# Supplementary Material: On the Faster Alternating Least-Squares for CCA

## 1 Theoretical Analysis

**Theorem 3.1** Given data matrices $(\mathbf{X},\mathbf{Y})\in\mathbb{R}^{d_x\times n}\times\mathbb{R}^{d_y\times n}$, if $\sigma_k^2>2\sqrt{\beta}=\sigma_{k+1}^2$, Algorithm 1 then computes $\mathbf{\Phi}_T$ and $\mathbf{\Psi}_T$ which are estimates of top-$k$ canonical subspaces $(\mathbf{U},\mathbf{V})$ such that $\sin\theta_T\le\epsilon$ and $\mathbf{\Phi}_T^\top\mathbf{C}_{xx}\mathbf{\Phi}_T=\mathbf{\Psi}_T^\top\mathbf{C}_{yy}\mathbf{\Psi}_T=\mathbf{I}$, in $T=O(\sqrt{\frac{\sigma_k^2}{\sigma_k^2-\sigma_{k+1}^2}}\log\frac{1}{\epsilon(\sigma_k^2-\sigma_{k+1}^2)\cos\theta_0})$ iterations. If accelerated gradient descent is used as the least-squares solver, the overall running time is at most

$$O((dk^2+k\mathrm{nnz}(\mathbf{X},\mathbf{Y})\kappa^{\frac{1}{2}}(\mathbf{X},\mathbf{Y})\log\frac{c_1c_2}{(\sigma_k^2-\sigma_{k+1}^2)\cos\theta_0})\sqrt{\frac{\sigma_k^2}{\sigma_k^2-\sigma_{k+1}^2}}\log\frac{1}{\epsilon(\sigma_k^2-\sigma_{k+1}^2)\cos\theta_0}),$$

where $d=\max\{d_x,d_y\}$, $\mathrm{nnz}(\mathbf{X},\mathbf{Y})=\mathrm{nnz}(\mathbf{X})+\mathrm{nnz}(\mathbf{Y})$, $\kappa(\mathbf{X},\mathbf{Y})=\max\{\kappa(\mathbf{C}_{xx}),\kappa(\mathbf{C}_{yy})\}$,

$$c_1=\max_t\frac{\tan\max\{\theta_t,\,\theta_{\max}(\widehat{\mathbf{\Phi}}_t,\mathbf{U}),\,\theta_{\max}(\widehat{\mathbf{\Psi}}_t,\mathbf{V})\}}{\tan\min\{\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}}),\,\theta_{\max}(\mathbf{Q}_t,\widetilde{\mathbf{V}})\}},\text{ and } c_2=\max\{\max_t\frac{\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{\theta_{\min}(\mathbf{P}_t,\widetilde{\mathbf{U}})},\max_t\frac{\theta_{\max}(\mathbf{Q}_t,\widetilde{\mathbf{V}})}{\theta_{\min}(\mathbf{Q}_t,\widetilde{\mathbf{V}})}\}\ .$$

**Proof** We now follow the proof sketch given in the main text to conduct the analysis. Note that $\theta_t\triangleq\max\{\theta_{\max}(\mathbf{\Phi}_t,\mathbf{U}),\theta_{\max}(\mathbf{\Psi}_t,\mathbf{V})\}$ and we need to show $\sin\theta_{\max}(\mathbf{\Phi}_t,\mathbf{U})\le\epsilon$ and $\sin\theta_{\max}(\mathbf{\Psi}_t,\mathbf{V})\le\epsilon$ hold simultaneously. Recall that our update is

$$\begin{cases}\mathbf{\Phi}_{t+1}\mathbf{R}_{t+1}=\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\big(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t+\boldsymbol{\xi}_t\big)-\beta\mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1}+\widehat{\boldsymbol{\xi}}_t\\ \mathbf{\Psi}_{t+1}\mathbf{S}_{t+1}=\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_t+\boldsymbol{\eta}_t\big)-\beta\mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1}+\widehat{\boldsymbol{\eta}}_t\end{cases}.$$

It can be equivalently written as

$$\begin{cases}\mathbf{\Phi}_{t+1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t=\Big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\left(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\mathbf{\Phi}_t+\boldsymbol{\xi}_t\right)-\beta\mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1}+\widehat{\boldsymbol{\xi}}_t\Big)\widetilde{\mathbf{R}}_t\\ \mathbf{\Psi}_{t+1}\mathbf{S}_{t+1}\widetilde{\mathbf{S}}_t=\big(\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\big(\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{\Psi}_t+\boldsymbol{\eta}_t\big)-\beta\mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1}+\widehat{\boldsymbol{\eta}}_t\big)\widetilde{\mathbf{S}}_t\end{cases},$$

where

$$\widetilde{\mathbf{R}}_t=\begin{cases}(\mathbf{I}+\mathbf{R}_t^{-\top}\mathbf{R}_t^{-1})^{-\frac{1}{2}}, & t>0\\ \mathbf{I}, & t=0\end{cases},\quad \widetilde{\mathbf{S}}_t=\begin{cases}(\mathbf{I}+\mathbf{S}_t^{-\top}\mathbf{S}_t^{-1})^{-\frac{1}{2}}, & t>0\\ \mathbf{I}, & t=0\end{cases}.$$

We now focus on the first equation. Together with $\mathbf{\Phi}_t\widetilde{\mathbf{R}}_t=\mathbf{\Phi}_t\widetilde{\mathbf{R}}_t$, it leads to the following augmented system:

$$\mathbf{P}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t=\mathbf{B}_\phi^{-1}\mathbf{A}_\phi\mathbf{P}_t+\boldsymbol{\delta}_t,$$

where

$$\mathbf{A}_\phi=\begin{pmatrix}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top & -\beta\mathbf{C}_{xx}\\ \mathbf{C}_{xx} & \mathbf{0}\end{pmatrix},\quad \mathbf{B}_\phi=\begin{pmatrix}\mathbf{C}_{xx} & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_{xx}\end{pmatrix},$$

$$\mathbf{P}_t=\begin{pmatrix}\mathbf{\Phi}_t\\ \mathbf{\Phi}_{t-1}\mathbf{R}_t^{-1}\end{pmatrix}\widetilde{\mathbf{R}}_t,\quad \boldsymbol{\delta}_t=\begin{pmatrix}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\boldsymbol{\xi}_t+\widehat{\boldsymbol{\xi}}_t\\ \mathbf{0}\end{pmatrix}\widetilde{\mathbf{R}}_t.$$

Since $\mathbf{\Phi}_t$ is $\mathbf{C}_{xx}$-orthonormal, it is easy to see that $\mathbf{P}_t$ is $\mathbf{B}_\phi$-orthonormal. To continue, we can write the SVD of $\mathbf{C}_{xy}$ as follows:

$$\mathbf{C}_{xy}=\mathbf{C}_{xx}\left(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top+\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{V}_\perp^\top\right)\mathbf{C}_{yy},$$

where $(\mathbf{U}_\perp, \mathbf{\Sigma}_\perp, \mathbf{V}_\perp)$ consists of the $(\text{rank}(\mathbf{C}_{xy}) - k)$ remaining triples of the left singular vector in metric $\mathbf{C}_{xx}$, singular value, and right singular vector in metric $\mathbf{C}_{yy}$, other than those in $(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V})$. It thus holds that

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top = \mathbf{C}_{xx}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{\Sigma}_\perp^2\mathbf{U}_\perp^\top\right)\mathbf{C}_{xx},$$

and accordingly, by Lemma I in Section 2, $\mathbf{A}_\phi$ has the following Schur decomposition in metric $\mathbf{B}_\phi$:

$$\mathbf{A}_\phi = \mathbf{B}_\phi \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \widetilde{\widetilde{\mathbf{\Sigma}}} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi,$$

where notations can be found in Lemma I. Further, since $\widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{\Sigma}}_\perp$ (see Lemma I in Section 2) don't share eigenvalues by assumption that $\sigma_k > \sigma_{k+1}$, there exists (Golub and Van Loan, 2013) a matrix $\mathbf{\Xi}$ such that $\widetilde{\mathbf{\Sigma}}\,\mathbf{\Xi} - \mathbf{\Xi}\,\widetilde{\mathbf{\Sigma}}_\perp = -\widetilde{\widetilde{\mathbf{\Sigma}}}$ and thus

$$\begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \widetilde{\widetilde{\mathbf{\Sigma}}} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1}.$$

Plugging the above equation and the Schur decomposition of $\mathbf{A}_\phi$ into the augmented system and then pre-multiply both sides by $\begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi$, results in the following equation:

$$\begin{aligned} &\begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t = \\ &\quad \begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_t + \begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \boldsymbol{\delta}_t. \end{aligned}$$

