Supplementary Material

Martingale

A filtration, defined on a measurable probability space, is an increasing sequence of sub-sigma algebras $\{\mathcal{F}_t\}$ for $t \geq 0$, meaning that $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s \leq t$. In our context, $\mathcal{F}_t$ encodes the set of all the random variables seen thus far (i.e., from 0 to t). For brevity, $\mathbb{E}[\cdot|\mathbf{X}_t] \triangleq \mathbb{E}[\cdot|\mathcal{F}_t]$. Let $H = \{H_t\}$ and $\mathcal{F} = \{\mathcal{F}_t\}$ be a stochastic process and a filtration on the same probability space, respectively. H then is called a martingale (super-martingale) with respect to $\mathcal{F}$ if for each t, H_t is $\mathcal{F}_t$-measurable, $\mathbb{E}[|H_t|] < \infty$, and $\mathbb{E}[H_{t+1}|\mathcal{F}_t] = H_t$ ($\mathbb{E}[H_{t+1}|\mathcal{F}_t] \leq H_t$). Given a random variable $X \geq 0$ and a constant $a > 0$, we have the Markov inequality $P(X \geq a) \leq \mathbb{E}[X]/a$. Let $X_0, X_1, \cdots, X_T$ be a martingale or supermartingale such that $|X_t - X_{t-1}| \leq d_t$ (i.e., bounded difference) where d_t is a deterministic function of t. We then have the Azuma-Hoeffding (Mitzenmacher and Upfal, 2005) inequality $P(X_t - X_0 \geq a) \leq \exp\{-a^2/(2\sum_{s=1}^{t} d_t^2)\}$ for all $t \geq 0$ and any $a > 0$.

von Neumann's Trace Inequality

For any two real and symmetric $n \times n$ matrices $\mathbf{B}$ and $\mathbf{C}$ with eigenvalues indexed in descending order, it holds that

$$\max\left\{\sum_{i=1}^{n}\lambda_{n-i+1}(\mathbf{B})\lambda_1(\mathbf{C}), \sum_{i=1}^{n}\lambda_i(\mathbf{B})\lambda_{n-i+1}(\mathbf{C})\right\} \leq \mathrm{tr}(\mathbf{BC}) \leq \sum_{i=1}^{n}\lambda_i(\mathbf{B})\lambda_i(\mathbf{C}).$$

Proof. See (Lewis, 1996) for the right inequality. The proof of the left one is completed by replacing $\mathbf{B}$ with $-\mathbf{B}$ or replacing $\mathbf{C}$ with $-\mathbf{C}$ in the right one. □

Lemma 5.1. Let $\tilde{\beta} = \max_i \left\|\tilde{\mathbf{A}}_i - \mathbf{A}\right\|_2$. Then

$$\begin{aligned} \|\mathbf{W}_t\|_2 \leq \mu &\triangleq 4\tilde{\beta}, \\ \|\mathbf{W}_t\|_F^2 \leq \nu_t^2 &\triangleq 24\tilde{\beta}^2\left(\Theta(\mathbf{X}_t, \mathbf{U}) + \Theta(\tilde{\mathbf{X}}_{s-1}, \mathbf{U})\right). \end{aligned}$$

Proof. For brevity, let's omit subscripts with $\mathbf{X}$, $\tilde{\mathbf{X}}$, $\mathbf{W}$ and $\mathbf{Q}$ herein. $\mathbf{W}$ then can be written as

$$\begin{aligned} \mathbf{W} &= \mathbf{X}_\perp\mathbf{X}_\perp^\top(\mathbf{A}_{t+1} - \mathbf{A})\left(\mathbf{X} - \tilde{\mathbf{X}}\mathbf{Q}\right) + \mathbf{X}_\perp\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q} \\ &\quad -\mathbf{X}\mathrm{skew}\left(\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\tilde{\mathbf{X}}_\perp^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right). \end{aligned}$$

