SI251 - Convex Optimization, Fall 2022
Final Exam

Note: We are interested in the reasoning underlying the solution, as opposed to simply the answer. Thus, solutions with the correct answer without adequate explanation will not receive full credit; on the other hand, partial solutions with an explanation will receive partial credit. Within a given problem, you can assume the results of previous parts in proving later parts (e.g., it is fine to solve part 3) first, assuming the results of parts 1) and 2)). The resources you use should be limited to printed lecture slides, lecture notes, homework, homework solutions, general resources, class reading and textbooks, and other related textbooks on optimization. You should not discuss the final exam problems with anyone or use electronic devices. Detected violations of this policy will be processed according to ShanghaiTech's code of academic integrity. Please hand in the exam papers and answer sheets at the end of the exam.

I. **Basic Knowledge**

1. (20 points) Determine the following statements are true or false and explain your reasons.

(a) Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $g : \mathbb{R} \to \mathbb{R}$ is concave and non-decreasing, then the function $g \circ f$ is convex.

(b) $f(x) = e^{-\|x\|^2}$ is convex, $x \in \mathbb{R}^n$..

(c) The set $\{x \in \mathbb{R}^n : x_n^2 \geq \sum_{i=1}^{n-1} x_i^2, x_n \geq 0\}$ is a self-dual convex cone.

(d) $f(x) = x_1^2 - 2x_1x_2 + \frac{1}{3}x_2^3 - 8x_2$ has no global maximizers or minimizers.

Solution:

(a) False.

(b) False. It can be shown that $\nabla^2 f(x) = 2f(x)\left[2xx^T - I\right]$. Therefore, if $n > 1$, there is a $u \in \mathbb{R}^n$ is such that $\|u\|_2 = 1$ and $\langle u, x\rangle = 0$ with $u^T\nabla^2 f(x)u = -2f(x) < 0$. If $n \geq 1$ and $0 < \|x\|_2 < 1$, then $x^T\nabla^2 f(x)x = -2f(x)\|x\|_2\left(1 - \frac{\|x\|_2}{2}\right) < 0$. So f is not convex.

(c) True. The Second-order cone is self-dual.

(d) True.

$$\nabla f(x) = \begin{bmatrix} 2(x_1 - x_2) \\ x_2^2 - 2x_1 - 8 \end{bmatrix} \quad \text{and} \quad \nabla^2 f(x) = \begin{bmatrix} 2 & -2 \\ -2 & 2x_2 \end{bmatrix}$$

Compute the critical points by setting $\nabla f(x) = 0$. Setting $\partial f(x)/\partial x_1 = 0$ gives $x_1 = x_2$. Plug this into the equation $\partial f(x)/\partial x_2 = 0$ to get $0 = x_2^2 - 2x_2 - 8 = (x_2 - 4)(x_2 + 2)$. This gives 2 critical points

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \begin{pmatrix} -2 \\ -2 \end{pmatrix},$$

with

$$\nabla^2 f(4,4) = \begin{bmatrix} 2 & -2 \\ -2 & 8 \end{bmatrix} \quad \text{and } \nabla^2 f(-2,-2) = \begin{bmatrix} 2 & -2 \\ -2 & -4 \end{bmatrix}.$$

It is easily shown that $\nabla^2 f(4,4)$ is positive definite and that $\nabla^2 f(-2,-2)$ has one positive and one negative eigenvalue. Hence $(4,4)$ is a local minimizer and $(-2,-2)$ is a saddle point. There are no global maximizers or minimizers since $f(0, x_2) = \frac{1}{3}x_2^3 - 8x_2$ which goes to $+\infty$ as $x_2 \uparrow +\infty$ and goes to $-\infty$ as $x_2 \downarrow -\infty$.

2. (20 points) Derive the conjugate function $f^*(y)$ for each of the following functions:

(a) $f(x) = -\ln x + 2$; (b) $f(x) = x^p$ on $[0, +\infty)$, where $p > 1$.

