Supplementary Materials to "A Non-Euclidean Gradient Descent Framework for Non-Convex Matrix Factorization"

Ya-Ping Hsieh, Yu-Chun Kao, Rabeeh Karimi Mahabadi, Alp Yurtsever, Anastasios Kyrillidis, *Member, IEEE,* and Volkan Cevher, *Senior Member, IEEE*

Appendix A
Proof of Sublinear Rate of Nuclear Gradient Descent

We first use following lemma to prove the sublinear rate.

Lemma 1. *For the sequence of the iterates* $\{U_i\}_{i=0}^{k}$*, we have*

$$f(U_iU_i^T) - f(U_{i+1}{U_{i+1}}^T) \geq \alpha_i \cdot \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \tag{A.1}$$

and

$$f(U_iU_i^T) - f(U^*{U^*}^T) \leq \beta_i \cdot \|\nabla f(X_i) \cdot U_i\|_{S_\infty} \tag{A.2}$$

where $\alpha_i = 1.117\ \eta_i$ *and* $\beta_i = (2 + \frac{19}{81})\|\Delta_{U_i}\|_* = (2 + \frac{19}{81})D_*(U_i, U^*)$.

Define $\delta_i = f(U_iU_i^T) - f(U^*{U^*}^T)$ and follow the previous lemma. We know $\{\delta_i\}$ is an positive decreasing sequence and

$$\begin{aligned}\delta_{i+1} &\leq \delta_i - \alpha_i \cdot \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \\ &\leq \delta_i - \frac{\alpha_i}{\beta_i^2} \cdot \delta_i^2\end{aligned}$$

Dividing both sides with $(\delta_i \cdot \delta_{i+1})$, we obtain, by assumption (III.5),

$$\frac{1}{\delta_{i+1}} - \frac{1}{\delta_i} \geq \frac{\alpha_i}{\beta_i^2} \cdot \frac{\delta_i}{\delta_{i+1}} \geq \frac{\alpha_i}{\beta_i^2} \geq \frac{\alpha_i}{\tilde{D}_*^2}.$$

Telescoping the inequality we get the desired result.
Now we prove (A.1) of lemma 1. The smoothness gives

$$\begin{aligned}&f(U_iU_i^T) - f(U_{i+1}{U_{i+1}}^T) \\ \geq&\langle \nabla f(X_i), X_i - X_{i+1}\rangle - \frac{L}{2}\|X_i - X_{i+1}\|_*^2 \\ =&\underbrace{\left\langle \nabla f(X_i), (U_i - U_{i+1})U_i^T + U_i(U_i - U_{i+1})^T \right\rangle}_{①} \\ &- \underbrace{\left\langle \nabla f(X_i), (U_i - U_{i+1})(U_i - U_{i+1})^T \right\rangle}_{②} \\ &- \frac{L}{2}\underbrace{\|X_i - X_{i+1}\|_*^2}_{③}.\end{aligned} \tag{A.3}$$

For ① we have

$$\begin{aligned}&\left\langle \nabla f(X_i), (U_i - U_{i+1})U_i^T + U_i(U_i - U_{i+1})^T \right\rangle \\ =&2\langle \nabla f(X_i)U_i, U_i - U_{i+1}\rangle \\ =&2\eta_i\left\langle \nabla f(X_i)U_i, [\nabla f(X_i)U_i]_\infty^\# \right\rangle \\ =&2\eta_i\|\nabla f(X_i)U_i\|_{S_\infty}^2.\end{aligned} \tag{A.4}$$

To upper bound ②, we use

$$\begin{aligned}
&\left\langle \nabla f(X_i), (U_i - U_{i+1})(U_i - U_{i+1})^T \right\rangle \\
=&\eta_i^2 \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \cdot Trace(\nabla f(X_i) A_1 A_1^T) \\
\leq&\eta_i^2 \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \cdot \|\nabla f(X_i)\|_{S_\infty} \\
\overset{(*)}{\leq}&\frac{1}{4}\eta_i \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2
\end{aligned} \tag{A.5}$$

in which A_1 is the first singular vector of $\nabla f(X_i) \cdot U_i$, and $(*)$ is by $\eta_i \leq \frac{1}{4\|\nabla f(X_i)\|_{S_\infty}}$.
To upper bound ③, we use

$$\begin{aligned}
&\|U_i U_i^T - U_{i+1} U_{i+1}^T\|_{S_\infty} \\
=&\|U_i (U_i - U_{i+1})^T + (U_i - U_{i+1}) U_i^T \\
&\qquad - (U_i - U_{i+1})(U_i - U_{i+1})^T\|_{S_\infty} \\
\leq&2\|U_i\|_{S_\infty} \|U_i - U_{i+1}\|_* + \|U_i - U_{i+1}\|_{S_\infty} \|U_i - U_{i+1}\|_* \\
=&2\eta_i \|U_i\|_{S_\infty} \|\nabla f(X_i) \cdot U_i\|_{S_\infty} + \eta_i^2 \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \\
=&\eta_i \ \|\nabla f(X_i) \cdot U_i\|_{S_\infty} \ [2\|U_i\|_{S_\infty} + \eta_i \|\nabla f(X_i) \cdot U_i\|_{S_\infty}] \\
\leq&\eta_i \ \|\nabla f(X_i) \cdot U_i\|_{S_\infty} \ [2\|U_i\|_{S_\infty} + \eta_i \|\nabla f(X_i)\|_{S_\infty} \|U_i\|_{S_\infty}] \\
\overset{(1)}{\leq}&\eta_i \ \|\nabla f(X_i) \cdot U_i\|_{S_\infty} \frac{9}{4} \|U_i\|_{S_\infty}
\end{aligned} \tag{A.6}$$

where (1) is by $\eta_i \leq \frac{1}{4\|\nabla f(X_i)\|_{S_\infty}}$.

Plugging above inequalities into (A.3), we obtain

$$\begin{aligned}
&f(U_i U_i^T) - f(U_{i+1} {U_{i+1}}^T) \\
\geq&2\eta_i \|\nabla f(X_i) U_i\|_{S_\infty}^2 - \frac{1}{4}\eta_i \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \\
&\quad - \frac{L}{2}(\frac{9}{4}\eta_i \ \|U_i\|_{S_\infty} \|\nabla f(X_i) \cdot U_i\|_{S_\infty})^2 \\
\geq&\eta_i \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2 \left[\frac{7}{4} - \frac{L}{2} \left(\frac{9}{4} \right)^2 \eta_i \|U_i\|_{S_\infty}^2 \right] \\
\overset{(*)}{\geq}&1.117 \ \eta_i \ \|\nabla f(X_i) \cdot U_i\|_{S_\infty}^2
\end{aligned} \tag{A.7}$$

where $(*)$ is by $\eta_i \leq \frac{1}{4L\|X_i\|_{S_\infty}} = \frac{1}{4L\|U_i\|_{S_\infty}^2}$. We have thus finished the first part of lemma 1.

