Supplementary Material: A Practical Riemannian Algorithm for Computing Dominant Generalized Eigenspace

Zhiqiang Xu, Ping Li
Cognitive Computing Lab, Baidu Research
No.10 Xibeiwang East Road, Beijing, 10085, China
10900 NE 8th St, Bellevue, WA 98004, USA
{xuzhiqiang04,liping11}@baidu.com

Proof of Theorem 6.1

First, Theorem 6.1 can be obtained from the following two statements:

i) Algorithm 1 with $\alpha_t = \frac{\mu}{t+\nu}$ for sufficiently large positive constants μ and ν will converge after

$$T = O\left(\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\,\log\frac{\lambda_1}{\Delta_\dagger \epsilon}\right)\left(\frac{\lambda_1}{\tilde{\Delta}}\right)^2 \frac{1}{\epsilon}\right)$$

iterations with high probability.

ii) Algorithm 1 with $\alpha_t \equiv O\left(\frac{\Delta_\dagger}{\lambda_1^2}\right)$ for Categories a)-c) and e) will converge after

$$T = O\left(\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\,\log\frac{\lambda_1}{\Delta_\dagger}\right)\left(\frac{\lambda_1}{\Delta_\dagger}\right)^2 \log\frac{1}{\epsilon}\right)$$

iterations with high probability. If $\mathbf{X}_0$ is sufficiently close to $\mathcal{U}$ then the complexity holds for Category d) as well.

The reason follows. The first statement i) shows that the convergence is global because of high probability, and it is globally sub-linear if diminishing step-sizes are used. The second ii) shows that the convergence is global and globally linear (more precisely, here linear convergence refers to the logarithmic dependence on accuracy ϵ) if constant step-sizes are used for all Categories except for d). For Category d), linear convergence is local. Theorem 6.1 then can be obtained by a two-stage process. The first stage follows i) until the iterate is sufficiently close to the solution space, while the second follows ii). Since the first stage is not dependent on the final accuracy ϵ, the overall complexity will be dominated by the second one.

The two statements are proven in what follows. To analyze $\psi(\mathbf{X},\mathcal{U})$, we focus on $-2\log\mathrm{Det}(\mathbf{X}^\top\mathbf{B}\mathbf{U}_{k''})$ and the other is analogous. To start, we have from Algorithm 1 and Lemma 6.3 that

$$-2\log\mathrm{Det}(\mathbf{X}^\top\mathbf{B}\mathbf{U}_{k''}) = -2\log\mathrm{Det}\left(\left(\mathbf{X}_t + \alpha_t\widehat{\tilde{\nabla} f}_t\right)^\top\mathbf{B}\mathbf{U}_{k''}\right) + \log\det\left(\mathbf{I} + \alpha_t^2 \widehat{\tilde{\nabla} f}_t^\top \mathbf{B}\widehat{\tilde{\nabla} f}_t\right), \quad (1)$$

where $\widehat{\tilde{\nabla} f}_t = \tilde{\nabla} f_t + (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top\mathbf{B})\xi_t(\widehat{\nabla f}_t)$. Letting $\mathbf{E}_t = (\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top\mathbf{B})\xi_t(\widehat{\nabla f}_t)$, we can write that $\mathbf{X}_t + \alpha_t\widehat{\tilde{\nabla} f}_t = \widehat{\mathbf{X}}_{t+1} + \alpha_t\mathbf{E}_t$, where notation $\widehat{\mathbf{X}}_{t+1}$ is defined in Lemma 6.2. Then

$$\left(\mathbf{X}_t + \alpha_t\widehat{\tilde{\nabla} f}_t\right)^\top \mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\left(\mathbf{X}_t + \alpha_t\widehat{\tilde{\nabla} f}_t\right) \succcurlyeq \mathbf{S}_1 + 2\alpha_t\mathbf{S}_2,$$

with

$$\mathbf{S}_1 = \widehat{\mathbf{X}}_{t+1}^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\widehat{\mathbf{X}}_{t+1} \quad \text{and} \quad \mathbf{S}_2 = \mathrm{sym}\left(\widehat{\mathbf{X}}_{t+1}^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{E}_t\right),$$

where $\mathrm{sym}(\mathbf{M}) = \frac{1}{2}(\mathbf{M} + \mathbf{M}^\top)$. By the Taylor expansion, we can get for some $\varsigma \in (0,1)$ that

$$\begin{aligned} -2\log\mathrm{Det}\left(\left(\mathbf{X}_t + \alpha_t\widehat{\tilde{\nabla} f}_t\right)^\top \mathbf{B}\mathbf{U}_{k''}\right) &\leq -2\log\mathrm{Det}\left(\widehat{\mathbf{X}}_{t+1}^\top \mathbf{B}\mathbf{U}_{k''}\right) - 2\alpha_t \mathrm{tr}\left(\left(\mathbf{S}_1 + 2\varsigma\alpha_t\mathbf{S}_2\right)^{-1}\mathbf{S}_2\right) \\ &\leq -2\log\mathrm{Det}\left(\widehat{\mathbf{X}}_{t+1}^\top \mathbf{B}\mathbf{U}_{k''}\right) + \frac{2k^{\frac{1}{2}}\alpha_t\|\mathbf{S}_2\|_F}{\sigma_{\min}\left(\mathbf{S}_1 + 2\varsigma\alpha_t\mathbf{S}_2\right)}. \end{aligned} \tag{2}$$

To proceed, we bound singular values of $\mathbf{S}_i$, $i = 1, 2$, as follows.

