SI251 - Convex Optimization, Fall 2021
Homework 2

Due on Nov. 21, 2021, 23:59 UTC+8

I. KKT:

1. State and solve the optimality conditions for the problem

$$\begin{array}{ll} \text{minimize} & \log\det\left(\begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \\ \boldsymbol{X}_2^T & \boldsymbol{X}_3 \end{bmatrix}^{-1}\right) \\ \text{subject to} & \operatorname{Tr}(\boldsymbol{X}_1) = \alpha \\ & \operatorname{Tr}(\boldsymbol{X}_2) = \beta \\ & \operatorname{Tr}(\boldsymbol{X}_3) = \gamma. \end{array} \tag{1}$$

The optimization variable is $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & \boldsymbol{X}_2 \\ \boldsymbol{X}_2^T & \boldsymbol{X}_3 \end{bmatrix}$ with $\boldsymbol{X}_1 \in \mathbb{S}^n, \boldsymbol{X}_2 \in \mathbb{R}^{n\times n}, \boldsymbol{X}_3 \in \mathbb{S}^n$. The domain of the objective function is $\mathbb{S}^{2n}_{++}$. We assume $\alpha > 0$, and $\alpha\gamma > \beta^2$ (15 points)

Solution: This is a convex problem with three equality constraints

$$\begin{array}{ll} \underset{\boldsymbol{X}}{\text{minimize}} & f_0(\boldsymbol{X}) \\ \text{subject to} & h_1(\boldsymbol{X}) = \alpha \\ & h_2(\boldsymbol{X}) = \beta \\ & h_3(\boldsymbol{X}) = \gamma, \end{array}$$

where $f_0(\boldsymbol{X}) = -\log\det\boldsymbol{X}$ and

$$h_1(\boldsymbol{X}) = \operatorname{Tr}\left(\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{X}\right), h_2(\boldsymbol{X}) = \frac{1}{2}\operatorname{Tr}\left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix}\boldsymbol{X}\right), h_3(\boldsymbol{X}) = \operatorname{Tr}\left(\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}\boldsymbol{X}\right)$$

. The general optimality condition for an equality constrained problem,

$$\nabla f_0(\boldsymbol{X}) + \sum_{i=1}^{3} v_i \nabla h_i(\boldsymbol{X}) = 0$$

reduces to

$$-\boldsymbol{X}^{-1} + v_1\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \frac{v_2}{2}\begin{bmatrix} \boldsymbol{0} & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{0} \end{bmatrix} + v_3\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix} = 0, \qquad \text{(Condition 1)}$$

along with the feasibility conditions

$$\operatorname{Tr}(\boldsymbol{X}_1) = \alpha,\ \operatorname{Tr}(\boldsymbol{X}_2) = \beta, \operatorname{Tr}(\boldsymbol{X}_3) = \gamma. \qquad \text{(Condition 2)}$$

From the first condition

$$\boldsymbol{X} = \begin{bmatrix} v_1\boldsymbol{I} & \frac{v_2}{2}\boldsymbol{I} \\ \frac{v_2}{2}\boldsymbol{I} & v_3\boldsymbol{I} \end{bmatrix}^{-1} = \begin{bmatrix} \lambda_1\boldsymbol{I} & \lambda_2\boldsymbol{I} \\ \lambda_2\boldsymbol{I} & \lambda_3\boldsymbol{I} \end{bmatrix}$$

where

$$\begin{bmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_3 \end{bmatrix} = \begin{bmatrix} v_1 & \frac{v_2}{2} \\ \frac{v_2}{2} & v_3 \end{bmatrix}^{-1}$$

From the feasibility conditions we see that we have to choose λ_i (and hence v_i), such that

$$\boldsymbol{X} = \frac{1}{n}\begin{bmatrix} \alpha\boldsymbol{I} & \beta\boldsymbol{I} \\ \beta\boldsymbol{I} & \gamma\boldsymbol{I} \end{bmatrix}.$$

II. CVX:

2. Consider the following compressive sensing problem via ℓ_1-minimization [1]:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad & \|\boldsymbol{x}\|_1 \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}. \end{aligned} \tag{2}$$