Now we give the proof of (A.2) of lemma 1.
We denote

$$R_{U_i} \equiv \underset{R \text{ is unitary}}{\arg\min_R} \|U_i - U^* R\|_*. \tag{A.8}$$

and define $\Delta_{U_i} \equiv U_i - U^* R_{U_i}$. We begin with

$$\begin{aligned}
&f(U_i U_i^T) - f(U^* {U^*}^T) \\
\leq&\langle \nabla f(X_i), X_i - X^* \rangle \\
=&\left\langle \nabla f(X_i), \Delta_{U_i} U_i^T \right\rangle + \left\langle \nabla f(X_i), U_i \Delta_{U_i}^T \right\rangle \\
&- \left\langle \nabla f(X_i), \Delta_{U_i} \Delta_{U_i}^T \right\rangle \\
=&2\langle \nabla f(X_i) U_i, \Delta_{U_i} \rangle - \left\langle \nabla f(X_i), \Delta_{U_i} \Delta_{U_i}^T \right\rangle \\
\leq&2\|\nabla f(X_i) \cdot U_i\|_{S_\infty} \|\Delta_{U_i}\|_* + \underbrace{\left| \left\langle \nabla f(X_i), \Delta_{U_i} \Delta_{U_i}^T \right\rangle \right|}_{①}.
\end{aligned} \tag{A.9}$$

To upper bound ①, we use

$$
\begin{aligned}
&\langle \nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T \rangle \\
=&\langle \nabla f(X_i)\Delta_{U_i}, \Delta_{U_i} \rangle \\
\leq&\|\nabla f(X_i)\Delta_{U_i}\|_{S_\infty}\|\Delta_{U_i}\|_* \\
=&\|\nabla f(X_i)P_{\Delta_{U_i}}\Delta_{U_i}\|_{S_\infty}\|\Delta_{U_i}\|_* \\
\leq&\|\nabla f(X_i)P_{\Delta_{U_i}}\|_{S_\infty}\|\Delta_{U_i}\|_{S_\infty}\|\Delta_{U_i}\|_* \\
\overset{(*)}{\leq}&\Big(\|\nabla f(X_i)P_{U_i}\|_{S_\infty} + \|\nabla f(X_i)P_{U^*}\|_{S_\infty}\Big)\|\Delta_{U_i}\|_*^2
\end{aligned}
\tag{A.10}
$$

in which P_U denotes the projection onto $\mathbf{Col}(U)$. $(*)$ is due to $\mathbf{Span}(\mathbf{Col}(\Delta_{U_i})) \subseteq \mathbf{Span}(\mathbf{Col}(U_i) \cup \mathbf{Col}(U_r^*))$ and $\|\Delta_{U_i}\|_{S_\infty} \leq \|\Delta_{U_i}\|_*$. Continuing, we get

$$
\begin{aligned}
\|\nabla f(X_i)P_{U_i}\|_{S_\infty} &= \|\nabla f(X_i)U_iU_i^\dagger\|_{S_\infty} \\
&\overset{(1)}{\leq} \|\nabla f(X_i)U_i\|_{S_\infty}\frac{1}{\sigma_r(U_i)} \\
&\overset{(2)}{\leq} \|\nabla f(X_i)U_i\|_{S_\infty}\frac{10}{9\sigma_r(U^*)}
\end{aligned}
\tag{A.11}
$$

in which $U_i^\dagger$ denotes the pseudoinverse of U_i. Here, (1) is due to $\sigma_1(U_i^\dagger) = \sigma_r(U_i)^{-1}$, and (2) is by assumption (III.5), Weyl's inequality and $\sigma_r(U^*R_{U_i}) = \sigma_r(U^*)$.

Similarly, we have

$$
\begin{aligned}
\|\nabla f(X_i)P_{U^*}\|_{S_\infty} &= \|\nabla f(X_i)U^*(U^*)^\dagger\|_{S_\infty} \\
&\leq \underbrace{\|\nabla f(X_i)U^*\|_{S_\infty}}_{Ⓐ}\frac{1}{\sigma_r(U^*)}.
\end{aligned}
\tag{A.12}
$$

To upper bound Ⓐ, we use the following inequality.

$$
\begin{aligned}
&\|\nabla f(X_i)U^*\|_{S_\infty} \\
=&\|\nabla f(X_i)U^*R_{U_i}\|_{S_\infty} \\
\leq&\|\nabla f(X_i)U_i\|_{S_\infty} + \|\nabla f(X_i)\Delta_{U_i}\|_{S_\infty} \\
=&\|\nabla f(X_i)U_i\|_{S_\infty} + \|\nabla f(X_i)P_{\Delta_{U_i}}\Delta_{U_i}\|_{S_\infty} \\
\leq&\|\nabla f(X_i)U_i\|_{S_\infty} + \|\nabla f(X_i)P_{\Delta_{U_i}}\|_{S_\infty}\|\Delta_{U_i}\|_{S_\infty} \\
\overset{(1)}{\leq}&\|\nabla f(X_i)U_i\|_{S_\infty} \\
&+ \Big(\|\nabla f(X_i)P_{U_i}\|_{S_\infty} + \|\nabla f(X_i)P_{U^*}\|_{S_\infty}\Big)\|\Delta_{U_i}\|_{S_\infty} \\
\overset{(2)}{\leq}&\|\nabla f(X_i)U_i\|_{S_\infty} \\
&+ \frac{10}{9}\Big(\|\nabla f(X_i)U_i\|_{S_\infty} + \|\nabla f(X_i)U^*\|_{S_\infty}\Big)\frac{\|\Delta_{U_i}\|_{S_\infty}}{\sigma_r(U^*)} \\
\overset{(3)}{\leq}&\|\nabla f(X_i)U_i\|_{S_\infty} \\
&+ \frac{1}{10}\Big(\frac{10}{9}\|\nabla f(X_i)U_i\|_{S_\infty} + \|\nabla f(X_i)U^*\|_{S_\infty}\Big) \\
=&\frac{10}{9}\|\nabla f(X_i)U_i\|_{S_\infty} + \frac{1}{10}\|\nabla f(X_i)U^*\|_{S_\infty}.
\end{aligned}
\tag{A.13}
$$