Letting

$$\begin{aligned} \begin{pmatrix} \widetilde{\mathbf{X}}_t \\ \widetilde{\mathbf{Y}}_t \end{pmatrix} &= \begin{pmatrix} \mathbf{I} & \mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^{-1} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_t \\ &= \begin{pmatrix} \mathbf{I} & -\mathbf{\Xi} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_t = \begin{pmatrix} \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t \\ \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t \end{pmatrix}, \end{aligned}$$

where $\mathbf{W}_{\mathbf{U}} = \widetilde{\mathbf{U}} - \widetilde{\mathbf{U}}_\perp\mathbf{\Xi}^{\mathrm{H}}$, the above equation then can be split into the following two equations:

$$\begin{cases} \widetilde{\mathbf{X}}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t = \widetilde{\mathbf{\Sigma}}\,\widetilde{\mathbf{X}}_t + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t \\ \widetilde{\mathbf{Y}}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t = \widetilde{\mathbf{\Sigma}}_\perp\widetilde{\mathbf{Y}}_t + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t \end{cases}.$$

Hence, we have that

$$\begin{aligned} \widetilde{\mathbf{Z}}_{t+1} = \ &\widetilde{\mathbf{Y}}_{t+1}\widetilde{\mathbf{X}}_{t+1}^{-1} = \left(\widetilde{\mathbf{Y}}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t\right)\left(\widetilde{\mathbf{X}}_{t+1}\widetilde{\mathbf{R}}_{t+1}^{-1}\mathbf{R}_{t+1}\widetilde{\mathbf{R}}_t\right)^{-1} \\ &= \left(\widetilde{\mathbf{\Sigma}}_\perp\widetilde{\mathbf{Y}}_t + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\right)\left(\widetilde{\mathbf{\Sigma}}\,\widetilde{\mathbf{X}}_t + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\right)^{-1} \\ &= \left(\widetilde{\mathbf{\Sigma}}_\perp\widetilde{\mathbf{Y}}_t\widetilde{\mathbf{X}}_t^{-1} + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\right)\left(\widetilde{\mathbf{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\right)^{-1} \\ &= \left(\widetilde{\mathbf{\Sigma}}_\perp\widetilde{\mathbf{Z}}_t + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\left(\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\widetilde{\mathbf{Z}}_t\right)^{-1}\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\widetilde{\mathbf{Z}}_t\right)\left(\widetilde{\mathbf{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\right)^{-1} \\ &= \left(\widetilde{\mathbf{\Sigma}}_\perp + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\left(\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\widetilde{\mathbf{Z}}_t\right)^{-1}\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\right)\widetilde{\mathbf{Z}}_t\left(\widetilde{\mathbf{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\right)^{-1}, \end{aligned}$$

and thus that

$$\widetilde{\mathbf{Z}}_T = \prod_{t=T-1}^{0}\left(\widetilde{\mathbf{\Sigma}}_\perp + \widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\left(\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\widetilde{\mathbf{Z}}_t\right)^{-1}\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\right)\widetilde{\mathbf{Z}}_0\prod_{t'=0}^{T-1}\left(\widetilde{\mathbf{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_{t'}\widetilde{\mathbf{X}}_{t'}^{-1}\right)^{-1}.$$

By Lemma 12 in Ge et al. (2016), $\sin\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}}) = \|\widetilde{\mathbf{U}}_\perp^\top \mathbf{B}_\phi \mathbf{P}_t\|_2$. We then can write that

$$\sin\theta_{\max}(\mathbf{P}_T, \widetilde{\mathbf{U}}) = \|\widetilde{\mathbf{Y}}_T\|_2 \le \|\widetilde{\mathbf{Z}}_T\|_2 \|\widetilde{\mathbf{X}}_T\|_2,$$

where

$$\begin{aligned}\|\widetilde{\mathbf{X}}_T\|_2 &= \left\| \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Xi} \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_t \right\|_2 \\ &\le \left\| \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Xi} \end{pmatrix} \right\|_2 \left\| \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \right\|_{\mathbf{B}_\phi, 2} \|\mathbf{P}_t\|_{\mathbf{B}_\phi, 2} = 1 + \|\boldsymbol{\Xi}\|_2,\end{aligned}$$

where the last equality with $\|\mathbf{P}_t\|_{\mathbf{B}_\phi} = \|\mathbf{B}_\phi^{\frac{1}{2}} \mathbf{P}_t\|_2 = 1$ is by the $\mathbf{B}_\phi$-orthonormality of both $\begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}$ and $\mathbf{P}_t$. What's more, we have permutation matrix $\boldsymbol{\Pi}$, constant $1+\gamma = \frac{8(1+\beta)}{\sigma_k^+ - \sqrt{\beta}}$, and diagonal matrix $\boldsymbol{\Gamma} = \operatorname{diag}\big(\operatorname{diag}(1, 1+\gamma), \operatorname{diag}(1, 1+\gamma), \mathbf{I}\big)$ such that

$$\boldsymbol{\Pi}\widetilde{\boldsymbol{\Sigma}}_\perp \boldsymbol{\Pi}^\top = \operatorname{diag}\big(\boldsymbol{\Sigma}_{k+1}, \cdots, \boldsymbol{\Sigma}_r, \mathbf{I}\big), \quad \boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\boldsymbol{\Sigma}}_\perp \boldsymbol{\Pi}^\top \boldsymbol{\Gamma}^{-1} = \operatorname{diag}\Big(\boldsymbol{\Sigma}_{k+1}^{(\gamma)}, \cdots, \boldsymbol{\Sigma}_r^{(\gamma)}, \mathbf{I}\Big),$$

where $\sigma_j^\pm$ is defined in Lemma I of Section 2, and

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} \sigma_j^+ & -(1 + (\sigma_j^+)^2) \\ 0 & \sigma_j^- \end{pmatrix}, \quad \boldsymbol{\Sigma}_j^{(\gamma)} = \begin{pmatrix} \sigma_j^+ & -\frac{1+(\sigma_j^+)^2}{1+\gamma} \\ 0 & \sigma_j^- \end{pmatrix}.$$

Thus, we can write that

$$\begin{aligned}\|\widetilde{\mathbf{Z}}_T\|_2 &= \Big\| \boldsymbol{\Pi}^\top \boldsymbol{\Gamma}^{-1} \prod_{t=T-1}^{0} \Big( \boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\boldsymbol{\Sigma}}_\perp \boldsymbol{\Pi}^\top \boldsymbol{\Gamma}^{-1} + \boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\mathbf{U}}_\perp^{\mathrm{H}} \mathbf{B}_\phi \boldsymbol{\delta}_t \\ &\qquad \widetilde{\mathbf{X}}_t^{-1} (\widetilde{\mathbf{Z}}_t^{\mathrm{H}} \widetilde{\mathbf{Z}}_t)^{-1} \widetilde{\mathbf{Z}}_t^{\mathrm{H}} \boldsymbol{\Pi}^\top \boldsymbol{\Gamma}^{-1} \Big) \boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\mathbf{Z}}_0 \prod_{t'=0}^{T-1} \Big( \widetilde{\boldsymbol{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}} \mathbf{B}_\phi \boldsymbol{\delta}_{t'} \widetilde{\mathbf{X}}_{t'}^{-1} \Big)^{-1} \Big\| \\ &\le \|\boldsymbol{\Gamma}\|_2 \|\boldsymbol{\Gamma}^{-1}\|_2 \|\widetilde{\mathbf{Z}}_0\|_2 \prod_{t=0}^{T-1} \big\| (\widetilde{\boldsymbol{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}} \mathbf{B}_\phi \boldsymbol{\delta}_t \widetilde{\mathbf{X}}_t^{-1})^{-1} \big\|_2 \Big( \|\boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\boldsymbol{\Sigma}}_\perp \boldsymbol{\Pi}^\top \boldsymbol{\Gamma}^{-1}\|_2 \\ &\qquad + \|\boldsymbol{\Gamma}\|_2 \|\boldsymbol{\Gamma}^{-1}\|_2 \|\widetilde{\mathbf{U}}_\perp\|_{\mathbf{B}_\phi, 2} \|\boldsymbol{\delta}_t\|_{\mathbf{B}_\phi, 2} \|\widetilde{\mathbf{X}}_t^{-1} (\widetilde{\mathbf{Z}}_t^{\mathrm{H}} \widetilde{\mathbf{Z}}_t)^{-1} \widetilde{\mathbf{Z}}_t^{\mathrm{H}}\|_2 \Big),\end{aligned}$$

where $\|\boldsymbol{\Gamma}\|_2 \|\boldsymbol{\Gamma}^{-1}\|_2 = 1 + \gamma$ and $\|\widetilde{\mathbf{U}}_\perp\|_{\mathbf{B}_\phi, 2} = 1$. The remaining factors above can be derived as follows. First, note that $\mathbf{W}_{\mathbf{U}}$ spans the top-$k$ generalized eigenspace of the matrix pair $(\mathbf{A}_\phi^{\mathbf{H}}, \mathbf{B}_\phi)$ which has $\mathbf{B}_\phi$-orthonormal basis $\widetilde{\widetilde{\mathbf{U}}}$ (see Lemma I in Section 2). In fact, by the Schur decomposition of $\mathbf{A}_\phi$, it holds that

$$\begin{aligned}\mathbf{A}_\phi^{\mathbf{H}} \mathbf{W}_{\mathbf{U}} &= \mathbf{B}_\phi \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{\Sigma}} & \mathbf{0}^{\mathrm{H}} \\ \widetilde{\widetilde{\boldsymbol{\Sigma}}}^{\mathrm{H}} & \widetilde{\boldsymbol{\Sigma}}_\perp^{\mathrm{H}} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\boldsymbol{\Xi} \end{pmatrix}^{\mathrm{H}} = \mathbf{B}_\phi \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{\Sigma}} \\ (\widetilde{\widetilde{\boldsymbol{\Sigma}}} - \boldsymbol{\Xi}\widetilde{\boldsymbol{\Sigma}}_\perp)^{\mathrm{H}} \end{pmatrix} \\ &= \mathbf{B}_\phi \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{\Sigma}} \\ (-\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\Xi})^{\mathrm{H}} \end{pmatrix} = \mathbf{B}_\phi \mathbf{W}_{\mathbf{U}} \widetilde{\boldsymbol{\Sigma}}.\end{aligned}$$

