Supplementary Material for Safe Subspace Screening for Nuclear Norm Regularized Least Squares Problems

Qiang Zhou ZHOUQIANG@U.NUS.EDU

Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

Qi Zhao ELEQIZ@NUS.EDU.SG

Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

A. The Derivation of Dual Problem

Here, we give the detailed derivation of the dual problem of Eq. (2). First, we rewrite it as the following equivalent constrained optimization problem

$$\min \frac{1}{2}\left\|\mathbf{Z}\right\|_F^2 + \lambda \left\|\mathbf{W}\right\|_* \ s.t.\ \mathbf{Z} = \mathbf{X}\mathbf{W} - \mathbf{Y} \tag{S1}$$

Let us introduce the dual variable $\lambda\mathbf{P} \in \mathbb{R}^{n\times m}$ for the equality constraint, then the Lagrangian of Eq. (S1) can be written as

$$L(\mathbf{W},\mathbf{Z},\mathbf{P}) = \frac{1}{2}\left\|\mathbf{Z}\right\|_F^2 + \lambda \left\|\mathbf{W}\right\|_* + \lambda \operatorname{Tr}\left[\mathbf{P}^T\left(\mathbf{Z} - (\mathbf{X}\mathbf{W} - \mathbf{Y})\right)\right] \tag{S2}$$

Then, the dual problem $g(\mathbf{P})$ is

$$g(\mathbf{P}) = \inf_{\mathbf{W},\mathbf{Z}} L(\mathbf{W},\mathbf{Z},\mathbf{P}) = \lambda \operatorname{Tr}\left[\mathbf{P}^T\mathbf{Y}\right] + \inf_{\mathbf{Z}}\left\{\frac{1}{2}\left\|\mathbf{Z}\right\|_F^2 + \lambda \operatorname{Tr}\left[\mathbf{P}^T\mathbf{Z}\right]\right\} + \lambda \inf_{\mathbf{W}}\left\{\left\|\mathbf{W}\right\|_* - \operatorname{Tr}\left[\mathbf{P}^T\mathbf{X}\mathbf{W}\right]\right\} \tag{S3}$$

In order to obtain $g(\mathbf{P})$, we need to solve the two optimization problems in Eq. (S3). For the first optimization problem, it can be solved by setting its derivative with respect to $\mathbf{Z}$ equal to $\mathbf{0}$ and we obtain the optimal solution $\mathbf{Z} = -\lambda\mathbf{P}$. Substituting $\mathbf{Z}$ into Eq. (S3), we obtain the optimal value for the first problem as $-\frac{1}{2}\lambda^2\left\|\mathbf{P}\right\|_F^2$. For the second problem, we have (Boyd & Vandenberghe, 2004)

$$\lambda \inf_{\mathbf{W}}\left\{\left\|\mathbf{W}\right\|_* - \operatorname{Tr}\left[\mathbf{P}^T\mathbf{X}\mathbf{W}\right]\right\} = -\lambda \sup_{\mathbf{W}}\left\{\operatorname{Tr}\left[\mathbf{P}^T\mathbf{X}\mathbf{W}\right] - \left\|\mathbf{W}\right\|_*\right\} = 0 \ s.t.\ \left\|\mathbf{X}^T\mathbf{P}\right\|_2 \le 1 \tag{S4}$$

Combing the optimal values of the first and second problems, we obtain the dual problem

$$\max_{\mathbf{P}} \frac{1}{2}\left\|\mathbf{Y}\right\|_F^2 - \frac{\lambda^2}{2}\left\|\mathbf{P} - \frac{\mathbf{Y}}{\lambda}\right\|_F^2 \ s.t.\ \left\|\mathbf{X}^T\mathbf{P}\right\|_2 \le 1 \tag{S5}$$

which can be rewritten as the following equivalent problem

$$\min_{\mathbf{P}} \frac{\lambda^2}{2}\left\|\mathbf{P} - \frac{\mathbf{Y}}{\lambda}\right\|_F^2 \ s.t.\ \left\|\mathbf{X}^T\mathbf{P}\right\|_2 \le 1 \tag{S6}$$

Let $\mathbf{W}_\lambda^*$ and $\mathbf{P}_\lambda^*$ be the primal and dual optimal solutions, respectively. By using the KKT condition, we have

$$\mathbf{Z}_\lambda^* = \mathbf{X}\mathbf{W}_\lambda^* - \mathbf{Y} \tag{S7}$$

$$\mathbf{Z}_\lambda^* + \lambda \mathbf{P}_\lambda^* = \mathbf{0} \tag{S8}$$

Then, the above two equations establish the following relationship between $\mathbf{W}_\lambda^*$ and $\mathbf{P}_\lambda^*$

$$\lambda \mathbf{P}_\lambda^* = \mathbf{Y} - \mathbf{X}\mathbf{W}_\lambda^* \tag{S9}$$

