A Proof of Main Results

In this section we provide the proofs of main results in Section 4. Note that in Theorem 4.2, Corollary 4.3 and Theorem 4.6, the only assumption on $\mathbf{U}^*$ is that it is an optimal solution. All quantities in these results that depend on $\mathbf{U}^*$, including $D(\mathbf{U}_0, \mathbf{U}^*)$, $\sigma_r(\mathbf{U}^*)$, $\|\mathbf{U}^*\|_2$ and $\mathcal{G}(\mathbf{U}^*)$, are all invariant to rotations in $\mathbb{R}^r$. Therefore throughout our proof, we can pick a specific $\mathbf{U}^*$ in $\Omega(\mathbf{U}_0)$, which can be easily obtained by setting $\mathbf{U}^* = \Pi_{\mathbf{U}_0}(\widehat{\mathbf{U}}^*)$, with $\widehat{\mathbf{U}}^*$ being an arbitrary optimal solution to the optimization problem (See Lemma B.3).

A.1 Proof of Proposition 3.1

Here we present the formal version of Proposition 3.1.

Lemma A.1 (Formal version of Proposition 3.1)**.** Suppose that $\mathbf{U}, \mathbf{V} \in \Omega(\mathbf{U}_0)$, $D(\mathbf{U}, \mathbf{U}^*) \leq C_{\text{convex}}$ and $D(\mathbf{V}, \mathbf{U}^*) \leq C_{\text{convex}}$, where $C_{\text{convex}} = \mu\sigma_r^7(\mathbf{U}^*)/[4200L\|\mathbf{U}^*\|_2^6 + 1000\|\mathbf{U}^*\|_2^4\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2]$, then

$$\mathcal{G}(\mathbf{U}) \geq \mathcal{G}(\mathbf{V}) + \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U} - \mathbf{V}\rangle + \frac{\bar{\mu}}{2}\|\mathbf{U} - \mathbf{V}\|_F^2,$$

where $\bar{\mu} = \mu\sigma_r^6(\mathbf{U}_*)/(200\|\mathbf{U}_*\|_2^4)$.

The proof of Lemma A.1 is provided in Section B.1.

A.2 Proof of Theorem 4.2

We first present the following key lemma. This lemma shows that all iterates $\{\mathbf{Y}_k\}$, $\{\mathbf{V}_k\}$, $\{\mathbf{X}_k\}$ stay in the neighborhood of $\mathbf{U}^*$. What's more, following the estimation sequence analysis for AGD in Nesterov (2004), we show that the Lyapunov function Φ_k enjoys the linear convergence with a certain amount of error.

Lemma A.2. With all parameters chosen the same as in Theorem 4.2, denote $C_{\text{R}} = c\mu\sigma_r^7(\mathbf{U}^*)/(\widehat{L}\|\mathbf{U}^*\|_2^6)$, where $\widehat{L} = L + \|\nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top)\|_2/\|\mathbf{U}^*\|_2^2$. Then for any $k \geq 0$, the following four statements hold:

1. $\mathbf{Y}_k \in \Omega(\mathbf{U}_0)$, $D(\mathbf{Y}_k, \mathbf{U}^*) \leq 1.1C_{\text{R}}$.
2. $\mathbf{V}_{k+1} \in \Omega(\mathbf{U}_0)$, $D(\mathbf{V}_{k+1}, \mathbf{U}^*) \leq C_{\text{R}}$.
3. $\mathbf{X}_{k+1} \in \Omega(\mathbf{U}_0)$, $D(\mathbf{X}_{k+1}, \mathbf{U}^*) \leq C_{\text{R}}$.
4. Define $\Phi_k := \mathcal{G}(\mathbf{X}_k) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2$, then it holds that $\Phi_{k+1} \leq (1-\alpha)\Phi_k + \alpha\epsilon/2$.

Proof of Theorem 4.2. By statement 4 in Lemma A.2, we have $\Phi_{k+1} - \epsilon/2 \leq (1-\alpha)(\Phi_k - \epsilon/2)$. Thus by induction, we have

$$\begin{aligned}
\mathcal{G}(\mathbf{X}_k) - \mathcal{G}(\mathbf{U}^*) &\leq (1-\alpha)^k\bigg[\mathcal{G}(\mathbf{X}_0) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_0\|_F^2 - \epsilon/2\bigg] + \epsilon/2 \\
&\leq (1-\alpha)^k\bigg[\mathcal{G}(\mathbf{U}_0) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{U}_0\|_F^2\bigg] + \epsilon/2
\end{aligned}$$

Since $\alpha = \sqrt{\eta\gamma}$, the result of Theorem 4.2 holds. □

A.3 Proof of Corollary 4.3

Proof of Corollary 4.3. Applying Theorem 4.2 and setting

$$K = \log(2\Delta_{\mathcal{G}}/\epsilon)/\sqrt{\eta\gamma} = O\bigg(\bigg[\frac{\|\mathbf{U}^*\|_2}{\sigma_r(\mathbf{U}^*)}\bigg]^3\sqrt{\frac{\widehat{L}}{\mu}}\log\frac{\Delta_{\mathcal{G}}}{\epsilon}\bigg),$$

we have $\mathcal{G}(\mathbf{X}_K) - \mathcal{G}(\mathbf{U}^*) \leq \epsilon/2 + \epsilon/2 = \epsilon$. This completes the proof. □

A.4 Proof of Theorem 4.6

We have the following lemma which gives the explicit formula for the projection of any matrix onto the positive definite matrix set.

Lemma A.3 (Li and Lin (2017)). Suppose that $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{r\times r}$ and $\widetilde{\mathbf{\Sigma}} = \operatorname{argmin}_{\mathbf{\Sigma}\succeq 0} \|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|_F$, then $\widetilde{\mathbf{\Sigma}}$ has the analytic formula $\widetilde{\mathbf{\Sigma}} = \mathbf{A}_0 \max\{\mathbf{D}_0, 0\}\mathbf{B}_0^\top$, where $[\mathbf{A}_0, \mathbf{D}_0, \mathbf{B}_0]$ is the SVD of $(\widehat{\mathbf{\Sigma}} + \widehat{\mathbf{\Sigma}}^\top)/2$.

We also need the following lemma to control the norms of any matrix $\mathbf{U}$ which is near to $\mathbf{U}^*$.

Lemma A.4. Suppose that $D(\mathbf{U}, \mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $D(\mathbf{V}, \mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, then we have the following inequalities:

$$0.9\|\mathbf{U}^*\|_2 \le \|\mathbf{U}\|_2 \le 1.1\|\mathbf{U}^*\|_2, \tag{A.1}$$
$$0.9\sigma_r(\mathbf{U}^*) \le \sigma_r(\mathbf{U}) \le 1.1\sigma_r(\mathbf{U}^*), \tag{A.2}$$
$$0.7\sigma_r^2(\mathbf{U}^*) \le \sigma_r(\mathbf{U}^\top\mathbf{V}) \le 1.3\sigma_r^2(\mathbf{U}^*). \tag{A.3}$$

Proof of Theorem 4.6. For simplicity, we use $\mathbf{V}'$ to denote $\mathbf{V}'_{k+1}$. Let $[\mathbf{A}_0, \mathbf{D}_0, \mathbf{B}_0]$ be the r-SVD of $\mathbf{U}_0 \in \mathbb{R}^{d\times r}$ such that $\mathbf{U}_0 = \mathbf{A}_0\mathbf{D}_0\mathbf{B}_0^\top$, where $\mathbf{A}_0 \in \mathbb{R}^{d\times r}, \mathbf{D}_0 \in \mathbb{R}^{r\times r}, \mathbf{B}_0 \in \mathbb{R}^{r\times r}$. Let

$$\mathbf{\Sigma}^* = \operatorname*{argmin}_{\mathbf{\Sigma}\succeq 0} \|\mathbf{D}_0^{-1}\mathbf{\Sigma} - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F$$

and $\bar{\mathbf{V}} = (\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)\mathbf{V}' + \mathbf{A}_0\mathbf{D}_0^{-1}\mathbf{\Sigma}^*\mathbf{B}_0^\top$. We now claim that $\bar{\mathbf{V}} = \operatorname{argmin}_{\mathbf{X}\in\Omega(\mathbf{U}_0)} \|\mathbf{X} - \mathbf{V}'\|_F$. To verify this claim, We first verify that $\bar{\mathbf{V}} \in \Omega(\mathbf{U}_0)$. It holds because

$$\mathbf{U}_0^\top\bar{\mathbf{V}} = \mathbf{B}_0\mathbf{D}_0\mathbf{A}_0^\top[(\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)\mathbf{V}' + \mathbf{A}_0\mathbf{D}_0^{-1}\mathbf{\Sigma}^*\mathbf{B}_0^\top] = \mathbf{B}_0\mathbf{\Sigma}^*\mathbf{B}_0^\top \succeq 0.$$

Then we note the following inequalities: for any $\mathbf{X}$ satisfying $\mathbf{X} \in \Omega(\mathbf{U}_0)$,

$$\begin{aligned}
\|\mathbf{X} - \mathbf{V}'\|_F &= \|\mathbf{X}\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0\|_F \\
&\ge \|\mathbf{A}_0^\top\mathbf{X}\mathbf{B}_0 - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F^2 \\
&\ge \inf_{\mathbf{\Sigma}\succeq 0} \|\mathbf{D}_0^{-1}\mathbf{\Sigma} - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F \\
&= \|\mathbf{D}_0^{-1}\mathbf{\Sigma}^* - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F,
\end{aligned} \tag{A.4}$$

where the first inequality holds since $\mathbf{A}_0$ is a column orthonomaral matrix, the second inequality holds since $\mathbf{D}_0\mathbf{A}_0^\top\mathbf{X}\mathbf{B}_0 \succeq 0$, which can be verified as

$$\mathbf{X} \in \Omega(\mathbf{U}_0) \Leftrightarrow \mathbf{X}^\top\mathbf{U}_0 \succeq 0 \Leftrightarrow \mathbf{X}^\top\mathbf{A}_0\mathbf{D}_0\mathbf{B}_0^\top \succeq 0 \Leftrightarrow \mathbf{D}_0\mathbf{A}_0^\top\mathbf{X}\mathbf{B}_0 \succeq 0.$$