Here, (1) is owing to the similar reason of (A.10), (2) is obtained by plugging in (A.11) and (A.12), and (3) is by assumption (III.5) and $\|\Delta_{U_i}\|_{S_\infty} \leq \|\Delta_{U_i}\|_*$. Thus we arrive at

$$
\|\nabla f(X_i)U^*\|_{S_\infty} \leq \left(\frac{10}{9}\right)^2\|\nabla f(X_i)U_i\|_{S_\infty}.
\tag{A.14}
$$

Plugging this into (A.12), we get

$$
\|\nabla f(X_i)P_{U^*}\|_{S_\infty} \leq \left(\frac{10}{9}\right)^2\|\nabla f(X_i)U_i\|_{S_\infty}\frac{1}{\sigma_r(U^*)}.
\tag{A.15}
$$

Combining (A.11) and (A.15) with (A.10), we obtain

$$\begin{aligned}
&\left\langle \nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T \right\rangle \\
\leq&(\|\nabla f(X_i)U_i\|_{S_\infty}\frac{10}{9\sigma_r(U^*)} \\
&+\left(\frac{10}{9}\right)^2\|\nabla f(X_i)U_i\|_{S_\infty}\frac{1}{\sigma_r(U^*)})\|\Delta_{U_i}\|_*^2 \\
=&\|\nabla f(X_i)U_i\|_{S_\infty}\frac{190}{81}\frac{\|\Delta_{U_i}\|_*}{\sigma_r(U^*)}\|\Delta_{U_i}\|_* \\
\overset{(*)}{\leq}&\frac{19}{81}\|\nabla f(X_i)U_i\|_{S_\infty}\|\Delta_{U_i}\|_*
\end{aligned} \tag{A.16}$$

where $(*)$ is by assumption (III.5). Now we plug (A.16) into (A.9) and obtain

$$f(U_iU_i^T) - f(U^*U^{*T}) \leq (2 + \frac{19}{81})\|\Delta_{U_i}\|_*\|\nabla f(X_i)\cdot U_i\|_{S_\infty}. \tag{A.17}$$

The last part is to prove $\min_i \gamma_i \geq \frac{1}{4}\overline{\eta}$ by showing $\|U_i\|_{S_\infty} \leq \frac{11}{9}\|U_0\|_{S_\infty}$ and

$$\|\nabla f(X_i)\|_{S_\infty} \leq \frac{40L}{81}\sigma_r(U_0)\sigma_1(U_0) + \|\nabla f(X_0)\|_{S_\infty}. \tag{A.18}$$

By assumption (III.5) and Weyl's inequality, we have for every $i \geq 0$

$$\begin{aligned}
&(1-\frac{1}{10})\sigma_1(U^*) \leq \sigma_1(U_i) \leq (1+\frac{1}{10})\sigma_1(U^*)\,, \text{ and thus} \\
&\frac{1+\frac{1}{10}}{1-\frac{1}{10}}\sigma_1(U_0) \geq \sigma_1(U_i).
\end{aligned} \tag{A.19}$$

For $\|\nabla f(X_i)\|_{S_\infty}$, we have

$$\begin{aligned}
\|\nabla f(X_i)\|_{S_\infty} &\leq \|\nabla f(X_i) - \nabla f(X_0)\|_{S_\infty} + \|\nabla f(X_0)\|_{S_\infty} \\
&\leq L_{S_1\to S_\infty}\|X_i - X_0\|_* + \|\nabla f(X_0)\|_{S_\infty} \\
&\leq L_{S_1\to S_\infty}\Big(\|X_i - X^*\|_* + \|X_0 - X^*\|_*\Big) \\
&\quad + \|\nabla f(X_0)\|_{S_\infty}.
\end{aligned} \tag{A.20}$$

Since

$$\begin{aligned}
\|X_i - X^*\|_* &= \|U_i(U_i - U^*R_{U_i})^T \\
&\qquad + (U_i - U^*R_{U_i})(U^*R_{U_i})^T\|_* \\
&\leq \|U_i - U^*R_{U_i}\|_*\Big(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty}\Big),
\end{aligned} \tag{A.21}$$

we have

$$\begin{aligned}
&\|X_i - X^*\|_* + \|X_0 - X^*\|_* \\
\leq&\|U_i - U^*R_{U_i}\|_*(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty}) \\
+&\|U_0 - U^*R_{U_0}\|_*(\|U_0\|_{S_\infty} + \|U^*\|_{S_\infty}) \\
\leq&\frac{\sigma_r(U^*)}{10}\sigma_1(U_0)(\frac{11}{9} + \frac{10}{9} + 1 + \frac{10}{9}) \\
\leq&\frac{1}{1-\frac{1}{10}}\frac{\sigma_r(U_0)}{10}\sigma_1(U_0)\frac{40}{9} \\
=&\frac{40}{81}\sigma_r(U_0)\sigma_1(U_0)
\end{aligned} \tag{A.22}$$

by applying inequality (A.19).

APPENDIX B
PROOF OF SUBLINEAR RATE FOR NUCLEAR GRADIENT DESCENT, TENSOR VERSION

We first define the action $(\cdot)$ on a bounded linear operator T of $H_1 \otimes H_2$ and H_1 where H_1 and H_2 are Hilbert spaces.

$$\begin{aligned}&\forall T \in L(H_1 \otimes H_2, \mathbb{R}),\ h_1 \in H_1,\ T = \sum_i \lambda_i (a_i \otimes b_i)\\ &T \cdot h_1 \triangleq \sum_i \lambda_i \langle a_i, h_1 \rangle b_i \in H_2.\end{aligned} \tag{B.1}$$

Immediately we have $\langle T, h_1 \otimes h_2 \rangle = \langle T \cdot h_1, h_2 \rangle$, since

$$\begin{aligned}\langle T, h_1 \otimes h_2 \rangle &= \left\langle \sum_i \lambda_i (a_i \otimes b_i), h_1 \otimes h_2 \right\rangle\\ &= \sum_i \lambda_i \langle a_i, h_1 \rangle \langle b_i, h_2 \rangle\\ &= \langle T \cdot h_1, h_2 \rangle.\end{aligned} \tag{B.2}$$

For the cases $H_i = \mathbb{R}^{n_i \times m_i}$, we define the norm to be injective cross norm with each H_i having the spectral norm $\|\|_{S_\infty}$ as primal norm and the consequent dual norm, nuclear norm $\|\|_*$.

$$\|x\| \triangleq \sup_{\|a_i\|_* \leq 1} \langle a_1 \otimes a_2, x \rangle \tag{B.3}$$

which satisfies

$$\begin{aligned}\|h_1 \otimes h_2\| &= \|h_1\|_{S_\infty} \|h_2\|_{S_\infty}\\ \|a_1 \otimes a_2\|_{dual} &= \|a_1\|_* \|a_2\|_*.\end{aligned} \tag{B.4}$$

We also use $\|h_1 \otimes h_2\|_{S_\infty}$ and $\|a_1 \otimes a_2\|_*$ to denote $\|h_1 \otimes h_2\|$ and $\|a_1 \otimes a_2\|_{dual}$.
We use following lemma to prove the sublinear rate.