We then have

$$\begin{aligned} \|\mathbf{W}\|_2 &\leq \left\|\mathbf{X}_\perp\mathbf{X}_\perp^\top(\mathbf{A}_{t+1} - \mathbf{A})(\mathbf{X} - \tilde{\mathbf{X}}\mathbf{Q})\right\|_2 + \left\|\mathbf{X}_\perp\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right\|_2 \\ &\quad + \left\|\mathbf{X}\mathrm{skew}\left(\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\tilde{\mathbf{X}}_\perp^\top(\mathbf{A}_{t+1} - \mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right)\right\|_2 \\ &\leq \tilde{\beta}\left(\|\mathbf{X}\|_2 + \left\|\tilde{\mathbf{X}}\mathbf{Q}\right\|_2\right) + 2\tilde{\beta} = 4\tilde{\beta}, \end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{W}\|_F^2 &\leq \left(\left\|\mathbf{X}_\perp\mathbf{X}_\perp^\top(\mathbf{A}_{t+1}-\mathbf{A})(\mathbf{X}-\tilde{\mathbf{X}}\mathbf{Q})\right\|_F + \left\|\mathbf{X}_\perp\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1}-\mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right\|_F \right. \\
&\quad \left. + \left\|\mathbf{X}\mathrm{skew}\left(\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\tilde{\mathbf{X}}_\perp^\top(\mathbf{A}_{t+1}-\mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right)\right\|_F \right)^2 \\
&\leq \left(\left\|\mathbf{X}_\perp\mathbf{X}_\perp^\top(\mathbf{A}_{t+1}-\mathbf{A})\right\|_2 \left\|\mathbf{X}-\tilde{\mathbf{X}}\mathbf{Q}\right\|_F + \|\mathbf{X}_\perp\|_2 \left\|\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\right\|_F \left\|\tilde{\mathbf{X}}^\top(\mathbf{A}_{t+1}-\mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right\|_2 \right. \\
&\quad \left. + \|\mathbf{X}\|_2 \left\|\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\right\|_F \left\|\tilde{\mathbf{X}}_\perp^\top(\mathbf{A}_{t+1}-\mathbf{A})\tilde{\mathbf{X}}\mathbf{Q}\right\|_2 \right)^2 \\
&\leq \tilde{\beta}^2\left(\left\|\mathbf{X}-\tilde{\mathbf{X}}\mathbf{Q}\right\|_F + \left\|\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\right\|_F + \left\|\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\right\|_F\right)^2 \\
&\leq 3\tilde{\beta}^2\left(\left\|\mathbf{X}-\tilde{\mathbf{X}}\mathbf{Q}\right\|_F^2 + \left\|\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\right\|_F^2 + \left\|\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\right\|_F^2\right).
\end{aligned}$$

From the proof of Lemma 11 in (Shamir, 2016), we have $\left\|\mathbf{X}-\tilde{\mathbf{X}}\mathbf{Q}\right\|_F^2 \leq 4\left(\Theta(\mathbf{X},\mathbf{U})+\Theta(\tilde{\mathbf{X}},\mathbf{U})\right)$. Further noting that $\mathbf{I}=\mathbf{U}\mathbf{U}^\top+\mathbf{U}_\perp\mathbf{U}_\perp^\top=\mathbf{X}\mathbf{X}^\top+\mathbf{X}_\perp\mathbf{X}_\perp^\top$, we have

$$\begin{aligned}
\left\|\mathbf{X}_\perp^\top\tilde{\mathbf{X}}\right\|_F^2 &= \left\|\mathbf{X}_\perp^\top(\mathbf{U}\mathbf{U}^\top+\mathbf{U}_\perp\mathbf{U}_\perp^\top)\tilde{\mathbf{X}}\right\|_F^2 \leq \left(\left\|\mathbf{X}_\perp^\top\mathbf{U}\mathbf{U}^\top\tilde{\mathbf{X}}\right\|_F + \left\|\mathbf{X}_\perp^\top\mathbf{U}_\perp\mathbf{U}_\perp^\top\tilde{\mathbf{X}}\right\|_F\right)^2 \\
&\leq \left(\|\mathbf{X}_\perp^\top\mathbf{U}\|_F\left\|\mathbf{U}^\top\tilde{\mathbf{X}}\right\|_2 + \left\|\mathbf{X}_\perp^\top\mathbf{U}_\perp\|_2\|\mathbf{U}_\perp^\top\tilde{\mathbf{X}}\right\|_F\right)^2 \leq \left(\|\mathbf{X}_\perp^\top\mathbf{U}\|_F + \left\|\mathbf{U}_\perp^\top\tilde{\mathbf{X}}\right\|_F\right)^2 \\
&\leq 2\left(\|\mathbf{X}_\perp^\top\mathbf{U}\|_F^2 + \left\|\mathbf{U}_\perp^\top\tilde{\mathbf{X}}\right\|_F^2\right) \\
&= 2\left(k-\|\mathbf{U}^\top\mathbf{X}\|_F^2 + k - \left\|\mathbf{U}^\top\tilde{\mathbf{X}}\right\|_F^2\right) = 2\left(\Theta(\mathbf{X},\mathbf{U})+\Theta(\tilde{\mathbf{X}},\mathbf{U})\right),
\end{aligned}$$