Now we substitute $\mathbf{X} = \bar{\mathbf{V}}$ into (A.4) and verify that each inequality in (A.4) is indeed equality, which suggests that $\bar{\mathbf{V}} = \operatorname{argmin}_{\mathbf{X}\in\Omega(\mathbf{U}_0)} \|\mathbf{X} - \mathbf{V}'\|_F$. We have

$$\begin{aligned}
\|\bar{\mathbf{V}} - \mathbf{V}'\|_F^2 &= \|\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0\|_F^2 \\
&= \|\mathbf{A}_0^\top\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F^2 + \operatorname{trace}\Big[(\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0)^\top \underbrace{(\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)(\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0)}_{I_1}\Big] \\
&= \|\mathbf{A}_0^\top\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F^2 \\
&= \|\mathbf{D}_0^{-1}\mathbf{\Sigma}^* - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0\|_F^2,
\end{aligned} \tag{A.5}$$

where the third equality holds since

$$\begin{aligned}
I_1 &= (\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)(\bar{\mathbf{V}}\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0) \\
&= (\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)((\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)\mathbf{V}'\mathbf{B}_0 + \mathbf{A}_0\mathbf{D}_0^{-1}\mathbf{\Sigma}^*\mathbf{B}_0^\top\mathbf{B}_0 - \mathbf{V}'\mathbf{B}_0) \\
&= (\mathbf{I} - \mathbf{A}_0\mathbf{A}_0^\top)\mathbf{A}_0[\mathbf{D}_0^{-1}\mathbf{\Sigma}^* - \mathbf{A}_0^\top\mathbf{V}'\mathbf{B}_0] \\
&= \mathbf{0},
\end{aligned}$$

the fourth equality holds since

$$(\mathbf{A}_0)^\top \bar{\mathbf{V}} \mathbf{B}_0 = \mathbf{A}_0^\top \Big[(\mathbf{I} - \mathbf{A}_0 \mathbf{A}_0^\top) \mathbf{V}' + \mathbf{A}_0 \mathbf{D}_0^{-1} \mathbf{\Sigma}^* \mathbf{B}_0^\top \Big] \mathbf{B}_0 = \mathbf{D}_0^{-1} \mathbf{\Sigma}^*.$$

Next we analyze the convergence rate of $\|\mathbf{\Sigma}_t^{(1)} - \mathbf{\Sigma}^*\|_F$. First, by Lemma A.3, we know that $\mathbf{\Sigma}_{t+1}^{(1)}$ is the projection of $\mathbf{\Sigma}'_{t+1}$ to convex set $\{\mathbf{\Sigma} : \mathbf{\Sigma} \succeq 0\}$, which implies that $\{\mathbf{\Sigma}_t^{(1)}, \mathbf{\Sigma}_t^{(2)}\}$ forms two sequences which are generated by proximal AGD with function $\|\mathbf{D}_0^{-1}\mathbf{\Sigma} - \mathbf{T}\|_F^2/2$. Meanwhile, we know that $\|\mathbf{D}_0^{-1}\mathbf{\Sigma} - \mathbf{T}\|_F^2/2$ is $\|\mathbf{D}_0^{-1}\|_2^2 = \sigma_r(\mathbf{U}_0)^{-2}$-smooth and $\lambda_{\min}(\mathbf{D}_0^{-2}) = \|\mathbf{U}_0\|_2^{-2}$-strongly convex. Thus by standard convergence result of proximal AGD (Nesterov, 2004), we have

$$\begin{aligned}
\|\mathbf{\Sigma}_t^{(1)} - \mathbf{\Sigma}^*\|_F^2 &\leq 2\bigg(1 - \frac{\sigma_r(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2}\bigg)^t \|\mathbf{U}_0\|_2^2 [\|\mathbf{D}_0^{-1}\mathbf{\Sigma}_0^{(1)} - \mathbf{T}\|_F^2 - \|\mathbf{D}_0^{-1}\mathbf{\Sigma}^* - \mathbf{T}\|_F^2] \\
&\leq 2\bigg(1 - \frac{\sigma_r(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2}\bigg)^t \|\mathbf{U}_0\|_2^2 \|\mathbf{D}_0^{-1}\mathbf{\Sigma}_0^{(1)} - \mathbf{T}\|_F^2 \\
&= 2\bigg(1 - \frac{\sigma_r(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2}\bigg)^t \|\mathbf{U}_0\|_2 \|\mathbf{T}\|_F \\
&\leq 2\bigg(1 - \frac{\sigma_r(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2}\bigg)^t \|\mathbf{U}_0\|_2 \|\mathbf{V}'\|_F, \qquad \text{(A.6)}
\end{aligned}$$

where the second inequality holds since $\|\mathbf{D}_0^{-1}\mathbf{\Sigma} - \mathbf{T}\|_F^2/2$ is $\|\mathbf{U}_0\|_2^{-2}$-strongly convex, the equality holds since $\mathbf{\Sigma}_0^{(1)} = \mathbf{0}$, the last inequality holds since $\|\mathbf{T}\|_F = \|\mathbf{A}_0^\top \mathbf{V}' \mathbf{B}_0\|_F \leq \|\mathbf{V}'\|_F$. Finally, by the definition of $\mathbf{V}_{k+1}$, $\bar{\mathbf{V}}$ and (A.6), we have

$$\begin{aligned}
\|\mathbf{V}_{k+1} - \bar{\mathbf{V}}\|_F &= \|\mathbf{A}_0 \mathbf{D}_0^{-1} \mathbf{\Sigma}_T^{(1)} \mathbf{B}_0^\top - \mathbf{A}_0 \mathbf{D}_0^{-1} \mathbf{\Sigma}^* \mathbf{B}_0^\top\|_F \\
&= \|\mathbf{D}_0^{-1}(\mathbf{\Sigma}_T^{(1)} - \mathbf{\Sigma}^*)\|_F \\
&\leq 2\bigg(1 - \frac{\sigma_r(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2}\bigg)^T \frac{\|\mathbf{U}_0\|_2}{\sigma_r(\mathbf{U}_0)} \|\mathbf{V}'\|_F \\
&\leq 3\bigg(1 - 2\frac{\sigma_r(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2}\bigg)^T \frac{\|\mathbf{U}^*\|_2}{\sigma_r(\mathbf{U}^*)} (\|\mathbf{U}^*\|_F + c^{1/2}\sigma_r(\mathbf{U}^*)) \\
&\leq 4\bigg(1 - 2\frac{\sigma_r(\mathbf{U}^*)}{\|\mathbf{U}^*\|_2}\bigg)^T \|\mathbf{U}^*\|_F. \qquad \text{(A.7)}
\end{aligned}$$

where the second inequality holds due to Lemma A.4 and

$$\begin{aligned}
\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F &= \big\|(1-\alpha)\mathbf{V}_k + \alpha\mathbf{Y}_k - \alpha/\gamma \nabla \mathcal{G}(\mathbf{Y}_k) - \mathbf{U}^*\big\|_F \\
&\leq 5\|\mathbf{U}^*\|_2^2/\sigma_r(\mathbf{U}^*)^2 \cdot [(1-\alpha)D(\mathbf{V}_k, \mathbf{U}^*) + \alpha D(\mathbf{Y}_k, \mathbf{U}^*)] + \alpha/\gamma \|\nabla \mathcal{G}(\mathbf{Y}_k)\|_F \\
&\leq C_{\mathrm{R}}[5.5\|\mathbf{U}^*\|_2^2/\sigma_r(\mathbf{U}^*)^2 + \alpha/\gamma \cdot (10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla \mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2)] \\
&\leq c^{1/2}\sigma_r(\mathbf{U}^*).
\end{aligned}$$

(A.7) directly implies that to $\|\mathbf{V}_{k+1} - \bar{\mathbf{V}}\|_F \leq \epsilon_S$ can be attained within iteration

$$T = O\bigg(\frac{\|\mathbf{U}^*\|_2}{\sigma_r(\mathbf{U}^*)} \log \frac{\|\mathbf{U}^*\|_F}{\epsilon_S}\bigg). \qquad \text{(A.8)}$$

Finally, substituting $\epsilon_S = \min\{\alpha\epsilon/[4c^{1/2}\gamma\sigma_r(\mathbf{U}^*)], c^{1/2}\sigma_r(\mathbf{U}^*)/2\}$ into (A.8), we have the conclusion. □

B Proofs of Lemmas in Appendix A

B.1 Proof of Lemma A.1

To prove Lemma A.1, we need the following lemmas:

Lemma B.1. Suppose that $D(\mathbf{U},\mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, $D(\mathbf{V},\mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $D(\mathbf{U}_0,\mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, $\mathbf{U},\mathbf{V} \in \Omega(\mathbf{U}_0)$, then

$$\|\mathbf{U}-\mathbf{V}\|_F \le \frac{5\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2} D(\mathbf{U},\mathbf{V}).$$

Lemma B.2. Suppose that $D(\mathbf{U},\mathbf{U}^*) \le \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$ and $D(\mathbf{V},\mathbf{U}^*) \le \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$, then

$$\begin{aligned}&\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V})\\ &\ge \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2 + 0.4\mu\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V}).\end{aligned}$$

Proof of Lemma A.1. We have

$$\begin{aligned}&\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V})\\ &\ge \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2 + 0.4\mu\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V})\\ &\ge \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2\\ &\quad + 0.4\mu\sigma_r^2(\mathbf{V})\frac{\sigma_r^4(\mathbf{U}^*)}{25\|\mathbf{U}^*\|_2^4}\|\mathbf{U}-\mathbf{V}\|_F^2,\end{aligned}$$

where the first inequality holds due to Lemma B.2, the second inequality holds due to Lemma B.1. With the fact $\sigma_r^2(\mathbf{V}) \ge 0.8\sigma_r^2(\mathbf{U}^*)$ from Lemma A.4, we can further give the bound

$$\begin{aligned}&\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V}) - \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle\\ &\ge \|\mathbf{U}-\mathbf{V}\|_F^2 \cdot \left(\frac{\mu\sigma_r^6(\mathbf{U}^*)}{100\|\mathbf{U}^*\|_2^4} - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\right)\\ &\ge \frac{\mu\sigma_r^6(\mathbf{U}^*)}{200\|\mathbf{U}^*\|_2^4}\|\mathbf{U}-\mathbf{V}\|_F^2,\end{aligned}$$

where the last inequality holds because

$$D(\mathbf{V},\mathbf{U}^*) \le C_{\mathrm{R}} \le \frac{\mu\sigma_r^7(\mathbf{U}^*)}{4200L\|\mathbf{U}^*\|_2^6 + 1000\|\mathbf{U}^*\|_2^4\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}.$$

This completes the proof. □

B.2 Proof of Lemma A.2

In this section we provide the proof of Lemma A.2. Let $\mathbf{X}_k, \mathbf{V}_k, \mathbf{Y}_k, \mathbf{V}'_k, \eta, \gamma, \mathbf{U}_0$ be defined as in Theorem 4.2.

Lemma B.3. For any $\mathbf{V} \in \mathbb{R}^{d\times r}$, we have $\Pi_{\mathbf{U}_0}(\mathbf{V}) \in \Omega(\mathbf{U}_0)$. Moreover, for any $\mathbf{X} \in \Omega(\mathbf{U}_0)$, $D(\mathbf{X},\mathbf{U}_0) = \|\mathbf{X}-\mathbf{U}_0\|_F$.

Next lemma provides some useful bounds in the proof of Lemma A.2.