To find the value of $\lambda_{\max}$ such that the solution $\mathbf{W}_\lambda^*$ is $\mathbf{0}$ for any $\lambda \geq \lambda_{\max}$, we substitute $\mathbf{W}_\lambda^* = \mathbf{0}$ into Eq. (S9) and obtain $\mathbf{P}_\lambda^* = \frac{\mathbf{Y}}{\lambda}$. Since $\mathbf{P}_\lambda^*$ is a dual feasible point and satisfies the constraints $\left\|\mathbf{X}^T \frac{\mathbf{Y}}{\lambda}\right\|_2 \leq 1$, which implies $\lambda_{\max} = \left\|\mathbf{X}^T \mathbf{Y}\right\|_2$.

B. The Screening Rule Based on KKT Condition is not Applicable for Subspace Screening

In this section, we derive why the screening rule based on KKT condition is not applicable for subspace screening. Suppose that the rank of $\mathbf{W}_\lambda^*$ is r and the SVD of $\mathbf{W}_\lambda^*$ is $\mathbf{U}_\lambda \mathbf{\Sigma}_\lambda \mathbf{V}_\lambda^T$ where $\mathbf{U}_\lambda \in \mathbb{R}^{d\times r}, \mathbf{\Sigma}_\lambda \in \mathbb{R}^{r\times r}$ and $\mathbf{V}_\lambda \in \mathbb{R}^{m\times r}$, then the subgradient of $\|\mathbf{W}_\lambda^*\|_*$ is (Watson, 1992)

$$\partial \|\mathbf{W}_\lambda^*\|_* = \{\mathbf{U}_\lambda \mathbf{V}_\lambda^T + \mathbf{Q} : \mathbf{U}_\lambda^T \mathbf{Q} = \mathbf{0}, \mathbf{Q}\mathbf{V}_\lambda = \mathbf{0}, \|\mathbf{Q}\|_2 \leq 1\} \tag{S10}$$

From the KKT condition, we have

$$0 \in \frac{\partial L(\mathbf{W}, \mathbf{Z}, \mathbf{P})}{\partial \mathbf{\Theta}_{ij}} = \mathrm{Tr}\left[\left(\frac{\partial L(\mathbf{W}, \mathbf{Z}, \mathbf{P})}{\partial \mathbf{W}}\right)^T \frac{\partial \mathbf{W}}{\partial \mathbf{\Theta}_{ij}}\right] = \mathbf{u}_i^T \mathbf{S}_{\mathrm{W}} \mathbf{v}_j - \mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j, \text{ where } \mathbf{S}_{\mathrm{W}} \in \partial \|\mathbf{W}_\lambda^*\|_* \tag{S11}$$

which implies that there exists $\mathbf{S}_{\mathrm{W}} \in \partial \|\mathbf{W}_\lambda^*\|_*$ such that

$$\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j = \mathbf{u}_i^T \mathbf{S}_{\mathrm{W}} \mathbf{v}_j \tag{S12}$$

Since $\{\mathbf{u}_i \mathbf{v}_j^T, i = 1, \ldots, d, j = 1, \ldots, m\}$ are orthogonal to each other, then the value of $\mathbf{u}_i^T \mathbf{U}_\lambda \mathbf{V}_\lambda^T \mathbf{v}_j$ will be zero if we have $(\mathbf{\Theta}_\lambda^*)_{ij} = 0$. In addition, if we know $(\mathbf{\Theta}_\lambda^*)_{ij} \neq 0$, then we obtain $\mathbf{u}_i \in \mathrm{span}(\mathbf{U}_\lambda^*)$ and $\mathbf{v}_j \in \mathrm{span}(\mathbf{V}_\lambda^*)$, which implies $\mathbf{u}_i^T \mathbf{Q} \mathbf{v}_j = 0$. As a result, we have

$$\mathbf{u}_i^T \mathbf{S}_{\mathrm{W}} \mathbf{v}_j \in \begin{cases} \mathbf{u}_i^T \mathbf{U}_\lambda \mathbf{V}_\lambda^T \mathbf{v}_j & \text{if } (\mathbf{\Theta}_\lambda^*)_{ij} \neq 0 \\ \mathbf{u}_i^T \mathbf{Q} \mathbf{v}_j, \|\mathbf{Q}\|_2 \leq 1 & \text{if } (\mathbf{\Theta}_\lambda^*)_{ij} = 0 \end{cases} \tag{S13}$$