$$\begin{aligned} \sigma(\mathbf{S}_1) &\geq \sigma_{\min}^2\left(\left(\mathbf{X}_t + \alpha_t\tilde{\nabla} f_t\right)^\top \mathbf{B}\mathbf{U}_{k''}\right) \geq \sigma_{\min}^2\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\right) - 2\alpha_t\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},2} \quad \text{(Lemma 6.6)} \\ &\geq \prod_{i=1}^{k}\sigma_i^2\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\right) - 2\alpha_t\lambda_1 = 1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathbf{U}_{k''}) - 2\alpha_t\lambda_1, \quad \text{(using notations in Lemma 6.4)} \end{aligned}$$

and

$$\begin{aligned} \sigma(\mathbf{S}_2) &\leq \left\|\widehat{\mathbf{X}}_{t+1}^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{E}_t\right\|_F \leq \left\|\widehat{\mathbf{X}}_{t+1}^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}(\mathbf{I} - \mathbf{X}_t\mathbf{X}_t^\top\mathbf{B})\mathbf{B}^{-\frac{1}{2}}\right\|_2 \left\|\mathbf{B}^{\frac{1}{2}}\xi_t(\widehat{\nabla f}_t)\right\|_F \\ &\leq \left(\|\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{X}_t^\perp\|_2 + \alpha_t\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},2}\right)\left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F}, \end{aligned}$$

where $\mathbf{X}_t^\perp$ represents the orthogonal complement of $\mathbf{X}_t$ in inner product $\langle,\rangle_\mathbf{B}$, i.e., $\mathbf{X}_t^\perp\mathbf{B}\mathbf{X}_t = \mathbf{0}$. Moreover, we have

$$\begin{aligned} \|\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{X}_t^\perp\|_2 &\leq \left\|(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{X}_t)^{-\frac{1}{2}}\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{X}_t^\perp\right\|_2 \\ &= \sqrt{\lambda_{\max}\left(\mathbf{I} - \mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\mathbf{U}_{k''}^\top\mathbf{B}\mathbf{X}_t\right)} \leq \sqrt{1 - \sigma_{\min}^2\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_{k''}\right)} \\ &\leq \mathrm{dist}_b(\mathbf{X}_t, \mathbf{U}_{k''}). \end{aligned}$$

Let $\mathrm{dist}_b(\mathbf{X}_t, \mathcal{U}) = \max\{\mathrm{dist}_b(\mathbf{X}_t, \mathbf{U}_{k'}), \mathrm{dist}_b(\mathbf{X}_t, \mathbf{U}_{k''})\}$, and assume that $0 < \alpha_t < \frac{1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})}{8\lambda_1}$ and

$$\left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F} = \frac{\Delta_\dagger}{4k^{\frac{1}{2}}}\frac{1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})}{1 + \psi(\mathbf{X}_t, \mathcal{U})}\mathrm{dist}_b(\mathbf{X}_t, \mathcal{U}).$$

By Lemma 6.6 and noting that $\Delta_\dagger \leq 2\lambda_1$, we get that $\sigma(\mathbf{S}_1) \geq \frac{1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})}{2}$ and

$$\sigma_{\min}(\mathbf{S}_1 + 2\varsigma\alpha_t\mathbf{S}_2) \geq \sigma_{\min}(\mathbf{S}_1) - 2\alpha_t\sigma_{\max}(\mathbf{S}_2) \geq \sigma_{\min}(\mathbf{S}_1) - \frac{\alpha_t(1 + \alpha_t\lambda_1)\Delta_\dagger}{2k^{\frac{1}{2}}} \geq \frac{1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})}{2}.$$

We thus have that

$$\frac{\|\mathbf{S}_2\|_F}{\sigma_{\min}\left(\mathbf{S}_1 + 2\varsigma\alpha_t\mathbf{S}_2\right)} \leq 2\left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F}\frac{\mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U}) + \alpha_t\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}}{1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})}. \tag{3}$$

By the Taylor expansion, we also have for some $\varsigma' \in (0,1)$ that

$$\begin{aligned} \log\det\left(\mathbf{I} + \alpha_t^2\widehat{\tilde{\nabla} f}_t^\top\mathbf{B}\widehat{\tilde{\nabla} f}_t\right) &= \alpha_t^2\mathrm{tr}\left(\left(\mathbf{I} + \varsigma'\alpha_t^2\widehat{\tilde{\nabla} f}_t^\top\mathbf{B}\widehat{\tilde{\nabla} f}_t\right)^{-1}\widehat{\tilde{\nabla} f}_t^\top\mathbf{B}\widehat{\tilde{\nabla} f}_t\right) \leq \alpha_t^2\mathrm{tr}\left(\widehat{\tilde{\nabla} f}_t^\top\mathbf{B}\widehat{\tilde{\nabla} f}_t\right) \\ &= \alpha_t^2\left\|\widehat{\tilde{\nabla} f}_t\right\|_{\mathbf{B},F}^2 \leq 2\alpha_t^2\left(\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2 + \left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F}^2\right). \end{aligned} \tag{4}$$

By Equations (1)-(4) and Lemma 6.2, we get that

$$\begin{aligned} \mathrm{dist}_m^2(\mathbf{X}_{t+1}, \mathbf{U}_{k''}) &\leq \mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k''}) - 2\alpha_t\mathrm{dist}_f(\mathbf{X}_t, \mathbf{U}_{k''}) + 32k\lambda_1^2\alpha_t^2\eta_{k''t}^2 \\ &\quad + 2\left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F}\frac{\mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U}) + \alpha_t\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}}{1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})} + 2\alpha_t^2\left(\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2 + \left\|\xi_t(\widehat{\nabla f}_t)\right\|_{\mathbf{B},F}^2\right). \end{aligned}$$

By Lemma 6.5,

$$\mathrm{dist}_f(\mathbf{X}_t, \mathbf{U}_{k''}) \geq \Delta_{k''} \mathrm{dist}_b^2(\mathbf{X}_t, \mathbf{U}_{k''}).$$

Further, by Lemma 6.4 and using inequality $x \geq \frac{-\log(1-x)}{1-\log(1-x)}$, we can write that

$$\mathrm{dist}_b^2(\mathbf{X}_0, \mathbf{U}_{k''}) \geq \mathrm{dist}_b^2(\mathbf{X}_t, \mathbf{U}_{k''}) \geq \frac{\mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k''})}{1 + \mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k''})} \geq \frac{\mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k''})}{1 + \psi(\mathbf{X}_t, \mathcal{U})}$$

and

$$\mathrm{dist}_b(\mathbf{X}_t, \mathbf{U}_{k''}) \leq \mathrm{dist}_b(\mathbf{X}_t, \mathcal{U}) \leq \psi^{\frac{1}{2}}(\mathbf{X}_t, \mathcal{U}) \leq \psi^{\frac{1}{2}}(\mathbf{X}_0, \mathcal{U}).$$

Simple algebraic manipulations then yield that

$$\begin{aligned} \mathrm{dist}_m^2(\mathbf{X}_{t+1}, \mathbf{U}_{k''}) \quad &\leq \quad \left(1 - \frac{2\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})}\right) \mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k''}) + \frac{\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})} \psi(\mathbf{X}_t, \mathcal{U}) \\ &\quad + 4\lambda_1^2 \alpha_t^2 \left(\frac{16k}{(1 - \mathrm{dist}_b^2(\mathbf{X}_0))^2} \psi(\mathbf{X}_t, \mathcal{U}) + \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\lambda_1^2} \right). \end{aligned}$$