Lemma B.4. Under the same assumptions as Theorem 4.2, taking constant $c \le 84000^{-1}$, the following inequalities hold:

$$\begin{aligned}&\gamma \le \bar{\mu},\\ &\eta \le \frac{1}{12L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top)\|_2},\\ &C_{\mathrm{R}}^2 \ge \max\left\{C_{\mathrm{ini}}^2, \frac{3}{\mu\sigma_r^2(\mathbf{U}^*)}\big(4\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2 + \epsilon\big), \frac{2}{\gamma}[(\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top}\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon]\right\},\\ &C_{\mathrm{R}} \le \min\{C_{\mathrm{convex}}/2, \sigma_r^2(\mathbf{U}^*)/(10\|\mathbf{U}^*\|_2)\}.\end{aligned}$$

Next lemma gives a bound for strong convexity at $\mathbf{U}^*$ that is tighter than the corresponding result of Lemma A.1.

Lemma B.5. Suppose $\mathbf{U}$ satisfy that $D(\mathbf{U}, \mathbf{U}^*) \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$, then

$$\mathcal{G}(\mathbf{U}) \geq \mathcal{G}(\mathbf{U}^*) + 0.4\mu\sigma_r^2(\mathbf{U}^*)D^2(\mathbf{U}, \mathbf{U}^*).$$

Next lemma shows that for any $\mathbf{V}$ which is located in the neighborhood of $\mathbf{U}^*$, if we perform one step gradient descent from $\mathbf{V}$, the function value will decrease sufficiently and the update is still located in the neighborhood of $\mathbf{U}^*$ with a slightly bigger radius.

Lemma B.6 (Gradient descent decrease)**.** Suppose that $D(\mathbf{V}, \mathbf{U}^*) \leq 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$. Let $\mathbf{V}' = \mathbf{V} - \eta\nabla\mathcal{G}(\mathbf{V})$ where $0 < \eta \leq (12L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*[\mathbf{U}^*]^\top)\|_2)^{-1}$, then

$$\mathcal{G}(\mathbf{V}') \leq \mathcal{G}(\mathbf{V}) - \frac{\eta}{2}\|\nabla\mathcal{G}(\mathbf{V})\|_F^2,$$

and $D(\mathbf{V}', \mathbf{U}^*) \leq 2D(\mathbf{V}, \mathbf{U}^*)$.

Next lemma shows that for any $\mathbf{U}$ which is located in the neighborhood of $\mathbf{U}^*$, the function value gap $\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{U}^*)$ can be bounded by the distance between $\mathbf{U}$ and $\mathbf{U}^*$.

Lemma B.7. Suppose that $D(\mathbf{U}, \mathbf{U}^*) \leq 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, then

$$\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{U}^*) \leq (\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 3L\|\mathbf{U}^*\|_2^2)D^2(\mathbf{U}, \mathbf{U}^*).$$

Next lemma gives a bound for $\|\nabla\mathcal{G}(\mathbf{U})\|_F$.

Lemma B.8. Suppose that $D(\mathbf{U}, \mathbf{U}^*) \leq 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, then

$$\|\nabla\mathcal{G}(\mathbf{U})\|_F \leq D(\mathbf{U}, \mathbf{U}^*)(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2).$$

Next lemma provides a upper bound for $\gamma\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2$, which will be used in our proof.

Lemma B.9. Under the same assumptions as Theorem 4.2, it holds that

$$\begin{aligned}\frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 \leq{}& \frac{\gamma(1-\alpha)}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \alpha\Big(\langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{U}^* - \mathbf{Y}_k\rangle + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{Y}_k\|_F^2\Big)\\ &+ \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2 - \alpha(1-\alpha)\langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{V}_k - \mathbf{Y}_k\rangle.\end{aligned}$$

Proof of Lemma A.2. We have $D(\mathbf{X}_0, \mathbf{U}^*) \leq C_{\mathrm{ini}} \leq C_{\mathrm{R}}$ and $D(\mathbf{V}_0, \mathbf{U}^*) \leq C_{\mathrm{ini}} \leq C_{\mathrm{R}}$ and $\mathbf{X}_0, \mathbf{V}_0 \in \Omega(\mathbf{U}_0)$ since $\mathbf{X}_0 = \mathbf{V}_0 = \mathbf{U}_0$. We prove Lemma A.2 by induction. Suppose that statement 1 to 4 hold for $k-1$. Next we show that statements 1 to 4 in Lemma A.2 still hold for k.

- We first show that statement 1 holds for k. Let $\xi = \alpha/(\alpha+1)$, then by the construction rule of $\mathbf{Y}_k$ in Algorithm 1, we have $\mathbf{Y}_k = \xi\mathbf{V}_k + (1-\xi)\mathbf{X}_k$, which indicates that $\mathbf{Y}_k \in \Omega(\mathbf{U}_0)$. Meanwhile, we have

$$\begin{aligned}D(\mathbf{Y}_k, \mathbf{U}^*) &\leq \|\mathbf{U}^* - \mathbf{Y}_k\|_F\\ &= \|\mathbf{U}^* - \mathbf{U}_0 + \mathbf{U}_0 - \mathbf{Y}_k\|_F\\ &\leq \|\mathbf{U}^* - \mathbf{U}_0\|_F + \|\mathbf{U}_0 - \mathbf{Y}_k\|_F\\ &\leq C_{\mathrm{ini}} + \xi\|\mathbf{U}_0 - \mathbf{V}_k\|_F + (1-\xi)\|\mathbf{U}_0 - \mathbf{X}_k\|_F, \qquad \text{(B.1)}\end{aligned}$$

  where the last inequality holds because of Lemma B.3 and $\|\mathbf{U}^* - \mathbf{U}_0\|_F = D(\mathbf{U}_0, \mathbf{U}^*) \leq C_{\mathrm{ini}}$. Since $\mathbf{V}_k, \mathbf{X}_k \in \Omega(\mathbf{U}_0)$, then we have

$$\|\mathbf{U}_0 - \mathbf{V}_k\|_F = D(\mathbf{U}_0, \mathbf{V}_k) \leq D(\mathbf{U}_0, \mathbf{U}^*) + D(\mathbf{U}^*, \mathbf{V}_k) \leq C_{\mathrm{ini}} + C_{\mathrm{R}}, \qquad \text{(B.2)}$$

$$\|\mathbf{U}_0 - \mathbf{X}_k\|_F = D(\mathbf{U}_0, \mathbf{X}_k) \leq D(\mathbf{U}_0, \mathbf{U}^*) + D(\mathbf{U}^*, \mathbf{X}_k) \leq C_{\mathrm{ini}} + C_{\mathrm{R}}. \qquad \text{(B.3)}$$

  Substituting (B.2) and (B.3) into (B.1), we have

$$D(\mathbf{Y}_k, \mathbf{U}^*) \leq C_{\mathrm{ini}} + \xi(C_{\mathrm{ini}} + C_{\mathrm{R}}) + (1-\xi)(C_{\mathrm{ini}} + C_{\mathrm{R}}) \leq 1.1C_{\mathrm{R}}, \qquad \text{(B.4)}$$

  where the second inequality holds due to Lemma B.4.

- Next we show statement 4 holds for $k+1$. We are going to bound $\mathcal{G}(\mathbf{X}_{k+1})$ and $\|\mathbf{U}^* - \mathbf{V}_{k+1}\|_F$ separately. To bound $\mathcal{G}(\mathbf{X}_{k+1})$, note that by (B.4) and Lemma B.4 we have $D(\mathbf{Y}_k, \mathbf{U}^*) \le 1.1C_{\mathrm{R}} \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$. Then by Lemma B.6, we have

$$\mathcal{G}(\mathbf{X}_{k+1}) = \mathcal{G}\Big(\mathbf{Y}_k - \eta\nabla\mathcal{G}(\mathbf{Y}_k)\Big) \le \mathcal{G}(\mathbf{Y}_k) - \frac{\eta}{2}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2 = \mathcal{G}(\mathbf{Y}_k) - \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2, \tag{B.5}$$

Next, we have

$$\mathcal{G}(\mathbf{Y}_k) = (1-\alpha)\mathcal{G}(\mathbf{Y}_k) + \alpha\mathcal{G}(\mathbf{Y}_k) \le (1-\alpha)(\mathcal{G}(\mathbf{X}_k) + \langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{Y}_k - \mathbf{X}_k\rangle) + \alpha\mathcal{G}(\mathbf{Y}_k), \tag{B.6}$$

where the second inequality holds due to Lemma A.1 with the condition that $D(\mathbf{Y}_k, \mathbf{U}^*) \le 1.1C_{\mathrm{R}} \le C_{\mathrm{convex}}, \mathbf{Y}_k \in \Omega(\mathbf{U}_0)$ by (B.4) and $D(\mathbf{X}_k, \mathbf{U}^*) \le C_{\mathrm{R}} \le C_{\mathrm{convex}}, \mathbf{X}_k \in \Omega(\mathbf{U}_0)$ by induction assumption. Substituting (B.6) into (B.5), we have

$$\mathcal{G}(\mathbf{X}_{k+1}) \le (1-\alpha)(\mathcal{G}(\mathbf{X}_k) + \langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{Y}_k - \mathbf{X}_k\rangle) + \alpha\mathcal{G}(\mathbf{Y}_k) - \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2. \tag{B.7}$$

Next we are going to bound $\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F$. Due to Lemma B.9, we have

$$\begin{aligned}\frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 \le{}& \frac{\gamma(1-\alpha)}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \alpha\Big(\langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{U}^* - \mathbf{Y}_k\rangle + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{Y}_k\|_F^2\Big)\\ &+ \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2 - \alpha(1-\alpha)\langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{V}_k - \mathbf{Y}_k\rangle.\end{aligned} \tag{B.8}$$

Adding (B.7) and (B.8) up, we have

$$\begin{aligned}&\mathcal{G}(\mathbf{X}_{k+1}) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2\\ &\le (1-\alpha)\Big[\mathcal{G}(\mathbf{X}_k) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2\Big] + \alpha\Big(\mathcal{G}(\mathbf{Y}_k) + \langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{U}^* - \mathbf{Y}_k\rangle + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{Y}_k\|_F^2\Big)\\ &\quad + (1-\alpha)\Big\langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{Y}_k - \mathbf{X}_k + \alpha(\mathbf{Y}_k - \mathbf{V}_k)\Big\rangle\\ &= (1-\alpha)\Big[\mathcal{G}(\mathbf{X}_k) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2\Big] + \alpha\Big(\mathcal{G}(\mathbf{Y}_k) + \langle\nabla\mathcal{G}(\mathbf{Y}_k), \mathbf{U}^* - \mathbf{Y}_k\rangle + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{Y}_k\|_F^2\Big)\\ &\le (1-\alpha)\Big[\mathcal{G}(\mathbf{X}_k) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2\Big] + \alpha\mathcal{G}(\mathbf{U}^*),\end{aligned} \tag{B.9}$$

where the equality holds due to the definition of $\mathbf{Y}_k$, the second inequality holds due to Lemma A.1 with the condition that $D(\mathbf{Y}_k, \mathbf{U}^*) \le 1.1C_{\mathrm{R}} \le C_{\mathrm{convex}}, \mathbf{Y}_k \in \Omega(\mathbf{U}_0)$ by (B.1) and $\gamma \le \bar{\mu}$. Rearranging (B.9), we have

$$\mathcal{G}(\mathbf{X}_{k+1}) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 \le (1-\alpha)\Big[\mathcal{G}(\mathbf{X}_k) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2\Big]. \tag{B.10}$$