(a) Equivalently reformulate (2) into a linear programming problem. (10 points)

Solution:

Suppose unknown signal is component-wise non-negative, ℓ_1 minimization problem is just

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad & \textstyle\sum_{i=1}^n \boldsymbol{x}_i \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z} \\ & \boldsymbol{x} \geq 0 \end{aligned}$$

The general case of real-valued signals, the key trick is to add additional variables to "linearize" the non-linear objective function. Use $\boldsymbol{y}_i$ to represent $\boldsymbol{x}_i$, then we have

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} & \sum_{i=1}^n \boldsymbol{y}_i \\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z} \\ & \boldsymbol{y}_i = |\boldsymbol{x}_i|, i = 1, 2, \ldots, n \end{aligned}$$

However, this problem is non-convex due to the second constraints. So we add "linear" inequalities, that is

$$\begin{aligned} \boldsymbol{y}_i - \boldsymbol{x}_i \geq 0, i = 1, 2, \ldots, n \\ \boldsymbol{y}_i + \boldsymbol{x}_i \geq 0, i = 1, 2, \ldots, n \end{aligned}$$

which is equivalent to

$$\boldsymbol{y}_i \geq \max\{\boldsymbol{x}_i, -\boldsymbol{x}_i\} = |\boldsymbol{x}_i|, i = 1, 2, \ldots, n$$

then we have the LP problem:

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} & \sum_{i=1}^n \boldsymbol{y}_i \\ \text{subject to } & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z} \\ & \boldsymbol{y}_i \geq \boldsymbol{x}_i, i = 1, 2, \ldots, n \\ & \boldsymbol{y}_i \geq -\boldsymbol{x}_i, i = 1, 2, \ldots, n. \end{aligned}$$

(b) This part describes the experiments that illustrate the empirical phase transition in compressed sensing via ℓ_1 minimization. In the compressed sensing example, we fix the ambient dimension $d = 20$. For each number of random measurements $m = 1, 2, ..., 20$, and each number of nonzero entries in $\boldsymbol{x}^\natural$ $s = 1, 2, ..., 20$, we repeat the following procedure 30 times:

- *Step 1:* Construct a vector $\boldsymbol{x}^\natural \in \mathbb{R}^d$ with s nonzero entries. The locations of the nonzero entries are selected at random; each nonzero entry equals ± 1 with equal probability.
- *Step 2:* Draw a standard normal matrix $\boldsymbol{A} \in \mathbb{R}^{m\times d}$ (i.e., each entry in $\boldsymbol{A}$ is drawn from a Gaussian random variable with zero mean and variance one), and form $\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}^\natural$.
- *Step 3:* Use CVX solve (2) to obtain an optimal point $\boldsymbol{x}^\star$.
- *Step 4:* Declare success if $\|\boldsymbol{x}^\star - \boldsymbol{x}^\natural\| \leq 10^{-5}$.

You need to program to implement this experiment and plot the *phase transition figure*, the simulation results can be referred to Figure 1.1 in [1]. (15 points)

Solution:

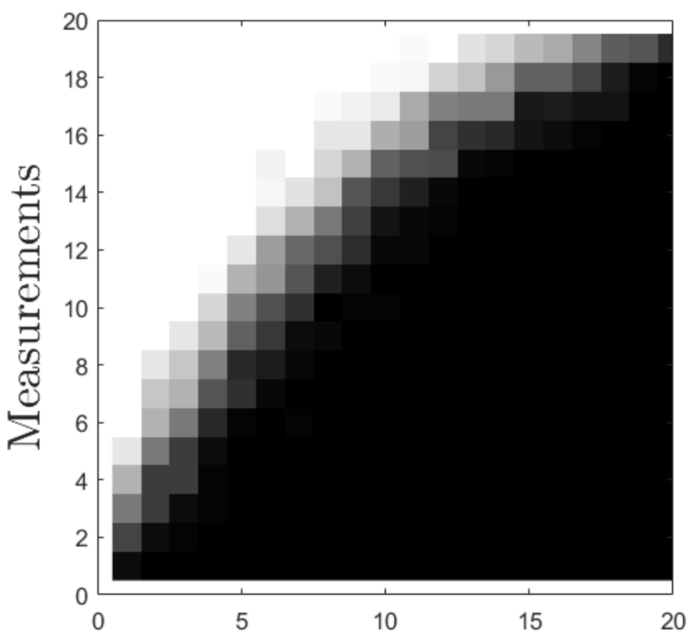
The following matlab code is just as an example.

```matlab
%xg ground truth
n=20;
d=20;
prod_m = rand(d,d);
loop = 30;
for m=1:1:d
    for s=1:1:d
        sum = 0;
        for i=0:1:loop
            i;
            xg = zeros(n,1);
            index = randperm(n);
            nonzeros = index(1:s);
            for j=1:s
                p=rand(1,1);
                if p>=0.5
                    xg(nonzeros(j)) = 1;
                else
                    xg(nonzeros(j)) = -1;
                end
            end
            A = randn(m,n);
            b = A*xg;

            cvx_begin quiet
            variable x(n,1)
            minimize(norm(x,1))
            subject to
                A*x == b;
            cvx_end

            if norm(x-xg,2)<=1e-5
                sum= sum+1;
            end
            sum;
        end
        prod = sum/loop;
        prod_m(m,s)=prod;
    end
end
prod_m;
```

```matlab
nonzeroLists = 1:1:d;
samplesize = 1:1:d;
colormap('gray');   % set colormap
imagesc(prod_m,[0,1]); % draw image and scale colormap to values range
hold on
axis xy
axis square
axis([0 d 0 d])
xlabel('Nonzeros','Interpreter','latex',  'FontSize',20)
ylabel('Measurements','Interpreter','latex','FontSize',20)
```



III. Gradient Methods:

3. Let f be differentiable, m-strongly convex, M-smooth and with minimizer x^*. In class, we proved geometric convergence of the error $\left\|x^l - x^*\right\|_2$. In this exercise, we explore how to prove convergence in the function value difference $f\left(x^l\right) - f\left(x^*\right)$ for gradient descent with step size $\alpha = 1/M$ Show the following characterizations are equivalent to L-smooth condition.

(1) Prove that:

$$f\left(x^{l+1}\right)-f\left(x^{*}\right) \leq f\left(x^{l}\right)-f\left(x^{*}\right)-\frac{1}{2 M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2}$$

This shows that we have a descent method(5 points)
Solution: We have by M-smoothness

$$f\left(x^{l+1}\right)-f\left(x^{l}\right)-\left\langle\nabla f\left(x^{l}\right), x^{l+1}-x^{l}\right\rangle \leq \frac{M}{2}\left\|x^{l+1}-x^{l}\right\|_{2}^{2}$$

Furthermore, note that by our gradient descent method, $x^{l+1}-x^{l}=-\frac{1}{M} \nabla f\left(x^{l}\right)$. Substituting this in, we see that

$$f\left(x^{l+1}\right)-f\left(x^{l}\right)+\frac{1}{M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2} \leq \frac{1}{2 M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2}$$

Rearranging, and adding $-f\left(x^{*}\right)$ to both sides, we obtain

$$f\left(x^{l+1}\right)-f\left(x^{*}\right) \leq f\left(x^{l}\right)-f\left(x^{*}\right)-\frac{1}{2 M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2}$$

(2) Prove that:

$$\frac{m}{M}\left(f\left(x^{l}\right)-f\left(x^{*}\right)\right) \leq \frac{1}{2 M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2}$$

(5 points)
Solution: From the characteristic of m-strong convexity, we have

$$f(y)-f(x)-\langle\nabla f(x), y-x\rangle \leq \frac{1}{2 m}\|\nabla f(x)-\nabla f(y)\|_{2}^{2} \quad \forall x, y \in \mathbb{R}^{d}$$

We substitute $y=x^{l}$, the output at step l, and $x=x^{*}$, the minimum. Since x^{*} is a minimum, $\nabla f\left(x^{*}\right)=0$. Therefore

$$f\left(x^{l}\right)-f\left(x^{*}\right) \leq \frac{1}{2 m}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2} \quad \forall x, y \in \mathbb{R}^{d}$$