Analogously, we also have that

$$\begin{aligned} \mathrm{dist}_m^2(\mathbf{X}_{t+1}, \mathbf{U}_{k'}) \quad &\leq \quad \left(1 - \frac{2\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})}\right) \mathrm{dist}_m^2(\mathbf{X}_t, \mathbf{U}_{k'}) + \frac{\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})} \psi(\mathbf{X}_t, \mathcal{U}) \\ &\quad + 4\lambda_1^2 \alpha_t^2 \left(\frac{16k}{(1 - \mathrm{dist}_b^2(\mathbf{X}_0))^2} \psi(\mathbf{X}_t, \mathcal{U}) + \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\lambda_1^2} \right). \end{aligned}$$

If $0 < \alpha_t < \frac{1+\psi(\mathbf{X}_t, \mathcal{U})}{2\Delta_\dagger}$, taking the maximum over $\mathbf{U}_{k'}$ and $\mathbf{U}_{k''}$ gives us

$$\begin{aligned} \psi(\mathbf{X}_{t+1}, \mathcal{U}) \quad &\leq \quad \left(1 - \frac{2\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})}\right) \psi(\mathbf{X}_t, \mathcal{U}) + \frac{\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_t, \mathcal{U})} \psi(\mathbf{X}_t, \mathcal{U}) \\ &\quad + 4\lambda_1^2 \alpha_t^2 \left(\frac{16k}{(1 - \mathrm{dist}_b^2(\mathbf{X}_0, \mathcal{U}))^2} \psi(\mathbf{X}_t, \mathcal{U}) + \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\lambda_1^2} \right) \\ &\leq \quad \left(1 - \frac{\alpha_t \Delta_\dagger}{1 + \psi(\mathbf{X}_0, \mathcal{U})}\right) \psi(\mathbf{X}_t, \mathcal{U}) + 4\lambda_1^2 \alpha_t^2 \left(\frac{16k}{(1 - \mathrm{dist}_b^2(\mathbf{X}_0, \mathcal{U}))^2} \psi(\mathbf{X}_t, \mathcal{U}) + \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\lambda_1^2} \right) \quad (5) \end{aligned}$$

Next, two different settings of step-sizes are considered.

- Consider $\alpha_t = \frac{\mu}{\nu + t}$. By Lemma 6.6, we have $\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2 < k\lambda_1^2$ and then can write

$$\psi(\mathbf{X}_{t+1}, \mathcal{U}) \leq \left(1 - \frac{\Delta_\dagger}{1 + \psi(\mathbf{X}_0, \mathcal{U})} \frac{\mu}{\nu + t}\right) \psi(\mathbf{X}_t, \mathcal{U}) + 4k \left(\frac{\mu \lambda_1}{\nu + t}\right)^2 \left(1 + \frac{16\psi(\mathbf{X}_0, \mathcal{U})}{(1 - \mathrm{dist}_b^2(\mathbf{X}_0, \mathcal{U}))^2}\right).$$

  Let $\mu = O\left(\frac{1}{\Delta_\dagger}\right)$ such that $a = \frac{\mu \Delta_\dagger}{1 + \psi(\mathbf{X}_0, \mathcal{U})} > 1$ and ν is sufficiently large. By Lemma 6.7, we get that $\psi(\mathbf{X}_t, \mathcal{U}) = O\left(\left(\frac{\lambda_1}{\Delta_\dagger}\right)^2 \frac{1}{t}\right)$ and thus $T = O\left(\left(\frac{\lambda_1}{\Delta_\dagger}\right)^2 \frac{1}{\epsilon}\right)$ such that $\psi(\mathbf{X}_T, \mathcal{U}) < \epsilon$. For $t < T$, we can assume that $\psi(\mathbf{X}_t, \mathcal{U}) \geq \epsilon$. Using inequality $\frac{x}{1+x} \leq \log(1 + x)$ for $x > -1$, we have that

$$\frac{\mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})}{1 - \mathrm{dist}_b^2(\mathbf{X}_t, \mathcal{U})} \geq \psi(\mathbf{X}_t, \mathcal{U}) \geq \epsilon.$$

Thus,

$$\begin{aligned}
\log \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\left\|\xi_t(\widehat{\nabla} f_t)\right\|_{\mathbf{B},F}^2} &= \log \frac{k\lambda_1^2}{\left(\frac{\Delta_\dagger}{4k^{\frac{1}{2}}} \frac{1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})}{1+\psi(\mathbf{X}_t,\mathcal{U})} \mathrm{dist}_b(\mathbf{X}_t,\mathcal{U})\right)^2} \\
&= O\left(\log \frac{k\lambda_1^2}{\left(\frac{\Delta_\dagger}{4k^{\frac{1}{2}}} \frac{1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})}{1+\psi(\mathbf{X}_t,\mathcal{U})} \left(1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})\right)\epsilon\right)^2}\right) \\
&= O\left(\log \frac{\lambda_1}{\Delta_\dagger} + \psi(\mathbf{X}_0,\mathcal{U}) + \log \frac{1}{\epsilon}\right) = O\left(\log \frac{\lambda_1}{\Delta_\dagger} + \log \frac{1}{\epsilon}\right),
\end{aligned}$$

where we have used that

$$\log\left(1+\psi(\mathbf{X}_0,\mathcal{U})\right) \le \psi(\mathbf{X}_0,\mathcal{U}) < -2k \log \frac{\eta\sqrt{\kappa(\mathbf{B})}}{k+\sqrt{nk}} < +\infty$$

with probability at least $1-\eta$ for any $\eta > 0$, by Lemma 6.9. By Lemma 6.3, the complexity for the subproblem then is

$$O\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})} \log \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B}}^2}{\left\|\xi_t(\widehat{\nabla} f_t)\right\|_{\mathbf{B}}^2}\right) = O\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\left(\log \frac{\lambda_1}{\Delta_\dagger} + \log \frac{1}{\epsilon}\right)\right).$$

Therefore, the total complexity is

$$O\left(\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\left(\log \frac{\lambda_1}{\Delta_\dagger} + \log \frac{1}{\epsilon}\right)\right)\left(\frac{\lambda_1}{\Delta_\dagger}\right)^2 \frac{1}{\epsilon}\right),$$

which completes the proof of the first statement.