and similarly $\left\|\mathbf{X}^\top\tilde{\mathbf{X}}_\perp\right\|_F^2 \leq 2\left(\Theta(\mathbf{X},\mathbf{U})+\Theta(\tilde{\mathbf{X}},\mathbf{U})\right)$. Therefore, we arrive at

$$\|\mathbf{W}\|_F^2 \leq 24\tilde{\beta}^2\left(\Theta(\mathbf{X},\mathbf{U})+\Theta(\tilde{\mathbf{X}},\mathbf{U})\right).$$

□

Lemma 5.2. If $\mathbf{C}_1 \succcurlyeq \mathbf{0}$, $\mathbf{C}_2 \succcurlyeq \mathbf{0}$ and $\mathbf{D}_2 \succcurlyeq \mathbf{D}_1 \succ \mathbf{0}$, then $\mathrm{tr}(\mathbf{C}_1\mathbf{D}_1^{-1}) \geq \mathrm{tr}((\mathbf{C}_1-\mathbf{C}_2)\mathbf{D}_2^{-1})$.

Proof.

$$\begin{aligned}
\mathbf{C}_1\mathbf{D}_1^{-1} &= \mathbf{C}_1\left(\mathbf{D}_2-(\mathbf{D}_2-\mathbf{D}_1)\right)^{-1} \\
&= \mathbf{C}_1\mathbf{D}_2^{-1/2}\left(\mathbf{I}-\mathbf{D}_2^{-1/2}(\mathbf{D}_2-\mathbf{D}_1)\mathbf{D}_2^{-1/2}\right)^{-1}\mathbf{D}_2^{-1/2}.
\end{aligned}$$

By Lemmas 2 and 3 in (Shamir, 2016), we have

$$\begin{aligned}
\mathrm{tr}\left(\mathbf{C}_1\mathbf{D}_1^{-1}\right) &= \mathrm{tr}\left(\mathbf{C}_1\mathbf{D}_2^{-1/2}(\mathbf{I}-\mathbf{D}_2^{-1/2}(\mathbf{D}_2-\mathbf{D}_1)\mathbf{D}_2^{-1/2})^{-1}\mathbf{D}_2^{-1/2}\right) \\
&= \mathrm{tr}\left(\mathbf{D}_2^{-1/2}\mathbf{C}_1\mathbf{D}_2^{-1/2}\left(\mathbf{I}-\mathbf{D}_2^{-1/2}(\mathbf{D}_2-\mathbf{D}_1)\mathbf{D}_2^{-1/2}\right)^{-1}\right) \\
&\geq \mathrm{tr}\left(\mathbf{D}_2^{-1/2}\mathbf{C}_1\mathbf{D}_2^{-1/2}\left(\mathbf{I}+\mathbf{D}_2^{-1/2}(\mathbf{D}_2-\mathbf{D}_1)\mathbf{D}_2^{-1/2}\right)\right) \\
&\geq \mathrm{tr}\left(\mathbf{D}_2^{-1/2}\mathbf{C}_1\mathbf{D}_2^{-1/2}\right) = \mathrm{tr}\left(\mathbf{C}_1\mathbf{D}_2^{-1}\right) \geq \mathrm{tr}\left((\mathbf{C}_1-\mathbf{C}_2)\mathbf{D}_2^{-1}\right).
\end{aligned}$$