Lemma 2. *For the sequence of the iterates $\{U_i\}_{i=0}^k$, we have*

$$f(U_i \otimes U_i) - f(U_{i+1} \otimes U_{i+1}) \geq \alpha_i \cdot \|[\nabla f(X_i) \cdot U_i]_\infty^\# \|_{S_\infty}^2 = \alpha_i \cdot \|\nabla f(X_i) \cdot U_i\|_*^2 \tag{B.5}$$

and

$$f(U_i \otimes U_i) - f(U^* \otimes U^*) \leq \beta_i \cdot \|[\nabla f(X_i) \cdot U_i]_\infty^\# \|_{S_\infty} \tag{B.6}$$

where $\alpha_i = 1.117\ \eta_i$ and $\beta_i = (2 + \frac{19}{81}) \|\Delta_{U_i}\|_{S_\infty} = (2 + \frac{19}{81}) D_\infty(U_i, U^*)$.

Define $\delta_i = f(U_i \otimes U_i) - f(U^* \otimes U^*)$ and follow the previous lemma. We know $\{\delta_i\}$ is an positive decreasing sequence and

$$\begin{aligned}\delta_{i+1} &\leq \delta_i - \alpha_i \cdot \|[\nabla f(X) \cdot U]_\infty^\# \|_{S_\infty}^2\\ &\leq \delta_i - \frac{\alpha_i}{\beta_i^2} \cdot \delta_i^2\end{aligned}$$

Dividing both sides with $(\delta_i \cdot \delta_{i+1})$, we obtain, by assumption (B.14),

$$\frac{1}{\delta_{i+1}} - \frac{1}{\delta_i} \geq \frac{\alpha_i}{\beta_i^2} \cdot \frac{\delta_i}{\delta_{i+1}} \geq \frac{\alpha_i}{\beta_i^2} \geq \frac{\alpha_i}{\tilde{D}_{S_\infty}^2}.$$

Telescoping the inequality we get the desired result.
Now we prove (B.5) of lemma 2. We assume $\nabla f(X)$ is symmetric, i.e. $\nabla f(X) = \sum_i \lambda_i (a_i \otimes a_i)$ throughout. The smoothness gives

$$\begin{aligned}f(U_i \otimes U_i) - f(U_{i+1} \otimes U_{i+1}) &\geq \langle \nabla f(X_i), X_i - X_{i+1} \rangle - \frac{L}{2} \|X_i - X_{i+1}\|_{S_\infty}^2\\ &= \underbrace{\langle \nabla f(X_i), (U_i - U_{i+1}) \otimes U_i + U_i \otimes (U_i - U_{i+1}) \rangle}_{①}\\ &\quad - \underbrace{\langle \nabla f(X_i), (U_i - U_{i+1}) \otimes (U_i - U_{i+1}) \rangle}_{②} - \frac{L}{2} \underbrace{\|X_i - X_{i+1}\|_{S_\infty}^2}_{③}\end{aligned} \tag{B.7}$$

For ① we have

$$\begin{aligned}\langle \nabla f(X_i), (U_i - U_{i+1}) \otimes U_i + U_i \otimes (U_i - U_{i+1})\rangle &\overset{(*)}{=} 2\langle \nabla f(X_i) \cdot U_i, U_i - U_{i+1}\rangle \\ &= 2\eta_i \big\langle \nabla f(X_i) \cdot U_i, [\nabla f(X_i) \cdot U_i]^{\#}_{\infty} \big\rangle \\ &= 2\eta_i \|\nabla f(X_i) \cdot U_i\|_*^2\end{aligned} \tag{B.8}$$

where $(*)$ is by the assumption that $\nabla f(X_i)$ is symmetric.
To upper bound ②, we use

$$\begin{aligned}\langle \nabla f(X_i), (U_i - U_{i+1}) \otimes (U_i - U_{i+1})\rangle &= \eta_i^2 \|\nabla f(X_i) \cdot U_i\|_*^2 \langle \nabla f(X_i), AB^T \otimes AB^T\rangle \\ &\le \eta_i^2 \|\nabla f(X_i) \cdot U_i\|_*^2 \cdot \|\nabla f(X_i)\|_* \|AB^T\|_{S_\infty}^2 \\ &\overset{(*)}{\le} \frac{1}{4}\eta_i \|\nabla f(X_i) \cdot U_i\|_*^2\end{aligned} \tag{B.9}$$

in which A and B are respectively the left and right singular vectors of $\nabla f(X_i) \cdot U_i$. $(*)$ is by $\eta_i \le \frac{1}{4\|\nabla f(X_i)\|_*}$.
To upper bound ③, we use

$$\begin{aligned}\|U_i \otimes U_i - U_{i+1} \otimes U_{i+1}\|_{S_\infty} &= \|U_i \otimes (U_i - U_{i+1}) + (U_i - U_{i+1}) \otimes U_i - (U_i - U_{i+1}) \otimes (U_i - U_{i+1})\|_{S_\infty} \\ &\le 2\|U_i\|_{S_\infty} \|U_i - U_{i+1}\|_{S_\infty} + \|U_i - U_{i+1}\|_{S_\infty}^2 \\ &= 2\eta_i \|U_i\|_{S_\infty} \|\nabla f(X_i) \cdot U_i\|_* + \eta_i^2 \|\nabla f(X_i) \cdot U_i\|_*^2 \\ &= \eta_i \ \|\nabla f(X_i) \cdot U_i\|_* \ [2\|U_i\|_{S_\infty} + \eta_i \|\nabla f(X_i) \cdot U_i\|_*] \\ &\overset{(1)}{\le} \eta_i \ \|\nabla f(X_i) \cdot U_i\|_* \ [2\|U_i\|_{S_\infty} + \eta_i \|\nabla f(X_i)\|_* \|U_i\|_{S_\infty}] \\ &\overset{(2)}{\le} \eta_i \ \|\nabla f(X_i) \cdot U_i\|_* \frac{9}{4}\|U_i\|_{S_\infty}\end{aligned} \tag{B.10}$$