Solution:

(a) $f^*(y) = \max_x x \cdot y + \ln(x) - 2$. If $y \geq 0$, then clearly $f^*(y) = \infty$. Otherwise, by stationarity $y + \frac{1}{x} = 0 \Longrightarrow x = -\frac{1}{y}$. Thus $f^*(y) = -3 + \ln\left(-\frac{1}{y}\right) = -3 - \ln(-y)$

(b) We'll use the standard notation: we define q by the equation $1/p + 1/q = 1$, i.e., $q = p/(p-1)$. Since $p > 1$, x^p is strictly convex on $\mathbf{R}_+$. For $y \leq 0$ the function $yx - x^p$ achieves its maximum for $x \geq 0$ at $x = 0$, so $f^*(y) = 0$. For $y > 0$ the function achieves its maximum at $x = (y/p)^{1/(p-1)}$, where it has value

$$y(y/p)^{1/(p-1)} - (y/p)^{p/(p-1)} = (p-1)(y/p)^q.$$

Therefore we have

$$f^*(y) = \begin{cases} 0 & y \leq 0 \\ (p-1)(y/p)^q & y > 0 \end{cases}$$

3. (10 points) Consider the function $f : \mathbb{R} \to (-\infty, \infty]$ given for any $x \in \mathbb{R}$ by

$$f(x) = \begin{cases} \mu x, & 0 \leq x \leq \alpha \\ \infty & \text{else} \end{cases}$$

where $\mu \in \mathbb{R}$ and $\alpha \in [0, \infty]$. Please represent and compute the proximal operator of f.

Solution:

f can be represented as

$$f(x) = \delta_{[0,\alpha] \cap \mathbb{R}}(x) + \mu x$$

Define $g(x) = \delta_{[0,\alpha]\cap\mathbb{R}}(x)$, we first prove $\mathrm{prox}_g(x) = \min\{\max\{x, 0\}, \infty\}$.

Assume that $\alpha < \infty$. Note that $\tilde{u} = \mathrm{prox}_g(x)$ is the minimizer of

$$w(u) = \frac{1}{2}(u - x)^2$$

over $[0, \alpha]$. The minimizer of w over $\mathbb{R}$ is $u = x$. Therefore, if $0 \leq x \leq \alpha$, then $\tilde{u} = x$. If $x < 0$, then w is increasing over $[0, \alpha]$, and hence $\tilde{u} = 0$. Finally, if $x > \alpha$, then w is decreasing over $[0, \alpha]$, and thus $\tilde{u} = \alpha$. To conclude,

$$\mathrm{prox}_g(x) = \tilde{u} = \begin{cases} x, & 0 \leq x \leq \alpha, \\ 0, & x < 0, \\ \alpha, & x > \alpha, \end{cases} = \min\{\max\{x, 0\}, \alpha\}.$$

For $\alpha = \infty, g(x) = \delta_{[0,\infty)}(x)$, $\mathrm{prox}_g(x) = [x]_+$, which can also be written as

$$\mathrm{prox}_g(x) = \min\{\max\{x, 0\}, \infty\}$$

Hence, $\mathrm{prox}_f(x) = \mathrm{prox}_g(x - \mu) = \min\{\max\{x - \mu, 0\}, \infty\}$

II. **Advanced Knowledge**

4. (15 points) Derive the KKT conditions for the problem

$$\begin{array}{ll} \text{minimize} & \mathrm{tr}\, X - \log\det X \\ \text{subject to} & Xs = y, \end{array}$$

with variable $X \in \mathbf{S}^n$ and domain $\mathbf{S}^n_{++} \cdot y \in \mathbf{R}^n$ and $s \in \mathbf{R}^n$ are given, with $s^T y = 1$. Verify that the optimal solution is given by

$$X^\star = I + yy^T - \frac{1}{s^T s} ss^T.$$

[hint: $\frac{\partial det(X)}{\partial X} = det(X)(X^{-1})^T$]

Solution:

We introduce a Lagrange multiplier $z \in \mathbf{R}^n$ for the equality constraint. The KKT optimality conditions are:

$$X \succ 0, \quad Xs = y, \quad X^{-1} = I + \frac{1}{2}\left(zs^T + sz^T\right).$$

We first determine z from the condition $Xs = y$. Multiplying the gradient equation on the right with y gives

$$s = X^{-1}y = y + \frac{1}{2}\left(z + \left(z^T y\right)s\right).$$

By taking the inner product with y on both sides and simplifying, we get $z^T y = 1 - y^T y$. Substituting in the above equation we get

$$z = -2y + \left(1 + y^T y\right)s,$$

and substituting this expression for z in the above formula gives

$$\begin{aligned} X^{-1} &= I + \frac{1}{2}\left(-2ys^T - 2sy^T + 2\left(1 + y^T y\right)ss^T\right) \\ &= I + \left(1 + y^T y\right)ss^T - ys^T - sy^T. \end{aligned}$$

5. (10 points) Prove: suppose that f is a convex function, $f(x)$ is L-Lipschitz if and only if the subgradient of f is bounded, i.e.,

$$\|g\| \leq L, \quad \forall g \in \partial f(x), \quad x \in R^n.$$

Solution:

(1) $\Rightarrow$: Suppose for all g, $\|g\| \leq L$, $g_x \in \partial f(x)$, $g_y \in \partial f(y)$, by the definition of subgradient, we have:

$$g_x^T(x-y) \geq f(x) - f(y) \geq g_y^T(x-y),$$

therefore, by Cauchy-Schwartz inequalities,

$$\begin{aligned} g_x^T(x-y) &\leq \|g_x\|\|(x-y)\| \leq L\|x-y\|, \\ g_y^T(x-y) &\geq -\|g_x\|\|(x-y)\| \geq -L\|x-y\|. \end{aligned}$$

Therefore,

$$|f(x) - f(y)| \leq L\|x-y\|.$$

$\Leftarrow$: Suppose $f(x)$ is Lipschitz continuous. Let there be x and $g \in \partial f(x)$ such that $\|g\| > L$, and make $y = x + \frac{g}{\|g\|}$. By the definition of subgradient, we have

$$\begin{aligned} f(y) &\geqslant f(x) + g^{\mathrm{T}}(y-x) \\ &= f(x) + \|g\| \\ &> f(x) + L, \end{aligned}$$

which contradicts that $f(x)$ is $L-$ Lipschitz continuous. So we prove the necessity.

6. (10 points) Convex-Concave Programming (CCP) is a well-known algorithm to handle the difference between two convex functions. Suppose f_i and g_i are convex functions with $i = 0, 1, \cdots, m$. Please show the monotone property of the CCP iteration, i.e. $f_0(x_{k+1}) - g_0(x_{k+1}) \leq f_0(x_k) - g_0(x_k)$.

DCP problems:

$$\begin{array}{ll} \text{minimize} & f_0(x) - g_0(x) \\ \text{subject to} & f_i(x) - g_i(x) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

Algorithm 1: CCP algorithm

1: given an initial feasible point x_0.
2: $k := 0$.
3: **repeat**
4: **Convexify**. Form $\hat{g}_i(x; x_k) \triangleq g_i(x_k) + \nabla g_i(x_k)^T (x - x_k)$ for $i = 0, \ldots, m$.
5: **Solve**. Set x_{k+1} as the solution of the convex problem

$$\begin{array}{l} \text{minimize } f_0(x) - \hat{g}_0(x; x_k) \\ \text{subject to } f_i(x) - \hat{g}_i(x; x_k) \leq 0, \quad i = 1, \ldots, m. \end{array}$$

6: **Update iteration**. $k := k + 1$.
7: **until** stopping criterion is satisfied.

Solution:

We will first observe that all of the iterates are feasible, and then show that CCP is a descent algorithm, i.e.,

$$f_0(x_{k+1}) - g_0(x_{k+1}) \le f_0(x_k) - g_0(x_k).$$

Assume x_k is a feasible point for (1). We know that x_k is a feasible point for the convexified subproblem because

$$f_i(x_k) - \hat{g}_i(x_k; x_k) = f_i(x_k) - g_i(x_k) \le 0,$$

so a feasible point x_{k+1} exists to the convexified subproblem (4). The convexity of g_i gives us $\hat{g}_i(x; x_k) \le g_i(x)$, for all x , so

$$f_i(x) - g_i(x) \le f_i(x) - \hat{g}_i(x; x_k)$$

It then follows that x_{k+1} must be a feasible point of the original problem since

$$f_i(x_{k+1}) - g_i(x_{k+1}) \le f_i(x_{k+1}) - \hat{g}_i(x_{k+1}; x_k) \le 0.$$

Thus, because x_0 was chosen as feasible, all iterations are feasible. We will now show that the objective value converges. Let $v_k = f_0(x_k) - g_0(x_k)$. Then

$$v_k = f_0(x_k) - g_0(x_k) = f_0(x_k) - \hat{g}_0(x_k; x_k) \ge f_0(x_{k+1}) - \hat{g}_0(x_{k+1}; x_k),$$

where the last inequality follows because at each iteration k we minimize the value of $f_0(x) - \hat{g}_0(x; x_k)$, and we know that we can achieve v_k by choosing $x_{k+1} = x_k$. Thus

$$v_k \ge f_0(x_{k+1}) - \hat{g}_0(x_{k+1}; x_k) \ge v_{k+1}.$$

Thus the sequence $\{v_i\}_{i=0}^{\infty}$ is nonincreasing and will converge, possible to negative infinity. The above analysis holds in the nondifferentiable case when the gradient is replaced by a subgradient, that is any $\delta g_i(x_k)$ such that for all x ,

$$g_i(x_k) + \delta g_i(x_k)(x - x_k) \le g_i(x)$$

7. (15 points) Consider the least squares regression problem with ℓ_2-norm regularization,

$$\underset{\boldsymbol{x} \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{A}_i \boldsymbol{x} - \boldsymbol{b}_i\|_2^2 + (\lambda/2)\|\boldsymbol{x}\|_2^2$$

where $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times d}$, $\boldsymbol{b}_i \in \mathbb{R}^{n_i}$, λ is the regularization parameter. Please write the **exact** ADMM update steps for this problem. (Note: the *argmin* operator should not be kept.)

