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 $\alpha_t = \mu_t + \nu$ for sufficiently large positive constants $\mu$ and $\nu$ will converge after

$$T = O \left( \left( \frac{\text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B)}}{\Delta_1} \right) \left( \frac{\lambda_1}{\Delta} \right)^2 \frac{1}{\epsilon} \right)$$

iterations with high probability.

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

$$T = O \left( \left( \frac{\text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B)}}{\Delta} \right) \left( \frac{\lambda_1}{\Delta} \right)^2 \log \frac{1}{\epsilon} \right)$$

iterations with high probability. If $X_0$ is sufficiently close to $U$ 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 $\epsilon$) 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 $\epsilon$, the overall complexity will be dominated by the second one.

The two statements are proven in what follows. To analyze $\psi(X, U)$, we focus on $-2 \log \det(X^T B U_{k^*})$ and the other is analogous. To start, we have from Algorithm 1 and Lemma 6.3 that

$$-2 \log \det(X^T B U_{k^*}) = -2 \log \det \left( (X_t + \alpha_t \hat{\nabla} f_t)^T B U_{k^*} \right) + \log \det \left( I + \alpha_t^2 \hat{\nabla} f_t B \hat{\nabla} f_t \right),$$

(1)

where $\hat{\nabla} f_t = \nabla f_t + (I - X_t X_t^T B) \xi_t (\nabla f_t)$. Letting $E_t = (I - X_t X_t^T B) \xi_t (\nabla f_t)$, we can write that $X_t + \alpha_t \hat{\nabla} f_t = \bar{X}_{t+1} + \alpha_t E_t$, where notation $\bar{X}_{t+1}$ is defined in Lemma 6.2. Then

$$\left( X_t + \alpha_t \hat{\nabla} f_t \right)^T B U_{k^*} U_{k^*}^T B \left( X_t + \alpha_t \hat{\nabla} f_t \right) \geq S_1 + 2 \alpha_t S_2,$$

with

$$S_1 = \bar{X}_{t+1}^T B U_{k^*} U_{k^*}^T B \bar{X}_{t+1}$$

and

$$S_2 = \text{sym} \left( \bar{X}_{t+1}^T B U_{k^*} U_{k^*}^T B E_t \right),$$
where \( \text{sym}(M) = \frac{1}{2}(M + M^T) \). By the Taylor expansion, we can get for some \( \varsigma \in (0, 1) \) that
\[
-2 \log \det \left( (X_t + \alpha_t \nabla f_t)^T BU_{k'} \right) \leq -2 \log \det \left( \tilde{X}_{t+1}^T BU_{k'} \right) - 2 \alpha_t \text{tr} \left( (S_1 + 2\varsigma \alpha_t S_2)^{-1} S_2 \right)
\]
\[
\leq -2 \log \det \left( \tilde{X}_{t+1}^T BU_{k'} \right) + \frac{2k^2 \alpha_t \|S_2\|_F}{\sigma_{\min}(S_1 + 2\varsigma \alpha_t S_2)}.
\]
(2)

To proceed, we bound singular values of \( S_i, i = 1, 2 \), as follows.
\[
\sigma(S_1) \geq \sigma_{\min}^2 \left( (X_t + \alpha_t \nabla f_t)^T BU_{k'} \right) \geq \sigma_{\min}^2 \left( X_t^T BU_{k'} \right) - 2\alpha_t \|\nabla f_t\|_{B,2} \quad \text{(Lemma 6.6)}
\]
\[
\geq \prod_{i=1}^k \sigma_i^2 \left( X_t^T BU_{k'} \right) - 2\alpha_t \lambda_1 = 1 - \text{dist}_b^2(X_t, U_{k'}) - 2\alpha_t \lambda_1,
\]
(3)

using notations in Lemma 6.4 and
\[
\sigma(S_2) \leq \left\| \tilde{X}_{t+1}^T BU_{k'} U_{k'}^T B \right\|_F \leq \left\| \tilde{X}_{t+1}^T BU_{k'} \right\|_F \leq \left\| (X_t^T BU_{k'} U_{k'}^T B)(I - X_t X_t^T B)B^{-\frac{1}{2}} \right\|_2 \left\| B^{\frac{1}{2}} \xi_t(\nabla f_t) \right\|_F \leq \left\| X_t^T BU_{k'} U_{k'}^T B \right\|_2 + \alpha_t \left\| \nabla f_t \right\|_{B,2} \left\| \xi_t(\nabla f_t) \right\|_{B,F}
\]