(2) is by $\eta_i \le \frac{1}{4\|\nabla f(X_i)\|_*}$ and (1) is due to

$$\begin{aligned}\|\nabla f(X_i) \cdot U_i\|_* &= \sup_{\|y\|_{S_\infty} \le 1} \langle \nabla f(X_i) \cdot U_i, y\rangle \\ &= \sup_{\|y\|_{S_\infty} \le 1} \langle \nabla f(X_i), U_i \otimes y\rangle \\ &\le \sup_{\|y\|_{S_\infty} \le 1} \|\nabla f(X_i)\|_* \|U_i\|_{S_\infty} \|y\|_{S_\infty} \\ &= \|\nabla f(X_i)\|_* \|U_i\|_{S_\infty}\end{aligned} \tag{B.11}$$

Plugging above inequalities into (B.7), we obtain

$$\begin{aligned}f(U_i \otimes U_i) - f(U_{i+1} \otimes U_{i+1}) &\ge 2\eta_i \|\nabla f(X_i) \cdot U_i\|_*^2 - \frac{1}{4}\eta_i \|\nabla f(X_i) \cdot U_i\|_*^2 \\ &\quad - \frac{L}{2}(\frac{9}{4}\eta_i \ \|U_i\|_{S_\infty} \|\nabla f(X_i) \cdot U_i\|_*)^2 \\ &\ge \eta_i \|\nabla f(X_i) \cdot U_i\|_*^2 \left[\frac{7}{4} - \frac{L}{2}\left(\frac{9}{4}\right)^2 \eta_i \|U_i\|_{S_\infty}^2\right] \\ &\overset{(*)}{\ge} 1.117 \ \eta_i \ \|\nabla f(X_i) \cdot U_i\|_*^2\end{aligned} \tag{B.12}$$

$(*)$ is by $\eta_i \le \frac{1}{4L\|X_i\|_{S_\infty}} = \frac{1}{4L\|U_i\|_{S_\infty}^2}$. We thus finish the first part of lemma 2.

We give the proof of (B.6) of lemma 2 which only holds for the phase retrieval case. We then use $\tilde{f}(\tilde{X})$ to denote the original objective function, i.e. $\tilde{X} = \tilde{U}\tilde{U}^T$, $\tilde{f}(\tilde{X}) = f(X) = f(U \otimes U)$ and $\tilde{U}$, a $p \times 1$ vector, is the vectorization of U, a $m \times n$ matrix where $p = m \cdot n$. We have the following equalities.

$$\begin{aligned}&\langle \nabla f(X_i), U_1 \otimes U_2\rangle = \left\langle \nabla \tilde{f}(\tilde{X}_i), \tilde{U_1}\tilde{U_2}^T \right\rangle \text{ and its immediate consequence} \\ &\langle \nabla f(X_i) \cdot U_1, U_2\rangle = \left\langle \nabla \tilde{f}(\tilde{X}_i)\tilde{U_1}, \tilde{U_2} \right\rangle\end{aligned} \tag{B.13}$$

in which both sides of the first equation are the first order term of the objective function.
The assumption made here, which corresponds to IV.5, would be

$$\tilde{D}_\infty \equiv \max_{\tilde{U}: f(\tilde{U}\tilde{U}^\top) \le f(\tilde{U}_0 \tilde{U}_0^\top)} D_\infty(\tilde{U}, \tilde{U}^*) \le \frac{\sigma_{min}(\tilde{U}^*)}{10} = \frac{\sigma_1(\tilde{U}^*)}{10} = \frac{\|\tilde{U}^*\|_{l_2}}{10}. \tag{B.14}$$

We denote

$$\begin{aligned} R_{\tilde{U}_i} &\equiv \underset{R \text{ unitary}}{\arg\min} \|\tilde{U}_i - \tilde{U}^* R\|_{S_\infty} \\ &= \underset{R \in \{1,-1\}}{\arg\min} \|\tilde{U}_i - \tilde{U}^* R\|_{S_\infty}. \end{aligned}$$

and define $\Delta_{\tilde{U}_i} \equiv \tilde{U}_i - \tilde{U}^* R_{\tilde{U}_i}$.

$$\begin{aligned} f(U_i \otimes U_i) - f(U^* \otimes U^*) &\le \langle \nabla f(X_i), X_i - X^* \rangle = \langle \nabla f(X_i), U_i \otimes U_i - U^* R_{\tilde{U}_i} \otimes U^* R_{\tilde{U}_i} \rangle \\ &= \left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i} \tilde{U}_i^T \right\rangle + \left\langle \nabla \tilde{f}(\tilde{X}_i), \tilde{U}_i \Delta_{\tilde{U}_i}^T \right\rangle - \left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i} \Delta_{\tilde{U}_i}^T \right\rangle \\ &= 2 \left\langle \nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i, \Delta_{\tilde{U}_i} \right\rangle - \left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i} \Delta_{\tilde{U}_i}^T \right\rangle \\ &\le 2 \|\nabla f(\tilde{X}_i) \tilde{U}_i\|_* \|\Delta_{\tilde{U}_i}\|_{S_\infty} + \underbrace{\left| \left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i} \Delta_{\tilde{U}_i}^T \right\rangle \right|}_{①} \end{aligned} \tag{B.15}$$