We now claim that $\|\mathbf{X} - \mathbf{V}_{k+1}\|_F \le \|\mathbf{X} - \mathbf{V}'_{k+1}\|_F + \epsilon_S$ for any $\mathbf{X} \in \Omega(\mathbf{U}_0)$. Denote $\bar{\mathbf{V}}$ to be the projection of $\mathbf{V}'_{k+1}$ to convex set $\Omega(\mathbf{U}_0)$, then we have

$$\|\mathbf{X} - \mathbf{V}_{k+1}\|_F \le \|\mathbf{X} - \bar{\mathbf{V}}\|_F + \|\mathbf{V}_{k+1} - \bar{\mathbf{V}}\|_F \le \|\mathbf{X} - \mathbf{V}'_{k+1}\|_F + \|\mathbf{V}_{k+1} - \bar{\mathbf{V}}\|_F \le \|\mathbf{X} - \mathbf{V}'_{k+1}\|_F + \epsilon_S,$$

where the second inequality holds due to the fact that $\mathbf{X} \in \Omega(\mathbf{U}_0)$, the last inequality holds since ACCPROJ outputs an ϵ_S-approximate projection for each k. Specifically, taking $\mathbf{X} = \mathbf{U}^*$, we have

$$\begin{aligned}\frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_{k+1}\|_F^2 &\le \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 + \frac{\gamma}{2}\epsilon_S^2 + \gamma\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F\epsilon_S\\ &\le \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 + 2c^{1/2}\gamma\sigma_r(\mathbf{U}^*)\epsilon_S\\ &\le \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 + \alpha\epsilon/2,\end{aligned} \tag{B.11}$$

where the second inequality holds since $\epsilon_S \leq c^{1/2}\sigma_r(\mathbf{U}^*)$ and

$$\begin{aligned}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F &= \big\|(1-\alpha)\mathbf{V}_k + \alpha\mathbf{Y}_k - \alpha/\gamma\nabla\mathcal{G}(\mathbf{Y}_k) - \mathbf{U}^*\big\|_F \\ &\leq 5\|\mathbf{U}^*\|_2^2/\sigma_r(\mathbf{U}^*)^2 \cdot [(1-\alpha)D(\mathbf{V}_k,\mathbf{U}^*) + \alpha D(\mathbf{Y}_k,\mathbf{U}^*)] + \alpha/\gamma\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F \\ &\leq C_{\mathrm{R}}[5.5\|\mathbf{U}^*\|_2^2/\sigma_r(\mathbf{U}^*)^2 + \alpha/\gamma\cdot(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2)] \\ &\leq c^{1/2}\sigma_r(\mathbf{U}^*), \end{aligned} \tag{B.12}$$

where the second inequality follows by Lemma B.1, Lemma B.8 and the facts $D(\mathbf{V}_k,\mathbf{U}^*), D(\mathbf{Y}_k,\mathbf{U}^*) \leq 1.1C_{\mathrm{R}}$, the last inequality holds due to Lemma B.4.

Substituting (B.11) into (B.10), we have

$$\begin{aligned}&\mathcal{G}(\mathbf{X}_{k+1}) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_{k+1}\|_F^2 \\ &\leq (1-\alpha)\bigg[\mathcal{G}(\mathbf{X}_k) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2\bigg] + \alpha\epsilon/2, \end{aligned} \tag{B.13}$$

which is $\Phi_{k+1} \leq (1-\alpha)\Phi_k + \alpha\epsilon/2$.

- Next we show statement 3 holds for $k+1$. By the construction rule of $\mathbf{X}_{k+1}$ in Algorithm 1 and Lemma B.3, clearly we have $\mathbf{X}_{k+1} \in \Omega(\mathbf{U}_0)$. To show $D(\mathbf{X}_{k+1},\mathbf{U}^*) \leq C_{\mathrm{R}}$, first by (B.13), we have

$$\begin{aligned}\mathcal{G}(\mathbf{X}_{k+1}) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_{k+1}\|_F^2 &\leq \mathcal{G}(\mathbf{X}_0) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_0\|_F^2 + \epsilon \\ &= \mathcal{G}(\mathbf{U}_0) - \mathcal{G}(\mathbf{U}^*) + \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{U}_0\|_F^2 + \epsilon \\ &\leq (\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 4L\|\mathbf{U}^*\|_2^2)\|\mathbf{U}^* - \mathbf{U}_0\|_F^2 + \epsilon \\ &\leq (\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon, \end{aligned} \tag{B.14}$$

where the second inequality holds due to Lemma B.7 with $\|\mathbf{U}^* - \mathbf{U}_0\|_F \leq C_{\mathrm{ini}} \leq 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $\gamma \leq L\|\mathbf{U}^*\|_2^2$. For simplicity, we denote $\mathbf{X}'_{k+1} = \mathbf{Y}_k - \eta\nabla\mathcal{G}(\mathbf{Y}_k)$. Then we also have

$$D(\mathbf{X}_{k+1},\mathbf{U}^*) = D(\mathbf{X}'_{k+1},\mathbf{U}^*) \leq 2D(\mathbf{Y}_k,\mathbf{U}^*) \leq 2.2C_{\mathrm{R}} \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2), \tag{B.15}$$

where the first inequality holds by Lemma B.6. Then we have

$$\begin{aligned}D^2(\mathbf{X}_{k+1},\mathbf{U}^*) &\leq \frac{3}{\mu\sigma_r^2(\mathbf{U}^*)}(\mathcal{G}(\mathbf{X}_{k+1}) - \mathcal{G}(\mathbf{U}^*)) \\ &\leq \frac{3}{\mu\sigma_r^2(\mathbf{U}^*)}[(\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon] \\ &\leq C_{\mathrm{R}}^2, \end{aligned}$$

where the first inequality holds due to Lemma B.5 with condition (B.15), the second inequality holds due to (B.14) and the last inequality holds due to Lemma B.4.

- Finally we show that statement 2 holds for $k+1$. $\mathbf{V}_{k+1} \in \Omega(\mathbf{U}_0)$ because of the construction rule of $\mathbf{V}_{k+1}$ in Algorithm 1. To prove $D(\mathbf{V}_{k+1},\mathbf{U}^*) \leq C_{\mathrm{R}}$, we have

$$D^2(\mathbf{V}_{k+1},\mathbf{U}^*) \leq \|\mathbf{U}^* - \mathbf{V}_{k+1}\|_F^2 \leq \frac{2}{\gamma}[(\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon] \leq C_{\mathrm{R}}^2,$$

where the second inequality holds due to (B.14) and the last inequality holds due to Lemma B.4. Thus, we have $D(\mathbf{X}_{k+1},\mathbf{U}^*) \leq C_{\mathrm{R}}$ and $D(\mathbf{V}_{k+1},\mathbf{U}^*) \leq C_{\mathrm{R}}$.

With induction, we conclude that the results of Lemma A.2 hold. □

C Proofs of Lemmas in Section B

C.1 Proof of Lemma B.1

We have the following lemma:

Lemma C.1 (Li (1995)). Let $\mathbf{A} \in \mathbb{R}^{r\times r}$ be of full rank and $[\mathbf{Q}, \mathbf{H}]$ be its unique polar decomposition, $\mathbf{A}+\Delta\mathbf{A}$ be of full rank and $(\mathbf{Q}+\Delta\mathbf{Q})(\mathbf{H}+\Delta\mathbf{H})$ be its unique polar decomposition, then $\|\Delta\mathbf{Q}\|_F \le \frac{2}{\sigma_r(\mathbf{A})}\|\Delta\mathbf{A}\|_F$.

Proof of Lemma B.1. Since $D(\mathbf{U},\mathbf{U}^*) \le 0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, $D(\mathbf{V},\mathbf{U}^*) \le 0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $D(\mathbf{U}_0,\mathbf{U}^*) \le 0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, then by Lemma A.4, we have

$$\sigma_r(\mathbf{U}^\top\mathbf{U}_0) \ge 0.7\sigma_r^2(\mathbf{U}^*),\ \sigma_r(\mathbf{V}^\top\mathbf{U}_0) \ge 0.7\sigma_r^2(\mathbf{U}^*),\ \sigma_r(\mathbf{U}^\top\mathbf{V}) \ge 0.7\sigma_r^2(\mathbf{U}^*),$$

which implies that $\mathbf{U}^\top\mathbf{V}$, $\mathbf{U}^\top\mathbf{U}_0$ and $\mathbf{V}^\top\mathbf{U}_0$ are of full rank. Note that $\Pi_{\mathbf{U}_0}(\mathbf{U}) = \Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U}))$. Next we are going to prove that

$$\|\Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U})) - \Pi_{\mathbf{U}_0}(\mathbf{V})\|_F \le D(\mathbf{U},\mathbf{V}).$$

First we have

$$\Pi_{\mathbf{U}_0}(\mathbf{V}) = \mathbf{V}\mathbf{P}_1, \Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U})) = \Pi_{\mathbf{V}}(\mathbf{U})\mathbf{P}_2,$$

where $\mathbf{P}_1$ is the orthogonal matrix of the unique polar decomposition of $\mathbf{V}^\top\mathbf{U}_0$, $\mathbf{P}_2$ is the orthogonal matrix of the unique polar decomposition of $(\Pi_{\mathbf{V}}(\mathbf{U}))^\top\mathbf{U}_0$[6]. Then by Lemma C.1, we have

$$\|\mathbf{P}_2-\mathbf{P}_1\|_F \le \frac{2}{\sigma_r(\mathbf{V}^\top\mathbf{U}_0)}\|(\Pi_{\mathbf{V}}(\mathbf{U})-\mathbf{V})^\top\mathbf{U}_0\|_F \le \frac{2\|\mathbf{U}_0\|_2}{\sigma_r(\mathbf{V}^\top\mathbf{U}_0)}D(\mathbf{U},\mathbf{V}). \tag{C.1}$$

Thus, we have

$$\begin{aligned}
\|\Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U})) - \Pi_{\mathbf{U}_0}(\mathbf{V})\|_F &= \|\Pi_{\mathbf{V}}(\mathbf{U})\mathbf{P}_2 - \mathbf{V}\mathbf{P}_1\|_F \\
&= \|\Pi_{\mathbf{V}}(\mathbf{U})\mathbf{P}_2 - \mathbf{V}\mathbf{P}_2 + \mathbf{V}\mathbf{P}_2 - \mathbf{V}\mathbf{P}_1\|_F \\
&\le D(\mathbf{U},\mathbf{V}) + \|\mathbf{V}\|_2\|\mathbf{P}_2-\mathbf{P}_1\|_F \\
&\le D(\mathbf{U},\mathbf{V})\left(1+\frac{2\|\mathbf{V}\|_2\|\mathbf{U}_0\|_2}{\sigma_r(\mathbf{V}^\top\mathbf{U}_0)}\right).
\end{aligned}$$