where \( X_t^\perp \) represents the orthogonal complement of \( X_t \) in inner product \( \langle \cdot, \cdot \rangle_B \), i.e., \( X_t^\perp BX_t = 0 \). Moreover, we have
\[
\left\| X_t^T BU_{k'} U_{k'}^T B \right\|_2 \leq \left\| (X_t^T BU_{k'} U_{k'}^T B) - \frac{1}{2} X_t^T BU_{k'} U_{k'}^T B \right\|_2
\]
\[
= \sqrt{\lambda_{\max}(I - X_t^T BU_{k'} U_{k'}^T B)} \leq \sqrt{1 - \sigma_{\min}^2(X_t^T BU_{k'})}
\]
\[
\leq \text{dist}_b(X_t, U_{k'}).
\]

Let \( \text{dist}_b(X_t, U) = \max\{\text{dist}_b(X_t, U_{k'}), \text{dist}_b(X_t, U_{k''})\} \), and assume that \( 0 < \alpha_t < \frac{1 - \text{dist}_b^2(X_t, U)}{8\lambda_1} \) and
\[
\left\| \xi_t(\nabla f_t) \right\|_{B,F} = \frac{\Delta_t}{4k^2 \alpha_t} \frac{1 - \text{dist}_b^2(X_t, U)}{1 + \psi(X_t, U)} \text{dist}_b(X_t, U).
\]

By Lemma 6.6 and noting that \( \Delta_t \leq 2\lambda_1 \), we get that \( \sigma(S_1) \geq \frac{1 - \text{dist}_b^2(X_t, U)}{2} \) and
\[
\sigma_{\min}(S_1 + 2\varsigma \alpha_t S_2) \geq \sigma_{\min}(S_1) - 2\alpha_t \sigma_{\max}(S_2) \geq \sigma_{\min}(S_1) - \frac{\alpha_t(1 + \alpha_t \lambda_1) \Delta_t}{2k^2} \geq 1 - \text{dist}_b^2(X_t, U) \]

We thus have that
\[
\frac{\|S_2\|_F}{\sigma_{\min}(S_1 + 2\varsigma \alpha_t S_2)} \leq 2 \frac{\left\| \xi_t(\nabla f_t) \right\|_{B,F}}{\text{dist}_b^2(X_t, U) + \alpha_t \left\| \nabla f_t \right\|_{B,F}}.
\]
(3)

By the Taylor expansion, we also have for some \( \varsigma' \in (0, 1) \) that
\[
\log \det \left( I + \alpha_t^2 \nabla f_t^T B \nabla f_t \right) = \alpha_t^2 \text{tr} \left( I + \varsigma' \alpha_t^2 \nabla f_t^T B \nabla f_t \right) \leq \alpha_t^2 \text{tr} \left( \nabla f_t^T B \nabla f_t \right) \leq \alpha_t^2 \left\| \nabla f_t \right\|_{B,F}^2 \leq 2\alpha_t^2 \left( \left\| \nabla f_t \right\|_{B,F}^2 + \left\| \xi_t(\nabla f_t) \right\|_{B,F}^2 \right).
\]
(4)