Letting $\mathbf{G} = (\mathbf{I} + \boldsymbol{\Xi}\boldsymbol{\Xi}^{\mathrm{H}})^{\frac{1}{2}}$, we have that $\|\mathbf{G}^{-\mathrm{H}}\|_2 \le 1$ due to $\mathbf{G} \succcurlyeq \mathbf{I}$, and $\mathbf{W}_{\mathbf{U}} \mathbf{G}^{-1} = \widetilde{\widetilde{\mathbf{U}}}\mathbf{O}$ for a certain unitary matrix $\mathbf{O} \in \mathbb{C}^{k \times k}$. Consequently,

$$\begin{aligned}\|\widetilde{\mathbf{Z}}_0\|_2 &= \|\widetilde{\mathbf{Y}}_0 \widetilde{\mathbf{X}}_0^{-1}\|_2 \le \|\widetilde{\mathbf{X}}_0^{-1}\|_2 = \big\| \big( (\mathbf{W}_{\mathbf{U}} \mathbf{G}^{-1})^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_0 \big)^{-1} \mathbf{G}^{-\mathrm{H}} \big\|_2 \\ &\le \big\| \big( \widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}} \mathbf{B}_\phi \mathbf{P}_0 \big)^{-1} \big\|_2 = \big\| \big( \mathbf{D}(\beta) \mathbf{U}^{\mathrm{H}} \mathbf{C}_{xx} \boldsymbol{\Phi}_0 \big)^{-1} \big\|_2 = \frac{1}{\sigma_{\min}\big( \mathbf{D}(\beta) \mathbf{U}^{\mathrm{H}} \mathbf{C}_{xx} \boldsymbol{\Phi}_0 \big)} \\ &\le \frac{1}{\sigma_{\min}(\mathbf{D}(\beta)) \sigma_{\min}\big( \mathbf{U}^{\mathrm{H}} \mathbf{C}_{xx} \boldsymbol{\Phi}_0 \big)} \le \frac{\sqrt{\beta^2 + (\sigma_k^+)^2}}{\sigma_k^+ \sigma_{\min}\big( \mathbf{U}^{\mathrm{H}} \mathbf{C}_{xx} \boldsymbol{\Phi}_0 \big)},\end{aligned}$$

where $\sigma_{\min}(\dot{)}$ represents the minimum singular value of a matrix. Similarly,

$$\begin{aligned}\big\|\widetilde{\mathbf{X}}_t^{-1}\big\|_2 &\leq \big\|\big(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big\|_2\\ &\leq \big\|\big(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big(\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)\big\|_2\big\|\big(\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big\|_2 = \frac{a_{\phi,t}}{\cos\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})},\end{aligned}$$

where $a_{\phi,t} = \big\|\big(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big(\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)\big\|_2$ and Lemma 12 in Ge et al. (2016) is used in the last equality. Noting that $\|\mathbf{C}\|_2 \leq \sigma_1 \leq 1$ and $\|\widetilde{\mathbf{R}}_t\|_2 \leq 1$ as $\mathbf{0} \prec \widetilde{\mathbf{R}}_t \prec \mathbf{I}$, it holds that

$$\begin{aligned}\|\boldsymbol{\delta}_t\|_{\mathbf{B}_\phi,2} &= \left\|\mathbf{B}_\phi^{\frac{1}{2}}\begin{pmatrix}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\boldsymbol{\xi}_t + \widehat{\boldsymbol{\xi}}_t\\ \mathbf{0}\end{pmatrix}\widetilde{\mathbf{R}}_t\right\|_2 \leq \big(\|\mathbf{C}\|_2\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2} + \big\|\widehat{\boldsymbol{\xi}}_t\big\|_{\mathbf{C}_{xx},2}\big)\|\widetilde{\mathbf{R}}_t\|_2\\ &\leq \|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2} + \big\|\widehat{\boldsymbol{\xi}}_t\big\|_{\mathbf{C}_{xx},2},\end{aligned}$$

$$\begin{aligned}\big\|\big(\widetilde{\boldsymbol{\Sigma}} + \mathbf{W}_{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\boldsymbol{\delta}_t\widetilde{\mathbf{X}}_t^{-1}\big)^{-1}\big\|_2 &\leq \frac{1}{\sigma_{\min}(\widetilde{\boldsymbol{\Sigma}}) - \|\mathbf{W}_{\mathbf{U}}\|_{\mathbf{B}_\phi,2}\|\boldsymbol{\delta}\|_{\mathbf{B}_\phi,2}\big\|\widetilde{\mathbf{X}}_t^{-1}\big\|_2}\\ &\leq \frac{1}{\sigma_k^+ - a_{\phi,t}(1+\|\boldsymbol{\Xi}\|_2)(\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2} + \big\|\widehat{\boldsymbol{\xi}}_t\big\|_{\mathbf{C}_{xx},2})/\cos\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}.\end{aligned}$$

Also, we have that

$$\begin{aligned}\|\boldsymbol{\Gamma}\boldsymbol{\Pi}\widetilde{\boldsymbol{\Sigma}}_\perp\boldsymbol{\Pi}^\top\boldsymbol{\Gamma}^{-1}\|_2 &= \max_{k+1\leq j\leq r}\|\boldsymbol{\Sigma}_j^{(\gamma)}\|_2 \leq \max_{k+1\leq j\leq r}(|\sigma_j^+| + \frac{1+\beta}{1+\gamma})\\ &= \sqrt{\beta} + \frac{1+\beta}{1+\gamma} = \sqrt{\beta} + \frac{\sigma_k^+ - \sqrt{\beta}}{8},\end{aligned}$$

and that

$$\begin{aligned}\big\|\widetilde{\mathbf{X}}_t^{-1}\big(\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\widetilde{\mathbf{Z}}_t\big)^{-1}\widetilde{\mathbf{Z}}_t^{\mathrm{H}}\big\|_2 &= \big\|\widetilde{\mathbf{X}}_t^{-1}\big((\widetilde{\mathbf{Y}}_t\widetilde{\mathbf{X}}_t^{-1})^{\mathrm{H}}(\widetilde{\mathbf{Y}}_t\widetilde{\mathbf{X}}_t^{-1})\big)^{-1}(\widetilde{\mathbf{Y}}_t\widetilde{\mathbf{X}}_t^{-1})^{\mathrm{H}}\big\|_2 = \big\|\big(\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\widetilde{\mathbf{Y}}_t\big)^{-1}\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\big\|_2\\ &= \big\|\big(\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\widetilde{\mathbf{Y}}_t\big)^{-\frac{1}{2}}\big\|_2\big\|\big(\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\widetilde{\mathbf{Y}}_t\big)^{-\frac{1}{2}}\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\big\|_2 = \big\|\big(\widetilde{\mathbf{Y}}_t^{\mathrm{H}}\widetilde{\mathbf{Y}}_t\big)^{-\frac{1}{2}}\big\|_2\\ &= \big\|\big((\widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t)^{\mathrm{H}}(\widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t)\big)^{-\frac{1}{2}}\big\|_2 = \big\|\big(\mathbf{I} - \mathbf{P}_t^{\mathrm{H}}\mathbf{B}_\phi\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-\frac{1}{2}}\big\|_2\\ &= \frac{1}{\sqrt{1-\sigma_{\max}^2(\mathbf{P}_t^{\mathrm{H}}\mathbf{B}_\phi\widetilde{\mathbf{U}})}} = \frac{1}{\sin\theta_{\min}(\mathbf{P}_t,\widetilde{\mathbf{U}})} = \frac{c_{\phi,t}}{c_{\phi,t}\sin\theta_{\min}(\mathbf{P}_t,\widetilde{\mathbf{U}})}\\ &\leq \frac{c_{\phi,t}}{\sin(c_{\phi,t}\theta_{\min}(\mathbf{P}_t,\widetilde{\mathbf{U}}))} = \frac{c_{\phi,t}}{\sin\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})},\end{aligned}$$

where $c_{\phi,t} = \frac{\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{\theta_{\min}(\mathbf{P}_t,\widetilde{\mathbf{U}})}$, the inequality is by the fact that $|\sin(nx)| \leq n\sin(x)$ for any natural number and real $x$, and the fifth equality is due to that $\begin{pmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{pmatrix}$ is unitary in non-Euclidean metric $\mathbf{B}_\phi$, i.e.,

$$\begin{pmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{pmatrix}^{\mathrm{H}}\mathbf{B}_\phi\begin{pmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{pmatrix} = \mathbf{B}_\phi^{\frac{1}{2}}\begin{pmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{pmatrix}\begin{pmatrix}\widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp\end{pmatrix}^{\mathrm{H}}\mathbf{B}_\phi^{\frac{1}{2}} = \mathbf{I},$$

and thus

$$\mathbf{B}_\phi^{\frac{1}{2}}\widetilde{\mathbf{U}}_\perp\widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi^{\frac{1}{2}} = \mathbf{I} - \mathbf{B}_\phi^{\frac{1}{2}}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi^{\frac{1}{2}}.$$