□

Lemma 5.3. For $i = 1, 2$ and any $\varsigma \in [0, 1]$,

$$\begin{aligned} \|b_i(\mathbf{W}_t)\|_F &\leq 2(1+\alpha\beta)\alpha\nu_t, \\ \|a_1(\mathbf{X}_t)\|_2 + \|b_1(\mathbf{W}_t)\|_2 &\leq (1+\alpha\beta)\left(1+\alpha\left(\beta+2\mu\right)\right), \\ a_2(\mathbf{X}_t) + \varsigma b_2(\mathbf{W}_t) &\succcurlyeq (1-2(1+\alpha\beta)\alpha\mu)\,\mathbf{I}. \end{aligned}$$

Proof. We omit subscripts for brevity herein. Since $\|\mathbf{Y}\|_2 = \left\|\left(\mathbf{I}+\alpha\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\right)\mathbf{X}\right\|_2 \leq 1+\alpha\beta$, we have

$$\begin{aligned} \|b_1(\mathbf{W})\|_F &= 2\alpha\left\|\mathrm{sym}(\mathbf{Y}^\top\mathbf{U}\mathbf{U}^\top\mathbf{W})\right\|_F \leq 2\alpha\left\|\mathbf{Y}^\top\mathbf{U}\mathbf{U}^\top\mathbf{W}\right\|_F \leq 2\alpha(1+\alpha\beta)\nu_t, \\ \|b_2(\mathbf{W})\|_F &= 2\alpha\left\|\mathrm{sym}(\mathbf{Y}^\top\mathbf{W})\right\|_F \leq 2\alpha\left\|\mathbf{Y}^\top\mathbf{W}\right\|_F \leq 2\alpha(1+\alpha\beta)\nu_t. \end{aligned}$$

In addition, we can write

$$\|a_1(\mathbf{X})\|_2 + \|b_1(\mathbf{W})\|_2 \leq \left\|\mathbf{Y}^\top\mathbf{U}\right\|_2\left(\left\|\mathbf{Y}^\top\mathbf{U}\right\|_2 + 2\alpha\left\|\mathbf{U}^\top\mathbf{W}\right\|_2\right) \leq (1+\alpha\beta)\left(1+\alpha\left(\beta+2\mu\right)\right).$$

Further we note that $\mathbf{Y}^\top\mathbf{Y} = \mathbf{I} + \alpha\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp^\top\mathbf{X}_\perp\mathbf{A}\mathbf{X}$.

$$a_2(\mathbf{X}) + \varsigma b_1(\mathbf{W}) \succcurlyeq (1 - \|b_1(\mathbf{W})\|_2)\,\mathbf{I} \succcurlyeq (1 - 2\alpha(1+\alpha\beta)\mu)\,\mathbf{I}.$$

□

Lemma 5.4. $\mathrm{tr}\left(\mathbf{X}_t^\top\mathbf{U}\mathbf{U}^\top\left(\mathbf{I}-\mathbf{X}_t\mathbf{X}_t^\top\right)\mathbf{A}\mathbf{X}\right) \geq \tau\left(\left\|\mathbf{X}_t^\top\mathbf{U}\right\|_F^2 - \left\|\mathbf{X}_t^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_t\right\|_F^2\right).$

Proof. Since $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{U}_\perp^\top$, then

$$\begin{aligned} & \mathrm{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right) \\ = \ & \mathrm{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{X}\right) + \mathrm{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{U}_\perp^\top\mathbf{X}\right) \\ = \ & \mathrm{tr}\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\mathbf{\Sigma}\right) + \mathrm{tr}\left(\mathbf{U}_\perp^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp\mathbf{\Sigma}_\perp\right). \end{aligned}$$

By von Neumann's trace inequality, we have

$$\begin{aligned} \mathrm{tr}\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\mathbf{\Sigma}\right) &\geq \sum_{i=1}^{k}\lambda_i\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\right)\lambda_{k-i+1}\left(\mathbf{\Sigma}\right), \\ \mathrm{tr}\left(\mathbf{U}_\perp^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp\mathbf{\Sigma}_\perp\right) &\geq \sum_{i=1}^{n-k}\lambda_i\left(\mathbf{U}_\perp^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp\right)\lambda_{n-k-i+1}\left(\mathbf{\Sigma}_\perp\right). \end{aligned}$$