Multiplying both sides by m / M we obtain the desired result

$$\frac{m}{M}\left(f\left(x^{l}\right)-f\left(x^{*}\right)\right) \leq \frac{1}{2 M}\left\|\nabla f\left(x^{l}\right)\right\|_{2}^{2}$$

(3) Conclude that:

$$f\left(x^{l+1}\right)-f\left(x^{*}\right) \leq\left(1-\frac{m}{M}\right)\left(f\left(x^{l}\right)-f\left(x^{*}\right)\right)$$

This shows that we have geometric convergence with parameter $1-\frac{m}{M}$(5 points)
Solution: Adding the previous results together, we obtain

$$f\left(x^{l+1}\right)-f\left(x^{*}\right)+\frac{m}{M}\left(f\left(x^{l}\right)-f\left(x^{*}\right)\right) \leq f\left(x^{l}\right)-f\left(x^{*}\right)$$

and thus

$$f\left(x^{l+1}\right)-f\left(x^{*}\right) \leq\left(1-\frac{m}{M}\right)\left(f\left(x^{l}\right)-f\left(x^{*}\right)\right)$$

as desired.

4. Consider a constrained optimization problem min ${}_{x \in \mathcal{C}} f(x)$ where $\mathcal{C}$ is a compact convex set, and f is convex and has a continuous derivative. The conditional gradient method with stepsizes $\left\{\alpha^{l}\right\}_{l=0}^{\infty}$ generates a sequence of the form

$$x^{l+1}=\left(1-\alpha^{l}\right) x^{l}+\alpha^{l} z^{l}$$

where $z^{l} \in \arg \min _{z \in \mathcal{C}}\left\langle\nabla f\left(x^{l}\right), z\right\rangle$. Compute the form of these updates forthe following cases:

(a) $\mathcal{C}=\left\{x \in \mathbb{R}^{d} \mid\|x\|_{1} \leq 1\right\}$ (10 points)
Solution: Let z be a vector with $\|z\|_{1}=1$. As mentioned in class, the problem is equivalent to maximizing $\left\langle-\nabla f\left(x^{l}\right), z\right\rangle$. Then we have

$$\left\langle -\nabla f\left(x^l\right), z\right\rangle \leq \left\|\nabla f\left(x^l\right)\right\|_\infty \|z\|_1$$

by Cauchy-Schwartz, where inequality is obtained when

$$z = -\operatorname{sign}\left(\left[\nabla f\left(x^l\right)\right]_{i^*}\right) e_{i^*}$$

with $i^* = \arg\min_{i=1,\ldots,d} -\left|\left[\nabla f\left(x^l\right)\right]_i\right|$ and e_i the standard basis vectors.
The update is therefore given by

$$x^{l+1} = \left(1-\alpha^l\right) x^l + \alpha^l z^l = \left(1-\alpha^l\right) x^l + \alpha^l \left(-\operatorname{sign}\left(\left[\nabla f\left(x^l\right)\right]_{i^*}\right) e_{i^*}\right)$$

(b) $\mathcal{C} = \left\{X \in \mathbb{R}^{d\times d} \mid \sum_{j=1}^d \sigma_j(X) \leq 1\right\}$ where $\sigma_j(X)$ is the j^{th} singular value(15 points)
Solution: We want to solve the following optimization problem

$$\arg\min_{Z\in\mathcal{C}} \left\langle \nabla f\left(X^l\right), Z\right\rangle = \arg\min_{Z\in\mathcal{C}} tr\left(Z^T \nabla f\left(X^l\right)\right)$$