If

$$\max\big\{\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},F}, \big\|\widehat{\boldsymbol{\xi}}_t\big\|_{\mathbf{C}_{xx},F}\big\} \leq \frac{\sigma_k^+ - \sqrt{\beta}}{16\max\big\{1+\gamma, 1+\|\boldsymbol{\Xi}\|_2\big\}}\min\Big\{\frac{\cos\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{a_{\phi,t}}, \frac{\sin\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{c_{\phi,t}}\Big\},$$

we then get from the results derived above that

$$
\begin{aligned}
\sin\theta_{\max}(\mathbf{P}_T, \widetilde{\mathbf{U}}) &\le (1+\gamma)(1+\|\mathbf{\Xi}\|_2)\frac{\sqrt{\beta^2+(\sigma_k^+)^2}/\sigma_k^+}{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big)} \times \\
&\quad \prod_{t=0}^{T-1} \frac{\sqrt{\beta}+\frac{\sigma_k^+-\sqrt{\beta}}{8}+\frac{2c_{\phi,t}(1+\gamma)\max\left\{\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2},\left\|\widehat{\boldsymbol{\xi}}_t\right\|_{\mathbf{C}_{xx},2}\right\}}{\sin\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}}{\sigma_k^+-\frac{2a_{\phi,t}(1+\|\mathbf{\Xi}\|_2)\max\left\{\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2},\left\|\widehat{\boldsymbol{\xi}}_t\right\|_{\mathbf{C}_{xx},2}\right\}}{\cos\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}} \\
&\le (1+\gamma)(1+\|\mathbf{\Xi}\|_2)\frac{\sqrt{\beta^2+(\sigma_k^+)^2}/\sigma_k^+}{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big)}\left(\frac{\sqrt{\beta}+\frac{\sigma_k^+-\sqrt{\beta}}{4}}{\sigma_k^+-\frac{\sigma_k^+-\sqrt{\beta}}{4}}\right)^T,
\end{aligned}
$$

where

$$
\begin{aligned}
\left(\frac{\sqrt{\beta}+\frac{\sigma_k^+-\sqrt{\beta}}{4}}{\sigma_k^+-\frac{\sigma_k^+-\sqrt{\beta}}{4}}\right)^T &= \left(\frac{\sigma_k^++3\sqrt{\beta}}{3\sigma_k^++\sqrt{\beta}}\right)^T \le \left(1-\frac{2(\sigma_k^+-\sqrt{\beta})}{3\sigma_k^++\sqrt{\beta}}\right)^T \\
&\le \left(1-\frac{\sigma_k^+-\sqrt{\beta}}{2\sigma_k^+}\right)^T \le \exp\{-\frac{\sigma_k^+-\sqrt{\beta}}{\sigma_k^+}\frac{T}{2}\},
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\sigma_k^+-\sqrt{\beta}}{\sigma_k^+} &= \frac{\sigma_k^2-2\sqrt{\beta}+\sqrt{\sigma_k^4-4\beta}}{\sigma_k^2+\sqrt{\sigma_k^4-4\beta}} = \frac{\sqrt{\sigma_k^2-2\sqrt{\beta}}\big(\sqrt{\sigma_k^2-2\sqrt{\beta}}+\sqrt{\sigma_k^2+2\sqrt{\beta}}\big)}{\sigma_k^2+\sqrt{\sigma_k^4-4\beta}} \\
&\ge \frac{\sqrt{\sigma_k^2-2\sqrt{\beta}}\cdot\sigma_k}{2\sigma_k^2} = \frac{1}{2}\sqrt{\frac{\sigma_k^2-2\sqrt{\beta}}{\sigma_k^2}} = \frac{1}{2}\sqrt{\frac{\sigma_k^2-\sigma_{k+1}^2}{\sigma_k^2}}.
\end{aligned}
$$

Thus, we have that

$$
\begin{aligned}
\sin\theta_{\max}(\mathbf{P}_T, \widetilde{\mathbf{U}}) &\le \frac{(1+\gamma)(1+\|\mathbf{\Xi}\|_2)\sqrt{\beta^2+(\sigma_k^+)^2}/\sigma_k^+}{\min\left\{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big),\sigma_{\min}\big(\mathbf{V}^{\mathrm{H}}\mathbf{C}_{yy}\mathbf{\Psi}_0\big)\right\}} \times \\
&\quad \exp\{-\sqrt{\frac{\sigma_k^2-\sigma_{k+1}^2}{\sigma_k^2}}\cdot\frac{T}{4}\} \triangleq \frac{\sigma_k^+}{k\sqrt{1+(\sigma_k^+)^2}}\epsilon,
\end{aligned}
$$

and thus $\sin\theta_{\max}(\mathbf{\Phi}_T, \mathbf{U}) < \epsilon$ by Lemma 3.4. Solving the last equation for $T$ yields that

$$
T = 4\sqrt{\frac{\sigma_k^2}{\sigma_k^2-\sigma_{k+1}^2}}\log\frac{k(1+\gamma)(1+\|\mathbf{\Xi}\|_2)\sqrt{\big(1+(\sigma_k^+)^2\big)\big(\beta^2+(\sigma_k^+)^2\big)}}{(\sigma_k^+)^2\epsilon\cos\theta_0},
$$

where $\theta_0 = \max\left\{\theta_{\max}\big(\mathbf{\Phi}_0, \mathbf{U}\big), \theta_{\max}\big(\mathbf{\Psi}_0, \mathbf{V}\big)\right\}$. By $\widetilde{\mathbf{\Sigma}}\mathbf{\Xi} - \mathbf{\Xi}\widetilde{\mathbf{\Sigma}}_\perp = -\widetilde{\widetilde{\mathbf{\Sigma}}}$, we have that (Golub and Van Loan, 2013)

$$
\|\widetilde{\widetilde{\mathbf{\Sigma}}}\|_F = \|\widetilde{\mathbf{\Sigma}}\mathbf{\Xi}-\mathbf{\Xi}\widetilde{\mathbf{\Sigma}}_\perp\|_F = \|\mathbf{\Xi}\|_F\frac{\|\widetilde{\mathbf{\Sigma}}\mathbf{\Xi}-\mathbf{\Xi}\widetilde{\mathbf{\Sigma}}_\perp\|_F}{\|\mathbf{\Xi}\|_F} \ge \|\mathbf{\Xi}\|_F\mathrm{sep}_F(\widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{\Sigma}}_\perp),
$$

and that

$$
\begin{aligned}
\mathrm{sep}_F(\widetilde{\mathbf{\Sigma}}, \widetilde{\mathbf{\Sigma}}_\perp) &\le \min|\lambda(\widetilde{\mathbf{\Sigma}})-\lambda(\widetilde{\mathbf{\Sigma}}_\perp)| \le \sigma_k^+-\sigma_k^- = \sqrt{\sigma_k^4-4\beta} \\
&= \sqrt{\sigma_k^4-\sigma_{k+1}^4} = \sqrt{(\sigma_k^2+\sigma_{k+1}^2)(\sigma_k^2-\sigma_{k+1}^2)},
\end{aligned}
$$

where $\mathrm{sep}_F(\cdot,\cdot)$ represents the separation between two matrices in Frobenious norm and $\lambda(\cdot)$ represent an arbitrary eigenvalue of a matrix. Thus, we have that $\|\mathbf{\Xi}\|_2 = O(\frac{1+\beta}{\sqrt{\sigma_k^2-\sigma_{k+1}^2}})$. Besides, $1+\gamma = \frac{1+\beta}{8(\sigma_k^+-\sqrt{\beta})} \leq \frac{1+\beta}{4\sigma_k\sqrt{\sigma_k^2-\sigma_{k+1}^2}}$. Thus, we can write that

$$T = O\left(\sqrt{\frac{\sigma_k^2}{\sigma_k^2-\sigma_{k+1}^2}}\log\frac{1}{(\sigma_k^2-\sigma_{k+1}^2)\epsilon\cos\theta_0}\right).$$

