PRATICAL WORK 3

CHAU Dang Minh

1 Compressed sensing and cardinality minimization problems

Consider the problem

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ \|x\|_0, \\ &\text{subject to } Ax = b, \end{aligned} \tag{Q_1}$$

where $A \in \mathbb{R}^{p\times d}$ (usually with $p \ll d$), $b \in \mathbb{R}^p$ and

$$\|x\|_0 = |\{i \in \{1,\dots,d\} : x_i \neq 0\}|.$$

1. Problem (Q_1) is not convex because the objective is not convex. A counterexample is with $x = \begin{bmatrix}0\\1\end{bmatrix}$ and $y = \begin{bmatrix}1\\0\end{bmatrix}$. We have $\|x\|_0 = \|y\|_0 = 1$ and $\|\lambda x + (1-\lambda)y\|_0 = 2, \forall\lambda \in (0,1)$. Hence

$$\forall\lambda \in (0,1), \lambda\|x\|_0 + (1-\lambda)\|y\|_0 < \|\lambda x + (1-\lambda)y\|_0.$$

   The function $\|\cdot\|_0$ is also not a norm because $2\|x\|_0 = 1$, instead of 2.

2. We will prove that

$$\text{aff}(\|\cdot\|_0 + \iota_{\overline{B}_\infty}) = \|\cdot\|_1 + \iota_{\overline{B}_\infty},$$

   where $\overline{B}_\infty = \{x \in \mathbb{R}^d : \|x\|_\infty \leq 1\}$ in the following steps.

   (a) For all $y \in \mathbb{R}^d, x \in \overline{B}_\infty$, we have

$$\begin{aligned} \langle y, x\rangle - \|x\|_0 &= \sum_{x_i\neq 0} y_i x_i - \sum_{x_i\neq 0} 1 = \sum_{x_i\neq 0}(y_i x_i - 1) \\ &\overset{(1)}{\leq} \sum_{x_i\neq 0}(|y_i|\cdot|x_i| - 1) \\ &\overset{(2)}{\leq} \sum_{x_i\neq 0}(|y_i| - 1) \\ &\overset{(3)}{\leq} \sum_{x_i\neq 0}\max\{0, |y_i| - 1\} \\ &\overset{(4)}{\leq} \sum_{i=1}^{d}\max\{0, |y_i| - 1\} \\ &= \sum_{i=1}^{d}(|y_i| - 1)^+. \end{aligned}$$

   In summary,

$$\langle y, x\rangle - \|x\|_0 \leq \sum_{i=1}^{d}(|y_i| - 1)^+.$$

   Let us find out the conditions to get equality.

   - Equality in (1) occurs when $\text{sign}(x_i)\text{sign}(y_i) \geq 0, \forall i \in \{1,\dots,d\}$.

- Equality in (2) occurs when $x \in \{-1, 0, 1\}^d$.
- Equality in (3) occurs when $|y_i| \geq 1$ when $x_i \neq 0$.
- Equality in (4) occurs when $|y_i| < 1$ when $x_i = 0$.

(b) Let $g = \|\cdot\|_0 + \iota_{\overline{B}_\infty}$. For any $x \in \mathbb{R}^d$, let $\hat{x} = \dfrac{x}{\|x\|_\infty}$, we have $\hat{x} \in \overline{B}_\infty$ and

$$\langle y, x\rangle - \|x\|_0 = \langle \|x\|_\infty y, \hat{x}\rangle - \|\hat{x}\|_0 .$$

Hence,

$$\sup_{x\in\mathbb{R}^d} (\langle y, x\rangle - \|x\|_0) = \sup_{x\in\overline{B}_\infty} (\langle y, x\rangle - \|x\|_0).$$

Moreover, according the equality analysis in the previous question, the supremum is attained at x such that

$$x_i = \begin{cases} 0, & \text{if} |y_i| < 1, \\ \operatorname{sign}(y_i), & \text{if} |y_i| \geq 1. \end{cases}$$

Therefore,

$$\begin{aligned} g^*(y) &= \sup_{x\in\mathbb{R}^d} (\langle y, x\rangle - \|x\|_0 - \iota_{\overline{B}_\infty}(x)) \\ &\leq \sup_{x\in\mathbb{R}^d} (\langle y, x\rangle - \|x\|_0) \\ &= \sup_{x\in\overline{B}_\infty} (\langle y, x\rangle - \|x\|_0) \\ &= \sum_{i=1}^d (|y_i| - 1)^+. \end{aligned}$$