According to Eq. (S12), the following holds

$$\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j \in \begin{cases} \mathbf{u}_i^T \mathbf{U}_\lambda \mathbf{V}_\lambda^T \mathbf{v}_j & \text{if } (\mathbf{\Theta}_\lambda^*)_{ij} \neq 0 \\ \mathbf{u}_i^T \mathbf{Q} \mathbf{v}_j, \|\mathbf{Q}\|_2 \leq 1 & \text{if } (\mathbf{\Theta}_\lambda^*)_{ij} = 0 \end{cases} \tag{S14}$$

Same as feature screening, we have $\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j \in [-1, 1]$ if $(\mathbf{\Theta}_\lambda^*)_{ij} = 0$. However, unlike feature screening, for $\left(\mathbf{\Theta}_{ij}^*\right)_{ij} \neq 0$, we usually do not have $\left|\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j\right| = 1$, which holds in feature screening since the absolute value of the subgradient of ℓ_1 norm at nonzero point is always 1. In particular, we also have $\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j \in [-1, 1]$ for nonzero $(\mathbf{\Theta}_\lambda^*)_{ij}$, thus the value of $\mathbf{u}_i^T \mathbf{X}^T \mathbf{P}_\lambda^* \mathbf{v}_j$ can not be used to determine whether $(\mathbf{\Theta}_\lambda^*)_{ij}$ is zero or not. Therefore, the screening rule based on KKT condition is not applicable for subspace screening.

C. Proof of Lemma 2

Proof. To prove this lemma, we first show that if there exists a γ such that $\mathbf{P}_{\lambda_0}^* = \gamma \mathbf{Y}$, then $\gamma = 1/\left\|\mathbf{X}^T \mathbf{Y}\right\|_2$. It is easy to check that $\mathbf{Y}/\left\|\mathbf{X}^T \mathbf{Y}\right\|_2$ satisfies the constraint in Eq. (13). Substituting $\mathbf{P} = \mathbf{Y}/\left\|\mathbf{X}^T \mathbf{Y}\right\|_2$ and $\mathbf{P}_{\lambda_0}^* = \gamma \mathbf{Y}$ into Eq. (13), we obtain

$$\mathrm{Tr}\left[\left(\gamma \mathbf{Y} - \frac{\mathbf{Y}}{\lambda_0}\right)^T \left(\frac{\mathbf{Y}}{\|\mathbf{X}^T \mathbf{Y}\|_2} - \gamma \mathbf{Y}\right)\right] = \left(\gamma - \frac{1}{\lambda_0}\right)\left(\frac{1}{\|\mathbf{X}^T \mathbf{Y}\|_2} - \gamma\right)\|\mathbf{Y}\|_F^2 \geq 0$$

Thus, we have $\gamma \in \left[1/\left\|\mathbf{X}^T \mathbf{Y}\right\|_2, 1/\lambda_0\right]$. Combining this with $\left\|\mathbf{X}^T \mathbf{P}_{\lambda_0}^*\right\|_2 = \left\|\mathbf{X}^T \gamma \mathbf{Y}\right\|_2 \leq 1$, we get $\gamma = 1/\left\|\mathbf{X}^T \mathbf{Y}\right\|_2$. Next we start to prove Lemma 2. We note that

$$\|\mathbf{A}\|_F^2 = \left\|\frac{\mathbf{Y}}{\lambda_0} - \mathbf{P}_{\lambda_0}^*\right\|_F^2 \geq 0 \tag{S15}$$

where the equality holds if and only if $\mathbf{Y}/\lambda_0 = \mathbf{P}^*_{\lambda_0}$. As shown above, $\mathbf{P}^*_{\lambda_0} = \mathbf{Y}/\lambda_0$ iff $\lambda_0 = \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$ while we know $\lambda_0 < \left\|\mathbf{X}^T\mathbf{Y}\right\|_2$, so we have $\|\mathbf{A}\|_F^2 > 0$ which implies $\mathbf{A} \neq \mathbf{0}$. The proof for $\mathbf{B} \neq \mathbf{0}$ is similar to the proof for $\mathbf{A} \neq \mathbf{0}$. □