Since $\|\Pi_{\mathbf{U}^*}(\mathbf{V}) - \mathbf{U}^*\|_F \le 0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $\|\Pi_{\mathbf{U}_0}(\mathbf{U}^*) - \mathbf{U}_0\|_F \le 0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2$, then by Lemma A.4, we have

$$\|\mathbf{U}_0\|_2 \le 1.1\|\mathbf{U}^*\|_2, \|\mathbf{V}\|_2 \le 1.1\|\mathbf{U}^*\|_2, \sigma_r(\mathbf{V}^\top\mathbf{U}_0) \ge 0.7\sigma_r(\mathbf{U}^*)^2.$$

Thus, we have

$$\|\Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U})) - \Pi_{\mathbf{U}_0}(\mathbf{V})\|_F \le \left(1+\frac{4\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2}\right)D(\mathbf{U},\mathbf{V}) \le \frac{5\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2}D(\mathbf{U},\mathbf{V}).$$

Since $\mathbf{U},\mathbf{V} \in \Omega(\mathbf{U}_0)$, then by Lemma B.3 we have $\Pi_{\mathbf{U}_0}(\Pi_{\mathbf{V}}(\mathbf{U})) = \mathbf{U}$ and $\Pi_{\mathbf{U}_0}(\mathbf{V}) = \mathbf{V}$. Thus we have

$$\|\mathbf{U}-\mathbf{V}\|_F \le \frac{5\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2}D(\mathbf{U},\mathbf{V}).$$

□

[6] In fact, for any matrix $\mathbf{X} \in \mathbb{R}^{r\times r}$, suppose $[\mathbf{Q},\mathbf{H}]$ to be the polar decomposition of $\mathbf{X}$ where $\mathbf{Q}$ is the unitary matrix and $\mathbf{H}$ is the positive-semidifinite matrix, and $[\mathbf{A},\mathbf{D},\mathbf{B}]$ to be the SVD of $\mathbf{X}$. We can verify that $\mathbf{Q} = \mathbf{A}\mathbf{B}^\top$.

C.2 Proof of Lemma B.2

To prove Lemma B.2, we need the following lemmas:

Lemma C.2 (Tu et al. (2016)). For any two matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d\times r}$, it holds that $\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2 \geq 0.8\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V})$.

Lemma C.3. Suppose that $D(\mathbf{U},\mathbf{U}^*) \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$ and $D(\mathbf{V},\mathbf{U}^*) \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$, then

$$
\begin{aligned}
&|\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top, (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle| \\
&\leq \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2.
\end{aligned}
$$

Proof of Lemma B.2. We have

$$
\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V}) = \mathcal{L}(\mathbf{U}\mathbf{U}^\top) - \mathcal{L}(\mathbf{V}\mathbf{V}^\top) \geq \langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\rangle + \frac{\mu}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2, \tag{C.2}
$$

where the inequality holds due to the μ-convexity of $\mathcal{L}$. The term $\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\rangle$ has the following equivalent formula:

$$
\begin{aligned}
&\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\rangle \\
&= \langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), [\mathbf{V} + (\mathbf{U}-\mathbf{V})][\mathbf{V} + (\mathbf{U}-\mathbf{V})]^\top - \mathbf{V}\mathbf{V}^\top\rangle \\
&= \langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), (\mathbf{U}-\mathbf{V})\mathbf{V}^\top + \mathbf{V}(\mathbf{U}-\mathbf{V})^\top + (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle \\
&= \langle [\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top]\mathbf{V}, \mathbf{U}-\mathbf{V}\rangle + \langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top), (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle \\
&= \langle [\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top]\mathbf{V}, \mathbf{U}-\mathbf{V}\rangle + \frac{1}{2}\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top, (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle \\
&= \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle + \frac{1}{2}\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top, (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle.
\end{aligned} \tag{C.3}
$$

Substituting (C.3) into (C.2), we have

$$
\begin{aligned}
\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V}) \geq{}& \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle + \frac{1}{2}\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top, (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle \\
&+ \frac{\mu}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2.
\end{aligned} \tag{C.4}
$$

Now, by Lemma C.2, we have

$$
\frac{\mu}{2}\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2 \geq 0.4\mu\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V}). \tag{C.5}
$$

Since we have $D(\mathbf{U},\mathbf{U}^*) \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$ and $D(\mathbf{V},\mathbf{U}^*) \leq \sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)$, then by Lemma C.3, we have

$$
\begin{aligned}
&\langle \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top, (\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle \\
&\geq -\frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2,
\end{aligned} \tag{C.6}
$$

Substituting (C.5) and (C.6) into (C.4), we have

$$
\begin{aligned}
\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V}) \geq{}& \langle \nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2 \\
&+ 0.4\mu\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V}),
\end{aligned} \tag{C.7}
$$

which completes the proof. □

C.3 Proof of Lemma B.3

Proof of Lemma B.3. By the definition of $\Pi_{\mathbf{U}_0}(\mathbf{V})$, we consider

$$\mathbf{R} = \mathop{\mathrm{argmin}}_{\mathbf{P}\in\mathbb{R}^{r\times r},\mathbf{P}^\top\mathbf{P}=\mathbf{I}} \|\mathbf{V}\mathbf{P}-\mathbf{U}_0\|_F = \mathop{\mathrm{argmin}}_{\mathbf{P}\in\mathbb{R}^{r\times r},\mathbf{P}^\top\mathbf{P}=\mathbf{I}} \|\mathbf{V}\mathbf{P}-\mathbf{U}_0\|_F = \mathop{\mathrm{argmax}}_{\mathbf{P}\in\mathbb{R}^{r\times r},\mathbf{P}^\top\mathbf{P}=\mathbf{I}} \mathrm{trace}(\mathbf{P}^\top\mathbf{V}^\top\mathbf{U}_0).$$

Suppose that $\mathbf{V}^\top\mathbf{U}_0$ has singular value decomposition $\mathbf{V}^\top\mathbf{U}_0 = \mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top$. Then we have

$$\mathbf{R} = \mathop{\mathrm{argmax}}_{\mathbf{P}\in\mathbb{R}^{r\times r},\mathbf{P}^\top\mathbf{P}=\mathbf{I}} \mathrm{trace}(\mathbf{P}^\top\mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top) = \mathop{\mathrm{argmax}}_{\mathbf{P}\in\mathbb{R}^{r\times r},\mathbf{P}^\top\mathbf{P}=\mathbf{I}} \mathrm{trace}(\mathbf{B}^\top\mathbf{P}^\top\mathbf{A}\mathbf{\Sigma}) = \mathbf{A}\mathbf{B}^\top.$$

Therefore

$$[\Pi_{\mathbf{U}_0}(\mathbf{V})]^\top\mathbf{U}_0 = \mathbf{B}\mathbf{A}^\top\mathbf{V}^\top\mathbf{U}_0 = \mathbf{B}\mathbf{A}^\top\mathbf{A}\mathbf{\Sigma}\mathbf{B}^\top = \mathbf{B}\mathbf{\Sigma}\mathbf{B}^\top \succeq 0.$$

This completes the proof that $\Pi_{\mathbf{U}_0}(\mathbf{V}) \in \Omega(\mathbf{U}_0)$. For any $\mathbf{X} \in \Omega(\mathbf{U}_0)$, it follows that $\mathbf{X} = \Pi_{\mathbf{U}_0}(\mathbf{X})$, and therefore $D(\mathbf{X},\mathbf{U}_0) = \|\Pi_{\mathbf{U}_0}(\mathbf{X}) - \mathbf{U}_0\|_F = \|\mathbf{X}-\mathbf{U}_0\|_F$. □

C.4 Proof of Lemma B.4

Proof of Lemma B.4. First, we have

$$\gamma = c\frac{\mu\sigma_r^6(\mathbf{U}_0)}{\|\mathbf{U}_0\|_2^4} \le \frac{\mu\sigma_r^6(\mathbf{U}^*)}{200\|\mathbf{U}^*\|_2^4} = \bar{\mu},$$

where the inequality holds due to Lemma A.4. Then we have the following inequalities:

$$\begin{aligned}
&\big\|\nabla\mathcal{L}(\mathbf{U}_0\mathbf{U}_0^\top) - \nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top)\big\|_2 \\
&\le \big\|\nabla\mathcal{L}(\mathbf{U}_0\mathbf{U}_0^\top) - \nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top)\big\|_F \\
&\le L\big\|\mathbf{U}_0\mathbf{U}_0^\top - \mathbf{U}^*(\mathbf{U}^*)^\top\big\|_F \\
&\le L(\|\mathbf{U}_0-\mathbf{U}^*\|_F^2 + \|\mathbf{U}_0-\mathbf{U}^*\|_F\|\mathbf{U}^*\|_2) \\
&\le 2L\|\mathbf{U}^*\|_2^2, \qquad \text{(C.8)}
\end{aligned}$$

where the first inequality holds due to the fact $\|\cdot\|_2 \le \|\cdot\|_F$, the second inequality holds due to the restricted smoothness assumption on $\mathcal{L}$, the third inequality holds due to triangle inequality, the fourth inequality holds due to the fact that $\|\mathbf{U}_0-\mathbf{U}^*\|_F = C_{\mathrm{ini}} \le \|\mathbf{U}^*\|_2$. Thus, we have

$$\eta = \frac{c}{L\|\mathbf{U}_0\|_2^2 + \|\nabla\mathcal{L}(\mathbf{U}_0\mathbf{U}_0^\top)\|_2} \le \frac{1}{12L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top)\|_2},$$

where the inequality holds due to Lemma A.4 and (C.8). Next, we have

$$\begin{aligned}
C_{\mathrm{R}} &= c\frac{\mu\sigma_r^7(\mathbf{U}^*)}{L\|\mathbf{U}^*\|_2^6 + \|\mathbf{U}^*\|_2^4\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2} \\
&\le \frac{\mu\sigma_r^7(\mathbf{U}^*)}{84000L\|\mathbf{U}^*\|_2^6 + 20000\|\mathbf{U}^*\|_2^4\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2} \\
&= C_{\mathrm{convex}}/2,
\end{aligned}$$

and

$$C_{\mathrm{R}} \le c\frac{\mu\sigma_r^7(\mathbf{U}^*)}{L\|\mathbf{U}^*\|_2^6} \le \frac{\sigma_r^2(\mathbf{U}^*)}{10\|\mathbf{U}^*\|_2},$$

since $c \le 84000^{-1}$. Finally, we have

$$C_{\mathrm{ini}} = c^2\frac{\mu^{3/2}\sigma_r^{10}(\mathbf{U}^*)}{\widehat{L}^{3/2}\|\mathbf{U}^*\|_2^9} \le c\frac{\mu\sigma_r^7(\mathbf{U}^*)}{\widehat{L}\|\mathbf{U}^*\|_2^6} = C_{\mathrm{R}},$$

since $\mu \leq L$ and $\sigma_r(\mathbf{U}^*) \leq \|\mathbf{U}^*\|_2$. We also have