Analogously, we also have $\sin\theta_{\max}(\mathbf{Q}_T, \widetilde{\mathbf{V}}) < \frac{\sigma_k^+}{k\sqrt{1+(\sigma_k^+)^2}}\epsilon$, where $\mathbf{Q}_t = \begin{pmatrix}\mathbf{\Psi}_t \\ \mathbf{\Psi}_{t-1}\mathbf{S}_t^{-1}\end{pmatrix}\widetilde{\mathbf{S}}_t$, and thus $\sin\theta_{\max}(\mathbf{\Psi}_T, \mathbf{V}) < \epsilon$. Therefore, we get that

$$\sin\theta_T = \sin\max\{\theta_{\max}(\mathbf{\Phi}_T, \mathbf{U}), \theta_{\max}(\mathbf{\Psi}_T, \mathbf{V})\} < \epsilon.$$

On the other hand, in each iteration, we need to solve four least-squares subproblems in order to make the error conditions on, e.g., $\boldsymbol{\xi}_t$ and $\widehat{\boldsymbol{\xi}}_t$, satisfied. This part of complexity can be obtained using Lemmas 3.2-3.3 as follows. First, by Lemma 3.2, the complexity of getting $\boldsymbol{\xi}_t$ as small as required previously is

$$O\left(\mathrm{nnz}(\mathbf{X}) + \mathrm{nnz}(\mathbf{Y})\sqrt{\kappa(\mathbf{C}_{yy})}\log\frac{\epsilon_t\big(\mathbf{\Psi}_t^{(0)}\big)}{\epsilon_t(\widehat{\mathbf{\Psi}}_t)}\right),$$

and the ratio of the initial to final error can be further written as

$$\begin{aligned}
\log\frac{\epsilon_t\big(\mathbf{\Psi}_t^{(0)}\big)}{\epsilon_t(\widehat{\mathbf{\Psi}}_t)} &= O(\log\frac{\tan^2\max\{\theta_{\max}(\mathbf{\Phi}_t, \mathbf{U}), \theta_{\max}(\mathbf{\Psi}_t, \mathbf{V})\}}{\|\boldsymbol{\xi}_t\|^2_{\mathbf{C}_{yy},F}}) \\
&= O\left(\log\frac{c_1^2\tan^2\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}})}{(\sigma_k^+-\sqrt{\beta})^2\min\left\{\frac{\cos^2\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{a_{\phi,t}^2}, \frac{\sin^2\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{c_{\phi,t}^2}\right\}}\right) \\
&= O\left(\log\Big(\frac{a_{\phi,t}c_{\phi,t}c_1}{\sigma_k^2-\sigma_{k+1}^2}\max\{\frac{\sin^2\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}})}{\cos^4\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}})}, \frac{1}{\cos^2\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}})}\}\Big)\right),
\end{aligned}$$

where the numerator in the second equation is by $c_1$'s definition in the theorem, and $O\left(\log\max\{\frac{\sin^2\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}{\cos^4\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}, \frac{1}{\cos^2\theta_{\max}(\mathbf{P}_t,\widetilde{\mathbf{U}})}\}\right) = O(\log\frac{1}{\cos\theta_{\max}(\mathbf{P}_0,\widetilde{\mathbf{U}})})$ when $\theta_{\max}(\mathbf{P}_t, \widetilde{\mathbf{U}})$ is large, otherwise $O(1)$. Also, we have that

$$\begin{aligned}
a_{\phi,t} &= \big\|\big(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big(\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)\big\|_2 \leq \big\|\big(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_t\big)^{-1}\big\|_2 \\
&= \frac{1}{\cos\theta_{\max}(\mathbf{P}_t, \widetilde{\widetilde{\mathbf{U}}})} \leq \frac{1}{\cos\theta_{\max}(\mathbf{P}_0, \widetilde{\widetilde{\mathbf{U}}})} = \frac{1}{\sigma_{\min}(\widetilde{\widetilde{\mathbf{U}}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_0)} \\
&\leq \frac{\sqrt{\beta^2+(\sigma_k^+)^2}/\sigma_k^+}{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big)} \quad (\text{ see Page 3 on } \|\widetilde{\mathbf{Z}}_0\|_2\ ),
\end{aligned}$$

and similarly,

$$\frac{1}{\cos\theta_{\max}(\mathbf{P}_0, \widetilde{\mathbf{U}})} = \frac{1}{\sigma_{\min}(\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_0)} \leq \frac{\sqrt{1+(\sigma_k^+)^2}/\sigma_k^+}{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big)}.$$

Thus, we can write that

$$\begin{aligned}
\log\frac{\epsilon_t\big(\mathbf{\Psi}_t^{(0)}\big)}{\epsilon_t(\widehat{\mathbf{\Psi}}_t)} &= O\left(\log\frac{c_1c_2}{(\sigma_k^2-\sigma_{k+1}^2)\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big)}\right) \\
&= O\left(\log\frac{c_1c_2}{(\sigma_k^2-\sigma_{k+1}^2)\min\big\{\sigma_{\min}\big(\mathbf{U}^{\mathrm{H}}\mathbf{C}_{xx}\mathbf{\Phi}_0\big), \sigma_{\min}\big(\mathbf{V}^{\mathrm{H}}\mathbf{C}_{yy}\mathbf{\Psi}_0\big)\big\}}\right),
\end{aligned}$$

where $c_2 = \max\{\max_t\{c_{\phi,t}\}, \max_t\{c_{\psi,t}\}\}$. Similarly, by Lemma 3.3, the complexity of getting $\widehat{\boldsymbol{\xi}}_t$ as small as required previously is

$$O\left(\text{nnz}(\mathbf{Y}) + \text{nnz}(\mathbf{X})\sqrt{\kappa(\mathbf{C}_{xx})}\log\frac{\widehat{\epsilon}_t\big(\widehat{\boldsymbol{\Phi}}_t^{(0)}\big)}{\widehat{\epsilon}_t(\widehat{\widehat{\boldsymbol{\Phi}}}_t)}\right),$$

and $\log\frac{\widehat{\epsilon}_t\big(\widehat{\boldsymbol{\Phi}}_t^{(0)}\big)}{\widehat{\epsilon}_t(\widehat{\widehat{\boldsymbol{\Phi}}}_t)}$ can be further written as

$$\begin{aligned}
& O(\log\frac{(\sigma_1^2 + \|\boldsymbol{\xi}_t\|^2_{\mathbf{C}_{yy},F})\tan^2\max\{\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t, \mathbf{U}), \theta_{\max}(\widehat{\boldsymbol{\Psi}}_t, \mathbf{V})\}}{\|\widehat{\boldsymbol{\xi}}_t\|^2_{\mathbf{C}_{xx},F}}) \\
= \quad & O(\log\frac{\tan^2\max\{\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t, \mathbf{U}), \theta_{\max}(\widehat{\boldsymbol{\Psi}}_t, \mathbf{V})\}}{\|\widehat{\boldsymbol{\xi}}_t\|^2_{\mathbf{C}_{xx},F}}),
\end{aligned}$$

where we have used that $\sigma_1 \leq 1$ and $\|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},F} \leq 1$. The cases for $\boldsymbol{\eta}_t$ and $\widehat{\boldsymbol{\eta}}_t$ are similar as well. We thus have the following overall complexity:

$$O\left(dk^2T + kT\text{nnz}(\mathbf{X}, \mathbf{Y})\kappa^{\frac{1}{2}}(\mathbf{X}, \mathbf{Y})\log\frac{c_1c_2}{(\sigma_k^2 - \sigma_{k+1}^2)\cos\theta_0}\right),$$

where $d = \max\{d_x, d_y\}$, $\text{nnz}(\mathbf{X}, \mathbf{Y}) = \text{nnz}(\mathbf{X}) + \text{nnz}(\mathbf{Y})$, and $\kappa(\mathbf{X}, \mathbf{Y}) = \max\{\kappa(\mathbf{C}_{xx}), \kappa(\mathbf{C}_{yy})\}$. □

**Lemma 3.3** Consider the least-squares subproblem $\min_{\boldsymbol{\Phi}} \widehat{l}_t(\boldsymbol{\Phi})$, for which the minimizer and the objective sub-optimality gap can be expressed as $\widehat{\boldsymbol{\Phi}}_t^\star = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t$ and $\widehat{\epsilon}_t(\boldsymbol{\Phi}) = \widehat{l}_t(\boldsymbol{\Phi}) - \widehat{l}_t(\widehat{\boldsymbol{\Phi}}_t^\star) = \frac{1}{2}\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}_t^\star\|^2_{\mathbf{C}_{xx},F}$. We have that

$$\widehat{\epsilon}_t(\widehat{\boldsymbol{\Phi}}_t^{(0)}) \leq 8k(\sigma_1^2 + \|\boldsymbol{\xi}_t\|^2_{\mathbf{C}_{yy},2})\tan^2\widehat{\theta}_t,$$

for the initial sub-optimality, and accelerated gradient descent takes $O\big(\text{nnz}(\mathbf{Y}) + \text{nnz}(\mathbf{X})\kappa^{\frac{1}{2}}(\mathbf{C}_{xx})\log\frac{\widehat{\epsilon}_t(\widehat{\boldsymbol{\Phi}}_t^{(0)})}{\widehat{\epsilon}_t(\widehat{\widehat{\boldsymbol{\Phi}}}_t)}\big)$ complexity to get the final sub-optimality $\widehat{\epsilon}_t(\widehat{\widehat{\boldsymbol{\Phi}}}_t)$, where $\widehat{\boldsymbol{\Phi}}_t^{(0)} = \widehat{\boldsymbol{\Phi}}_t(\widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t)^{-1}(\widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t)$, $\|\mathbf{A}\|_{\mathbf{B},2} = \|\mathbf{B}^{\frac{1}{2}}\mathbf{A}\|_2$, and $\widehat{\theta}_t = \max\{\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t, \mathbf{U}), \theta_{\max}(\widehat{\boldsymbol{\Psi}}_t, \mathbf{V})\}$. Parallel results hold for $\min_{\boldsymbol{\Psi}} \widehat{h}_t(\boldsymbol{\Psi})$ as well.