To upper bound ①,

$$\begin{aligned} \left| \left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i} \Delta_{\tilde{U}_i}^T \right\rangle \right| &= \left| \left\langle \nabla \tilde{f}(\tilde{X}_i) \Delta_{\tilde{U}_i}, \Delta_{\tilde{U}_i} \right\rangle \right| \\ &\le \|\nabla \tilde{f}(\tilde{X}_i) \Delta_{\tilde{U}_i}\|_* \|\Delta_{\tilde{U}_i}\|_{S_\infty} \\ &= \|\nabla \tilde{f}(\tilde{X}_i) (P_{\Delta_{\tilde{U}_i}} \Delta_{\tilde{U}_i})\|_* \|\Delta_{\tilde{U}_i}\|_{S_\infty} \\ &\le \|\nabla \tilde{f}(\tilde{X}_i) P_{\Delta_{\tilde{U}_i}}\|_* \|\Delta_{\tilde{U}_i}\|_{S_\infty}^2 \\ &\overset{(*)}{\le} (\|\nabla \tilde{f}(\tilde{X}_i) P_{\tilde{U}_i}\|_* + \|\nabla \tilde{f}(\tilde{X}_i) P_{\tilde{U}^*}\|_*) \|\Delta_{\tilde{U}_i}\|_{S_\infty}^2 \end{aligned} \tag{B.16}$$

in which P_U denotes the projection onto $\mathbf{Col}(U)$. $(*)$ is due to $\mathbf{Span}(\mathbf{Col}(\Delta_{\tilde{U}_i})) \subseteq \mathbf{Span}(\mathbf{Col}(\tilde{U}_i) \cup \mathbf{Col}(\tilde{U}_r^*))$.

$$\|\nabla \tilde{f}(\tilde{X}_i) P_{U_i}\|_* = \|\nabla \tilde{f}(\tilde{X}_i) U_i U_i^\dagger\|_* \overset{(1)}{\le} \|\nabla \tilde{f}(\tilde{X}_i) U_i\|_* \frac{1}{\sigma_{min}(U_i)} \overset{(2)}{\le} \|\nabla \tilde{f}(\tilde{X}_i) U_i\|_* \frac{10}{9\sigma_{min}(U^*)} \tag{B.17}$$

in which $U_i^\dagger$ denotes the pseudoinverse of U_i and σ_{min} denotes the smallest non-zero singular value. (1) is due to $\sigma_1(U_i^\dagger) = \sigma_r(U_i)^{-1}$. (2) is by assumption (B.14) and Weyl's inequality.
Similarly, we have

$$\|\nabla \tilde{f}(\tilde{X}_i) P_{\tilde{U}^*}\|_* = \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^* (\tilde{U}^*)^\dagger\|_* \le \underbrace{\|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^*\|_*}_{Ⓐ} \frac{1}{\sigma_{min}(\tilde{U}^*)}. \tag{B.18}$$

To upper bound Ⓐ, we use the following inequality.

$$\begin{aligned} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^*\|_* &= \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^* R_{\tilde{U}_i}\|_* \\ &\le \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + \|\nabla \tilde{f}(\tilde{X}_i) \Delta_{\tilde{U}_i}\|_* \\ &= \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + \|\nabla \tilde{f}(\tilde{X}_i) P_{\Delta_{\tilde{U}_i}} \Delta_{\tilde{U}_i}\|_* \\ &\le \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + \|\nabla \tilde{f}(\tilde{X}_i) P_{\Delta_{\tilde{U}_i}}\|_* \|\Delta_{\tilde{U}_i}\|_{S_\infty} \\ &\overset{(1)}{\le} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + (\|\nabla \tilde{f}(\tilde{X}_i) P_{\tilde{U}_i}\|_* + \|\nabla \tilde{f}(\tilde{X}_i) P_{\tilde{U}^*}\|_*) \|\Delta_{\tilde{U}_i}\|_{S_\infty} \\ &\overset{(2)}{\le} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + (\|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* \frac{10}{9} + \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^*\|_*) \frac{\|\Delta_{\tilde{U}_i}\|_{S_\infty}}{\sigma_{min}(\tilde{U}^*)} \\ &\overset{(3)}{\le} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + (\|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* \frac{10}{9} + \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^*\|_*) \frac{1}{10} \\ &= \frac{10}{9} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}_i\|_* + \frac{1}{10} \|\nabla \tilde{f}(\tilde{X}_i) \tilde{U}^*\|_*. \end{aligned} \tag{B.19}$$

(1) is owing to the similar reason of (B.16). (2) is obtained by plugging in (B.17) and (B.18). (3) is by assumption (B.14). Thus we arrive

$$\|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}^*\|_* \leq \left(\frac{10}{9}\right)^2 \|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_*. \tag{B.20}$$

Plugging this into (B.18), we get

$$\|\nabla \tilde{f}(\tilde{X}_i)P_{\tilde{U}^*}\|_* \leq \left(\frac{10}{9}\right)^2 \|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_* \frac{1}{\sigma_{min}(\tilde{U}^*)}. \tag{B.21}$$

Combining (B.17) and (B.21) with (B.16), we obtain

$$\begin{aligned}\left\langle \nabla \tilde{f}(\tilde{X}_i), \Delta_{\tilde{U}_i}\Delta_{\tilde{U}_i}^T \right\rangle &\leq (\|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_* \frac{10}{9\sigma_{min}(U^*)} + \left(\frac{10}{9}\right)^2 \|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_* \frac{1}{\sigma_{min}(U^*)})\|\Delta_{\tilde{U}_i}\|_{S_\infty}^2 \\ &= \|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_* \frac{190}{81}\frac{\|\Delta_{\tilde{U}_i}\|_{S_\infty}}{\sigma_{min}(U^*)}\|\Delta_{\tilde{U}_i}\|_{S_\infty} \\ &\overset{(*)}{\leq} \frac{19}{81}\|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_*\|\Delta_{\tilde{U}_i}\|_{S_\infty}\end{aligned} \tag{B.22}$$

where $(*)$ is by assumption (B.14). Now we plug (B.22) into (B.15) and obtain

$$\begin{aligned}f(U_i \otimes U_i) - f(U^* \otimes U^*) &\leq (2 + \frac{19}{81})\|\Delta_{\tilde{U}_i}\|_{S_\infty}\|\nabla \tilde{f}(\tilde{X}_i)\tilde{U}_i\|_* \\ &\overset{(**)}{\leq} (2 + \frac{19}{81})C_0\|\Delta_{\tilde{U}_i}\|_{S_\infty}\|\nabla f(X_i)\cdot U_i\|_*\end{aligned} \tag{B.23}$$

where C_0 is a constant and $(**)$ is obtained by the connection between $\nabla f(X_i)\cdot U_i$ and $\tilde{f}(\tilde{X}_i)\tilde{U}_i$ (see (B.13)) and the equivalence of norms of finite dimensional Banach space.
The last part is to prove $\min_i \gamma_i \geq \frac{1}{4}\overline{\eta}$ by showing $\|U_i\|_{S_\infty} \leq \frac{11}{9}\|U_0\|_{S_\infty}$ and

$$\|\nabla f(X_i)_r\|_* \leq \frac{40L}{81}\sigma_{min}(U_0)\sigma_1(U_0) + \|\nabla f(X_0)_r\|_*. \tag{B.24}$$