Note that both matrices above, i.e., $\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}$ and $\mathbf{U}_\perp^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp$, are symmetric and thus von Neumann's trace inequality can be applied. In fact,

$$\begin{aligned} \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U} &= \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\left(\mathbf{I}-\mathbf{X}\mathbf{X}^\top\right)\mathbf{U} \\ &= \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} - \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} \\ &= \mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} - \left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right)^2, \end{aligned}$$

which is symmetric. Furthermore, it is positive semi-definite, because

$$\left\|\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right\|_2 \leq \left(\|\mathbf{X}^\top\|_2\|\mathbf{U}\|_2\right)^2 = 1$$

and thus

$$\lambda_i\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\right) = \lambda_i\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right) - \lambda_i^2\left(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right) \geq 0.$$

Likewise,

$$\begin{aligned} \mathbf{U}_\perp^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp &= \mathbf{U}_\perp^\top\left(\mathbf{I}-\mathbf{X}_\perp\mathbf{X}_\perp^\top\right)\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp \\ &= -\mathbf{U}_\perp^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\left(\mathbf{I}-\mathbf{U}_\perp\mathbf{U}_\perp^\top\right)\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp \\ &= -\left(\mathbf{U}_\perp^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp - \left(\mathbf{U}_\perp^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}_\perp\right)^2\right), \end{aligned}$$

which is symmetric but negative semi-definite now, because

$$\left\|\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right\|_2 \leq \left(\|\mathbf{X}_\perp^\top\|_2 \|\mathbf{U}_\perp\|_2\right)^2 = 1$$

and thus

$$\lambda_i\left(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) = -\left(\lambda_i(\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp) - \lambda_i^2\left(\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right)\right) \leq 0.$$

We now can write

$$\begin{aligned}
& \operatorname{tr}\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\boldsymbol{\Sigma}\right) + \operatorname{tr}(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp \boldsymbol{\Sigma}_\perp) \\
\geq & \sum_{i=1}^{k} \lambda_i\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\right) \lambda_{k-i+1}\left(\boldsymbol{\Sigma}\right) + \sum_{i=1}^{n-k} \lambda_i\left(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) \lambda_{n-k-i+1}\left(\boldsymbol{\Sigma}_\perp\right) \\
\geq & \sum_{i=1}^{k} \lambda_i\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\right) \lambda_k\left(\mathbf{A}\right) + \sum_{i=1}^{n-k} \lambda_i\left(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) \lambda_{k+1}\left(\mathbf{A}\right),
\end{aligned}$$

where we find that

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_i\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\right) &= \sum_{i=1}^{k} \lambda_i\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right) - \sum_{i=1}^{k} \lambda_i^2\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right) \\
&= \operatorname{tr}\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right) - \operatorname{tr}\left(\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right)^2\right) \\
&= \left\|\mathbf{X}^\top \mathbf{U}\right\|_F^2 - \left\|\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right\|_F^2
\end{aligned}$$

and similarly

$$\sum_{i=1}^{k} \lambda_i\left(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) = -\left(\left\|\mathbf{X}_\perp^\top \mathbf{U}_\perp\right\|_F^2 - \left\|\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right\|_F^2\right).$$

Note that

$$\begin{aligned}
\left\|\mathbf{X}_\perp^\top \mathbf{U}_\perp\right\|_F^2 &= \operatorname{tr}\left(\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) = \operatorname{tr}\left(\mathbf{U}_\perp \mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top\right) = \operatorname{tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\right) \\
&= \operatorname{tr}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right) \\
&= n - 2k + \left\|\mathbf{X}^\top \mathbf{U}\right\|_F^2
\end{aligned}$$

and

$$\begin{aligned}
& \left\|\mathbf{U}_\perp^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right\|_F^2 \\
= & \operatorname{tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\right) \\
= & \operatorname{tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right)\right) \\
= & \operatorname{tr}\left(\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top\right)\left(\mathbf{I} - \mathbf{X}\mathbf{X}^\top\right)\left(\mathbf{I} + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right)\right) \\
= & \operatorname{tr}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top + \left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right)\mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right) \\
= & \operatorname{tr}\left(\mathbf{I} - \mathbf{U}\mathbf{U}^\top - \mathbf{X}\mathbf{X}^\top + \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top\right) \\
= & n - 2k + \left\|\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right\|_F^2.
\end{aligned}$$