- Consider $\alpha_t = \alpha > 0$ and note for Categories a)-c) and e) that by Lemma 6.8, it holds

  $$\psi(\mathbf{X}_t,\mathcal{U}) = \min_{\mathbf{U}\in\mathcal{U}} \mathrm{dist}_m^2(\mathbf{X}_t,\mathbf{U}),$$

  which holds for Category d) as well if $\psi(\mathbf{X}_0,\mathcal{U})$ is sufficiently close to $\mathcal{U}$. Accordingly, by Lemma 6.6, we get that

  $$\left\|\tilde{\nabla} f(\mathbf{X}_t)\right\|_{\mathbf{B},F}^2 \le 4k\lambda_1^2 \psi(\mathbf{X}_t,\mathcal{U}).$$

  Plugging into Equation (5), we arrive at

  $$\psi(\mathbf{X}_{t+1},\mathcal{U}) \le \left(1 - \frac{\alpha\Delta_\dagger}{1+\psi(\mathbf{X}_0,\mathcal{U})}\right)\psi(\mathbf{X}_t,\mathcal{U}) + 16k\lambda_1^2\alpha^2\left(1 + \left(\frac{1-\mathrm{dist}_b^2(\mathbf{X}_0,\mathcal{U})}{2}\right)^{-2}\right)\psi(\mathbf{X}_t,\mathcal{U}).$$

  If $0 < \alpha < \frac{\Delta_\dagger}{32k\lambda_1^2(1+\psi(\mathbf{X}_0,\mathcal{U}))\left(1+\left(\frac{1-d_b^2(\mathbf{X}_0,\mathcal{U})}{2}\right)^{-2}\right)}$, one can write that

  $$\psi(\mathbf{X}_T,\mathcal{U}) \le \left(1 - \frac{\alpha\Delta_\dagger}{2(1+\psi(\mathbf{X}_0,\mathcal{U}))}\right)\psi(\mathbf{X}_{T-1},\mathcal{U}) \le \cdots \le \left(1 - \frac{\alpha\Delta_\dagger}{2(1+\psi(\mathbf{X}_0,\mathcal{U}))}\right)^T \psi(\mathbf{X}_0,\mathcal{U}).$$

  Setting $\left(1 - \frac{\alpha\Delta_\dagger}{2(1+\psi(\mathbf{X}_0,\mathcal{U}))}\right)^T \psi(\mathbf{X}_0,\mathcal{U}) = \epsilon$ yields that

  $$\begin{aligned}
  T &= O\left(\frac{1}{-\log\left(1 - \frac{\alpha\Delta_\dagger}{2(1+\psi(\mathbf{X}_0,\mathcal{U}))}\right)} \log \frac{\psi(\mathbf{X}_0,\mathcal{U})}{\epsilon}\right) = O\left(\frac{1+\psi(\mathbf{X}_0,\mathcal{U})}{\alpha\Delta_\dagger} \log \frac{\psi(\mathbf{X}_0,\mathcal{U})}{\epsilon}\right) \\
  &= O\left(\left(\frac{\lambda_1}{\Delta_\dagger}\right)^2 \log \frac{\psi(\mathbf{X}_0,\mathcal{U})}{\epsilon}\right).
  \end{aligned}$$

For the subproblem, we now have that

$$
\begin{aligned}
\log \frac{\left\|\tilde{\nabla} f_t\right\|_{\mathbf{B},F}^2}{\left\|\xi_t(\widehat{\nabla} f_t)\right\|_{\mathbf{B},F}^2} &= O\left(\log \frac{2k\lambda_1^2\psi(\mathbf{X}_t,\mathcal{U})}{\left(\frac{\Delta_\dagger}{4k^{\frac{1}{2}}}\frac{1-\mathrm{dist}_b^2(\mathbf{X}_t,\mathcal{U})}{1+\psi(\mathbf{X}_t,\mathcal{U})}\mathrm{dist}_b(\mathbf{X}_t,\mathcal{U})\right)^2}\right) = O\left(\log\frac{\lambda_1}{\Delta_\dagger} + \psi(\mathbf{X}_0,\mathcal{U})\right) \\
&= O\left(\log \frac{\lambda_1}{\Delta_\dagger}\right).
\end{aligned}
$$

Therefore, the total complexity is

$$
O\left(\left(\mathrm{nnz}(\mathbf{A}) + \mathrm{nnz}(\mathbf{B})\sqrt{\kappa(\mathbf{B})}\log\frac{\lambda_1}{\tilde{\Delta}}\right)\left(\frac{\lambda_1}{\tilde{\Delta}}\right)^2 \log\frac{1}{\epsilon}\right),
$$

which completes the proof of the second statement.

□

Proof of Lemma 6.2

Let $j \geq k$ and denote $\nabla f_t \triangleq \tilde{\nabla} f(\mathbf{X}_t)$ and $-2\log \mathrm{Det}\left((\widehat{\mathbf{X}}_{t+1})^\top \mathbf{B}\mathbf{U}_j\right) \triangleq -\log\det(\mathbf{S})$. Note that

$$
\mathbf{S} \succcurlyeq \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}_t + 2\alpha_{t+1}\mathrm{sym}\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\nabla f_t\right) \triangleq \mathbf{H}_1 + \mathbf{H}_2.
$$

Hence, we have that $-\log\det(\mathbf{S}) \leq -\log\det(\mathbf{H}_1+\mathbf{H}_2)$. By Taylor expansion, we can write for certain $\varsigma \in (0,1)$ that

$$
-\log\det(\mathbf{S}) \quad\leq\quad -\log\det(\mathbf{H}_1) - \mathrm{tr}\left(\mathbf{H}_1^{-1}\mathbf{H}_2\right) + \frac{1}{2}\mathrm{tr}\left(\left(\left(\mathbf{H}_1+\varsigma\mathbf{H}_2\right)^{-1}\mathbf{H}_2\right)^2\right),
$$

where $-\log\det(\mathbf{H}_1) = \mathrm{dist}_m^2(\mathbf{X}_t,\mathbf{U}_j)$. Noting that $\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j = \mathbf{P}_j\mathbf{\Sigma}_j\mathbf{Q}_j^\top$ (subscripts t on the right-hand side are omitted for brevity), we can write that

