g, y - x \rangle$.

$\Rightarrow f(y) - f(x) \geq \langle g/\alpha, y - x \rangle$. $\Rightarrow g/\alpha \in \partial f(x)$

$\Rightarrow \partial(\alpha f(x)) \subseteq \alpha \partial f(x)$. □.

(b). Pf. $\forall\ g_1 \in \partial f_1(x)$, $\forall\ g_2 \in \partial f_2(x)$.

$f_1(y) - f_1(x) \geq \langle g_1, y - x \rangle$.

$f_2(y) - f_2(x) \geq \langle g_2, y - x \rangle$.

$\Rightarrow (f_1 + f_2)(y) - (f_1 + f_2)(x) \geq \langle g_1 + g_2, y - x \rangle$.

$g_1 + g_2 \in \partial(f_1 + f_2)(x)$. $\Rightarrow \partial f_1 + \partial f_2 \subseteq \partial(f_1 + f_2)$.

Let $h(x) = f_1(x) + f_2(x)$.

$\forall\ g \in h(x)$, consider the following optimization.

$$\min_{y, r} \quad r - h(x) - \langle g, y - x \rangle$$
$$\text{s.t.} \quad f_1(y) + f_2(y) = r.$$

The Lagrangian is

$$\mathcal{L}(y, r, \lambda) = r - h(x) - \langle g, y - x \rangle - \lambda(f_1(y) + f_2(y) - r).$$

We know that the optimal $y^* = x$ as $g \in \partial h(x)$. Thus taking the gradient of L w.r.t. y at x gives.

$$\lambda^* (\partial f_1(x) + \partial f_2(x)) = g.$$

Take the gradient of L w.r.t. Γ gives $\lambda^* = 1$.

i.e. $\exists\ g_1 \in \partial f_1(x),\ g_2 \in \partial f_2(x).$ s.t. $g_1 + g_2 = g.$

So $\partial(f_1 + f_2) \subseteq \partial f_1 + \partial f_2$. □

(c) Pf. $\forall z \in \partial f(Ax+b)$. $f(Ay+b) \geq f(Ax+b) + z^T(Ay+b-Ax-b)$.

i.e. $g(y) \geq g(x) + z^T A(y-x)$.

$\Rightarrow A^T z \in \partial g(x)$. $A^T \partial f(Ax+b) \subseteq \partial g(x)$.

For each $z \in \partial g(x)$, we have that

$f(Ay+b) \geq f(Ax+b) + z^T(y-x)$. or

$f(Ay+b) - z^T y \geq f(Ax+b) - z^T x$. $\forall y$.

So the minimization problem

$$\min_{u,y} f(u) - z^T y \quad \text{s.t. } u = Ay + b$$

has an optimum $y = x$, $u = Ax+b$.

Assume f is convex, then the optimality condition tells us that $\exists\ \lambda$. s.t.

$0 \in \partial(f(u) - z^T y - \lambda^T(Ay+b-u))$. with $y = x$, $u = Ax+b$

Since $\partial_y(f(u) - z^T y - \lambda^T(Ay+b-u)) = -z^T - \lambda^T A$.

$$\partial_u(f(u) - z^T y - \lambda^T(Ay+b-u)) = \partial f(u) + \lambda.$$

So $-\lambda \in \partial f(u) = \partial f(Ax+b)$. and $z = -A^T\lambda$.

So $\partial g(x) \subseteq A^T \partial f(Ax+b)$. □

(d). Pf. Let $z \in \text{conv}\left(\bigcup_{i: f_i(x)=f(x)} \partial f_i(x)\right)$. Denote $I = \{i : f_i(x) = f(x)\}$.

Then $\exists \{\lambda_i\}_{i\in I}$ s.t. $\lambda_i \geq 0$. $\sum_{i\in I} \lambda_i = 1$ and

$z = \sum_{i\in I} \lambda_i z_i$, $z_i \in \partial f_i(x)$. $i \in I$.