Therefore, we arrive at

$$\begin{aligned}
& \operatorname{tr}\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\boldsymbol{\Sigma}\right) + \operatorname{tr}(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp \boldsymbol{\Sigma}_\perp) \\
\geq & \lambda_k \sum_{i=1}^{k} \lambda_i\left(\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}\right) + \lambda_{k+1} \sum_{i=1}^{n-k} \lambda_i\left(\mathbf{U}_\perp^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}_\perp \mathbf{X}_\perp^\top \mathbf{U}_\perp\right) \\
= & \left(\lambda_k - \lambda_{k+1}\right)\left(\left\|\mathbf{X}^\top \mathbf{U}\right\|_F^2 - \left\|\mathbf{U}^\top \mathbf{X}\mathbf{X}^\top \mathbf{U}\right\|_F^2\right) \\
= & \tau\left(\left\|\mathbf{X}^\top \mathbf{U}\right\|_F^2 - \left\|\mathbf{X}^\top \mathbf{U}\mathbf{U}^\top \mathbf{X}\right\|_F^2\right).
\end{aligned}$$

□

Lemma 5.5. If $\alpha^2\left(\beta^2+48k\tilde{\beta}^2\right)<1$, then

$$\begin{aligned}\operatorname{tr}\left(a_1\left(\mathbf{X}_t\right)a_2^{-1}\left(\mathbf{X}_t\right)\right) &\geq \left\|\mathbf{X}_t^\top\mathbf{U}\right\|_F^2+2\alpha\tau\left(\left\|\mathbf{X}_t^\top\mathbf{U}\right\|_F^2-\left\|\mathbf{U}^\top\mathbf{X}_t\mathbf{X}_t^\top\mathbf{U}\right\|_F^2\right)\\ &\quad -2\left(1+2\alpha\beta\right)\alpha^2\beta^2\Theta(\mathbf{X}_t,\mathbf{U})-k(1+2\alpha\beta)\alpha^2\nu_t^2.\end{aligned}$$

Proof. First, we have $a_1\left(\mathbf{X}\right)\succcurlyeq\mathbf{0}$ and $a_2\left(\mathbf{X}\right)\succ\mathbf{0}$. By Lemma 3 in (Shamir, 2016), we then get

$$\operatorname{tr}\left(a_1\left(\mathbf{X}\right)a_2^{-1}\left(\mathbf{X}\right)\right)\geq\operatorname{tr}\left(a_1\left(\mathbf{X}\right)\left(2\mathbf{I}-a_2\left(\mathbf{X}\right)\right)\right),$$

where if $\alpha^2\left(\beta^2+48k\tilde{\beta}^2\right)<1$ then

$$\begin{aligned}2\mathbf{I}-a_2\left(\mathbf{X}\right) &= \mathbf{I}-\alpha^2\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}-\alpha^2\nu_t^2\mathbf{I}\\ &\succcurlyeq \left(1-\alpha^2\left\|\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right\|_2-\alpha^2\nu_t^2\right)\mathbf{I}\succcurlyeq\left(1-\alpha^2\left(\beta^2+48k\tilde{\beta}^2\right)\right)\mathbf{I}\succ\mathbf{0}.\end{aligned}$$

Then by Lemma 2 in (Shamir, 2016) we have

$$\begin{aligned}\operatorname{tr}\left(a_1\left(\mathbf{X}\right)a_2^{-1}(\mathbf{X})\right) &\geq \operatorname{tr}\left(\left(a_1\left(\mathbf{X}\right)-\alpha^2\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)\left(2\mathbf{I}-a_2\left(\mathbf{X}\right)\right)\right)\\ &= \operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\right)-\alpha^2\operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)-\alpha^2\nu_t^2\operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\right),\end{aligned}$$

where $\tilde{a}_1\left(\mathbf{X}\right)=\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}+2\alpha\operatorname{sym}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)$. We now further lower bound the right hand side of the above inequality. By Lemma 5.4., we get

$$\begin{aligned}\operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\right) &= \operatorname{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}\right)+2\alpha\operatorname{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)\\ &\geq \left\|\mathbf{X}^\top\mathbf{U}\right\|_F^2+2\alpha\tau\left(\left\|\mathbf{X}^\top\mathbf{U}\right\|_F^2-\left\|\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}\right\|_F^2\right).\end{aligned}$$