$$\begin{aligned}
C_{\mathrm{R}}^2 &= c^{-2}\frac{\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2}{\mu\sigma_r^2(\mathbf{U}^*)} \cdot \frac{\|\mathbf{U}^*\|_2^4}{\sigma_r^4(\mathbf{U}^*)} \\
&\geq c^{-2}\frac{\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2}{\mu\sigma_r^2(\mathbf{U}^*)} \\
&\geq \frac{3}{\mu\sigma_r^2(\mathbf{U}^*)}\big(4\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2 + \epsilon\big) \\
&\geq \frac{3}{\mu\sigma_r^2(\mathbf{U}^*)}[(\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon],
\end{aligned}$$

where the first inequality holds since $\|\mathbf{U}^*\|_2 \geq \sigma_r(\mathbf{U}^*)$, the second inequality holds due to $c \leq 84000^{-1}$ and the assumption of ϵ. We also have

$$\begin{aligned}
C_{\mathrm{R}}^2 &= c^{-1}\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2 \frac{\|\mathbf{U}^*\|_2^4}{c\mu\sigma_r^6(\mathbf{U}^*)} \\
&\geq \frac{c^{-1}}{\gamma}\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2 \\
&\geq \frac{2}{\gamma}\big(4\widehat{L}\|\mathbf{U}^*\|_2^2 C_{\mathrm{ini}}^2 + \epsilon\big) \\
&\geq \frac{2}{\gamma}[(\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top}\|_2 + 4L\|\mathbf{U}^*\|_2^2)C_{\mathrm{ini}}^2 + \epsilon],
\end{aligned}$$

where the first inequality holds due to $c \leq 84000^{-1}$ and Lemma A.4, the second inequality holds due to the assumption of ϵ, the third inequality holds due to the definition of $\widehat{L}$. That completes our proof. □

C.5 Proof of Lemma B.5

Proof of Lemma B.5. By Lemma B.2, we have

$$\begin{aligned}
\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{V}) \geq \langle\nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle - \frac{21L\|\mathbf{U}^*\|_2^2 + 5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)} D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2 \\
+ 0.4\mu\sigma_r^2(\mathbf{V})D^2(\mathbf{U},\mathbf{V}). \qquad \text{(C.9)}
\end{aligned}$$

Note that $\nabla\mathcal{G}(\mathbf{U}^*) = 0$ and $D(\mathbf{U}^*,\mathbf{U}^*) = 0$. Taking $\mathbf{V} = \mathbf{U}^*$ in (C.9), we have

$$\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{U}^*) \geq 0.4\mu\sigma_r^2(\mathbf{U}^*)D(\mathbf{U},\mathbf{U}^*)^2,$$

This finishes the proof. □

C.6 Proof of Lemma B.6

Lemma C.4 (Li and Lin (2017))**.** For any $\mathbf{U}$ and $\mathbf{V}$, we have

$$\mathcal{G}(\mathbf{U}) \leq \mathcal{G}(\mathbf{V}) + \langle\nabla\mathcal{G}(\mathbf{V}), \mathbf{U}-\mathbf{V}\rangle + \frac{2\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2 + L(\|\mathbf{V}\|_2 + \|\mathbf{U}\|_2)^2}{2}\|\mathbf{U}-\mathbf{V}\|_F^2. \qquad \text{(C.10)}$$

Proof of Lemma B.6. Let $\mathbf{U} = \mathbf{V} - \eta\nabla\mathcal{G}(\mathbf{V})$, then by Lemma C.4, we have

$$\begin{aligned}
&\mathcal{G}(\mathbf{V} - \eta\nabla\mathcal{G}(\mathbf{V})) \\
&\leq \mathcal{G}(\mathbf{V}) - \eta\|\nabla\mathcal{G}(\mathbf{V})\|_F^2 + \frac{2\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2 + L(\|\mathbf{V}\|_2 + \|\mathbf{V} - \eta\nabla\mathcal{G}(\mathbf{V})\|_2)^2}{2}\eta^2\|\nabla\mathcal{G}(\mathbf{V})\|_F^2 \\
&\leq \mathcal{G}(\mathbf{V}) - \eta\|\nabla\mathcal{G}(\mathbf{V})\|_F^2 + \frac{2\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2 + L(2\|\mathbf{V}\|_2 + \eta\|\nabla\mathcal{G}(\mathbf{V})\|_2)^2}{2}\eta^2\|\nabla\mathcal{G}(\mathbf{V})\|_F^2. \qquad \text{(C.11)}
\end{aligned}$$

Next we bound $\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2$, $\|\mathbf{V}\|_2$ and $\eta\|\nabla\mathcal{G}(\mathbf{V})\|_2$ separately. By Lemma A.4, we have $\|\mathbf{V}\|_2 \le 1.1\|\mathbf{U}^*\|_2$. By Lemma B.8, we have

$$\eta\|\nabla\mathcal{G}(\mathbf{V})\|_2 \le \eta D(\mathbf{V},\mathbf{U}^*)(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2) \le D(\mathbf{V},\mathbf{U}^*) \le 0.1\|\mathbf{U}^*\|_2, \tag{C.12}$$

where the second inequality holds since $\eta(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2) < 1$. To bound $\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2$, we have

$$\begin{aligned}
\|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)\|_2 &\le \|\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) - \nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)[\Pi_{\mathbf{V}}(\mathbf{U}^*)]^\top)\|_2 + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 \\
&\le L\|\mathbf{V}\mathbf{V}^\top - \Pi_{\mathbf{V}}(\mathbf{U}^*)[\Pi_{\mathbf{V}}(\mathbf{U}^*)]^\top\|_F + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 \\
&\le L\|\mathbf{V}(\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*))^\top\|_F + L\|(\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*))[\Pi_{\mathbf{V}}(\mathbf{U}^*)]^\top\|_F + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 \\
&\le L(\|\mathbf{V}\|_2 + \|\mathbf{U}^*\|_2)D(\mathbf{V},\mathbf{U}^*) + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 \\
&\le 0.21L\|\mathbf{U}^*\|_2^2 + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2,
\end{aligned} \tag{C.13}$$

where the second inequality holds due to the L-smoothness of $\mathcal{L}$, the fourth inequality holds due to $D(\mathbf{V},\mathbf{U}^*) = \|\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F$, the last inequality holds due to (C.12) and $D(\mathbf{V},\mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2 \le 0.1\|\mathbf{U}^*\|_2$. Substituting (C.12) and (C.13) into (C.11), we have

$$\begin{aligned}
&\mathcal{G}(\mathbf{V} - \eta\nabla\mathcal{G}(\mathbf{V})) \\
&\le \mathcal{G}(\mathbf{V}) - \eta\|\nabla\mathcal{G}(\mathbf{V})\|_F^2 \\
&\quad + \frac{1}{2}\eta^2\|\nabla\mathcal{G}(\mathbf{V})\|_F^2 \cdot (0.42L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + L(2.2\|\mathbf{U}^*\|_2 + 0.1\|\mathbf{U}^*\|_2)^2) \\
&\le \mathcal{G}(\mathbf{V}) - \|\nabla\mathcal{G}(\mathbf{V})\|_F^2[\eta - \eta^2(6L\|\mathbf{U}^*\|_2^2 + \|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2)] \\
&= \mathcal{G}(\mathbf{V}) - \eta\|\nabla\mathcal{G}(\mathbf{V})\|_F^2/2,
\end{aligned} \tag{C.14}$$

where the last equality holds because of the definition of η. Finally, by (C.12), we have

$$D(\mathbf{V}',\mathbf{U}^*) \le D(\mathbf{V},\mathbf{U}^*) + \|\eta\nabla\mathcal{G}(\mathbf{V})\|_F \le D(\mathbf{V},\mathbf{U}^*) + \|\eta\nabla\mathcal{G}(\mathbf{V})\|_2 \le 2D(\mathbf{V},\mathbf{U}^*),$$

where the second inequality holds due to $\|\eta\nabla\mathcal{G}(\mathbf{V})\|_2 \le D(\mathbf{V},\mathbf{U}^*)$ from (C.12). This completes the proof.

□

C.7 Proof of Lemma B.7

Proof of Lemma B.7. Let $\mathbf{U}' = \Pi_{\mathbf{U}^*}(\mathbf{U})$ and $\mathbf{V} = \mathbf{U}^*$, then by Lemma C.4, we have

$$\begin{aligned}
\mathcal{G}(\mathbf{U}) - \mathcal{G}(\mathbf{U}^*) &\le \frac{2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2 + L(\|\mathbf{U}^*\|_2 + \|\mathbf{U}'\|_2)^2}{2}\|\mathbf{U}' - \mathbf{U}^*\|_F^2 \\
&\le (\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top}) + 3L\|\mathbf{U}^*\|_2^2)D^2(\mathbf{U},\mathbf{U}^*),
\end{aligned}$$

where the second inequality follows by Lemmas B.3 and A.4. This completes the proof. □

C.8 Proof of Lemma B.8

Proof of Lemma B.8. We have

$$
\begin{aligned}
&\|\nabla\mathcal{G}(\mathbf{V})\|_F \\
&= \|\nabla\mathcal{G}(\mathbf{V}) - \nabla\mathcal{G}(\Pi_{\mathbf{V}}(\mathbf{U}^*))\|_F \\
&= \|(\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top)\mathbf{V} \\
&\quad - (\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top) + \nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)^\top)\Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F \\
&\le \|[\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top) + \nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top - \nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top) - \nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)^\top]\mathbf{V}\|_F \\
&\quad + \|(\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top) + \nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)^\top)(\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*))\|_F \\
&\le 4L\|\mathbf{V}\mathbf{V}^\top - \Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top\|_F\|\mathbf{V}\|_2 + 2\|\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)\|_2\|\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F \\
&\le 4L\|\mathbf{V}(\mathbf{V} - \mathbf{U}^*)^\top\|_F\|\mathbf{V}\|_2 + 4L\|(\mathbf{V} - \mathbf{U}^*)\mathbf{U}^{*\top}\|_F\|\mathbf{V}\|_2 \\
&\quad + 2\|\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)\|_2\|\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F \\
&\le 4L(\|\mathbf{U}^*\|_2 + \|\mathbf{V}\|_2)\|\mathbf{V}\|_2\|\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F + 2\|\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)\|_2\|\mathbf{V} - \Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F \\
&\le D(\mathbf{V}, \mathbf{U}^*)(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\Pi_{\mathbf{V}}(\mathbf{U}^*)\Pi_{\mathbf{V}}(\mathbf{U}^*)^\top)\|_2) \\
&= D(\mathbf{V}, \mathbf{U}^*)(10L\|\mathbf{U}^*\|_2^2 + 2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2),
\end{aligned}
$$

where the first equality holds due to $\nabla\mathcal{G}(\Pi_{\mathbf{V}}(\mathbf{U}^*)) = \mathbf{0}$, the second inequality holds due to the L-smoothness of $\mathcal{L}$, the last inequality holds due to Lemma A.4 with the condition that $D(\mathbf{V}, \mathbf{U}^*) \le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$. □