Solution:

We first rewrite the above problem with local variables $\{\boldsymbol{x}_i\}$ and a common global variable $\boldsymbol{z}$ as

$$\begin{aligned} \underset{\boldsymbol{x}_i, \boldsymbol{z} \in \mathbb{R}^d}{\text{minimize}} \quad & \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{A}_i \boldsymbol{x}_i - \boldsymbol{b}_i\|_2^2 + (\lambda/2)\|\boldsymbol{z}\|_2^2 \\ \text{subject to} \quad & \boldsymbol{x}_i - \boldsymbol{z} = 0, \end{aligned}$$

which can be further rewritten as

$$\begin{aligned} \underset{\boldsymbol{x}_i, \boldsymbol{z} \in \mathbb{R}^d}{\text{minimize}} \quad & \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{A}_i \boldsymbol{x}_i - \boldsymbol{b}_i\|_2^2 + (\lambda/2)\|\boldsymbol{z}\|_2^2 \\ \text{subject to} \quad & \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_N \end{bmatrix} - \begin{bmatrix} \boldsymbol{z} \\ \vdots \\ \boldsymbol{z} \end{bmatrix} = \boldsymbol{0}, \end{aligned}$$

By introducing Lagrange multipliers $\{\boldsymbol{y}_i\}$, the augmented Lagrangian $L_\rho(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)$ is given by

$$L_\rho(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{z}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \frac{1}{2} \sum_{i=1}^{N} \|\boldsymbol{A}_i \boldsymbol{x}_i - \boldsymbol{b}_i\|_2^2 + g(\boldsymbol{z}) + \sum_{i=1}^{N} \left(\boldsymbol{y}_i^\mathsf{T} (\boldsymbol{x}_i - \boldsymbol{z}) + (\rho/2) \|\boldsymbol{x}_i - \boldsymbol{z}\|_2^2 \right).$$

The ADMM update rules can be given by

$$
\begin{aligned}
\boldsymbol{x}_i^{k+1} &:= \underset{\boldsymbol{x}_i}{\operatorname{argmin}} \left(\|\boldsymbol{A}_i \boldsymbol{x}_i - \boldsymbol{b}_i\|_2^2 + (\boldsymbol{y}_i^k)^\mathsf{T} \left(\boldsymbol{x}_i - \boldsymbol{z}^k\right) + (\rho/2) \left\|\boldsymbol{x}_i - \boldsymbol{z}^k\right\|_2^2 \right) \\
\boldsymbol{z}^{k+1} &:= \underset{\boldsymbol{z}}{\operatorname{argmin}} \left((\lambda/2)\|\boldsymbol{z}\|_2^2 + \sum_{i=1}^{N} \left(-(\boldsymbol{y}_i^k)^\mathsf{T} \boldsymbol{z} + (\rho/2)\|\boldsymbol{x}_i^{k+1} - \boldsymbol{z}\|_2^2 \right) \right) \\
\boldsymbol{y}_i^{k+1} &:= \boldsymbol{y}_i^k + \boldsymbol{x}_i^{k+1} - \boldsymbol{z}^{k+1}.
\end{aligned}
$$

Since the objective function is convex, we can exploit the optimality condition to obtain the exact update rules as follows,

$$
\begin{aligned}
\boldsymbol{x}_i^{k+1} &= (\boldsymbol{A}_i^\mathsf{T} \boldsymbol{A}_i + \rho I)^{-1} (\boldsymbol{A}_i^\mathsf{T} \boldsymbol{b}_i - \boldsymbol{y}_i^k + \rho \boldsymbol{z}^k) \\
\boldsymbol{z}^{k+1} &= \frac{1}{\lambda + \rho N} \sum_{i=1}^{N} (\rho \boldsymbol{x}_i^{k+1} + \boldsymbol{y}_i^k) \\
\boldsymbol{y}_i^{k+1} &:= \boldsymbol{y}_i^k + \boldsymbol{x}_i^{k+1} - \boldsymbol{z}^{k+1}.
\end{aligned}
$$