On the other hand, by the Cauchy-Schwarz inequality, we can obtain

$$\begin{aligned}\operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\right) &= \operatorname{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}\right)+2\alpha\operatorname{tr}\left(\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)\\ &\leq \left\|\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\right\|_F\|\mathbf{X}\|_F+2\alpha\left\|\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\right\|_F\|\mathbf{X}\|_F\\ &\leq \left\|\mathbf{X}^\top\right\|_F\left\|\mathbf{U}\mathbf{U}^\top\right\|_2\|\mathbf{X}\|_F+2\alpha\left\|\mathbf{X}^\top\right\|_F\left\|\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\right\|_2\|\mathbf{X}\|_F\\ &\leq (1+2\alpha\beta)\|\mathbf{X}\|_F^2=(1+2\alpha\beta)\,k.\end{aligned}$$

In addition,

$$\begin{aligned}&\operatorname{tr}\left(\tilde{a}_1\left(\mathbf{X}\right)\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right)\\ \leq\ &\left\|\tilde{a}_1(\mathbf{X})\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\right\|_F\left\|\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right\|_F\leq\|\tilde{a}_1(\mathbf{X})\|_2\left\|\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\right\|_F^2\\ \leq\ &\left(\left\|\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}\right\|_2+2\alpha\left\|\mathbf{X}^\top\mathbf{U}\mathbf{U}^\top\mathbf{X}_\perp\mathbf{X}_\perp^\top\mathbf{A}\mathbf{X}\right\|_2\right)\left\|\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\right\|_F^2\\ \leq\ &(1+2\alpha\beta)\left\|\mathbf{X}^\top\mathbf{A}\mathbf{X}_\perp\right\|_F^2=(1+2\alpha)\left\|\mathbf{X}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top+\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{U}_\perp^\top)\mathbf{X}_\perp\right\|_F^2\\ \leq\ &(1+2\alpha\beta)\left(\left\|\mathbf{X}^\top\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top\mathbf{X}_\perp\right\|_F+\left\|\mathbf{X}^\top\mathbf{U}_\perp\mathbf{\Sigma}_\perp\mathbf{U}_\perp^\top\mathbf{X}_\perp\right\|_F\right)^2\\ \leq\ &(1+2\alpha\beta)\left(\left\|\mathbf{X}^\top\mathbf{U}\mathbf{\Sigma}\right\|_2\left\|\mathbf{U}^\top\mathbf{X}_\perp\right\|_F+\left\|\mathbf{X}^\top\mathbf{U}_\perp\right\|_F\left\|\mathbf{\Sigma}_\perp\mathbf{U}_\perp^\top\mathbf{X}_\perp\right\|_2\right)^2\\ \leq\ &(1+2\alpha\beta)\left(\|\mathbf{A}\|_2\left\|\mathbf{U}^\top\mathbf{X}_\perp\right\|_F+\left\|\mathbf{X}^\top\mathbf{U}_\perp\right\|_F\|\mathbf{A}\|_2\right)^2\\ \leq\ &\beta^2(1+2\alpha\beta)\left(\left\|\mathbf{U}^\top\mathbf{X}_\perp\right\|_F+\left\|\mathbf{X}^\top\mathbf{U}_\perp\right\|_F\right)^2=2\beta^2(1+2\alpha\beta)\,\Theta(\mathbf{X},\mathbf{U}).\end{aligned}$$

Therefore, we arrive at

$$\begin{aligned}\operatorname{tr}\left(a_1\left(\mathbf{X}\right)a_2^{-1}\left(\mathbf{X}\right)\right) &\geq \left\|\mathbf{X}^\top\mathbf{U}\right\|_F^2+2\alpha\tau\left(\left\|\mathbf{X}^\top\mathbf{U}\right\|_F^2-\left\|\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U}\right\|_F^2\right)\\ &\quad -2\left(1+2\alpha\beta\right)\alpha^2\beta^2\Theta(\mathbf{X},\mathbf{U})-k\left(1+2\alpha\beta\right)\alpha^2\nu_t^2.\end{aligned}$$

□

Lemma 5.6. For any $\iota \in (0,1)$, if α and m satisfy that $\alpha < 2\xi\tau/\gamma$, $\min\{2(1+\alpha\beta)\alpha\mu, \alpha^2(\beta^2+48k\tilde{\beta}^2)\} < 1$ and $\Theta_0 + km\rho + \theta\sqrt{2m\log(1/\iota)} < \delta$, then $\Theta_t < \delta$ holds for all $t = 1, 2, \cdots, m$ with probability at least $1-\iota$, where $\theta = \frac{4k(\beta+4\tilde{\beta})\alpha}{1-(\beta+4\tilde{\beta})\alpha} + k\rho$.

