SI251 - Convex Optimization Homework 4

Deadline: 2024-06-15 23:59:59

1. You can use Word, Latex or handwriting to complete this assignment. If you want to submit a handwritten version, scan it clearly.
2. The **report** has to be submitted as a PDF file to Gradescope, other formats are not accepted.
3. The submitted file name is **student_id+your_student_name.pdf**.
4. Late policy: You have 4 free late days for the quarter and may use up to 2 late days per assignment with no penalty. Once you have exhausted your free late days, we will deduct a late penalty of 25% per additional late day. Note: The timeout period is recorded in days, even if you delay for 1 minute, it will still be counted as a 1 late day.
5. You are required to follow ShanghaiTech's academic honesty policies. You are not allowed to copy materials from other students or from online or published resources. Violating academic honesty can result in serious sanctions.

Any plagiarism will get Zero point.

1 Proximal Operator

For each of the following convex functions, compute the proximal operator prox_f.

(1) (10 pts) $f(x) = \lambda\|x\|_1$, where $x \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}_+$ is the regularization parameter.

(2) (20 pts) $f(X) = \lambda\|X\|_*$, where $X \in \mathbb{R}^{d\times m}$ is a matrix, $\|X\|_*$ denotes the nuclear norm, and $\lambda \in \mathbb{R}_+$ is the regularization parameter.

Solution:
(1) By definition of the prox,

$$\text{prox}_f(\mathbf{x}) = \arg\min_{\mathbf{u}\in\mathbb{R}^d} \left\{ \lambda\|\mathbf{u}\|_1 + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2 \right\}$$

This formulation takes the form as Lasso. Then we see the minimizer $\mathbf{u}^*$ is a soft-thresholding operator of x at λ

$$\left[\text{prox}_f(\mathbf{x})\right]_i = [\mathbf{u}^*]_i = \mathcal{S}_\lambda(\mathbf{u}) = \begin{cases} \mathbf{u}_i - \lambda & \text{if} \quad \mathbf{u}_i > \lambda \\ 0 & \text{if} \quad |\mathbf{u}_i| \leq \lambda \\ \mathbf{u}_i + \lambda & \text{if} \quad \mathbf{u}_i < -\lambda \end{cases}$$

which is

$$\text{prox}_f(\mathbf{v}_i) = \text{sgn}(\mathbf{v}_i)\cdot\max\{|\mathbf{v}_i| - \lambda, 0\}$$

(2) Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the SVD. Since the Frobenius norm is rotation-invariant (i.e., given a matrix A, for any orthogonal matrix U and V, we have $||A||_F = ||UAV||_F$), we write

$$\begin{aligned} \text{prox}_f(\mathbf{X}) &= \arg\min_{\mathbf{Y}\in\mathbb{R}^{d\times m}} \left\{ f(\mathbf{Y}) + \frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|_F^2 \right\} \\ &= \arg\min_{\mathbf{Y}\in\mathbb{R}^{d\times m}} \left\{ f(\mathbf{Y}) + \frac{1}{2}\|\mathbf{\Sigma}-\mathbf{U}^T\mathbf{Y}\mathbf{V}\|_F^2 \right\} \\ &= \mathbf{U}\left(\arg\min_{\tilde{\mathbf{Y}}\in\mathbb{R}^{d\times m}} \left\{ f(\tilde{\mathbf{Y}}) + \frac{1}{2}\|\mathbf{\Sigma}-\tilde{\mathbf{Y}}\|_F^2 \right\} \right)\mathbf{V}^T \\ &= \mathbf{U}\left(\text{prox}_f(\mathbf{\Sigma})\right)\mathbf{V}^T \end{aligned}$$

where we used rotational invariance of the Frobenius norm in the second equality of (1) and reparameterized the problem using $\tilde{\mathbf{Y}} = \mathbf{U}^T\mathbf{Y}\mathbf{V}$ in third equality of (1) since $f(\tilde{\mathbf{Y}}) = f(\mathbf{Y})$ We still have to show how to compute $\text{prox}_f(\mathbf{\Sigma})$ for a non-negative diagonal matrix $\mathbf{\Sigma}$ of singular values (Generally, the singular values obtained from the SVD are non-negative values.). Clearly the minimizer $\mathbf{Y}$ is diagonal, since every non-zero off-diagonal term in $\mathbf{Y}$ gets a positive penalty from the Frobenius norm. This yields