(c) For $x \in \overline{B}_\infty$, we have

$$\begin{aligned} \langle x, y\rangle - g^*(y) &= \sum_{i=1}^d \big(x_i y_i - (|y_i| - 1)^+\big) \\ &\leq \sum_{i=1}^d \big(|x_i| \cdot |y_i| - (|y_i| - 1)^+\big) \\ &= \sum_{|y_i|<1} |x_i| \cdot |y_i| + \sum_{|y_i|\geq 1} (|x_i| \cdot |y_i| - |y_i| + 1) \\ &\leq \sum_{|y_i|<1} |x_i| + \sup_{|y_i|\geq 1} (|y_i|(|x_i| - 1) + 1) \\ &\leq \sum_{|y_i|<1} |x_i| + \sum_{|y_i|\geq 1} (|x_i| - 1 + 1) && (|x_i| - 1 \leq 0) \\ &= \|x\|_1. \end{aligned}$$

In summary,

$$\langle x, y\rangle - g^*(y) \leq \|x\|_1.$$

Equality occurs when $y_i = \operatorname{sign}(x_i)$.

(d) From the previous question, $\sup_{y\in\mathbb{R}^d} (\langle x, y\rangle - g^*(y))$ is attainable and equal to $\|x\|_1$ for any $x \in \overline{B}_\infty$. Therefore,

$$\forall x \in \overline{B}_\infty, g^{**}(x) = \|x\|_1.$$

(e) Choose $x_i = n\text{sign}(y_i)$, we have $\lim_{n\to\infty} (\langle x, y\rangle - g^*(y)) = \infty$. Hence,

$$\forall x \in \mathbb{R}^d \setminus \overline{B}_\infty, g^{**}(x) = \infty.$$

(f) To this point, we deduce that $g^{**} = \|\cdot\|_1 + \iota_{\overline{B}_\infty}$.

(g) From the biconjugate theorem, $g^{**} = \text{aff}(g)$, we have

$$\|\cdot\|_1 + \iota_{\overline{B}_\infty} = \text{aff}(\|\cdot\|_0 + \iota_{\overline{B}_\infty}).$$

Therefore,

$$\forall x \in \overline{B}_\infty, \|\cdot\|_1 = \text{aff}(\|\cdot\|_0).$$

(h) From the previous question, we have

$$\forall x \in \overline{B}_\infty, \|x\|_1 = \sup\{g(x) : g \text{ is affine and } g \leq \|\cdot\|_0\}.$$

The following sentences are equivalent to this sentence. Given $M > 0$,

$$\forall x \in \overline{B}_\infty, \frac{1}{M}\|x\|_1 = \sup\{g(x) : g \text{ is affine and } g \leq \|M\cdot\|_0\}.$$

$$\forall x \in \overline{B}_\infty^M, \frac{1}{M}\|x\|_1 = \sup\{g(x) : g \text{ is affine and } g \leq \|\cdot\|_0\}.$$

$$\forall x \in \overline{B}_\infty^M, \frac{1}{M}\|\cdot\|_1 = \text{aff}(\|\cdot\|_0).$$

3. Thanks to the affine hull calculation, we can relax the problem into minimizing $\|\cdot\|_1$ instead of $\|\cdot\|_0$. As there is noise in reality, we consider $\theta > 0$ and solve the following problem instead

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \|x\|_1, \\ &\text{subject to } \|Ax - b\|_2 \leq \theta. \end{aligned} \tag{Q_3}$$

We will prove that for an appropriate choice of parameters θ and μ, (Q_2) is equivalent to

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \ \mu\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2. \tag{Q_4}$$

Indeed, if x^* is a solution to (Q_3), then there exists $\lambda \geq 0$ such that

$$0 \in \partial\left(\|\cdot\|_1 + \frac{\lambda}{2}(\|A\cdot - b\|_2^2 - \theta^2)\right)(x^*) = \partial\left(\frac{1}{\lambda}\|\cdot\|_1 + \frac{1}{2}\|A\cdot - b\|_2^2\right)(x^*).$$

Let $\mu = \frac{1}{\lambda}$. Since $\mu\|\cdot\|_1 + \frac{1}{2}\|A\cdot - b\|_2^2$ is convex, we have

$$x^* = \underset{x\in\mathbb{R}^d}{\text{argmin}} \left(\mu\|x\|_1 + \frac{1}{2}\|Ax - b\|_2^2\right).$$

Hence x^* is a solution to (Q_4). Conversely, let x^* be a solution to (Q_4). Since the objective of (Q_4) is convex, this is the unique solution. Let $\theta = \|Ax^* - b\|$. Suppose that there exists $\hat{x} \neq x^*$ to be a solution to (Q_3), i.e. $\|\hat{x}\|_1 < \|x^*\|$ and $\|A\hat{x} - b\| \leq \theta$. Then $\hat{x}$ is another solution to (Q_4), which is a contradiction. Therefore, x^* is a solution to (Q_3).