**Proof** Noting that $\widehat{\boldsymbol{\Phi}}_t^\star = \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t$ and

$$\widehat{l}_t(\widehat{\boldsymbol{\Phi}}_t^\star) = -\frac{1}{2}\text{tr}\left(\widehat{\boldsymbol{\Psi}}_t^\top\mathbf{C}_{xy}^\top\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t\right) + \frac{1}{2n}\|\mathbf{Y}^\top\widehat{\boldsymbol{\Psi}}_t\|_F^2.$$

we have that

$$\begin{aligned}
\frac{1}{2}\|\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}_t^\star\|^2_{\mathbf{C}_{xx},F} &= \frac{1}{2}\text{tr}\left((\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}_t^\star)^\top\mathbf{C}_{xx}(\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}_t^\star)\right) \\
&= \text{tr}\left(\frac{1}{2}\boldsymbol{\Phi}^\top\mathbf{C}_{xx}\boldsymbol{\Phi} - \boldsymbol{\Phi}^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t^\star + \frac{1}{2}(\widehat{\boldsymbol{\Phi}}_t^\star)^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t^\star\right) \\
&= \text{tr}\left(\frac{1}{2}\boldsymbol{\Phi}^\top\mathbf{C}_{xx}\boldsymbol{\Phi} - \boldsymbol{\Phi}^\top\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t + \frac{1}{2}\widehat{\boldsymbol{\Psi}}_t^\top\mathbf{C}_{xy}^\top\mathbf{C}_{xx}\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t\right) \\
&= \widehat{l}_t(\boldsymbol{\Phi}) - \widehat{l}_t(\widehat{\boldsymbol{\Phi}}_t^\star) = \widehat{\epsilon}_t(\boldsymbol{\Phi}).
\end{aligned}$$

Setting $\frac{\partial}{\partial\boldsymbol{\Gamma}}\widehat{l}_t(\widehat{\boldsymbol{\Phi}}_t\boldsymbol{\Gamma}) = \widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\boldsymbol{\Gamma} - \widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t = 0$ yields the optimal

$$\boldsymbol{\Gamma}^\star = \left(\widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\right)^{-1}\widehat{\boldsymbol{\Phi}}_t^\top\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t.$$

Thus, $\widehat{\boldsymbol{\Phi}}_t^{(0)} = \widehat{\boldsymbol{\Phi}}_t \boldsymbol{\Gamma}^\star$. Noting that $\mathbf{C}_{xy} = \mathbf{C}_{xx}(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top + \mathbf{U}_\perp \boldsymbol{\Sigma}_\perp \mathbf{V}_\perp^\top)\mathbf{C}_{yy}$, it holds that

$$
\begin{aligned}
2\widehat{\epsilon}_t\big(\widehat{\boldsymbol{\Phi}}_t^{(0)}\big) &\le 2\widehat{\epsilon}_t\big(\widehat{\boldsymbol{\Phi}}_t\widetilde{\boldsymbol{\Gamma}}\big) = \big\|\widehat{\boldsymbol{\Phi}}_t\widetilde{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Phi}}_t^\star\big\|_{\mathbf{C}_{xx},F}^2 \\
&= \big\|\mathbf{U}^\top\mathbf{C}_{xx}(\widehat{\boldsymbol{\Phi}}_t\widetilde{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Phi}}_t^\star)\big\|_F^2 + \big\|\mathbf{U}_\perp^\top\mathbf{C}_{xx}(\widehat{\boldsymbol{\Phi}}_t\widetilde{\boldsymbol{\Gamma}} - \widehat{\boldsymbol{\Phi}}_t^\star)\big\|_F^2 \\
&= \big\|\mathbf{U}_\perp^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\widetilde{\boldsymbol{\Gamma}} - \mathbf{U}_\perp^\top\mathbf{C}_{xy}\widehat{\boldsymbol{\Psi}}_t\big\|_F^2 \quad \Big(\text{ let } \widetilde{\boldsymbol{\Gamma}} = \big(\mathbf{U}^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\big)^{-1}\mathbf{U}^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t^\star\Big) \\
&= \big\|\mathbf{U}_\perp^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\big(\mathbf{U}^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\big)^{-1}\boldsymbol{\Sigma}\mathbf{V}^\top\mathbf{C}_{yy}\widehat{\boldsymbol{\Psi}}_t - \boldsymbol{\Sigma}_\perp^\top\mathbf{V}_\perp^\top\mathbf{C}_{yy}\widehat{\boldsymbol{\Psi}}_t\big\|_F^2 \\
&\le 2k\Big(\big\|\mathbf{U}_\perp^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\big(\mathbf{U}^\top\mathbf{C}_{xx}\widehat{\boldsymbol{\Phi}}_t\big)^{-1}\big\|_2^2\|\boldsymbol{\Sigma}\|_2^2\|\mathbf{V}^\top\|_{\mathbf{C}_{yy},2}^2\big\|\widehat{\boldsymbol{\Psi}}_t\big\|_{\mathbf{C}_{yy},2}^2 \\
&\quad + \|\boldsymbol{\Sigma}_\perp\|_2^2\big\|\mathbf{V}_\perp^\top\mathbf{C}_{yy}\widehat{\boldsymbol{\Psi}}_t(\widehat{\boldsymbol{\Psi}}_t^\top\mathbf{C}_{yy}\widehat{\boldsymbol{\Psi}}_t)^{-\frac{1}{2}}\big\|_2^2\big\|\big(\widehat{\boldsymbol{\Psi}}_t^\top\mathbf{C}_{yy}\widehat{\boldsymbol{\Psi}}_t\big)^{\frac{1}{2}}\big\|_2^2\Big) \\
&= 2k\Big(\sigma_1^2\tan^2\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t,\mathbf{U}) + \sigma_{k+1}^2\sin^2\theta_{\max}(\widehat{\boldsymbol{\Psi}}_t,\mathbf{V})\Big)\big\|\widehat{\boldsymbol{\Psi}}_t\big\|_{\mathbf{C}_{yy},2}^2 \\
&\le 4k\big\|\widehat{\boldsymbol{\Psi}}_t\big\|_{\mathbf{C}_{yy},2}^2\tan^2\max\big\{\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t,\mathbf{U}),\theta_{\max}(\widehat{\boldsymbol{\Psi}}_t,\mathbf{V})\big\},
\end{aligned}
$$

where we have used that $\sigma_1 \le 1$ and additionally,

$$
\begin{aligned}
\big\|\widehat{\boldsymbol{\Psi}}_t\big\|_{\mathbf{C}_{yy},2}^2 &= \big\|\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top\boldsymbol{\Phi}_t + \boldsymbol{\xi}_t\big\|_{\mathbf{C}_{yy},2}^2 \\
&\le 2\Big(\|\mathbf{C}\|_2^2\|\boldsymbol{\Phi}_t\|_{\mathbf{C}_{xx},2}^2 + \|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2}^2\Big) = 2\Big(\sigma_1^2 + \|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2}^2\Big).
\end{aligned}
$$

Thus, we can write that

$$
\widehat{\epsilon}_t\big(\widehat{\boldsymbol{\Phi}}_t^{(0)}\big) \le 8k\Big(\sigma_1^2 + \|\boldsymbol{\xi}_t\|_{\mathbf{C}_{yy},2}^2\Big)\tan^2\max\big\{\theta_{\max}(\widehat{\boldsymbol{\Phi}}_t,\mathbf{U}),\theta_{\max}(\widehat{\boldsymbol{\Psi}}_t,\mathbf{V})\big\}.
$$

The proof completes by noting that $\widehat{l}_t(\boldsymbol{\Phi})$ is $\lambda_{\max}(\mathbf{C}_{xx})$-smooth and $\lambda_{\min}(\mathbf{C}_{xx})$-strongly convex and using the complexity of Nesterov's accelerated gradient descent (Nesterov, 2014). The case for the least-squares subproblem $\min_{\boldsymbol{\Psi}}\widehat{h}_t(\boldsymbol{\Psi})$ is analogous. □

**Lemma 3.4** If $\sin\max\{\theta_{\max}(\mathbf{P}_T,\widetilde{\mathbf{U}}),\theta_{\max}(\mathbf{Q}_T,\widetilde{\mathbf{V}})\} < \frac{\sigma_k^+\epsilon}{k\sqrt{1+(\sigma_k^+)^2}}$ where $\sigma_k^+ = \frac{\sigma_k^2+\sqrt{\sigma_k^4-4\beta}}{2}$, it holds that $\sin\max\{\theta_{\max}(\boldsymbol{\Phi}_T,\mathbf{U}),\theta_{\max}(\boldsymbol{\Psi}_T,\mathbf{V})\} < \epsilon$.