$$\begin{aligned} \text{prox}_f(\mathbf{D}) &= \arg\min_{\mathbf{Y}\in\mathbb{R}^{d\times m}} \left\{ f(\mathbf{Y}) + \frac{1}{2}\|\mathbf{D}-\mathbf{Y}\|_F^2 \right\} \\ &= \arg\min_{\mathbf{Y}\in\mathbb{R}^{d\times m}} \left\{ \lambda \sum_{j=1}^{\min(d,m)} \sigma_j(\mathbf{Y}) + \frac{1}{2}\left(\sigma_j(\mathbf{D}) - \sigma_j(\mathbf{Y})\right)^2 \right\} \\ &= \arg\min_{\mathbf{y}\in\mathbb{R}^{\min(d,m)}} \lambda\|\mathbf{y}\|_1 + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \end{aligned}$$

where $\mathbf{y}, \mathbf{x}$ are the vectors of singular values of $\mathbf{Y}$ and $\mathbf{X}$ respectively and the last equality of (2) follows by $\sigma_j(\mathbf{X}) = \sigma_j(\mathbf{D})$ for all j. After the problem reduced to a ℓ_1 regularized problem in the vector space, we see that the minimizer $\mathbf{y}^*$ is a soft-thresholding version of $\mathbf{x}$ at λ and thus we have

$$\left(\operatorname{prox}_f(\mathbf{D})\right)_{ij} = \begin{cases} 0 & i \neq j \\ \max\{\sigma_i(\mathbf{X}) - \lambda, 0\} & i = j \end{cases}$$

Thus the whole procedure for computing the proximal operator involves taking the SVD and softthresholding the singular values.

2 Alternating Direction Method of Multipliers

(35 pts) Consider the following problem.

$$\begin{array}{ll} \text{minimize} & -\log\det X + \operatorname{Tr}(XC) + \rho\|X\|_1 \\ \text{subject to} & X \succeq 0 \end{array} \tag{1}$$

In (1), $\|\cdot\|_1$ is the entrywise ℓ_1-norm. This problem arises in estimation of sparse undirected graphical models. C is the empirical covariance matrix of the observed data. The goal is to estimate a covariance matrix with sparse inverse for the observed data. In order to apply ADMM we rewrite (1) as

$$\begin{array}{ll} \text{minimize} & -\log\det X + \operatorname{Tr}(XC) + \mathbb{I}_{X\succeq 0}(X) + \rho\|Y\|_1 \\ \text{subject to} & X = Y \end{array} \tag{2}$$

where $\mathbb{I}_{X\succeq 0}(\cdot)$ is the indicator function associated with the set $X \succeq 0$. Please provide the ADMM update (the derivation process is required) for each variable at the t-th iteration.

Solution:

The augmented Lagrangian with penalty parameter μ for (2) is

$$\mathcal{L}_\mu(X, Y, Z) = -\log\det X + \operatorname{Tr}(XC) + \mathbb{I}_{X\succeq 0}(X) + \rho\|Y\|_1 + \mu\langle Z, X - Y\rangle + \frac{\mu}{2}\|X - Y\|_F^2.$$

Based on this, we derive the update rules for the ADMM algorithm. We have

$$X_k = \arg\min_X \mathcal{L}_\mu(X, Y_{k-1}, Z_{k-1})$$

$$= \arg\min_{X\succeq 0} \left(-\log\det X + \langle X, C\rangle + \rho\|Y_{k-1}\|_1 + \mu\langle Z_{k-1}, X - Y_{k-1}\rangle + \frac{\mu}{2}\|X - Y_{k-1}\|_F^2\right)$$

$$= \arg\min_{X\succeq 0} \left(-\log\det X + \frac{\mu}{2}\left\|X + \left(Z_{k-1} - Y_{k-1} + \frac{1}{\mu}C\right)\right\|_F^2\right).$$

Thus, letting

$$\hat{C}_k = -Z_{k-1} + Y_{k-1} - \frac{1}{\mu}C,$$

we have

$$\nabla_X \mathcal{L}_\mu(X_k, Y_{k-1}, Z_{k-1}) = -X^{-1} + \mu X - \mu\hat{C}_{k-1},$$

It is clear that if for some $X^* \geq 0$, $\nabla_X \mathcal{L}_\mu(X^*, Y_{k-1}, Z_{k-1}) = 0$; then X^* is $\arg\min_X \mathcal{L}_\mu(X, Y_{k-1}, Z_{k-1})$. Now if

$$\hat{C}_{k-1} = U\Lambda U^T, \quad \Lambda = \operatorname{diag}(\{\lambda_i\})$$

is the eigenvalue decomposition of $\hat{C}_{k-1}$ with $\lambda_i \geq 0$, then it is easy to see that for

$$X^* = F_\mu(\hat{C}_{k-1}) = UF_\mu(\Lambda)U^T, \quad F_\mu(\Lambda) = \frac{1}{2}\text{diag}\left(\{\lambda_i + \sqrt{\lambda_i^2 + \frac{4}{\mu}}\}\right),$$