D. Proof of Theorem 1

Proof. Let α and β denote the dual variable for the two constraints in Eq. (26), then the Lagrangian can be written as

$$L(\mathbf{R}, \alpha, \beta) = e\left(\mathbf{S}_{\mathrm{R}}\right)_{ij} + \alpha \operatorname{Tr}\left[\mathbf{A}^T(\mathbf{R}+\mathbf{B})\right] + \beta\left(\|\mathbf{R}\|_F^2 - \|\mathbf{B}\|_F^2\right) \tag{S16}$$

If we only consider the constraint $\|\mathbf{R}\|_F^2 \leq \|\mathbf{B}\|_F^2$ in Eq. (26), the optimal objective value is $-\|\mathbf{D}_{\cdot i}\|_2 \|\mathbf{B}\|_F$, which implies that the optimal value of Eq. (26) is lower bounded. Thus, the optimal dual variable β should be greater than 0, otherwise, the Lagrangian $L(\mathbf{R}, \alpha, \beta)$ is unbound below in $\mathbf{R}$ Setting the derivative of $L(\mathbf{R}, \alpha, \beta)$ with respect to $\mathbf{R}$ equal to $\mathbf{0}$, we obtain

$$e\mathbf{D}_{\cdot i}\left(\mathbf{V}_{\cdot j}\right)^T + \alpha\mathbf{A} + 2\beta\mathbf{R} = \mathbf{0} \Rightarrow \mathbf{R} = \frac{-e\mathbf{D}_{\cdot i}\left(\mathbf{V}_{\cdot j}\right)^T - \alpha\mathbf{A}}{2\beta} \tag{S17}$$

Substituting $\mathbf{R}$ into Eq. (S16), we obtain the dual problem of Eq. (26)

$$\begin{aligned} \max\ & -\frac{\alpha^2}{4\beta}\|\mathbf{A}\|_F^2 + \left(\operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right] - \frac{e\left(\mathbf{D}_{\cdot i}\right)^T\mathbf{A}\mathbf{V}_{\cdot j}}{2\beta}\right)\alpha - \frac{1}{4\beta}\|\mathbf{D}_{\cdot i}\|_2^2 - \beta\|\mathbf{B}\|_F^2 \\ s.t.\ & \alpha \geq 0, \beta > 0 \end{aligned} \tag{S18}$$

For $\mathbf{D}^T\mathbf{A}\mathbf{V}$, it can be expressed as

$$\mathbf{D}^T\mathbf{A}\mathbf{V} = \frac{\mathbf{U}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{W}^*_{\lambda_0}\mathbf{V}}{\lambda_0} = \frac{\mathbf{U}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Sigma}}{\lambda_0} = \frac{\widehat{\mathbf{\Sigma}}}{\lambda_0}$$

where $\widehat{\mathbf{\Sigma}} = \mathbf{U}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{\Sigma}$. Then, we have $\left(\mathbf{D}_{\cdot i}\right)^T\mathbf{A}\mathbf{V}_{\cdot j} = \frac{\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0}$. Hence, the dual problem can be rewritten as

$$\begin{aligned} \max\ & -\frac{\alpha^2}{4\beta}\|\mathbf{A}\|_F^2 + \left(\operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right] - \frac{e\widehat{\mathbf{\Sigma}}_{ij}}{2\lambda_0\beta}\right)\alpha - \frac{1}{4\beta}\|\mathbf{D}_{\cdot i}\|_2^2 - \beta\|\mathbf{B}\|_F^2 \\ & \alpha \geq 0, \beta > 0 \end{aligned} \tag{S19}$$

Maximizing the dual problem leads to a closed form solution for α with given β

$$\alpha = \max\left(\frac{2\lambda_0 \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\beta - e\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0\|\mathbf{A}\|_F^2}, 0\right) \tag{S20}$$

Here we need to consider two cases: $\alpha = 0$ and $\alpha \neq 0$.
If $\alpha = 0$, which means

$$2\lambda_0 \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\beta - e\widehat{\mathbf{\Sigma}}_{ij} \leq 0 \tag{S21}$$

In addition, substituting $\alpha = 0$ into Eq. (S17) gives

$$\mathbf{R} = -\frac{e\mathbf{D}_{\cdot i}\left(\mathbf{V}_{\cdot j}\right)^T}{2\beta} \tag{S22}$$

Since we have $\beta \neq 0$, by using the complementary slackness condition, we have

$$\|\mathbf{R}\|_F^2 = \frac{1}{4\beta^2}\left\|\left(\mathbf{D}_{\cdot i}\right)^T\mathbf{V}_{\cdot j}\right\|_F^2 = \|\mathbf{B}\|_F^2 \tag{S23}$$