By Equations (1)-(4) and Lemma 6.2, we get that
\[
\text{dist}_b^2(X_{t+1}, U_{k''}) \leq \text{dist}_b^2(X_t, U_{k'}) - 2\alpha_t \text{dist}_f(X_t, U_{k'}) + 32k^2 \alpha_t^2 \sigma_{\min}^2(U_{k''})
\]
\[
+ 2 \left\| \xi_t(\nabla f_t) \right\|_{B,F} \frac{\text{dist}_b^2(X_t, U) + \alpha_t \left\| \nabla f_t \right\|_{B,F}}{1 - \text{dist}_b^2(X_t, U)} + 2\alpha_t^2 \left( \left\| \nabla f_t \right\|_{B,F}^2 + \left\| \xi_t(\nabla f_t) \right\|_{B,F}^2 \right).
By Lemma 6.5, 
\[
\text{dist}_f (X_t, U_{k'}) \geq \Delta_{k'} \text{dist}_m^2 (X_t, U_{k'}).
\]
Further, by Lemma 6.4 and using inequality \( x \geq \frac{-\log(1-x)}{1 - \log(1-x)} \), we can write that 
\[
\text{dist}_m^2 (X_0, U_{k'}) \geq \text{dist}_m^2 (X_t, U_{k'}) \geq \frac{\text{dist}_m^2 (X_t, U_{k'})}{1 + \text{dist}_m^2 (X_t, U_{k'})} \geq \frac{\text{dist}_m^2 (X_t, U_{k'})}{1 + \psi (X_t, U)}
\]
and 
\[
\text{dist}_b (X_t, U_{k'}) \leq \text{dist}_b (X_t, U) \leq \psi^\frac{1}{2} (X_t, U) \leq \psi^\frac{1}{2} (X_0, U).
\]
Simple algebraic manipulations then yield that 
\[
\text{dist}_m^2 (X_{t+1}, U_{k'}) \leq \left( 1 - \frac{2\alpha_t \Delta_t}{1 + \psi (X_t, U)} \right) \text{dist}_m^2 (X_t, U_{k'}) + \frac{\alpha_t \Delta_t}{1 + \psi (X_t, U)} \psi (X_t, U)
\]
\[
+ 4\lambda_t^2 \alpha_t^2 \left( \frac{16k}{(1 - \text{dist}_b^2 (X_0))^2} \psi (X_t, U) + \frac{\| \nabla f_t \|_{B,F}^2}{\lambda_t^2} \right).
\]
Analogously, we also have that 
\[
\text{dist}_m^2 (X_{t+1}, U_{k'}) \leq \left( 1 - \frac{2\alpha_t \Delta_t}{1 + \psi (X_t, U)} \right) \text{dist}_m^2 (X_t, U_{k'}) + \frac{\alpha_t \Delta_t}{1 + \psi (X_t, U)} \psi (X_t, U)
\]
\[
+ 4\lambda_t^2 \alpha_t^2 \left( \frac{16k}{(1 - \text{dist}_b^2 (X_0))^2} \psi (X_t, U) + \frac{\| \nabla f_t \|_{B,F}^2}{\lambda_t^2} \right).
\]
If \( 0 < \alpha_t < \frac{1 + \psi (X_t, U)}{2\Delta_t} \), taking the maximum over \( U_{k'} \) and \( U_{k''} \) gives us 
\[
\psi (X_{t+1}, U) \leq \left( 1 - \frac{2\alpha_t \Delta_t}{1 + \psi (X_0, U)} \right) \psi (X_t, U) + \frac{\alpha_t \Delta_t}{1 + \psi (X_t, U)} \psi (X_t, U)
\]
\[
+ 4\lambda_t^2 \alpha_t^2 \left( \frac{16k}{(1 - \text{dist}_b^2 (X_0))^2} \psi (X_t, U) + \frac{\| \nabla f_t \|_{B,F}^2}{\lambda_t^2} \right)
\]
\[
\leq \left( 1 - \frac{\alpha_t \Delta_t}{1 + \psi (X_0, U)} \right) \psi (X_t, U) + 4\lambda_t^2 \alpha_t^2 \left( \frac{16k}{(1 - \text{dist}_b^2 (X_0))^2} \psi (X_t, U) + \frac{\| \nabla f_t \|_{B,F}^2}{\lambda_t^2} \right).
\]
Next, two different settings of step-sizes are considered.

- Consider \( \alpha_t = \frac{\mu}{\nu + t} \). By Lemma 6.6, we have \( \| \nabla f_t \|_{B,F}^2 < k\lambda_t^2 \) and then can write 
\[
\psi (X_{t+1}, U) \leq \left( 1 - \frac{\Delta_t}{1 + \psi (X_0, U)} \frac{\mu}{\nu + t} \right) \psi (X_t, U) + 4k \left( \frac{\mu \lambda_t}{\nu + t} \right)^2 \left( 1 + \frac{16\psi (X_0, U)}{1 - \text{dist}_b^2 (X_0, U)} \right).
\]
Let \( \mu = O \left( \frac{1}{\Delta_t} \right) \) such that \( a = \frac{\mu \lambda_t}{\nu + t} > 1 \) and \( \nu \) is sufficiently large. By Lemma 6.7, we get that 
\[
\psi (X_t, U) = O \left( \frac{\lambda_t^2}{\Delta_t^2} \right) \right) \text{ and thus } T = O \left( \frac{\lambda_t^2}{\Delta_t^2} \right) \text{ such that } \psi (X_T, U) < \epsilon. \text{ For } t < T, \text{ we can assume that } \psi (X_t, U) \geq \epsilon. \text{ Using inequality } \frac{1}{1+x} \leq \log(1+x) \text{ for } x > -1, \text{ we have that }
\[
\frac{\text{dist}_b^2 (X_t, U)}{1 - \text{dist}_b^2 (X_t, U)} \geq \psi (X_t, U) \geq \epsilon.
\]
Thus,