where $F_\mu(\cdot)$ is given in (25.8). We also have

$$\begin{aligned} Y_k &= \arg\min_Y \mathcal{L}_\mu(X_k, Y, Z_{k-1}) \\ &= \arg\min_Y \left(-\log\det X_k + \langle X_k, C\rangle + \rho\|Y\|_1 + \mu\langle Z_{k-1}, X_k - Y\rangle + \frac{\mu}{2}\|X_k - Y\|_F^2\right) \\ &= \arg\min_Y \left(\frac{\mu}{2}\|Y - (X_k + Z_{k-1})\|_F^2 + \rho\|Y\|_1\right) \\ &= ET_{\rho/\mu}(X_k + Z_{k-1}), \end{aligned}$$

where $ET_{\rho/\mu}(\cdot)$ is given as in (25.4). Finally, we can write the ADMM steps for graphical Lasso which can be seen in Algorithm 1.

Algorithm 1 ADMM for solving the graphical Lasso problem

```
$Y_0 \leftarrow \hat{Y}, \quad Z_0 \leftarrow \hat{Z}, \quad k \leftarrow 1$ //initialize
$\mu \leftarrow \hat{\mu} > 0$
while convergence criterion is not satisfied do
    $X_k \leftarrow F_\mu(Y_{k-1} - Z_{k-1} - \frac{1}{\mu}C)$
    $Y_k \leftarrow ET_{\rho/\mu}(X_k + Z_{k-1})$
    $Z_k \leftarrow Z_{k-1} + \mu(X_k - Y_k)$
    $k \leftarrow k + 1$
end while
```

3 Monotone Operators and Base Splitting Schemes

(35 pts) Proof the theorem [1] below:

Theorem 1. *For $v \in \mathbb{R}^n$, the solution of the equation*

$$u^* = (I - JW)^{-T}v \tag{3}$$

is given by

$$u^* = v + W^T\tilde{u}^* \tag{4}$$

where I is the identity matrix and $\tilde{u}^*$ is a zero of the operator splitting problem $0 \in (F + G)(u^*)$, with operators defined as

$$F(\tilde{u}) = (I - W^T)(\tilde{u}), \quad G(\tilde{u}) = D\tilde{u} - v \tag{5}$$

where D is a diagonal matrix defined by $J = (I + D)^{-1}$ (where $J_{ii} > 0$).

[1] Details can refer to Winston, Ezra, and J. Zico Kolter. ”Monotone operator equilibrium networks.” Advances in neural information processing systems 33 (2020): 10718-10728. `https://arxiv.org/abs/2006.08591`

(Hint-1, please refer to Monotone Operators-note.pdf)
(Hint-2, $I = (I - JW)^{-T}(I - JW)^T$)

Proof. We begin with the case where $J_{ii} \neq 0$ and thus $D_{ii} < \infty$. As above, because proximal operators are themselves monotone non-expansive operators, we always have $0 \leq J_{ii} \leq 1$, so that $D_{ii} \geq 0$. Now, first assuming that $J_{ii} > 0$, and hence $D_{ii} < \infty$, we have

$$u = (I - JW)^{-T} v$$

$$\iff (I - W^T(I + D)^{-1})u = v$$

$$\iff W^{-T}u - (I + D)^{-1}u = W^{-T}v$$

$$\iff (I + D)W^{-T}u - u = (I + D)W^{-T}v$$

$$\iff W^{-T}u - u + DW^{-T}u = (I + D)W^{-T}v$$

$$\hat{u} - W^T\hat{u} + D\hat{u} = (I + D)W^{-T}v$$

where we define $\hat{u} = W^{-T}u$. To simplify the right hand side of this equation and remove the explicit $W^{-T}v$ terms[1] we note that

$$(I - JW)^{-T} = (I - W^TJ)^{-1} = I + (I - W^TJ)^{-1}W^TJ.$$

Thus, we can always solve the above equation with the v term of the form W^TJv, giving

$$(I + D)W^{-T}W^TJv = (I + D)Jv = v.$$

This gives us a (linear) operator splitting problem with the $\mathcal{F}$ and $\mathcal{G}$ operators given in (5) [2].

[2] Although we could solve this operator splitting problem directly, the presence of the $W^{-T}v$ term has two notable downsides: 1) even if the W matrix itself is nonsingular, it may be arbitrarily close to a singular matrix, thus making direct solutions with this matrix introduce substantial numerical errors; and 2) for operator splitting methods that do not require an inverse of W (e.g., forward-backward splitting), it would be undesirable to require an explicit inverse in the backward pass.