C.9 Proof of Lemma B.9

Proof of Lemma B.9. We have

$$
\begin{aligned}
&\frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}'_{k+1}\|_F^2 \\
&= \frac{\gamma}{2}\bigg\|\mathbf{U}^* - (1-\alpha)\mathbf{V}_k - \alpha\mathbf{Y}_k + \frac{\alpha}{\gamma}\nabla\mathcal{G}(\mathbf{Y}_k)\bigg\|_F^2 \\
&= \frac{\gamma}{2}\bigg\|(\mathbf{U}^* - \mathbf{V}_k) + \alpha(\mathbf{V}_k - \mathbf{Y}_k) + \frac{\alpha}{\gamma}\nabla\mathcal{G}(\mathbf{Y}_k)\bigg\|_F^2 \\
&= \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \frac{\alpha^2\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2 + \alpha\gamma\langle\mathbf{U}^* - \mathbf{V}_k, \mathbf{V}_k - \mathbf{Y}_k\rangle \\
&\quad + \alpha\langle\mathbf{U}^* - \mathbf{V}_k, \nabla\mathcal{G}(\mathbf{Y}_k)\rangle + \alpha^2\langle\mathbf{V}_k - \mathbf{Y}_k, \nabla\mathcal{G}(\mathbf{Y}_k)\rangle \\
&= \frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \frac{\alpha^2\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + \frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2 + \alpha\gamma\langle\mathbf{U}^* - \mathbf{V}_k, \mathbf{V}_k - \mathbf{Y}_k\rangle \\
&\quad + \alpha\langle\mathbf{U}^* - \mathbf{Y}_k, \nabla\mathcal{G}(\mathbf{Y}_k)\rangle - \alpha(1-\alpha)\langle\mathbf{V}_k - \mathbf{Y}_k, \nabla\mathcal{G}(\mathbf{Y}_k)\rangle.
\end{aligned} \tag{C.15}
$$

where the first equality follows by the definition of $\mathbf{V}'_{k+1}$ in (3.4), the third equality is by expanding the square, and the last equality is obtained by rearranging terms. Furthermore, we have

$$
\begin{aligned}
&\frac{\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \frac{\alpha^2\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + \alpha\gamma\langle\mathbf{U}^* - \mathbf{V}_k, \mathbf{V}_k - \mathbf{Y}_k\rangle \\
&= \frac{\gamma(1-\alpha) + \alpha\gamma}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \frac{\alpha^2\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + \alpha\gamma\langle\mathbf{U}^* - \mathbf{V}_k, \mathbf{V}_k - \mathbf{Y}_k\rangle \\
&= \frac{\gamma(1-\alpha)}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 - \frac{\alpha(1-\alpha)\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 \\
&\quad + \frac{\alpha\gamma}{2}\Big[\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 + \|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + 2\langle\mathbf{U}^* - \mathbf{V}_k, \mathbf{V}_k - \mathbf{Y}_k\rangle\Big] \\
&= \frac{\gamma(1-\alpha)}{2}\|\mathbf{U}^* - \mathbf{V}_k\|_F^2 - \frac{\alpha(1-\alpha)\gamma}{2}\|\mathbf{V}_k - \mathbf{Y}_k\|_F^2 + \frac{\alpha\gamma}{2}\|\mathbf{U}^* - \mathbf{Y}_k\|_F^2.
\end{aligned} \tag{C.16}
$$

Substituting (C.16) into (C.15) and rearranging it, we have

$$\frac{\gamma}{2}\|\mathbf{U}^*-\mathbf{V}'_{k+1}\|_F^2 \le \frac{\gamma(1-\alpha)}{2}\|\mathbf{U}^*-\mathbf{V}_k\|_F^2+\alpha\Big(\langle\nabla\mathcal{G}(\mathbf{Y}_k),\mathbf{U}^*-\mathbf{Y}_k\rangle+\frac{\gamma}{2}\|\mathbf{U}^*-\mathbf{Y}_k\|_F^2\Big)$$
$$+\frac{\alpha^2}{2\gamma}\|\nabla\mathcal{G}(\mathbf{Y}_k)\|_F^2-\alpha(1-\alpha)\langle\nabla\mathcal{G}(\mathbf{Y}_k),\mathbf{V}_k-\mathbf{Y}_k\rangle.$$

The conclusion holds. □

C.10 Proof of Lemma A.4

Proof of Lemma A.4. Since that $D(\mathbf{U},\mathbf{U}^*)\le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2\le 0.1\sigma_r(\mathbf{U}^*)$, then we have (A.1) and (A.2) by the fact that $\|\mathbf{A}\|_2-\|\mathbf{B}\|_F\le\|\mathbf{A}+\mathbf{B}\|_2\le\|\mathbf{A}\|_2+\|\mathbf{B}\|_F$ and $\sigma_r(\mathbf{A})-\|\mathbf{B}\|_F\le\sigma_r(\mathbf{A}+\mathbf{B})\le\sigma_r(\mathbf{A})+\|\mathbf{B}\|_F$ with $\mathbf{A}=\mathbf{U}^*$ and $\mathbf{B}=\Pi_{\mathbf{U}^*}(\mathbf{U})-\mathbf{U}^*$. To prove (A.3), note that

$$\begin{aligned}\|\Pi_{\mathbf{U}^*}(\mathbf{U})^\top\Pi_{\mathbf{U}^*}(\mathbf{V})-\mathbf{U}^{*\top}\mathbf{U}^*\|_F &= \|\Pi_{\mathbf{U}^*}(\mathbf{U})^\top\Pi_{\mathbf{U}^*}(\mathbf{V})-\Pi_{\mathbf{U}^*}(\mathbf{U})^\top\mathbf{U}^*+\Pi_{\mathbf{U}^*}(\mathbf{U})^\top\mathbf{U}^*-\mathbf{U}^{*\top}\mathbf{U}^*\|_F\\ &\le \|\Pi_{\mathbf{U}^*}(\mathbf{U})\|_2\|\Pi_{\mathbf{U}^*}(\mathbf{V})-\mathbf{U}^*\|_F+\|\mathbf{U}^*\|_2\|\Pi_{\mathbf{U}^*}(\mathbf{U})-\mathbf{U}^*\|_F\\ &= \|\Pi_{\mathbf{U}^*}(\mathbf{U})\|_2 D(\mathbf{V},\mathbf{U}^*)+\|\mathbf{U}^*\|_2 D(\mathbf{U},\mathbf{U}^*)\\ &\le 2.1\|\mathbf{U}^*\|_2(0.1\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)/\|\mathbf{U}^*\|_2)\\ &\le 0.3\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*),\end{aligned}$$

where the equality holds due to $D(\mathbf{V},\mathbf{U}^*)=\|\Pi_{\mathbf{U}^*}(\mathbf{V})-\mathbf{U}^*\|_F$ and $D(\mathbf{U},\mathbf{U}^*)=\|\Pi_{\mathbf{U}^*}(\mathbf{U})-\mathbf{U}^*\|_F$. Thus, we have $0.7\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)\le\sigma_r(\Pi_{\mathbf{U}^*}(\mathbf{U})^\top\Pi_{\mathbf{U}^*}(\mathbf{V}))\le 1.3\sigma_r(\mathbf{U}^{*\top}\mathbf{U}^*)$, which implies (A.3). □

D Proofs of Lemmas in Section C

D.1 Proof of Lemma C.3

We need the following lemma:

Lemma D.1 (Li and Lin (2017))**.** Suppose that $\mathbf{Z}\in\mathbb{R}^{d\times r}$ is of full rank, $\|\mathbf{U}-\mathbf{Z}\|_F\le 0.01\sigma_r(\mathbf{Z})$ and $\|\mathbf{V}-\mathbf{Z}\|_F\le 0.01\sigma_r(\mathbf{Z})$, then

$$|\langle\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)+\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top,(\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle|\le 2.08\frac{\|\nabla\mathcal{G}(\mathbf{V})\|_2}{\sigma_r(\mathbf{Z})}\|\mathbf{U}-\mathbf{V}\|_F^2, \tag{D.1}$$

Proof of Lemma C.3. We have $D(\mathbf{U},\mathbf{U}^*)\le\sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)\le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ and $D(\mathbf{V},\mathbf{U}^*)\le\sigma_r^3(\mathbf{U}^*)/(500\|\mathbf{U}^*\|_2^2)\le 0.1\sigma_r^2(\mathbf{U}^*)/\|\mathbf{U}^*\|_2$ with $\mathbf{U},\mathbf{V},\mathbf{U}^*\in\Omega(\mathbf{U}_0)$. Then by Lemma B.1, we have

$$\|\mathbf{U}-\mathbf{U}^*\|_F\le\frac{5\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2}D(\mathbf{U},\mathbf{U}^*)\le 0.01\sigma_r(\mathbf{U}^*),$$
$$\|\mathbf{V}-\mathbf{U}^*\|_F\le\frac{5\|\mathbf{U}^*\|_2^2}{\sigma_r(\mathbf{U}^*)^2}D(\mathbf{V},\mathbf{U}^*)\le 0.01\sigma_r(\mathbf{U}^*).$$

Thus, by Lemma D.1 with $\mathbf{Z}=\mathbf{U}^*$, we have

$$|\langle\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)+\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top,(\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle|\le 2.08\frac{\|\nabla\mathcal{G}(\mathbf{V})\|_2}{\sigma_r(\mathbf{U}^*)}\|\mathbf{U}-\mathbf{V}\|_F^2. \tag{D.2}$$

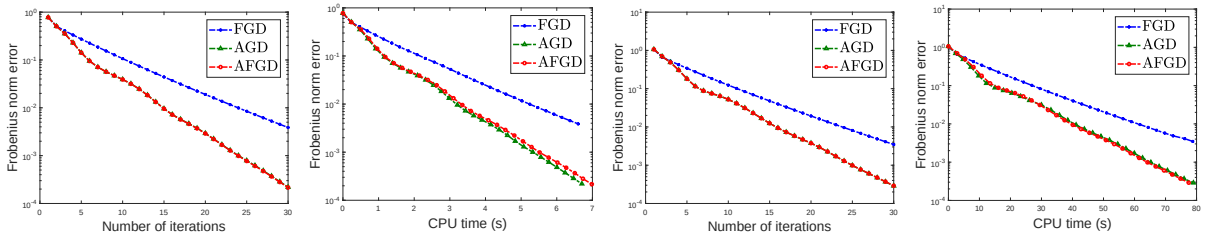
By Lemma B.8, we have

$$\|\nabla\mathcal{G}(\mathbf{V})\|_2\le\|\mathbf{V}-\Pi_{\mathbf{V}}(\mathbf{U}^*)\|_F(10L\|\mathbf{U}^*\|_2^2+2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2). \tag{D.3}$$

Substituting (D.3) into (D.2), we have

$$\begin{aligned}&|\langle\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)+\nabla\mathcal{L}(\mathbf{V}\mathbf{V}^\top)^\top,(\mathbf{U}-\mathbf{V})(\mathbf{U}-\mathbf{V})^\top\rangle|\\ &\le 2.08\cdot\frac{10L\|\mathbf{U}^*\|_2^2+2\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)}\cdot D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2\\ &\le\frac{21L\|\mathbf{U}^*\|_2^2+5\|\nabla\mathcal{L}(\mathbf{U}^*\mathbf{U}^{*\top})\|_2}{\sigma_r(\mathbf{U}^*)}\cdot D(\mathbf{V},\mathbf{U}^*)\|\mathbf{U}-\mathbf{V}\|_F^2,\end{aligned} \tag{D.4}$$

which completes the proof. □



(a) $d = 100, r = 3$ (b) $d = 100, r = 3$ (c) $d = 200, r = 5$ (d) $d = 200, r = 5$

Figure 2: Comparison of FGD, AGD and AFGD for matrix regression. Plots of squared error $\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}^*\|_F^2$ versus number of iterations ((a) and (c)) and CPU time ((b) and (d)).