**Proof** We only show that if $\sin\theta_{\max}(\mathbf{P}_T,\widetilde{\mathbf{U}}) < \frac{\sigma_k^+\epsilon}{k\sqrt{1+(\sigma_k^+)^2}}$ then it holds that $\sin\theta_{\max}(\boldsymbol{\Phi}_T,\mathbf{U}) < \epsilon$, because the case of $\theta_{\max}(\mathbf{Q}_T,\widetilde{\mathbf{V}})$ is analogous and then it is easy to see the lemma holds.

Note that

$$
\begin{aligned}
\|\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_F^2 &= k - \|\widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_F^2 \ge k - k\|\widetilde{\mathbf{U}}_\perp^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_2^2 \\
&= k - k\sin^2\theta_{\max}(\mathbf{P}_T,\widetilde{\mathbf{U}}) \ge k - \frac{(\sigma_k^+)^2\epsilon^2}{k(1+(\sigma_k^+)^2)},
\end{aligned}
$$

and

$$
\begin{aligned}
\|\widetilde{\mathbf{u}}_j^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_2^2 &= \left\|\begin{pmatrix}\mu_j(1)\mathbf{u}_j \\ \nu_j(1)\mathbf{u}_j\end{pmatrix}^\top \mathbf{B}_\phi \begin{pmatrix}\boldsymbol{\Phi}_T \\ \boldsymbol{\Phi}_{T-1}\mathbf{R}_T^{-1}\end{pmatrix}\widetilde{\mathbf{R}}_T\right\|_2^2 \\
&= \left\|\begin{pmatrix}\mu_j(1)\boldsymbol{\Phi}_T^\top\mathbf{C}_{xx}\mathbf{u}_j \\ \nu_j(1)\boldsymbol{\Phi}_{T-1}^\top\mathbf{C}_{xx}\mathbf{u}_j\end{pmatrix}^\top \begin{pmatrix}\mathbf{I} \\ \mathbf{R}_T^{-1}\end{pmatrix}\widetilde{\mathbf{R}}_T\right\|_2^2 \\
&\le \left\|\begin{pmatrix}\mu_j(1)\boldsymbol{\Phi}_T^\top\mathbf{C}_{xx}\mathbf{u}_j \\ \nu_j(1)\boldsymbol{\Phi}_{T-1}^\top\mathbf{C}_{xx}\mathbf{u}_j\end{pmatrix}^\top\right\|_2^2 = \mu_j(1)^2\left\|\boldsymbol{\Phi}_T^\top\mathbf{C}_{xx}\mathbf{u}_j\right\|_2^2 + \nu_j(1)^2\left\|\boldsymbol{\Phi}_{T-1}^\top\mathbf{C}_{xx}\mathbf{u}_j\right\|_2^2,
\end{aligned}
$$

for any $j = 1, \cdots, k$, where $\widetilde{\mathbf{u}}_j$ is the $j$-th column of $\widetilde{\mathbf{U}}$, $\mu_j(\alpha)$ and $\mu_j(\alpha)$ are given in Lemma I of Section 2. If $\left\|\mathbf{\Phi}_T^\top \mathbf{C}_{xx}\mathbf{u}_{j'}\right\|_2^2 < 1 - \frac{\epsilon^2}{k}$ for some $j'$, there must be

$$\begin{aligned}
\|\widetilde{\mathbf{u}}_{j'}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_2^2 &\leq \mu_{j'}(1)^2 \left\|\mathbf{\Phi}_T^\top \mathbf{C}_{xx}\mathbf{u}_{j'}\right\|_2^2 + \nu_{j'}(1)^2 \left\|\mathbf{\Phi}_{T-1}^\top \mathbf{C}_{xx}\mathbf{u}_{j'}\right\|_2^2 \\
&< \mu_{j'}(1)^2(1 - \frac{\epsilon^2}{k}) + \nu_{j'}(1)^2 = \mu_{j'}(1)^2 + \nu_{j'}(1)^2 - \mu_{j'}(1)^2\frac{\epsilon^2}{k} \\
&< 1 - \mu_k(1)^2\frac{\epsilon^2}{k} = 1 - \frac{(\sigma_k^+)^2\epsilon^2}{k(1+(\sigma_k^+)^2)},
\end{aligned}$$

and then

$$\|\widetilde{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_F^2 = \sum_{j=1}^k \|\widetilde{\mathbf{u}}_j^{\mathrm{H}}\mathbf{B}_\phi\mathbf{P}_T\|_2^2 < 1 - \frac{(\sigma_k^+)^2\epsilon^2}{k(1+(\sigma_k^+)^2)} + k - 1 = k - \frac{(\sigma_k^+)^2\epsilon^2}{k(1+(\sigma_k^+)^2)},$$

contradiction. We thus have $\left\|\mathbf{\Phi}_T^\top \mathbf{C}_{xx}\mathbf{u}_j\right\|_2^2 > 1 - \frac{\epsilon^2}{k}$ for all $j = 1, \cdots, k$ and then

$$\left\|\mathbf{\Phi}_T^\top \mathbf{C}_{xx}\mathbf{U}\right\|_F^2 = \sum_{j=1}^k \left\|\mathbf{\Phi}_T^\top \mathbf{C}_{xx}\mathbf{u}_j\right\|_2^2 > k(1 - \frac{\epsilon^2}{k}) = k - \epsilon^2.$$

We thus get that

$$\sin\theta_{\max}(\mathbf{\Phi}_T, \mathbf{U}) < \|\sin\boldsymbol{\theta}(\mathbf{\Phi}_T, \mathbf{U})\|_2 = \|\mathbf{U}_\perp^\top \mathbf{C}_{xx}\mathbf{\Phi}_T\|_F \leq \sqrt{k - \|\mathbf{U}^\top \mathbf{C}_{xx}\mathbf{\Phi}_T\|_F^2} = \epsilon.$$

□

## 2 Auxiliary Lemma

**Lemma I.** $\mathbf{A}_\phi$ and $\mathbf{A}_\phi^{\mathrm{H}}$ have the following Schur decompositions in non-Euclidean metric $\mathbf{B}_\phi$:

$$\begin{aligned}
\mathbf{A}_\phi &= \mathbf{B}_\phi \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \widetilde{\widetilde{\mathbf{\Sigma}}} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi, \\
\mathbf{A}_\phi^{\mathrm{H}} &= \mathbf{B}_\phi \begin{pmatrix} \widetilde{\widetilde{\mathbf{U}}} & \widetilde{\widetilde{\mathbf{U}}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\mathbf{\Lambda}} & \widetilde{\widetilde{\mathbf{\Lambda}}} \\ \mathbf{0} & \widetilde{\mathbf{\Lambda}}_\perp \end{pmatrix} \begin{pmatrix} \widetilde{\widetilde{\mathbf{U}}} & \widetilde{\widetilde{\mathbf{U}}}_\perp \end{pmatrix}^{\mathrm{H}} \mathbf{B}_\phi,
\end{aligned}$$

respectively, where both $\begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}$ and $\begin{pmatrix} \widetilde{\widetilde{\mathbf{U}}} & \widetilde{\widetilde{\mathbf{U}}}_\perp \end{pmatrix}$ are $\mathbf{B}_\phi$-unitary, $\widetilde{\mathbf{U}}_\perp$ represents $\widetilde{\mathbf{U}}$'s $\mathbf{B}_\phi$-orthogonal complement, and $\sigma_j^\pm = \frac{\sigma_j^2 \pm \sqrt{\sigma_j^4 - 4\beta}}{2}$,

$$\widetilde{\mathbf{U}} = \begin{pmatrix} \mathbf{U}\mathbf{D}(1) \\ \mathbf{U}\mathbf{J}(1) \end{pmatrix}, \quad \widetilde{\widetilde{\mathbf{U}}} = \begin{pmatrix} \mathbf{U}\mathbf{D}(\beta) \\ -\mathbf{U}\mathbf{J}(\beta) \end{pmatrix}, \quad \mu_j(\alpha) = \frac{\sigma_j^+}{\sqrt{\alpha^2 + (\sigma_j^+)^2}}, \quad \nu_j(\alpha) = \frac{\alpha}{\sqrt{\alpha^2 + (\sigma_j^+)^2}},$$

$$\mathbf{D}(\alpha) = \mathrm{diag}(\mu_1(\alpha), \cdots, \mu_k(\alpha)), \qquad \mathbf{J}(\alpha) = \mathrm{diag}(\nu_1(\alpha), \cdots, \nu_k(\alpha)),$$

$$\widetilde{\mathbf{\Sigma}}_\perp = \begin{pmatrix} \mathbf{\Sigma}_\perp^+ & \mathbf{L}(-1) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}^- \end{pmatrix}, \qquad \widetilde{\mathbf{\Lambda}}_\perp = \begin{pmatrix} \mathbf{\Sigma}_\perp^+ & \mathbf{L}(\beta) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_\perp^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma}^- \end{pmatrix},$$

$$\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{\Lambda}} = \mathbf{\Sigma}^+, \qquad \widetilde{\widetilde{\mathbf{\Sigma}}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & -(1+\beta)\mathbf{I} \end{pmatrix}, \quad \widetilde{\widetilde{\mathbf{\Lambda}}} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & (1+\beta)\mathbf{I} \end{pmatrix},$$

$$\mathbf{\Sigma}^+ = \mathrm{diag}(\sigma_1^+, \cdots, \sigma_k^+), \quad \mathbf{\Sigma}_\perp^+ = \mathrm{diag}(\sigma_{k+1}^+, \cdots, \sigma_{2d_x}^+), \quad \mathbf{L}(\alpha) = \alpha\mathbf{I} + \frac{1}{\alpha}(\mathbf{\Sigma}_\perp^-)^2,$$

$$\mathbf{\Sigma}^- = \mathrm{diag}(\sigma_1^-, \cdots, \sigma_k^-), \quad \mathbf{\Sigma}_\perp^- = \mathrm{diag}(\sigma_{k+1}^-, \cdots, \sigma_{2d_x}^-).$$