$$
\begin{aligned}
\mathrm{tr}\left(\mathbf{H}_1^{-1}\mathbf{H}_2\right) &= 2\alpha_t\mathrm{tr}\left(\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}_t\right)^{-1}\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\left(\mathbf{B}^{-1}-\mathbf{X}_t\mathbf{X}_t^\top\right)\mathbf{A}\mathbf{X}_t\right) \\
&= 2\alpha_t\left(\mathrm{tr}\left(\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}_t\right)^{-1}\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{A}\mathbf{X}_t\right) - \mathrm{tr}\left(\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t\right)\right) \\
&= 2\alpha_t\left(\mathrm{tr}\left(\left(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}_t\right)^{-1}\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j\mathbf{\Lambda}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}_t\right) - \mathrm{tr}\left(\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t\right)\right) \\
&= 2\alpha_t\left(\mathrm{tr}\left(\left(\mathbf{P}_j\mathbf{\Lambda}_j^2\mathbf{P}_j^\top\right)^{-1}\mathbf{P}_j\mathbf{\Sigma}_j\mathbf{Q}_j^\top\mathbf{\Lambda}_j\mathbf{Q}_j\mathbf{\Sigma}_j\mathbf{P}_j^\top\right) - \mathrm{tr}\left(\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t\right)\right) \\
&= 2\alpha_t\left(\mathrm{tr}(\mathbf{Q}_j^\top\mathbf{\Lambda}_j\mathbf{Q}_j) - \mathrm{tr}(\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t)\right) = 2\alpha_t\left(f(\mathbf{U}_j\mathbf{Q}_j) - f(\mathbf{X}_t)\right).
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\mathrm{tr}\left(\left(\left(\mathbf{H}_1+\varsigma\mathbf{H}_2\right)^{-1}\mathbf{H}_2\right)^2\right) &\leq \left\|\left(\mathbf{H}_1+\varsigma\mathbf{H}_2\right)^{-1}\mathbf{H}_2\right\|_F^2 \leq \left(\left\|\left(\mathbf{H}_1+\varsigma\mathbf{H}_2\right)^{-1}\right\|_2\|\mathbf{H}_2\|_F\right)^2 \\
&= \left(\frac{\|\mathbf{H}_2\|_F}{\sigma_{\min}(\mathbf{H}_1+\varsigma\mathbf{H}_2)}\right)^2,
\end{aligned}
$$

where we need to lower bound $\sigma_{\min}(\mathbf{H}_1+\varsigma\mathbf{H}_2)$ and upper bound $\|\mathbf{H}_2\|_F$. To this end, notice that

$$
\sigma_{\min}(\mathbf{H}_1) = \sigma_{\min}^2(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j)\mathbf{I} \geq \prod_{i=1}^k \sigma_i^2(\mathbf{X}_t^\top\mathbf{B}\mathbf{U}_j) = 1 - \mathrm{dist}_b^2(\mathbf{X}_t,\mathbf{U}_j).
$$

Letting $\mathbf{\Omega} = \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j$, we have that

$$\begin{aligned}\mathbf{H}_2 &= 2\alpha_t \text{sym}(\mathbf{\Omega}\mathbf{U}_j^\top \mathbf{B}\nabla f_t) = 2\alpha_t \text{sym}\left(\mathbf{\Omega}\mathbf{U}_j^\top \mathbf{B}(\mathbf{B}^{-1} - \mathbf{X}_t\mathbf{X}_t^\top)\mathbf{A}\mathbf{X}_t)\right) \\ &= 2\alpha_t \mathbf{\Omega}\mathbf{\Lambda}_j\mathbf{\Omega}^\top - 2\alpha_t \text{sym}(\mathbf{\Omega}\mathbf{\Omega}^\top\mathbf{\Omega}\mathbf{\Lambda}_j\mathbf{\Omega}^\top) - 2\alpha_t \text{sym}\left(\mathbf{\Omega}\mathbf{\Omega}^\top \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp \mathbf{\Lambda}_j^\perp (\mathbf{U}_j^\perp)^\top \mathbf{B}\mathbf{X}_t\right) \\ &= 2\alpha_t \text{sym}\left((\mathbf{I} - \mathbf{\Omega}\mathbf{\Omega}^\top)\mathbf{\Omega}\mathbf{\Lambda}_j\mathbf{\Omega}^\top\right) - 2\alpha_t \text{sym}\left(\mathbf{\Omega}\mathbf{\Omega}^\top \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp \mathbf{\Lambda}_j^\perp (\mathbf{U}_j^\perp)^\top \mathbf{B}\mathbf{X}_t\right).\end{aligned}$$

Thus, we have that

$$\begin{aligned}\|\mathbf{H}_2\|_2 &\leq 2\alpha_t \left(\left\|(\mathbf{I} - \mathbf{\Omega}\mathbf{\Omega}^\top)\mathbf{\Omega}\mathbf{\Lambda}_j\mathbf{\Omega}^\top\right\|_2 + \left\|\mathbf{\Omega}\mathbf{\Omega}^\top \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp \mathbf{\Lambda}_j^\perp (\mathbf{U}_j^\perp)^\top \mathbf{B}\mathbf{X}_t\right\|_2\right) \\ &\leq 2\alpha_t \left(\|\mathbf{I} - \mathbf{\Omega}\mathbf{\Omega}^\top\|_2 \|\mathbf{\Lambda}_j\|_2 + \|\mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp\|_2^2 \|\mathbf{\Lambda}_j^\perp\|_2\right) \\ &\leq 2\alpha_t\lambda_1 \left(\|\mathbf{I} - \mathbf{\Omega}\mathbf{\Omega}^\top\|_2 + \|\mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp\|_2^2\right) = 4\alpha_t\lambda_1\|\mathbf{I} - \mathbf{\Omega}\mathbf{\Omega}^\top\|_2 \\ &= 4\alpha_t\lambda_1\|\mathbf{I} - \mathbf{P}_j\mathbf{\Sigma}_j^2\mathbf{P}_j^\top\|_2 = 4\alpha_t\lambda_1\|\mathbf{I} - \mathbf{\Sigma}_j^2\|_2 \leq 4\alpha_t\lambda_1 \text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j),\end{aligned}$$

where we have used for the first equality that

$$\|\mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp\|_2^2 = \lambda_{\max}\left(\mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j^\perp (\mathbf{U}_j^\perp)^\top \mathbf{B}\mathbf{X}_t\right) = \lambda_{\max}(\mathbf{I} - \mathbf{X}_t^\top \mathbf{B}\mathbf{U}_j\mathbf{U}_j^\top \mathbf{B}\mathbf{X}_t).$$