Then, we have

$$\beta = \frac{\|\mathbf{D}_{\cdot i}\|_2}{2\|\mathbf{B}\|_F} \tag{S24}$$

Substituting β into Eq. (S21) gives

$$\lambda_0 \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right] \|\mathbf{D}_{\cdot i}\|_2 \leq e \|\mathbf{B}\|_F \widehat{\mathbf{\Sigma}}_{ij} \tag{S25}$$

In addition, we also have

$$(\mathbf{S}_{\mathrm{R}})_{ij} = -e \|\mathbf{D}_{\cdot i}\|_2 \|\mathbf{B}\|_F \tag{S26}$$

Next, we consider the case that $\alpha \neq 0$. In other words,

$$\alpha = \frac{2\lambda_0 \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right] \beta - e\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0 \|\mathbf{A}\|_F^2} \tag{S27}$$

Substituting α into Eq. (S17) gives

$$\mathbf{R} = \frac{e}{2\beta}\left(\frac{\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0 \|\mathbf{A}\|_F^2}\mathbf{A} - \mathbf{D}_{\cdot i}\left(\mathbf{V}_{\cdot j}\right)^T\right) - \frac{\operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]}{\|\mathbf{A}\|_F^2}\mathbf{A} \tag{S28}$$

Similar to the last case, we can use the complementary slackness condition since $\beta \neq 0$. By $\|\mathbf{R}\|_F^2 = \|\mathbf{B}\|_F^2$, gives

$$\beta = \frac{\sqrt{\lambda_0^2 \|\mathbf{A}\|_F^2 \|\mathbf{D}_{\cdot i}\|_2 - \widehat{\mathbf{\Sigma}}_{ij}^2}}{2\lambda_0\sqrt{\|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 - \left(\operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\right)^2}} \tag{S29}$$

Then, we obtain $\mathbf{S}_{\mathrm{R}ij}$

$$(\mathbf{S}_{\mathrm{R}})_{ij} = \frac{-e\sqrt{\left(\|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2 - \left(\operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\right)^2\right)\left(\lambda_0^2 \|\mathbf{A}\|_F^2 \|\mathbf{D}_{\cdot i}\|_2^2 - \widehat{\mathbf{\Sigma}}_{ij}^2\right)} - \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0 \|\mathbf{A}\|_F^2} \tag{S30}$$

This ends of the proof. □

E. Proof of Corollary 1

Proof. We first show the proof for $\mathbf{\Phi}_{ij}$, which is equal to $0.5\lambda \max\left((\mathbf{S}_{\mathrm{R}})_{ij} - (\mathbf{S}_{\mathrm{C}})_{ij}\right)$. Therefore, we need to minimize Eq. (26) by setting $e = -1$. Here we need to consider two cases: Eq. (28) holds or does not. If Eq. (28) holds, the optimal value for $(\mathbf{S}_{\mathrm{R}})_{ij}$ is $\|\mathbf{B}\|_F \|\mathbf{D}_{\cdot i}\|_2$. So the value for $\mathbf{\Phi}_{ij}$ is $0.5\lambda\left(\|\mathbf{B}\|_F \|\mathbf{D}_{\cdot i}\|_2 - (\mathbf{S}_{\mathrm{C}})_{ij}\right)$. Otherwise, the optimal value for $(\mathbf{S}_{\mathrm{R}})_{ij}$ is equal to

$$\eta = \frac{\mathbf{G}_{ij} - \operatorname{Tr}\left[\mathbf{A}^T\mathbf{B}\right]\widehat{\mathbf{\Sigma}}_{ij}}{\lambda_0 \|\mathbf{A}\|_F^2} \tag{S31}$$

As a result, the value for $\mathbf{\Phi}_{ij}$ is $0.5\lambda\left(\eta - (\mathbf{S}_{\mathrm{C}})_{ij}\right)$.

Next, we consider the proof for $\mathbf{\Psi}_{ij}$ that is $0.5\lambda \max\left(-(\mathbf{S}_{\mathrm{R}})_{ij} + (\mathbf{S}_{\mathrm{C}})_{ij}\right)$. The proof is similar to the proof for $\mathbf{\Phi}_{ij}$ except we need to set $e = 1$ in this case. This ends the overall proof. □

References

Boyd, Stephen P. and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Watson, G. Alistair. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Its Applications*, 170:33–45, 1992.