4. To apply the ADMM algorithm, we finally rewrite (Q_4) as

$$\begin{aligned} &\underset{x\in\mathbb{R}^d}{\text{minimize}}\ \mu\|y\|_1 + \frac{1}{2}\|Ax-b\|_2^2. \\ &\text{subject to } x - y = 0. \end{aligned} \tag{Q_3}$$

The augmented Lagrangian $L^\lambda : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is

$$L^\lambda(x,y,z) = \mu\|y\|_1 + \frac{1}{2}\|Ax-b\|_2^2 + \langle z, x-y\rangle + \frac{\lambda}{2}\|x-y\|_2^2.$$

Updates in ADMM follow

$$\begin{cases} x_{k+1} \in \underset{x}{\operatorname{argmin}}\, \partial_y L^\lambda(x, y_k, z_k) \\ y_{k+1} \in \underset{y}{\operatorname{argmin}}\, L^\lambda(x_k, y, z_k) \\ z = z + \lambda(x-y). \end{cases}$$

Note that L^λ is convex in terms of x and y the minimum is attained at the points where subderivative contains zero. We have

$$\nabla_x L^\lambda = A^\top(Ax-b) + z + \lambda(x-y).$$

Therefore, the update for (x_k) satisfies $A^\top(Ax_{k+1}-b) + z + \lambda(x_{k+1}-y_k) = 0$. Equivalently,

$$x_{k+1} = (\lambda I + A^\top A)^{-1}(A^\top b + \lambda y_k - z).$$

On the other hand,

$$\partial_y L^\lambda = \partial_y(\mu\|y\|_1) - z + \lambda(y-x).$$

The update for (y_k) satisfies $0 \in \partial_y(\mu\|\cdot\|_1)(y_{k+1}) - z + \lambda(y_{k+1}-x_k)$. Equivalently,

$$x_k + \frac{1}{\lambda}z \in \left(\partial_y\left(\frac{\mu}{\lambda}\|\cdot\|_1\right) + \mathrm{Id}\right)(y_{k+1}),$$

$$y_{k+1} \in \operatorname{prox}_{\frac{\mu}{\lambda}\|\cdot\|_1}\left(x_k + \frac{1}{\lambda}z\right).$$

2 Going back to the low-rank nonnegative matrix completion problem

1. Let us recall the proof of SVD theorem. Let $X \in \mathbb{R}^{m\times n}$. Since $X^\top X$ is positive semidefinite, we can diagonalize it as

$$X^\top X = V\Lambda V^\top,$$

where $V \in \mathbb{R}^n$ is orthogonal and $V = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Let $r = \operatorname{rank}(X)$, we can assume that $\lambda_i \geq 0$ for $1 \leq i \leq \operatorname{rank}(X)$ and $\lambda_i = 0$, otherwise. Let $\Sigma = (\Sigma_{ij}) \in \mathbb{R}^{m\times n}$ defined as

$$\Sigma_{ij} = \begin{cases} \sqrt{\lambda_i}, & \text{if } 1 \leq i = j \leq r \\ 0, & \text{otherwise.} \end{cases}$$

Next, we partition $\begin{bmatrix} V_1 & V_2 \end{bmatrix}$, where $V_1 \in \mathbb{R}^{n\times r}$ and $V_2 \in \mathbb{R}^{n\times(n-r)}$. Then

$$X^\top X = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}^2 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} = V_1\tilde{\Sigma}^2 V_1^\top,$$

where $\tilde{\Sigma}^2 = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_r^2)$. Hence $(X^\top X)V_2 = V_1\tilde{\Sigma}^2(V_1^\top V_2) = 0$, which means $(X^\top X)v = X^\top(Xv) = 0$ for all column vectors of V_2. Therefore,