Hence,if $0 < \alpha_t < \frac{1-\text{dist}_b^2(\mathbf{X}_t,\mathbf{U}_j)}{8\lambda_1\text{dist}_b^2(\mathbf{X}_t,\mathbf{U}_j)}$ then

$$\begin{aligned}\sigma_{\min}(\mathbf{H}_1 + \varsigma\mathbf{H}_2) &\geq \sigma_{\min}(\mathbf{H}_1) - \sigma_{\max}(\mathbf{H}_2) \geq (1 - \text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)) - 4\alpha_t\lambda_1\text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j) \\ &\geq \frac{1 - \text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)}{2},\end{aligned}$$

and

$$\|\mathbf{H}_2\|_F \leq k^{\frac{1}{2}}\|\mathbf{H}_2\|_2 \leq 4k^{\frac{1}{2}}\alpha_t\lambda_1\text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j).$$

We thus get that

$$\text{tr}\left(\left((\mathbf{H}_1 + \varsigma\mathbf{H}_2)^{-1}\mathbf{H}_2\right)^2\right) \leq 64k\lambda_1^2\alpha_t^2\left(\frac{\text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)}{1 - \text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)}\right)^2$$

and consequently,

$$-2\log\text{Det}\left(\widehat{\mathbf{X}}_{t+1}^\top \mathbf{B}\mathbf{U}_j\right) \leq \text{dist}_m^2(\mathbf{X}_t, \mathbf{U}_j) - 2\alpha_t(f(\mathbf{U}_j\mathbf{Q}_j) - f(\mathbf{X}_t)) + 32k\lambda_1^2\alpha_t^2\left(\frac{\text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)}{1 - \text{dist}_b^2(\mathbf{X}_t, \mathbf{U}_j)}\right)^2.$$

The case that $j \leq k$ is similar and thus omitted. □

Proof of Lemma 6.3

$l_t(\mathbf{X})$ reaches its minimum at

$$\begin{aligned}l_t(\mathbf{X}_t^\star) &= \frac{1}{2}\text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star\right) - \text{tr}((\mathbf{X}_t^\star)^\top \mathbf{A}\mathbf{X}_t) = \frac{1}{2}\text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star\right) - \text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{B}^{-1}\mathbf{A}\mathbf{X}_t\right) \\ &= -\frac{1}{2}\text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star\right).\end{aligned}$$

Thus, we have that

$$\begin{aligned}\epsilon_t(\mathbf{X}) &= l_t(\mathbf{X}) - l_t(\mathbf{X}_t^\star) = \frac{1}{2}\text{tr}(\mathbf{X}^\top \mathbf{B}\mathbf{X}) - \text{tr}(\mathbf{X}^\top \mathbf{A}\mathbf{X}_t) + \frac{1}{2}\text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star\right) \\ &= \frac{1}{2}\text{tr}(\mathbf{X}^\top \mathbf{B}\mathbf{X}) - \text{tr}(\mathbf{X}^\top \mathbf{B}\mathbf{B}^{-1}\mathbf{A}\mathbf{X}_t) + \frac{1}{2}\text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star\right) \\ &= \frac{1}{2}(\text{tr}(\mathbf{X}^\top \mathbf{B}\mathbf{X}) - 2\text{tr}(\mathbf{X}^\top \mathbf{B}\mathbf{X}_t^\star) + \text{tr}\left((\mathbf{X}_t^\star)^\top \mathbf{B}\mathbf{X}_t^\star)\right) \\ &= \frac{1}{2}\text{tr}\left((\mathbf{X} - \mathbf{X}_t^\star)^\top \mathbf{B}(\mathbf{X} - \mathbf{X}_t^\star)\right) = \frac{1}{2}\|\xi_t(\mathbf{X})\|_{\mathbf{B},F}^2.\end{aligned}$$

In particular,

$$\xi_t(\mathbf{X}_t^{(0)}) = \mathbf{X}_t^{(0)} - \mathbf{B}^{-1}\mathbf{A}\mathbf{X}_t = \mathbf{X}_t(\mathbf{X}_t^\top\mathbf{B}\mathbf{X}_t)^{-1}\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{X}_t$$
$$= \mathbf{X}_t\mathbf{X}_t^\top\mathbf{A}\mathbf{X}_t - \mathbf{B}^{-1}\mathbf{A}\mathbf{X}_t = -\tilde{\nabla}f(\mathbf{X}_t).$$

The complexity of Nesterov's accelerated gradient descent for the least squares subproblem can be found in Nesterov (2014); Bubeck (2015); Ge et al. (2016), given that $l_t(\mathbf{X})$ is $\lambda_{\min}(\mathbf{B})$-strongly convex and $\lambda_{\max}(\mathbf{B})$-smooth, where $\lambda_{\max}(\mathbf{B})$ and $\lambda_{\min}(\mathbf{B})$ represent the largest and smallest eigenvalue of $\mathbf{B}$, respectively. □

Proof of Lemma 6.4

Let $x = \mathrm{dist}_b^2(\mathbf{X}, \mathbf{Y})$. We then have that $\mathrm{dist}_b^2(\mathbf{X}, \mathbf{Y}) = x \le -\log(1-x) = \mathrm{dist}_m^2(\mathbf{X}, \mathbf{Y})$. We next prove by induction that $\mathrm{dist}_b(\mathbf{X}, \mathbf{Y}) \le \mathrm{dist}_c(\mathbf{X}, \mathbf{Y})$. Let $r = \min\{k, l\}$ and θ_i be the i-th principal angle between $\mathbf{X}$ and $\mathbf{Y}$, $i = 1, \cdots, r$. That is, $\cos\theta_i = \sigma_i(\mathbf{X}^\top\mathbf{B}\mathbf{Y})$, where $\sigma_i(\cdot)$ represents the i-th largest singular value of a matrix. First, we have for $r = 1$ that

$$\mathrm{dist}_b^2(\mathbf{X}, \mathbf{Y}) = 1 - \prod_{i=1}^r \cos^2\theta_i = r - \sum_{i=1}^r \cos^2\theta_i = \mathrm{dist}_c^2(\mathbf{X}, \mathbf{Y}).$$