E Additional Experiments

E.1 Matrix Regression

In matrix regression, we aim to estimate a low-rank positive semidefinite matrix $\mathbf{M}^* \in \mathbb{R}^{d\times d}$ from a set of measurements $\mathbf{y} = \mathcal{A}(\mathbf{M}^*) \in \mathbb{R}^n$, where $\mathcal{A}$ is a linear operator defined as $\mathcal{A}(\mathbf{M}) = (\langle \mathbf{A}_1, \mathbf{M}\rangle, \ldots, \langle \mathbf{A}_n, \mathbf{M}\rangle)^\top$. Then matrix regression is formulated as:

$$\min_{\mathbf{M}\succeq \mathbf{0},\ \mathrm{rank}(\mathbf{M})\leq r} \frac{1}{2}\|\mathcal{A}(\mathbf{M}) - \mathbf{y}\|_2^2.$$

By using the idea of matrix factorization, we rewrite $\mathbf{M}$ as $\mathbf{M} = \mathbf{U}\mathbf{U}^\top$ and solve the following low-rank matrix factorization problem:

$$\min_{\mathbf{U}\in\mathbb{R}^{d\times r}} \frac{1}{2}\|\mathcal{A}(\mathbf{U}\mathbf{U}^\top) - \mathbf{y}\|_2^2.$$

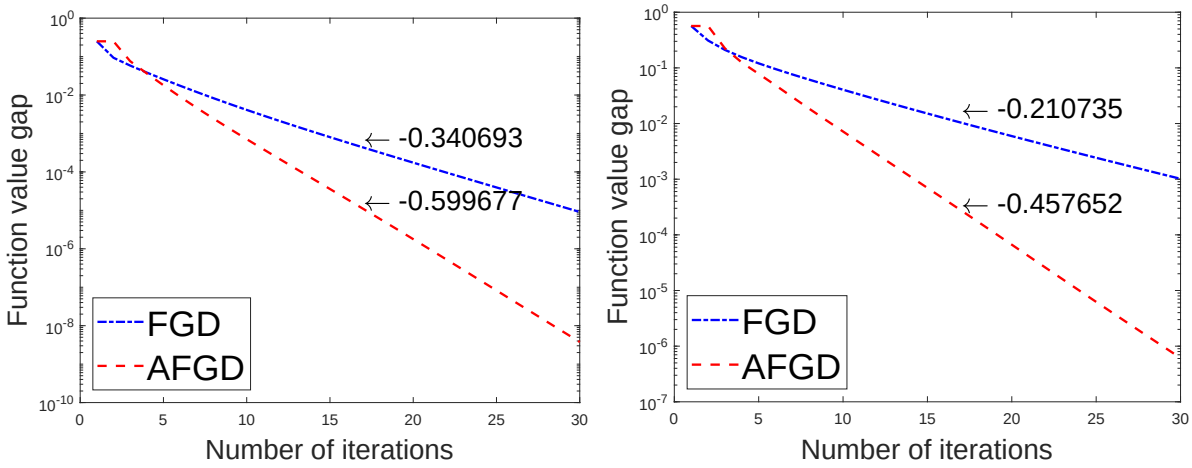
We consider the following two settings: $d = 100, r = 3$ and $d = 200, r = 10$. We set the number of measurements $n = 10dr$. We first generate $\mathbf{U}^* \in \mathbb{R}^{d\times r}$ from standard Gaussian distribution, then we generate the unknown matrix $\mathbf{M}^* = \mathbf{U}^*(\mathbf{U}^*)^\top$. Next, we generate noiseless measurements based on the observation model $y_i = \langle \mathbf{A}_i, \mathbf{M}^*\rangle$, where each entry of $\mathbf{A}_i$ follows i.i.d. standard Gaussian distribution. All algorithms use the same initialization method proposed in Bhojanapalli et al. (2015) (See Appendix F) for more details. All results are averaged over 20 trials. To illustrate the linear convergence rate, we report the squared error $\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}^*\|_F^2$ in logarithmic scale. We compare different algorithms in terms of the squared error with respect to number of iterations and CPU time, and plot them in Figure 2. The results for $d = 100, r = 3$ are presented in Figures 2(a) and 2(b), while the results for $d = 200, r = 10$ are shown in Figures 2(c) and 2(d).

It can be seen that AFGD converges faster than FGD in both number of iterations and CPU time, which validates our theory. We also observe that vanilla AGD performs almost the same as AFGD in both settings. This suggests while theoretically not justified, vanilla AGD might still have an accelerated convergence rate for solving (1.2). This leaves an open question that whether and how we can prove the accelerated convergence rate of vanilla AGD for low-rank matrix factorization. In sharp contrast, our proposed AFGD enjoys an accelerated convergence rate both in theory and practice, and its projection step (solved by Algorithm 2) does not introduce significant computational overhead.

We plot the function value gap $\mathcal{G}(\mathbf{X}_k) - \mathcal{G}^*$ in logarithmic scale v.s. iteration number in Figure 3 for AFGD and FGD. We also highlight the slopes for different lines in the plot. For Figure 3(a), by calculation we have $L/\mu = 2.793$. The ratio between the slopes of AFGD line and FGD line is $0.6/0.34 \approx 1.73 \approx \sqrt{2.793}$, which validates that the acceleration is indeed by a factor of $\sqrt{L/\mu}$. Similar calculation for Figure 3(b) also validates the $\sqrt{L/\mu}$ acceleration. This verifies our theoretical result, since under our experimental setup we expect $\nabla\mathcal{L}(\mathbf{U}^*(\mathbf{U}^*)^\top) = 0$ and therefore $\widehat{L} = L$.

E.2 Matrix Completion

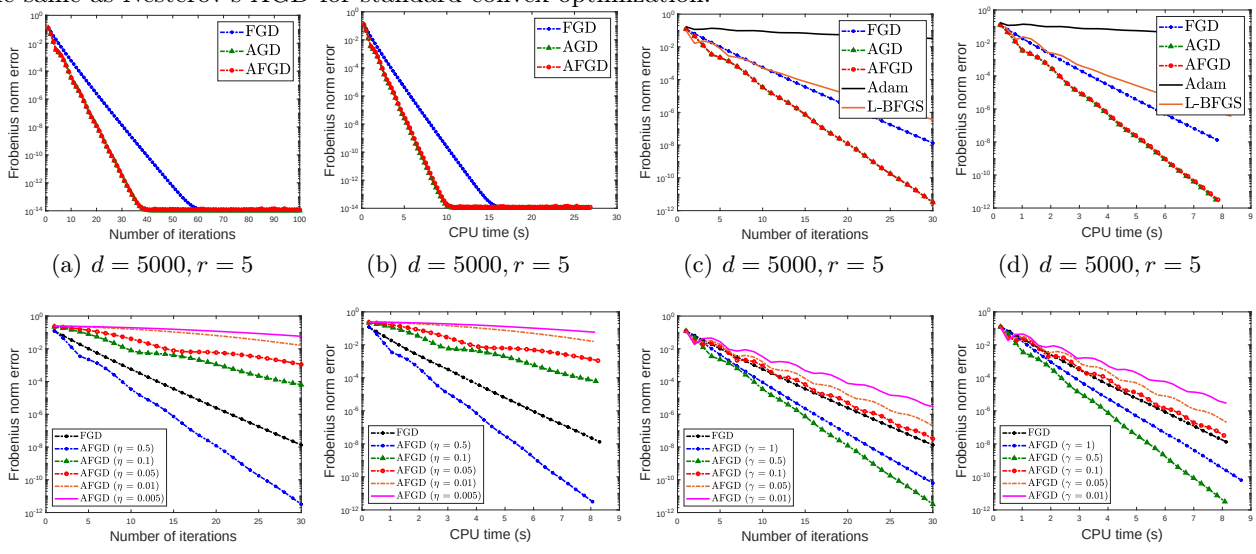
We have additional experiments for matrix completion where $d = 5000, r = 5$. We report the squared error $\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}^*\|_F^2$ in logarithmic scale and plot them in Figure 4. Figures 4(a) and 4(b) show the convergence



(a) $d = 100, r = 3$ (b) $d = 200, r = 5$

Figure 3: Plots of logarithmic function value gap v.s. iteration number for matrix regression with slope calculation.

results when $T = 100$. We can see that AGD and AFGD still outperform FGD. We also compare with Adam (Kingma and Ba, 2014) and L-BFGS (Byrd et al., 1994) in Figures 4(c) and 4(d). For Adam, we set its parameters $\beta_1 = 0.9, \beta_2 = 0.999, \eta = 0.001$, as suggested by Kingma and Ba (2014). For L-BFGS, we use the recent 10 iterates and gradients to construct an approximate Hessian. It can be seen that L-BFGS performs similar to FGD, while Adam performs the worst. Figures 4(e) - 4(h) show the convergence results for AFGD with different choices of η and γ. It can be seen that the performance of AFGD is sensitive on the choice of η and γ, which is the same as Nesterov's AGD for standard convex optimization.



(a) $d = 5000, r = 5$ (b) $d = 5000, r = 5$ (c) $d = 5000, r = 5$ (d) $d = 5000, r = 5$

(e) $d = 5000, r = 5$ (f) $d = 5000, r = 5$ (g) $d = 5000, r = 5$ (h) $d = 5000, r = 5$

Figure 4: Plots of squared error $\|\mathbf{U}\mathbf{U}^\top - \mathbf{M}^*\|_F^2$ v.s. iteration number and CPU time for matrix completion.

F Initialization Method

For the self-containedness, we present the initialization method used in our experiments in Algorithm 3, which is originally proposed in Bhojanapalli et al. (2015). Here $\mathcal{L}$ is the loss function defined in (1.1), L is the smoothness parameter and $\mathcal{P}_+(\cdot)$ denotes the projection of a given matrix onto the PSD cone.

Algorithm 3 Initialization

Require: Function $\mathcal{L}$, smoothness parameter L.

1: $\mathbf{M}_0 = \mathbf{0}$

2: $\mathbf{M}^+ = L^{-1}\mathcal{P}_+(-\nabla\mathcal{L}(\mathbf{M}_0))$

Ensure: $\mathbf{M}^+$