\[
\log \frac{\|\nabla f_t\|_{B,F}^2}{\|\nabla (\nabla f_t)\|^2} = \log \frac{k\lambda_t^2}{\left(\frac{\Delta_t}{4k^2} \frac{1-\text{dist}_t^2(X_t, U)}{1+\psi(X_t, U)} \text{dist}_t^b(X_t, U)\right)^2}
\]

\[
= O\left(\log \frac{k\lambda_t^2}{\left(\frac{\Delta_t}{4k^2} \frac{1-\text{dist}_t^2(X_t, U)}{1+\psi(X_t, U)} \epsilon\right)^2}\right)
\]

\[
= O\left(\log \frac{\lambda_t}{\Delta_t} + \psi(X_0, U) + \log \frac{1}{\epsilon}\right) = O\left(\log \frac{\lambda_t}{\Delta_t} + \log \frac{1}{\epsilon}\right),
\]

where we have used that

\[
\log (1 + \psi(X_0, U)) \leq \psi(X_0, U) < -2k \log \frac{\eta \sqrt{\kappa(B)}}{k + \sqrt{nk}} < +\infty
\]

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

\[
O\left(\text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B)} \log \frac{\|\nabla f_t\|_{B,F}^2}{\|\nabla (\nabla f_t)\|^2} \right) = O\left(\text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B)} \left(\log \frac{\lambda_t}{\Delta_t} + \log \frac{1}{\epsilon}\right)\right).
\]

Therefore, the total complexity is

\[
O\left(\left(\text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B)} \left(\log \frac{\lambda_t}{\Delta_t} + \log \frac{1}{\epsilon}\right)\right) \left(\frac{\lambda_t}{\Delta_t}\right)^2 \frac{1}{\epsilon}\right),
\]

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

\[
\psi(X_t, U) = \min_{U \in \mathcal{U}} \text{dist}_m^2(X_t, U),
\]

which holds for Category d) as well if \(\psi(X_0, U)\) is sufficiently close to \(\mathcal{U}\). Accordingly, by Lemma 6.6, we get that

\[
\|\nabla f(X_t)\|_{B,F}^2 \leq 4k\lambda_t^2 \psi(X_t, U).
\]

Plugging into Equation (5), we arrive at

\[
\psi(X_{t+1}, U) \leq 1 - \frac{\alpha \Delta_t}{1 + \psi(X_0, U)} \psi(X_t, U) + 16k\lambda_t^2 \left(1 + \frac{1 - \text{dist}_t^2(X_0, U)}{2}\right)^2 \psi(X_t, U).
\]

If \(0 < \alpha < \frac{\Delta_t}{32k\lambda_t^2(1 + \psi(X_0, U))(1 + \frac{1 - \text{dist}_t^2(X_0, U)}{2})^{-2}}\), one can write that

\[
\psi(X_{T}, U) \leq 1 - \frac{\alpha \Delta_t}{2(1 + \psi(X_0, U))} \psi(X_{T-1}, U) \leq \cdots \leq \left(1 - \frac{\alpha \Delta_t}{2(1 + \psi(X_0, U))}\right)^T \psi(X_0, U).
\]

Setting \(1 - \frac{\alpha \Delta_t}{2(1 + \psi(X_0, U))}\) \(T\) \(\psi(X_0, U) = \epsilon\) yields that

\[
T = O\left(\frac{1}{\alpha \Delta_t} \log \frac{\psi(X_0, U)}{\epsilon}\right) = O\left(\frac{1 + \psi(X_0, U)}{\alpha \Delta_t} \log \frac{\psi(X_0, U)}{\epsilon}\right)
\]