Assuming that it holds for r, one then has for $r+1$ that

$$\begin{aligned}
\mathrm{dist}_c^2 &= r + 1 - \sum_{i=1}^{r+1}\cos^2\theta_i = r - \sum_{i=1}^{r}\cos^2\theta_i + 1 - \cos^2\theta_{r+1} \\
&\ge 1 - \prod_{i=1}^{r}\cos^2\theta_i + 1 - \cos^2\theta_{r+1} - (1 - \prod_{i=1}^{r+1}\cos^2\theta_i) + 1 - \prod_{i=1}^{r+1}\cos^2\theta_i \\
&= (1 - \cos^2\theta_{r+1})(1 - \prod_{i=1}^{r}\cos^2\theta_i) + 1 - \prod_{i=1}^{r+1}\cos^2\theta_i \ge 1 - \prod_{i=1}^{r+1}\cos^2\theta_i = \mathrm{dist}_b^2,
\end{aligned}$$

which completes the proof. The last inequality can be shown by the generalized mean inequality as follows:

$$\sum_{i=1}^r \cos^2\theta_i = r\left(\frac{\sum_{i=1}^r \cos^2\theta_i}{r}\right)^{\frac{1}{2}\cdot 2} \ge r\left(\prod_{i=1}^r \cos\theta_i\right)^{\frac{2}{r}} \ge r\left(\prod_{i=1}^r \cos\theta_i\right)^2.$$

It then holds that $\mathrm{dist}_c^2 \le r - r(\prod_{i=1}^r \cos\theta_i)^2 = r(1 - \prod_{i=1}^r \cos^2\theta_i) = r\mathrm{dist}_b^2$. □

Proof of Lemma 6.5

Suppose that $j \le k$, $\Lambda_j = \mathrm{diag}(\lambda_1, \cdots, \lambda_j)$ and $\Lambda_j^\perp = \mathrm{diag}(\lambda_{j+1}, \cdots, \lambda_n)$. We have that

$$\begin{aligned}
\mathrm{dist}_f(\mathbf{X}, \mathbf{U}_j) &= f(\mathbf{U}_j) - f(\mathbf{X}\mathbf{P}_j) = \mathrm{tr}(\mathbf{\Lambda}_j) - \mathrm{tr}\left(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{A}\mathbf{X}\mathbf{P}_j\right) \\
&= \mathrm{tr}(\mathbf{\Lambda}_j) - \mathrm{tr}\left(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{B}\mathbf{U}_j\mathbf{\Lambda}_j\mathbf{U}_j^\top\mathbf{B}\mathbf{X}\mathbf{P}_j\right) - \mathrm{tr}\left(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{B}\mathbf{U}_j^\perp\mathbf{\Lambda}_j^\perp(\mathbf{U}_j^\perp)^\top\mathbf{B}\mathbf{X}\mathbf{P}_j\right) \\
&= \mathrm{tr}(\mathbf{\Lambda}_j) - \mathrm{tr}(\mathbf{\Sigma}_j\mathbf{Q}_j^\top\mathbf{\Lambda}_j\mathbf{Q}_j\mathbf{\Sigma}_j) - \mathrm{tr}(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{B}\mathbf{U}_j^\perp\mathbf{\Lambda}_j^\perp(\mathbf{U}_j^\perp)^\top\mathbf{B}\mathbf{X}\mathbf{P}_j) \\
&= \mathrm{tr}(\mathbf{\Lambda}_j\mathbf{Q}_j(\mathbf{I} - \mathbf{\Sigma}_j^2)\mathbf{Q}_j^\top) - \mathrm{tr}(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{B}\mathbf{U}_j^\perp\mathbf{\Lambda}_j^\perp(\mathbf{U}_j^\perp)^\top\mathbf{B}\mathbf{X}\mathbf{P}_j) \\
&\ge \lambda_j\mathrm{tr}(\mathbf{Q}_j(\mathbf{I} - \mathbf{\Sigma}_j^2)\mathbf{Q}_j^\top) - \lambda_{j+1}\mathrm{tr}(\mathbf{P}_j^\top\mathbf{X}^\top\mathbf{B}\mathbf{U}_j^\perp(\mathbf{U}_j^\perp)^\top\mathbf{B}\mathbf{X}\mathbf{P}_j) \\
&= (\lambda_j - \lambda_{j+1})\mathrm{tr}(\mathbf{Q}_j(\mathbf{I} - \mathbf{\Sigma}_j^2)\mathbf{Q}_j^\top) = \Delta_j\left(j - \|\mathbf{X}^\top\mathbf{B}\mathbf{U}_j\|_F^2\right) = \Delta_j\mathrm{dist}_c^2(\mathbf{X}, \mathbf{U}_j) \\
&\ge \Delta_j\mathrm{dist}_b^2(\mathbf{X}, \mathbf{U}_j),
\end{aligned}$$

where the last inequality is by Lemma 6.4. The case that $j \ge k$ is similar and thus omitted. □

Proof of Lemma 6.6

Note that $\mathbf{X}$'s orthogonal complement $\mathbf{X}_\perp \in \mathrm{gSt}_{\mathbf{B}}(n, n-k)$ and $\mathbf{X}_\perp^\top \mathbf{B}\mathbf{X} = \mathbf{0}$. Thus,

$$\begin{aligned}\left\|\mathbf{B}^{1/2}\tilde{\nabla} f(\mathbf{X})\right\|_2 &= \left\|\left(\mathbf{I} - \mathbf{B}^{1/2}\mathbf{X}\mathbf{X}^\top\mathbf{B}^{1/2}\right)\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{X}\right\|_2 \\ &= \left\|\mathbf{B}^{1/2}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{B}^{1/2}\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}\mathbf{B}^{1/2}\mathbf{X}\right\|_2 \leq \left\|\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-\frac{1}{2}}\right\|_2 = \lambda_1.\end{aligned}$$

Accordingly, $\left\|\tilde{\nabla} f(\mathbf{X})\right\|_{\mathbf{B},F}^2 = \left\|\mathbf{B}^{1/2}\tilde{\nabla} f(\mathbf{X})\right\|_F^2 \leq k\lambda_1^2$.