By assumption (B.14) and Weyl's inequality, we have for every $i \geq 0$

$$(1 - \frac{1}{10})\sigma_1(U^*) \leq \sigma_1(U_i) \leq (1 + \frac{1}{10})\sigma_1(U^*)\,, \text{ and thus } \frac{1+\frac{1}{10}}{1-\frac{1}{10}}\sigma_1(U_0) \geq \sigma_1(U_i). \tag{B.25}$$

Since $\|\nabla f(X_i)_r\|_*$ is the Ky Fan r-norm of $\nabla f(X_i)$, we have

$$\begin{aligned}\|\nabla f(X_i)_r\|_* &\leq \|(\nabla f(X_i) - \nabla f(X_0))_r\|_* + \|\nabla f(X_0)_r\|_* \\ &\leq \|\nabla f(X_i) - \nabla f(X_0)\|_* + \|\nabla f(X_0)_r\|_* \\ &\leq L_{S_\infty\to S_1}\|X_i - X_0\|_{S_\infty} + \|\nabla f(X_0)_r\|_* \\ &\leq L_{S_\infty\to S_1}(\|X_i - X^*\|_{S_\infty} + \|X_0 - X^*\|_{S_\infty}) + \|\nabla f(X_0)_r\|_*.\end{aligned}$$

Since

$$\begin{aligned}\|X_i - X^*\|_{S_\infty} &= \|U_i \otimes (U_i - U^*R_{U_i}) + (U_i - U^*R_{U_i}) \otimes (U^*R_{U_i})\|_{S_\infty} \\ &\leq \|U_i - U^*R_{U_i}\|_{S_\infty}(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty})\end{aligned}$$

we have

$$\begin{aligned}\|X_i - X^*\|_{S_\infty} + \|X_0 - X^*\|_{S_\infty} &\leq \|U_i - U^*R_{U_i}\|_{S_\infty}(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty}) \\ &\quad + \|U_0 - U^*R_{U_0}\|_{S_\infty}(\|U_0\|_{S_\infty} + \|U^*\|_{S_\infty}) \\ &\leq \frac{\sigma_{min}(U^*)}{10}\sigma_1(U_0)(\frac{11}{9} + \frac{10}{9} + 1 + \frac{10}{9}) \\ &\leq \frac{1}{1-\frac{1}{10}}\frac{\sigma_{min}(U_0)}{10}\sigma_1(U_0)\frac{40}{9} \\ &= \frac{40}{81}\sigma_{min}(U_0)\sigma_1(U_0)\end{aligned}$$

by applying inequality (B.25).

Appendix C
Proof of Linear Rate for Nuclear Gradient Descent

We use U and U^+ to denotes the current state and the updated state. Let $X^* = U^*(U^*)^T$ be the optimum, $X = UU^T$ and $X^+ = U^+(U^+)^T$.

$$U^+ = U - \eta_U \nabla f(UU^T)^\# \cdot U \tag{C.1}$$

where $\eta_U = \frac{1}{16(L\|X\|_{S_\infty} + \|\nabla f(X)^\# Q_U Q_U^T\|_{S_\infty})}$ which also denoted as η for simplicity.

We now start to prove the following key lemma.

Lemma 3. *Given* $D_F(U,U^*) \leq \rho\sigma_r(U_r^*)$ *and* $D_*(U,U^*) \leq \frac{1}{81\kappa}\frac{\sigma_r(X^*)}{\sigma_1(U^*)}$,

$$\begin{aligned}&\frac{1}{\eta}\langle U - U^+, U - U_r^* R_U\rangle\\ =&\langle \nabla f(X)^\# U, U - U_r^* R_U\rangle\\ \geq& 0.86\eta\|\nabla f(X)^\# U\|_F^2 + \frac{0.7\,\mu}{4}\sigma_r(X^*)D_F(U,U_r^*)^2\\ &-\frac{L}{4}\|X^* - X_r^*\|_F^2,\end{aligned} \tag{C.2}$$

in which $R_U = \underset{R \text{ is unitary}}{\arg\min_R}\|U - U_r^* R\|_F$.

First, we define $\Delta \equiv U - U_r^* R_U$ and thus

$$\begin{aligned}&\langle \nabla f(X)^\# U, U - U_r^* R_U\rangle\\ =&\frac{1}{2}\langle \nabla f(X)^\#, X - X_r^*\rangle + \frac{1}{2}\langle \nabla f(X)^\#, \Delta\Delta^T\rangle.\end{aligned} \tag{C.3}$$

First, we lower bound $\langle \nabla f(X)^\#, X - X_r^*\rangle$:

$$\begin{aligned}f(X) &\geq f(X^+) - \langle \nabla f(X), X^+ - X\rangle - \frac{L}{2}\|X^+ - X\|_*^2\\ &\geq f(X^*) - \langle \nabla f(X), X^+ - X\rangle - \frac{L}{2}\|X^+ - X\|_*^2,\end{aligned} \tag{C.4}$$

and

$$f(X_r^*) \geq f(X) + \langle \nabla f(X), X_r^* - X\rangle + \frac{\mu}{2}\|X_r^* - X\|_*^2. \tag{C.5}$$

Noticing PSD matrices form a convex cone, we obtain $\langle \nabla f(X^*), X^*\rangle = 0$ and consequently $\langle \nabla f(X^*), X_r^*\rangle = 0$. Therefore we have

$$\begin{aligned}f(X_r^*) &\leq f(X^*) + \langle \nabla f(X^*), X_r^* - X^*\rangle + \frac{L}{2}\|X_r^* - X^*\|_*^2\\ &= f(X^*) + \frac{L}{2}\|X_r^* - X^*\|_*^2.\end{aligned} \tag{C.6}$$

Summing up previous three inequalities, we have

$$\begin{aligned}&\langle \nabla f(X), X - X_r^*\rangle\\ \geq&\langle \nabla f(X), X - X^+\rangle - \frac{L}{2}\|X^+ - X\|_*^2\\ &+\frac{\mu}{2}\|X_r^* - X\|_*^2 - \frac{L}{2}\|X_r^* - X^*\|_*^2.\end{aligned} \tag{C.7}$$