$$Xv \in \ker X^\top = (\mathrm{Im}X)^\perp.$$

Hence, $Xv = 0$ for each column vector of V_2, or $XV_2 = 0$. Let

$$U_1 = XV_1\tilde{\Sigma}^{-1}.$$

We have $U_1^\top U_1 = I_r$ and $U_1^\top XV_1 = \tilde{\Sigma}$. Using the Gram-Schmidt process, we can extend U_1 to $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{m\times m}$ such that U is orthonormal. Then

$$U^\top XV = \begin{bmatrix} U_1^\top \\ U_2^\top \end{bmatrix} X \begin{bmatrix} V_1 & V_2 \end{bmatrix} = \begin{bmatrix} U_1^\top XV_1 & U_1^\top XV_2 \\ U_2^\top XV_1 & U_2^\top XV_2 \end{bmatrix}.$$

By our construction, $U_1^\top XV_1 = \tilde{\Sigma}$. Since $XV_2 = 0$, we have $U_1^\top XV_2 = U_2^\top XV_2 = 0$. Since $U_2^\top U_1 = 0$ and $XV_1 = U_1\tilde{\Sigma}$, we have

$$U_2^\top XV_1 = (U_2^\top U_1)\tilde{\Sigma} = 0.$$

Thus, $U^\top XV = \Sigma$ or $X = U\Sigma V^\top$.

2. We have $(X^\top X)_{ii} = \sum\limits_{j=1}^{m} X_{ij}^2$ for $1 \le i \le n$. Hence,

$$\mathrm{Trace}(X^\top X) = \sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}^2 = \|X\|_F^2.$$

On the other hand,

$$\begin{aligned}
\mathrm{Trace}(X^\top X) &= \mathrm{Trace}(V\Sigma V^top) \\
&= \mathrm{Trace}(\sum_{i=1}^{n} \Sigma_{ii} v_i v_i^\top) \\
&= \sum_{i=1}^{n} (\Sigma_{ii}\mathrm{Trace}(v_i v_i^\top)) \\
&= \sum_{i=1}^{n} \left(\Sigma_{ii}\sum_{j=1}^{m} v_{ij}^2\right) \\
&= \sum_{i=1}^{n} \Sigma_{ii} \\
&= \|\sigma\|_2^2.
\end{aligned}$$

Thus, $\|A\|_F = \sqrt{\mathrm{Trace}(X^\top X)} = \|\sigma\|_2$.

3. Let $u \in \ker(X^\top X)$, then

$$\|Xu\|_2^2 = u^\top(X^\top Xu) = 0.$$

Hence $Xu = 0$ or $u \in \ker(X)$. Conversely, if $u \in \ker(X)$, then

$$X^\top Xu = X^\top(Xu) = 0,$$

or $u \in \ker(X^\top X)$. Thus, $\ker(X) = \ker(X^\top X)$. By the rank-nullity theorem, we have

$$\mathrm{rank}(X) = \mathrm{rank}(X^\top X).$$

Also,

$$\mathrm{rank}(X^\top X) = \mathrm{rank}(V\Sigma V^\top) = \mathrm{rank}(\Sigma) = \|\sigma\|_0,$$

since V is orthonormal.

Thus, $\mathrm{rank}(X) = \mathrm{rank}(X^\top X) = \|\sigma\|_0$.

4. We are provided the nuclear norm $\|X\|_N = \|\sigma\|_1$ and the operator norm $\|X\|_O = \|\sigma\|_\infty$. These norms are proved to be dual. Furthermore, let $\overline{B}_O = \{X \in \mathbb{R}^{m\times n} : \|X\|_O \leq 1\}$, we have

$$\forall X \in \overline{B}_O, \|\cdot\|_N = \text{aff}(\text{rank}(\cdot)).$$

Finally,

$$\forall \gamma > 0, \forall X \in \mathbb{R}^{m\times n}, \text{prox}_{\gamma\|\cdot\|_N}(X) = U\text{diag}\left(\text{prox}_{\gamma\|\cdot\|_1}(\sigma)\right)V^\top. \tag{1}$$

5. Following section 1, we arrive at the problem

$$\begin{aligned} &\underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}}\ \mu\|X\|_N + \frac{1}{2}\|P_\Omega(X) - P_\Omega(B)\|_F^2, \\ &\text{subject to } X \geq 0, \end{aligned} \tag{P_3}$$

where $\Omega \subset \mathbb{N}^2$ is an index set and $P_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ such that

$$\forall X \in \mathbb{R}^{m\times n}, P_\Omega(X)_{ij} = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases}$$