\[
= O\left(\frac{\lambda_t}{\Delta_t}\right)^2 \log \frac{\psi(X_0, U)}{\epsilon}.
\]
For the subproblem, we now have that
\[
\log \frac{\| \nabla f_t \|^2_{B,F}}{\| \xi_t (\nabla f_t) \|^2_{B,F}} = O \left( \log \frac{2k \lambda_1^2 \psi(X_t, U)}{\left( \frac{1 - \text{dist}_2^2(X_t, U)}{1 + \psi(X_t, U)} \right)^2 \text{dist}_b(X_t, U)} \right) = O \left( \log \frac{\lambda_1}{\Delta_1} + \psi(X_0, U) \right)
\]
Therefore, the total complexity is
\[
O \left( \left( \text{nnz}(A) + \text{nnz}(B) \sqrt{\kappa(B) \log \frac{\lambda_1}{\Delta}} \right) \left( \frac{\lambda_1}{\Delta} \right)^2 \log \frac{1}{\epsilon} \right),
\]
which completes the proof of the second statement.

\[\square\]

**Proof of Lemma 6.2**

Let \( j \geq k \) and denote \( \nabla f_t \triangleq \nabla f(X_t) \) and \(-2 \log \det \left( (\bar{X}_{t+1})^\top B U_j \right) \triangleq - \log \det(S) \). Note that
\[
S \succ X_t^\top B U_j U_j^\top B X_t + 2\alpha_{t+1} \text{sym} (X_t^\top B U_j U_j^\top B \nabla f_t) \triangleq H_1 + H_2.
\]
Hence, we have that \(- \log \det(S) \leq - \log \det(H_1 + H_2) \). By Taylor expansion, we can write for certain \( \varsigma \in (0, 1) \) that
\[
- \log \det(S) \leq - \log \det(H_1) - \text{tr} (H_1^{-1} H_2) + \frac{1}{2} \text{tr} \left( \left( H_1 + \varsigma H_2 \right)^{-1} H_2 \right)^2,
\]
where \(- \log \det(H_1) = \text{dist}^2_m(X_t, U_j) \). Noting that \( X_t^\top B U_j = P_j \Sigma_j Q_j^\top \) (subscripts \( t \) on the right-hand side are omitted for brevity), we can write that
\[
\text{tr} \left( H_1^{-1} H_2 \right) = 2\alpha_t \text{tr} \left( (P_j^\top P_j)^{-1} P_j \Sigma_j Q_j^\top A_j Q_j \Sigma_j P_j^\top \right) - \text{tr}(X_t^\top A X_t).
\]
On the other hand,
\[
\text{tr} \left( (H_1 + \varsigma H_2)^{-1} H_2 \right)^2 \leq \left\| (H_1 + \varsigma H_2)^{-1} H_2 \right\|_F \leq \left\| (H_1 + \varsigma H_2)^{-1} \right\|_2 \left\| H_2 \right\|_F.
\]
where we need to lower bound \( \sigma_{\min}(H_1 + \varsigma H_2) \) and upper bound \( \|H_2\|_F \). To this end, notice that
\[
\sigma_{\min}(H_1) = \sigma_{\min}^2(X_t^\top B U_j) I \geq \prod_{i=1}^k \sigma_i^2(X_t^\top B U_j) = 1 - \text{dist}_b^2(X_t, U_j).
\]
Thus, we have that
\[ H_2 = 2\alpha t \text{sym}(\Omega U_j^TB \nabla f_t) = 2\alpha t \text{sym} \left( \Omega U_j^T B(B^{-1} - X_t X_t^T)AX_t \right) \]
\[ = 2\alpha t \Omega A_j \Omega^\top - 2\alpha t \text{sym}(\Omega \Omega^\top \Omega A_j \Omega^\top) - 2\alpha t \text{sym} \left( \Omega \Omega^\top X_t^T B U_j^\perp \Lambda_j^\perp (U_j^\perp)^\top BX_t \right) \]
\[ = 2\alpha t \text{sym} \left( (I - \Omega \Omega^\top) \Omega A_j \Omega^\top \right) - 2\alpha t \text{sym} \left( \Omega \Omega^\top X_t^T B U_j^\perp \Lambda_j^\perp (U_j^\perp)^\top BX_t \right). \]