Using SVD, we can write $\nabla f\left(X^l\right) = U\Lambda V^T$, where U and V are unitary, and Λ is a diagonal matrix; also define $\hat{Z} = U^T Z V$. We have

$$\begin{aligned}
\arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(Z^T \nabla f\left(X^l\right)\right) &= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(Z^T\left(U\Lambda V^T\right)\right) \\
&= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(V^T Z^T U \Lambda\right) \\
&= \arg\min_{Z\in\mathcal{C}} \operatorname{tr}\left(\left(U^T Z V\right)^T \Lambda\right) \\
&= U\left[\arg\min_{\hat{Z}\in\mathcal{C}} \operatorname{tr}\left(\hat{Z}^T \Lambda\right)\right] V^T
\end{aligned}$$

where we used invariance of the trace operator under cyclic permutations. Now we seek to find a Z^* such that the above is minimized. Note that the diagonal entries do not affect the trace since Λ is diagonal. Hence, w.l.o.g. we take Z^* to be diagonal, and the problem reduces to part (a), i.e. the problem of finding the diagonal vector $\hat{z}$ of $\hat{Z}$ given the diagonal vector of Λ which is equivalent to the singular vector v of $\nabla f\left(X^l\right)$, i.e.

$$\arg\min_{\hat{Z}\in\mathcal{C}} \operatorname{tr}\left(\hat{Z}^T \Lambda\right)\right) = \operatorname{diag}\left(\arg\min_{\|\hat{z}_1 \leq 1\|} \langle v, \hat{z}\rangle\right)$$

As a consequence, defining $i^* = \arg\max_i |\Lambda_{ii}|$, the minimum is attained by defining $\hat{Z}^*$ as:

$$\left[\hat{Z}^*\right]_{ii} = \begin{cases} 0 & \text{if } i \neq i^* \\ -\operatorname{sign}\left(\left[\Lambda_{ii}\right]\right) & \text{if } i = i^* \end{cases}$$

Therefore, in our update, Z^l is given by $Z^l = U\hat{Z}^* V^T$

IV. Subgradient Methods:
5. For a convex function $f : \mathbb{R}^d \to \mathbb{R}$, a subgradient at $x \in \mathbb{R}^d$ is a vector $g \in \mathbb{R}^d$ such that

$$f(y) \geq f(x) + \langle g, y - x\rangle \quad \text{for all } y \in \mathbb{R}^d$$

In this case, we write $g \in \partial f(x)$. We say that f is sub-differentiable when $\partial f(x)$ is non-empty for each $x \in \mathbb{R}^d$.

(1) Show that x^* is a minimizer of f if $0 \in \partial f\left(x^*\right)$. (5 points)

Solution: Suppose $0 \in \partial f(x^*)$. Then for all $y \in \mathbb{R}^d$,

$$f(y) \geq f(x^*) + \langle 0, y - x^*\rangle = f(x^*)$$

so $f(x^*)$ is a minimizer.

(2) Show that f is Lipschitz with parameter L (i.e., $|f(x) - f(y)| \leq L\|x - y\|_2$ for all $x, y \in \mathbb{R}^d$ if and only if $\|g\|_2 \leq L$ for all subgradient vectors g. (15 points)

Solution:
($\Rightarrow$). Suppose f is L-Lipschitz; choose $x, y \in \mathbb{R}^d$, and let g be a subgradient vector at x. Then,

$$\langle g, y - x \rangle \leq f(y) - f(x) \leq L\|y - x\|$$

Rearranging,

$$\frac{\langle g, y - x \rangle}{\|y - x\|} \leq L$$

Since inequality in fact holds for all $y \in \mathbb{R}^d$, we choose $y = g + x$ to obtain the desired result.
($\Leftarrow$). Suppose $\|g\|_2 \leq L$ for all subgradient vectors g. Choose $x, y \in \mathbb{R}^d$. The subgradient condition tells us

$$f(y) - f(x) \geq \langle g, y - x \rangle$$

where g is a subgradient at x. Multiplying both sides by (-1),

$$\begin{aligned} f(x) - f(y) &\leq \langle g, x - y \rangle \\ &\leq \|g\|\|x - y\| \quad \text{(Cauchy-Schwarz)} \\ &\leq L\|x - y\| \end{aligned}$$

In a similar fashion, by switching the roles of x and y, and letting h be a subgradient vector at y, we have

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

Hence, we conclude that $|f(y) - f(x)| \leq L\|x - y\|$ and f is L-Lipschitz.

References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Inf. Inference*, vol. 3, pp. 224–294, Jun 2014.