6. To apply ADMM, we rewrite the equivalence of (P_3)

$$\begin{aligned} &\underset{X\in\mathbb{R}^{m\times n}}{\text{minimize}}\ \mu\|Y\|_N + \frac{1}{2}\|P_\Omega(X) - P_\Omega(B)\|_F^2 + \iota_{X\geq 0}(X), \\ &\text{subject to } X - Y = 0. \end{aligned} \tag{P_3}$$

   (a) For each $\lambda > 0$, the augmented Lagrangian is given by

$$\mathcal{L}^\lambda(X,Y,Z) = \mu\|Y\|_N + \frac{1}{2}\|P_\Omega(X) - P_\Omega(B)\|_F^2 + +\iota_{X\geq 0}(X) + \langle Z, X-Y\rangle + \frac{\lambda}{2}\|X-Y\|_F^2.$$

Recall that for f convex differentiable and $g \in \Gamma_0$, we have

$$\forall \lambda > 0, \text{Argmin}(f+g) = \text{Fix}\left(\text{prox}_{\lambda g} \circ (\text{Id} - \lambda\nabla f)\right).$$

In our case,

$$f(X) = \frac{1}{2}\|P_\Omega(X) - P_\Omega(B)\|_F^2 + \langle Z, X-Y\rangle + \frac{\lambda}{2}\|X-Y\|_F^2$$

and

$$g(X) = \iota_{P_+}(X),$$

where P_+ is the convex set of matrices whose all entries are nonnegative. We slightly abuse notation by denoting $P_+ = \text{proj}_{P_+}$. Choose $\lambda = 1$, we can see that if $\nabla f(X) = 0$, then $P_+(X)$ is a fix point of $\text{prox}_{\iota_{P_+}} \circ (\text{Id} - \nabla f)$. Indeed,

$$\text{prox}_{\iota_{P_+}} \circ (\text{Id} - \nabla f)(X) = \text{prox}_{\iota_{P_+}}(X) = P_+(P_+(X)) = P_+(X).$$

Therefore, $P_+(X) = \text{Armin}(f+g)$. We have

$$\nabla f(X) = P_\Omega(X) - P_\Omega(B) + Z + \lambda(X-Y).$$

By setting $f(X) = 0$, we have following equivalences

$$P_\Omega((1+\lambda)X - B - \lambda Y + Z) + P_{\Omega^c}(\lambda X - \lambda Y + Z) = 0,$$

$$\begin{cases} P_\Omega(X) = \dfrac{1}{1+\lambda} P_\Omega(B + \lambda Y - Z) \\ P_{\Omega^c}(X) = P_{\Omega^c}\left(Y - \dfrac{1}{\lambda} Z\right) \end{cases}.$$

In terms of the update rule, we have

$$\begin{aligned} X_{k+1} &= P_+(P_\Omega(X_{k+1}) + P_{\Omega^c}(X_{k+1})) = P_+(P_\Omega(X_{k+1})) + P_+(P_{\Omega^c}(X_{k+1})) \\ &= P_+\left(\frac{1}{1+\lambda} P_\Omega(B + \lambda Y_k - Z_k)\right) + P_+\left(P_{\Omega^c}\left(Y_k - \frac{1}{\lambda} Z_k\right)\right). \end{aligned}$$

Note that we can split the projection because the elements of $P_\Omega(X_{k+1})$ and $P_{\Omega^c}(X_{k+1})$ do not affect the other's. To get Y_{k+1}, we need

$$\begin{aligned} & 0 \in \partial_Y L^\lambda(Y_{k+1}) \\ \Leftrightarrow\ & 0 \in \partial(\mu\|Y_{k+1}\|_N) - Z_k + \lambda(Y_{k+1} - X_k) \\ \Leftrightarrow\ & 0 \in \partial\left(\frac{\mu}{\lambda}\|Y_{k+1}\|_N + Y_{k+1}\right) - \frac{1}{\lambda} Z_k - X_k \\ \Leftrightarrow\ & X_k + \frac{1}{\lambda} Z_k \in \partial\left(\frac{\mu}{\lambda}\|\cdot\|_N + \mathrm{Id}\right)(Y_{k+1}) \\ \Leftrightarrow\ & Y_{k+1} \in \mathrm{prox}_{\frac{\mu}{\lambda}\|\cdot\|_N}\left(X_k + \frac{1}{\lambda} Z_k\right). \end{aligned}$$

We can use Equality 1 to further expand the rule.