Thus, we have that
\[
\|H_2\|_2 \leq 2\alpha t \left( \| (I - \Omega \Omega^\top) \Omega A_j \Omega^\top \|_2 + \| \Omega \Omega^\top X_t^T B U_j^\perp \Lambda_j^\perp (U_j^\perp)^\top BX_t \|_2 \right)
\]
\[
\leq 2\alpha t \left( \|I - \Omega \Omega^\top\|_2 \|A_j\|_2 + \|X_t^T BU_j^\perp \|_2^2 \|\Lambda_j^\perp\|_2 \right)
\]
\[
\leq 2\alpha t \lambda_1 \left( \|I - \Omega \Omega^\top\|_2 + \|X_t^T BU_j^\perp \|_2^2 \right) = 4\alpha t \lambda_1 \|I - \Omega \Omega^\top\|_2
\]
\[
= 4\alpha t \lambda_1 \|I - P_j \Sigma_j^2 P_j^\top\|_2 = 4\alpha t \lambda_1 \|I - \Sigma_j^2\|_2 \leq 4\alpha t \lambda_1 \text{dist}_b^2(X_t, U_j),
\]
where we have used for the first equality that
\[
\|X_t^T BU_j^\perp\|_2^2 = \lambda_{\max}(X_t^T BU_j^\perp (U_j^\perp)^\top BX_t) = \lambda_{\max}(I - X_t^T BU_j U_j^\top BX_t).
\]

Hence if \( 0 \lt \alpha t \lt \frac{1 - \text{dist}_b^2(X_t, U_j)}{2\lambda_1 \text{dist}_b^2(X_t, U_j)} \) then
\[
\sigma_{\min}(H_1 + \varsigma H_2) \geq \sigma_{\min}(H_1) - \sigma_{\max}(H_2) \geq (1 - \text{dist}_b^2(X_t, U_j)) - 4\alpha t \lambda_1 \text{dist}_b^2(X_t, U_j)
\]
\[
\geq \frac{1 - \text{dist}_b^2(X_t, U_j)}{2},
\]
and
\[
\|H_2\|_F \leq k^{\frac{1}{2}} \|H_2\|_2 \leq 4k^{\frac{1}{2}} \alpha t \lambda_1 \text{dist}_b^2(X_t, U_j).
\]

We thus get that
\[
\text{tr} \left( \left( (H_1 + \varsigma H_2)^{-1} H_2 \right)^2 \right) \leq 64k^{\frac{1}{2}} \alpha t^2 \left( \frac{\text{dist}_b^2(X_t, U_j)}{1 - \text{dist}_b^2(X_t, U_j)} \right)^2
\]
and consequently,
\[
-2 \log \text{Det} \left( X_t^{j+1} B U_j \right) \leq \text{dist}_m^2(X_t, U_j) - 2\alpha t (f(U_j Q_j) - f(X_t)) + 32k^{\frac{1}{2}} \alpha t^2 \left( \frac{\text{dist}_b^2(X_t, U_j)}{1 - \text{dist}_b^2(X_t, U_j)} \right)^2.
\]

The case that \( j \leq k \) is similar and thus omitted.

\[ \square \]

**Proof of Lemma 6.3**

\( l_t(X) \) reaches its minimum at
\[ l_t(X^*_t) = \frac{1}{2} \text{tr} \left( (X_t^*)^\top BX_t^* \right) - \text{tr}((X_t^*)^\top AX_t) = \frac{1}{2} \text{tr} \left( (X_t^*)^\top BX_t^* \right) - \text{tr}((X_t^*)^\top BB^{-1}AX_t) \]
\[ = -\frac{1}{2} \text{tr} \left( (X_t^*)^\top BX_t^* \right). \]

Thus, we have that
\[
\epsilon_t(X) = l_t(X) - l_t(X^*_t) = \frac{1}{2} \text{tr}(X^\top BX) - \text{tr}(X^\top AX_t) + \frac{1}{2} \text{tr}((X_t^*)^\top BX_t^*)
\]
\[ = \frac{1}{2} \text{tr}(X^\top BX) - \text{tr}(X^\top BB^{-1}AX_t) + \frac{1}{2} \text{tr}((X_t^*)^\top BX_t^*) \]
\[ = \frac{1}{2} \text{tr}(X^\top BX) - 2\text{tr}(X^\top BX t^*) + \text{tr}((X_t^*)^\top BX_t^*)) \]
\[ = \frac{1}{2} \text{tr}((X - X^*_t)^\top B(X - X^*_t)) = \frac{1}{2} \|\xi_t(X)\|_{B,F}^2. \]
In particular, 
\[
\hat{\xi}_t(X_t^{(0)}) = X_t^{(0)} - B^{-1}AX_t = X_t(X_t^\top BX_t)^{-1}X_t^\top AX_t - B^{-1}AX_t
\]
\[
= X_tX_t^\top AX_t - B^{-1}AX_t = -\tilde{v}_t(X_t).
\]