So for any y,

$f(y) = \max_{1\leq i\leq m} f_i(y) \geq f_i(y) \geq f_i(x) + z_i^T(y-x)$. for $1 \leq i \leq m$.

$\Rightarrow \sum_{i\in I} \lambda_i f(y) \geq \sum_{i\in I} \lambda_i f_i(x) + \sum_{i\in I} \lambda_i z_i^T(y-x)$.

i.e. $f(y) \geq f(x) + z^T(y-x)$. $z \in \partial f(x)$.

$\text{conv}\left(\bigcup_{i: f_i(x)=f(x)} \partial f_i(x)\right) \subseteq \partial f(x)$.

On the other hand, suppose $z \in \partial f(x)$. then

$f(y) - z^T y \geq f(x) - z^T x$. $\forall y$.

Consider the minimization problem.

$$\min_{u.y.} \quad u - z^T y.$$
$$\text{s.t.} \quad f_i(y) \leq u. \quad \forall\, 1 \leq i \leq m.$$

The problem has an optimum $y = x$. $u = f(x)$.

Assume f_i's are convex, then the optimality condition tells us that

$\exists \{\mu_i\}_{i=1}^m$, $\mu_i \geq 0$. s.t.

$0 \in \partial\left(u - z^T y + \sum_{i=1}^m \mu_i (f_i(y) - u)\right)$.

for $y = x$. $u = f(x)$.

Since $\partial_y\left(u - z^T y + \sum_{i=1}^m \mu_i (f_i(y) - u)\right) = -z^T + \sum_{i=1}^m \mu_i z_i$

where $z_i \in \partial f_i(x)$.

$$\partial_u\left(u - z^T y + \sum_{i=1}^m \mu_i (f_i(y) - u)\right) = 1 - \sum_{i=1}^m \mu_i$$

Therefore $z = \sum_{i=1}^{m} \mu_i z_i$, $z_i \in \partial f_i(x)$ and $\sum_{i=1}^{m} \mu_i = 1$, $\mu_i \geq 0$.

The complimentary slackness also indicates that $\mu_i = 0$ if $f_i(x) \neq f(x)$.

So $\partial f(x) \subseteq \text{conv}\left(\bigcup_{i: f_i(x) = f(x)} \partial f_i(x) \right)$. □

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

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

Sol. Let $f_i(M) = \|M_i\|_2$

$$\partial f_i(M) = \begin{cases} \|M_i\|_2^{-1} [0, 0, \cdots, M_i, \cdots, 0]. & \text{if } \|M_i\|_2 \neq 0. \\ \{ [0, \cdots, g, \cdots, 0] \mid \|g\|_2 \leq 1 & \text{if } x = 0. \end{cases}$$

↑ the ith column

Then $\partial \|M\|_{2,1} = \partial f_1(M) + \partial f_2(M) + \cdots + \partial f_n(M)$.

7. (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.

Sol. $f(x) = \sup_{\|y\|_2=1} y^T A(x) y.$

$$\partial f(x) = \text{Conv} \bigcup_{\|y\|_2=1} \{\partial f_y(x) \mid f_y(x) = f(x)\}.$$

where $f_y(x) = y^T A(x) y$

$$= y^T A_0 y + x_1 y^T A_1 y + \cdots + x_n y^T A_n y$$

, which is differentiable in x.

$$\nabla_x f_y(x) = (y^T A_1 y, y^T A_2 y, \cdots, y^T A_n y)^T.$$

Therefore

$$\partial f(x) = \text{Conv}\{(y^T A_1 y, \cdots, y^T A_n y)^T \mid \|y\|_2 = 1, \ y^T A(x) y = \lambda_{\max}(A(x))\}.$$

$$= \text{Conv}\{(y^T A_1 y, \cdots, y^T A_n y)^T \mid \|y\|_2 = 1, \ A(x) y = \lambda_{\max}(A(x)) y.\}.$$