If $\sigma_j = 2\sqrt{\beta}$ then corresponding entries in blocks $\mathbf{L}(-1)$ and $\mathbf{L}(\beta)$ are replaced with $(1+\beta)$.

**Proof** We have that

$$\mathbf{A}_\phi = \begin{pmatrix} \mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top & -\beta\mathbf{C}_{xx} \\ \mathbf{C}_{xx} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B}_\phi = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{xx} \end{pmatrix},$$

where $\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{xy}^\top = \mathbf{C}_{xx}\left(\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{\Sigma}_\perp^2\mathbf{U}_\perp^\top\right)\mathbf{C}_{xx}$. We can assume that $\begin{pmatrix} \mathbf{U} & \mathbf{U}_\perp \end{pmatrix} = (\mathbf{u}_1, \cdots, \mathbf{u}_{d_x})$ are orthogonal in metric $\mathbf{C}_{xx}$ with $\mathbf{\Sigma}_\perp = \mathrm{diag}(\sigma_{k+1}, \cdots, \sigma_r, 0, \cdots, 0) \in \mathbb{R}^{(d_x-k)\times(d_x-k)}$, where $r = \mathrm{rank}(\mathbf{C}_{xy})$. Similar to Proposition 9 in Xu et al. (2018), $(\mathbf{A}_\phi, \mathbf{B}_\phi)$'s generalized eigenpairs can be written as $(\sigma_j^\pm, \mathbf{u}_j^\pm)$ where $\sigma_j^\pm = \frac{\sigma_j^2 \pm \sqrt{\sigma_j^4 - 4\beta}}{2}$ and $\mathbf{u}_j^\pm = \begin{pmatrix} \sigma_j^\pm \mathbf{u}_j \\ \mathbf{u}_j \end{pmatrix}$. Note that when $\sigma_j^2 = 2\sqrt{\beta}$, i.e., $\sigma_j^+$ with algebraic multiplicity of 2, $\mathbf{u}_j^\pm$ supplies only one generalized eigenvector $\mathbf{u}_j^+$, i.e., $\sigma_j^+$ with geometric multiplicity of only 1. One "generalized" generalized eigenvector of $(\mathbf{A}_\phi, \mathbf{B}_\phi)$ corresponding to $\sigma_j^+ = \sqrt{\beta}$, i.e., the solution to the equation $(\mathbf{A}_\phi - \sqrt{\beta}\mathbf{B}_\phi)\mathbf{z}_j = \mathbf{B}_\phi \begin{pmatrix} \sqrt{\beta}\mathbf{u}_j \\ \mathbf{u}_j \end{pmatrix}$ in $\mathbf{z}_j$, is needed. It is easy to see $\mathbf{z}_j = \begin{pmatrix} \mathbf{u}_j \\ \mathbf{0} \end{pmatrix}$. For notational convenience, denote $\mathbf{u}_j^- = \mathbf{z}_j$ in this case. $(\mathbf{A}_\phi, \mathbf{B}_\phi)$'s $2d_x$ generalized eigenvectors or "generalized" generalized eigenvectors now can span $\mathbb{C}^{2d_x}$ in metric $\mathbf{B}_\phi$ and thus we can write that

$$\mathbf{A}_\phi\left(\mathbf{u}_1^+, \mathbf{u}_1^-, \cdots, \mathbf{u}_n^+, \mathbf{u}_n^-\right) = \mathbf{B}_\phi\left(\mathbf{u}_1^+, \mathbf{u}_1^-, \cdots, \mathbf{u}_n^+, \mathbf{u}_n^-\right)\mathrm{diag}\left(\begin{pmatrix} \sigma_1^+ & \delta_1 \\ 0 & \sigma_1^- \end{pmatrix}, \cdots, \begin{pmatrix} \sigma_n^+ & \delta_n \\ 0 & \sigma_n^- \end{pmatrix}\right),$$

where $\delta_j = 0$ if $\sigma_n^+ \neq \sigma_n^-$ otherwise 1. Letting $(\mathbf{u}_j^+, \mathbf{u}_j^-) = (\hat{\mathbf{u}}_{2j-1}, \hat{\mathbf{u}}_{2j})\widehat{\mathbf{R}}_j$ representing $\mathbf{B}_\phi$-orthonormalization of $(\mathbf{u}_j^+, \mathbf{u}_j^-)$ in $\mathbb{C}^{2n}$, we then can write that

$$\begin{aligned}\mathbf{A}_\phi =& \mathbf{B}_\phi\left(\hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_{2d_x}\right)\mathrm{diag}\left(\widehat{\mathbf{R}}_1\begin{pmatrix} \sigma_1^+ & \delta_1 \\ 0 & \sigma_1^- \end{pmatrix}\widehat{\mathbf{R}}_1^{-1}, \cdots, \widehat{\mathbf{R}}_n\begin{pmatrix} \sigma_n^+ & \delta_n \\ 0 & \sigma_n^- \end{pmatrix}\widehat{\mathbf{R}}_n^{-1}\right)\left(\hat{\mathbf{u}}_1, \cdots, \hat{\mathbf{u}}_{2d_x}\right)^{\mathrm{H}}\mathbf{B}_\phi \\ \triangleq& \mathbf{B}_\phi\widehat{\mathbf{U}}\widehat{\widehat{\mathbf{\Sigma}}}\widehat{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi,\end{aligned}$$

where $\widehat{\mathbf{U}}$ is $\mathbf{B}_\phi$-unitary and thus $\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{H}}\mathbf{B}_\phi = \mathbf{I}$. After some algebraic manipulations and permutations, we will arrive at $\widehat{\mathbf{U}}\widehat{\widehat{\mathbf{\Sigma}}}\widehat{\mathbf{U}}^{\mathrm{H}} = (\widetilde{\mathbf{U}}, \widetilde{\mathbf{U}}_\perp)\widehat{\widehat{\mathbf{\Sigma}}}(\widetilde{\mathbf{U}}, \widetilde{\mathbf{U}}_\perp)^{\mathrm{H}}$, where $(\widetilde{\mathbf{U}}, \widetilde{\mathbf{U}}_\perp)$ is $\mathbf{B}_\phi$-unitary and

$$\widetilde{\mathbf{U}} = \begin{pmatrix} \mathbf{U}\mathbf{D}(1) \\ \mathbf{U}\mathbf{J}(1) \end{pmatrix}, \quad \widehat{\widehat{\mathbf{\Sigma}}} = \begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \widetilde{\widetilde{\mathbf{\Sigma}}} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix},$$

as described in the lemma. Thus, we have that

$$\mathbf{A}_\phi = \mathbf{B}_\phi\begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}\begin{pmatrix} \widetilde{\mathbf{\Sigma}} & \widetilde{\widetilde{\mathbf{\Sigma}}} \\ \mathbf{0} & \widetilde{\mathbf{\Sigma}}_\perp \end{pmatrix}\begin{pmatrix} \widetilde{\mathbf{U}} & \widetilde{\mathbf{U}}_\perp \end{pmatrix}^{\mathrm{H}}\mathbf{B}_\phi.$$

It is analogous for $(\mathbf{A}_\phi^{\mathrm{H}}, \mathbf{B}_\phi)$, except for the generalized eigenpair being $(\sigma_j^\pm, \mathbf{u}_j^\pm)$ with $\mathbf{u}_j^\pm = \begin{pmatrix} \sigma_j^\pm \mathbf{u}_j \\ -\beta\mathbf{u}_j \end{pmatrix}$. □

## References


Rong Ge, Chi Jin, Sham M. Kakade, Praneeth Netrapalli, and Aaron Sidford. Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 2741–2750, New York City, NY, 2016.

G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. 2014.

Peng Xu, Bryan D. He, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 58–67, Playa Blanca, Lanzarote, Canary Islands, Spain, 2018.