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 \(\ell_t(X) = \lambda_{\min}(B)\)-strongly convex and \(\lambda_{\max}(B)\)-smooth, where \(\lambda_{\max}(B)\) and \(\lambda_{\min}(B)\) represent the largest and smallest eigenvalue of \(B\), respectively. \(\square\)

**Proof of Lemma 6.4**

Let \(x = \text{dist}_b^2(X, Y)\). We then have that \(\text{dist}_b^2(X, Y) = x \leq -\log(1 - x) = \text{dist}_m^2(X, Y)\). We next prove by induction that \(\text{dist}_b(X, Y) \leq \text{dist}_c(X, Y)\). Let \(r = \min\{k, l\}\) and \(\theta_i\) be the \(i\)-th principal angle between \(X\) and \(Y\), \(i = 1, \ldots, r\). That is, \(\cos \theta_i = \sigma_i(X^\top BY)\), where \(\sigma_i(\cdot)\) represents the \(i\)-th largest singular value of a matrix. First, we have for \(r = 1\) that
\[
\text{dist}_b^2(X, Y) = 1 - \prod_{i=1}^r \cos^2 \theta_i = r - \sum_{i=1}^r \cos^2 \theta_i = \text{dist}_c^2(X, Y).
\]
Assuming that it holds for \(r\), one then has for \(r + 1\) that
\[
\text{dist}_c^2 = r + 1 - \sum_{i=1}^{r+1} \cos^2 \theta_i = r - \sum_{i=1}^r \cos^2 \theta_i + 1 - \cos^2 \theta_{r+1}
\]
\[
\geq 1 - \prod_{i=1}^r \cos^2 \theta_i + 1 - \cos^2 \theta_{r+1} - (1 - \prod_{i=1}^r \cos^2 \theta_i) + 1 - \prod_{i=1}^{r+1} \cos^2 \theta_i
\]
\[
= (1 - \cos^2 \theta_{r+1})(1 - \prod_{i=1}^r \cos^2 \theta_i) + 1 - \prod_{i=1}^r \cos^2 \theta_i \geq 1 - \prod_{i=1}^r \cos^2 \theta_i = \text{dist}_b^2,
\]

which completes the proof. The last inequality can be shown by the generalized mean inequality as follows:
\[
\sum_{i=1}^r \cos^2 \theta_i = r\left(\frac{\sum_{i=1}^r \cos^2 \theta_i}{r}\right)^2 \geq r\left(\prod_{i=1}^r \cos \theta_i\right)^2 \geq r\left(\prod_{i=1}^r \cos \theta_i\right) = r \text{dist}_b^2.
\]

It then holds that \(\text{dist}_c^2 \leq r - r(\prod_{i=1}^r \cos \theta_i)^2 = r(1 - \prod_{i=1}^r \cos^2 \theta_i) = r \text{dist}_b^2\). \(\square\)

**Proof of Lemma 6.5**

Suppose that \(j \leq k\), \(A_j = \text{diag}(\lambda_1, \ldots, \lambda_j)\) and \(\Lambda_j^\perp = \text{diag}(\lambda_{j+1}, \ldots, \lambda_n)\). We have that
\[
\text{dist}(X, U_j) = f(U_j) - f(XP_j) = \text{tr}(A_j) - \text{tr}(P_j^\top X^\top AXP_j)
\]
\[
= \text{tr}(A_j) - \text{tr}(P_j^\top X^\top BU_jA_j^\perp U_j^\top B X P_j) - \text{tr}(P_j^\top X^\top BU_j^\perp A_j^\perp(U_j^\perp)^\top B X P_j)
\]
\[
= \text{tr}(A_j) - \text{tr}(\Sigma_j Q_j^\perp A_j Q_j^\perp) - \text{tr}(P_j^\top X^\top BU_j^\perp A_j^\perp(U_j^\perp)^\top B X P_j)
\]
\[
= \text{tr}(A_j Q_j(I - \Sigma_j^2 Q_j^\top)) - \text{tr}(P_j^\top X^\top BU_j^\perp A_j^\perp(U_j^\perp)^\top B X P_j)
\]
\[
\geq \lambda_j \text{tr}(Q_j(I - \Sigma_j^2 Q_j^\top)) - \lambda_{j+1} \text{tr}(P_j^\top X^\top BU_j^\perp(U_j^\perp)^\top B X P_j)
\]
\[
= (\lambda_j - \lambda_{j+1}) \text{tr}(Q_j(I - \Sigma_j^2 Q_j^\top)) = \Delta_j (j - \|X^\top BU_j\|_F) = \Delta_j \text{dist}_b^2(X, U_j)
\]
\[
\geq \Delta_j \text{dist}_b^2(X, U_j),
\]
where the last inequality is by Lemma 6.4. The case that \(j \geq k\) is similar and thus omitted. \(\square\)
Proof of Lemma 6.6