Proof. Consider the stochastic process $\{\Theta_0, \Theta_1, \ldots, \Theta_m\}$ and the filtration $\mathcal{F} = \{\mathcal{F}_t\}$ about random draws y_t. Since $\mathbb{E}[\Theta_{t+1}|\mathbf{X}_t] \leq (1-\alpha(2\xi\tau-\gamma\alpha))\Theta_t + \rho\tilde{\Theta}_{s-1}$, we have $\mathbb{E}[\Theta_{t+1}|X^{(t)}] \leq \Theta_t + k\rho$ if $\alpha < 2\xi\tau/\gamma$. Define $\Psi_t = \Theta_t - k\rho t$ for $t = 0, 1, 2, \cdots, m$. Note that$\{\Psi_t : t = 0, 1, 2, \cdots, m\}$ is a finite sequence of random variables and thus the natural continuation can be applied to get an infinite sequence such that $|\Psi_t| \leq \Theta_t + k\rho m \leq k(1+\rho m)$ for all t including $t > m$. In addition,

$$\mathbb{E}[\Psi_t|\mathbf{X}_{t-1}] = \mathbb{E}[\Theta_t|\mathbf{X}_{t-1}] - k\rho t \leq \Theta_{t-1} + k\rho - k\rho t = \Theta_{t-1} - k\rho(t-1) = \Psi_{t-1}.$$

$\{\Psi_t\}$ thus is a super-martingale. Next, we prove that it has bounded differences. Note that the intermediate update in the tangent space is $\mathbf{X}_t + \alpha_{t+1}\mathrm{Grad}f(\mathbf{X}_t) + \alpha_{t+1}\mathbf{W}_t$, for which we have

$$\|\mathrm{Grad}f(\mathbf{X}_t) + \mathbf{W}_t\|_2 \leq \|\mathrm{Grad}f(\mathbf{X}_t)\|_2 + \|\mathbf{W}_t\|_2 \leq \beta + 4\tilde{\beta}$$

using Lemma 5.1. Based on the proof of Lemma 9 in (Shamir, 2016), we have that

$$\left| \left\|\mathbf{X}_{t+1}^\top \mathbf{U}\right\|_F^2 - \left\|\mathbf{X}_t^\top \mathbf{U}\right\|_F^2 \right| \leq \frac{4k(\beta+4\tilde{\beta})\alpha}{1-(\beta+4\tilde{\beta})\alpha}.$$

Thus, we get the bounded difference

$$|\Psi_{t+1} - \Psi_t| \leq \left| \left\|\mathbf{X}_{t+1}^\top \mathbf{U}\right\|_F^2 - \left\|\mathbf{X}_t^\top \mathbf{U}\right\|_F^2 \right| + k\rho \leq \frac{4k(\beta+4\tilde{\beta})\alpha}{1-(\beta+4\tilde{\beta})\alpha} + k\rho \triangleq \theta.$$

We now are able to apply the Azuma-Hoeffding inequality, i.e., for any $0 \leq t \leq m$ and $a > 0$

$$P(\Psi_t - \Psi_0 \geq a) \leq \exp\{-\frac{a^2}{2\sum_{s=1}^t \theta^2}\} = \exp\{-\frac{a^2}{2t\theta^2}\} \leq \exp\{-\frac{a^2}{2m\theta^2}\} \triangleq \iota \in (0,1).$$

Solving $\iota = \exp\{-\frac{a^2}{2m\theta^2}\}$ with respect to a yields $a = \theta\sqrt{2m\log(1/\iota)}$. Therefore, we get that $\Psi_t - \Psi_0 < a$, i.e.,

$$\Theta_t \leq \Theta_0 + k\rho t + a \leq \Theta_0 + km\rho + \theta\sqrt{2m\log(1/\iota)} < \delta$$

for all $t = 1, 2, \cdots, m$, with probability at least $1-\iota$. □