Let $(j_1, \cdots, j_k)$ be an arbitrary k-combination chosen from $\{1, 2, \cdots, n\}$. Then for any $\mathbf{V} = (\mathbf{u}_{j_1}, \cdots, \mathbf{u}_{j_k})$ and corresponding $\Lambda = (\lambda_{j_1}, \cdots, \lambda_{j_k})$, we have that

$$\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \mathbf{B}^{1/2}(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{V}_\perp\mathbf{\Lambda}_\perp\mathbf{V}_\perp^\top)\mathbf{B}^{1/2}.$$

Plugging in this equation to the above derivation and using Lemma 6.4, we can write that

$$\begin{aligned}\left\|\tilde{\nabla} f(\mathbf{X})\right\|_{\mathbf{B},F}^2 &= \left\|\mathbf{X}_\perp^\top\mathbf{B}\left(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{V}_\perp\mathbf{\Lambda}_\perp\mathbf{V}_\perp^\top\right)\mathbf{B}\mathbf{X}\right\|_F^2 \\ &\leq \left(\|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{V}\|_F\|\mathbf{\Lambda}\|_2 + \|\mathbf{\Lambda}_\perp\|_2\|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{X}\|_F\right)^2 \\ &\leq \lambda_1^2\left(\|\mathbf{X}_\perp^\top\mathbf{B}\mathbf{V}\|_F + \|\mathbf{V}_\perp^\top\mathbf{B}\mathbf{X}\|_F\right)^2 \\ &= \lambda_1^2\left((k - \|\mathbf{X}^\top\mathbf{B}\mathbf{V}\|_F^2)^{1/2} + (k - \|\mathbf{V}^\top\mathbf{B}\mathbf{X}\|_F^2)^{1/2}\right)^2 \\ &= 4\lambda_1^2\mathrm{dist}_c^2(\mathbf{X}, \mathbf{V}) \leq 4k\lambda_1^2\mathrm{dist}_b^2(\mathbf{X}, \mathbf{V}) \leq 4k\lambda_1^2\mathrm{dist}_m^2(\mathbf{X}, \mathbf{V}).\end{aligned}$$

The proof completes by noting that any $\mathbf{U} \in \mathcal{U}$ is such a $\mathbf{V}$ up to an orthogonal matrix. □

Proof of Lemma 6.9

For any $\mathbf{U} \in \mathcal{U}$, by the above Lemma 6.8 we have that

$$\mathrm{dist}_m^2(\mathbf{X}_0, \mathbf{U}_j) \leq \mathrm{dist}_m^2(\mathbf{X}_0, \mathbf{U}) = -2\sum_{i=1}^{k}\log\sigma_i(\mathbf{X}_0^\top\mathbf{B}\mathbf{U}) \leq -2k\log\sigma_{\min}(\mathbf{X}_0^\top\mathbf{B}\mathbf{U}),$$

and

$$\begin{aligned}\sigma_{\min}(\mathbf{X}_0^\top\mathbf{B}\mathbf{U}) &= \sigma_{\min}\left((\mathbf{W}^\top\mathbf{B}\mathbf{W})^{-\frac{1}{2}}\mathbf{W}^\top\mathbf{B}\mathbf{U}\right) \geq \sigma_{\min}\left((\mathbf{W}^\top\mathbf{B}\mathbf{W})^{-\frac{1}{2}}\right)\sigma_{\min}(\mathbf{W}^\top\mathbf{B}\mathbf{U}) \\ &= \frac{\sigma_{\min}(\mathbf{W}^\top\mathbf{B}\mathbf{U})}{\sigma_{\max}(\mathbf{B}^{\frac{1}{2}}\mathbf{W})} \geq \frac{\sigma_{\min}(\mathbf{W}^\top\mathbf{B}\mathbf{U})}{\sigma_{\max}(\mathbf{B}^{\frac{1}{2}})\|\mathbf{W}\|_2},\end{aligned}$$

where $\|\mathbf{W}\|_2 \sim O(n^{\frac{1}{2}} + k^{\frac{1}{2}})$ with high probability. Let $\widehat{\mathbf{U}} \in \mathbb{R}^{n\times k}$ be the left singular vectors of $\mathbf{B}\mathbf{U}$. One then can write $\mathbf{W}^\top\mathbf{B}\mathbf{U} = \mathbf{W}^\top\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{B}\mathbf{U}$ and thus

$$\sigma_{\min}(\mathbf{W}^\top\mathbf{B}\mathbf{U}) \geq \sigma_{\min}(\mathbf{W}^\top\widehat{\mathbf{U}})\sigma_{\min}(\widehat{\mathbf{U}}^\top\mathbf{B}\mathbf{U}) = \sigma_{\min}(\mathbf{W}^\top\widehat{\mathbf{U}})\sigma_{\min}(\mathbf{B}\mathbf{U}) \geq \sigma_{\min}(\mathbf{W}^\top\widehat{\mathbf{U}})\sigma_{\min}(\mathbf{B}^{\frac{1}{2}}),$$

where the last inequality is because that

$$\begin{aligned}\sigma_{\min}^2(\mathbf{B}\mathbf{U}) &= \lambda_{\min}(\mathbf{U}^\top\mathbf{B}^2\mathbf{U}) = \min_{\|\mathbf{x}\|_2=1}\mathbf{x}^\top\mathbf{U}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{B}\mathbf{B}^{\frac{1}{2}}\mathbf{U}\mathbf{x} \\ &\geq \lambda_{\min}(\mathbf{B})\min_{\|\mathbf{x}\|_2=1}\mathbf{x}^\top\mathbf{U}^\top\mathbf{B}^{\frac{1}{2}}\mathbf{B}^{\frac{1}{2}}\mathbf{U}\mathbf{x} = \lambda_{\min}(\mathbf{B})\sigma_{\min}^2(\mathbf{B}^{\frac{1}{2}}\mathbf{U}) \\ &= \lambda_{\min}(\mathbf{B})\min_{\|\mathbf{x}\|_2=1}\mathbf{x}^\top\mathbf{x} = \lambda_{\min}(\mathbf{B}) = \sigma_{\min}(\mathbf{B}).\end{aligned}$$

We thus get that

$$\sigma_{\min}(\mathbf{X}_0^\top \mathbf{B}\mathbf{U}) \geq \frac{\sqrt{\kappa(\mathbf{B})}}{n^{\frac{1}{2}} + k^{\frac{1}{2}}} \sigma_{\min}(\mathbf{W}^\top \widehat{\mathbf{U}}).$$

Since $\mathbf{W}$ are entry-wise i.i.d. standard normal and $\widehat{\mathbf{U}}$ is orthonormal, $\mathbf{W}^\top \widehat{\mathbf{U}}$ are entry-wise i.i.d. standard normal as well. By Equation (3.2) in Rudelson and Vershynin (2010), we have that for $\eta \geq 0$, $\sigma_{\min}(\mathbf{W}^\top \widehat{\mathbf{U}}) > \eta k^{-\frac{1}{2}}$ with probability at least $1 - \eta$. The proof completes. □

References

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

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, NY, 2016.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1st edition, 2014.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv preprint arXiv:1003.2990*, 2010.