Note that $X$’s orthogonal complement $X_\perp \in \text{gSt}_B(n, n-k)$ and $X_\perp^T BX = 0$. Thus,

$$\left\| B^{1/2} \tilde{\nabla} f(X) \right\|_2 = \left\| (I - B^{1/2}XX^T B^{1/2}) B^{-1/2} AB^{-1/2} B^{1/2} X \right\|_2$$

$$= \left\| B^{1/2} X_\perp B^{1/2} B^{-1/2} A B^{-1/2} B^{1/2} X \right\|_2 \leq \left\| B^{-1/2} A B^{-1/2} \right\|_2 = \lambda_1.$$

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

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

$$B^{-1/2} A B^{-1/2} = B^{1/2} (V \Lambda V^T + V \Lambda_\perp V_\perp^T) B^{1/2}.$$

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

$$\left\| \tilde{\nabla} f(X) \right\|_{B,F}^2 = \left\| X_\perp^T B (V \Lambda V^T + V \Lambda_\perp V_\perp^T) BX \right\|_F^2$$

$$\leq (\|X_\perp^T BV\|_F \|\Lambda\|_2 + \|\Lambda_\perp\|_2 \|V_\perp^T BX\|_F)^2$$

$$\leq \lambda_1^2 \left( (k - \|X^T BV\|_F^2)^{1/2} + (k - \|V^T BX\|_F^2)^{1/2} \right)^2$$

$$= 4 \lambda_1^2 \text{dist}_m^2(X, V) \leq 4k \lambda_1^2 \text{dist}_m^2(X, V) \leq 4k \lambda_1^2 \text{dist}_m^2(X, V).$$

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

\[ \square \]

Proof of Lemma 6.9

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

$$\text{dist}_m^2(X_0, U_j) \leq \text{dist}_m^2(X_0, U) = -2 \sum_{i=1}^k \log \sigma_i(X_0^T BU) \leq -2k \log \sigma_{\min}(X_0^T BU),$$

and

$$\sigma_{\min}(X_0^T BU) = \sigma_{\min} \left( (W^T BW)^{-\frac{1}{2}} W^T BU \right) \geq \sigma_{\min} \left( (W^T BW)^{-\frac{1}{2}} \right) \sigma_{\min}(W^T BU)$$

$$= \frac{\sigma_{\min}(W^T BU)}{\sigma_{\max}(B^{\frac{1}{2}} W)} \geq \frac{\sigma_{\min}(W^T BU)}{\sigma_{\max}(B^{\frac{1}{2}}) \|W\|_2},$$

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

$$\sigma_{\min}(W^T BU) \geq \sigma_{\min}(W^T \tilde{U}) \sigma_{\min}(\tilde{U}^T BU) = \sigma_{\min}(W^T \tilde{U}) \sigma_{\min}(BU) \geq \sigma_{\min}(W^T \tilde{U}) \sigma_{\min}(B^{\frac{1}{2}}),$$

where the last inequality is because that

$$\sigma_{\min}^2(BU) = \lambda_{\min}(U^T B^{\frac{1}{2}} U) = \min_{\|x\|_2=1} x^T U^T B^{\frac{1}{2}} B^{\frac{1}{2}} U x = \lambda_{\min}(B) \min_{\|x\|_2=1} x^T x = \lambda_{\min}(B) = \sigma_{\min}(B).$$
We thus get that
\[
\sigma_{\min}(X_0^TBU) \geq \frac{\sqrt{\kappa(B)}}{n^2 + k^2} \sigma_{\min}(W^T\hat{U}).
\]

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

References