Let $A \equiv I - \frac{\eta}{2}Q_U Q_U^T \nabla f(X)^\#$, we have

$$\begin{aligned}X^+ - X &= (U - \eta\nabla f(X)^\# U)(U - \eta\nabla f(X)^\# U)^T - UU^T\\ &= -\eta\nabla f(X)^\# XA - \eta A^T X \nabla f(X)^\#\end{aligned} \tag{C.8}$$

where we have used the property of $\nabla f(X)^{\#}$ being symmetric.
Plugging the previous equality into (C.7), we achieve

$$\begin{aligned}&\langle \nabla f(X), X - X_r^* \rangle - \frac{\mu}{2}\|X_r^* - X\|_*^2 + \frac{L}{2}\|X_r^* - X^*\|_*^2 \\ \geq&\big\langle \nabla f(X), X - X^+ \big\rangle - \frac{L}{2}\|X^+ - X\|_*^2 \\ \geq&2\eta\big\langle \nabla f(X), \nabla f(X)^{\#} X A\big\rangle - \frac{L}{2}(2\|\eta \nabla f(X)^{\#} X A\|_*)^2 \end{aligned} \tag{C.9}$$

where we have used $\|Y + Y^T\|_* \leq 2\|Y\|_*$.

For the two terms on the RHS of (C.9) we have the following bounds.

$$\begin{aligned}&\big\langle \nabla f(X), \nabla f(X)^{\#} X A\big\rangle \\ =&\Big[\big\langle \nabla f(X), \nabla f(X)^{\#} U U^T\big\rangle \\ &- \frac{\eta}{2}\big\langle \nabla f(X), \nabla f(X)^{\#} U U^T Q_U Q_U^T \nabla f(X)^{\#}\big\rangle\Big] \\ \geq&\|\nabla f(X)^{\#} U\|_F^2 - \frac{\eta}{2}\|\nabla f(X)^{\#} U\|_F^2 \|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} \\ \geq&(1 - \frac{1}{32})\|\nabla f(X)^{\#} U\|_F^2 \end{aligned} \tag{C.10}$$

where the last inequality is due to the choice of the step size. Continuing, we compute

$$\begin{aligned}|\nabla f(X)^{\#} X A\|_* &\leq \|\nabla f(X)^{\#} U\|_* \|U\|_{S_\infty} \|A\|_{S_\infty} \\ &\leq \|\nabla f(X)^{\#} U\|_* \|U\|_{S_\infty} \Big(1 + \frac{1}{32}\Big). \end{aligned} \tag{C.11}$$

Now plugging these two bounds into (C.9), we have

$$\begin{aligned}&\langle \nabla f(X), X - X_r^* \rangle - \frac{\mu}{2}\|X_r^* - X\|_*^2 + \frac{L}{2}\|X_r^* - X^*\|_*^2 \\ \geq&2\eta\ \|\nabla f(X)^{\#} U\|_F^2 \left[1 - \frac{1}{32} - \eta L \|U\|_{S_\infty}^2 \left(\frac{33}{32}\right)^2\right] \\ \geq&2\eta\ \|\nabla f(X)^{\#} U\|_F^2 \left[1 - \frac{1}{32} - \frac{1}{16} \left(\frac{33}{32}\right)^2\right] \\ \geq&\frac{18\eta}{10}\ \|\nabla f(X)^{\#} U\|_F^2 \end{aligned}$$

That is,

$$\begin{aligned}&\langle \nabla f(X), X - X_r^* \rangle \\ \geq&\frac{18\eta}{10}\ \|\nabla f(X)^{\#} U\|_F^2 + \frac{\mu}{2}\|X_r^* - X\|_*^2 - \frac{L}{2}\|X_r^* - X^*\|_*^2. \end{aligned} \tag{C.12}$$

We now lower bound $\big\langle \nabla f(X)^{\#}, \Delta\Delta^T \big\rangle$, the second term of (C.3):

$$\begin{aligned}&\big\langle \nabla f(X)^{\#}, \Delta\Delta^T \big\rangle \\ =&\big\langle Q_\Delta Q_\Delta^T \nabla f(X)^{\#}, \Delta\Delta^T \big\rangle \\ \geq&-|\mathbf{Trace}(\Delta\Delta^T Q_\Delta Q_\Delta^T \nabla f(X)^{\#})| \\ \geq&-\|Q_\Delta Q_\Delta^T \nabla f(X)^{\#}\|_{S_\infty} \langle \Delta, \Delta \rangle \\ \geq&-\Big[\|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} \\ &\qquad + \|Q_{U_r^*} Q_{U_r^*}^T \nabla f(X)^{\#}\|_{S_\infty}\Big] \cdot D_F(U, U_r^*)^2 \end{aligned} \tag{C.13}$$

where the last inequality is owing to $\mathbf{Span}(\mathbf{Col}(\Delta)) \subseteq \mathbf{Span}(\mathbf{Col}(U) \cup \mathbf{Col}(U_r^*))$.

$$\begin{aligned}\|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty}\ D_F(U, U_r^*)^2 &= \eta\ 16(L\|X\|_{S_\infty} + \|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}) \|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty}\ D_F(U, U_r^*)^2 \\ &= \underline{16\eta L \|X\|_{S_\infty} \|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} D_F(U, U_r^*)^2} + \\ &\quad 16\eta \|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 D_F(U, U_r^*)^2 \end{aligned} \tag{C.14}$$

We bound the underlined term by considering two possible conditions, $\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty} \leq \frac{\mu\sigma_r(X)}{40}$ and $\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty} > \frac{\mu\sigma_r(X)}{40}$.

$$\begin{aligned}
16\eta L\|X\|_{S_\infty}\|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} D^2 &\leq \max\left\{\frac{16\eta L\|X\|_{S_\infty}\mu\sigma_r(X)}{40}D^2, 16\eta 40\kappa\tau(X)\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 D^2\right\} \\
&\leq 16\eta L\|X\|_{S_\infty}\frac{\mu\sigma_r(X)}{40}D^2 + 16\eta 40\kappa\tau(X)\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 D^2 \\
&\leq \frac{\mu\sigma_r(X)}{40}D^2 + 16\eta 40\kappa\tau(X)\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 D^2 \qquad \text{(C.15)}
\end{aligned}$$

in which D denotes $D_F(U, U_r^*)$. Combining the previous inequality with inequality (C.14), we get

$$\begin{aligned}
\|Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} D^2 &\leq \frac{\mu\sigma_r(X)}{40}D^2 + (40\kappa\tau(X)+1)16\eta\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 D^2 \\
&\overset{(i)}{\leq} \frac{\mu\sigma_r(X)}{40}D^2 + (40(\frac{101}{99})^2\kappa\tau(X_r^*)+1)16\eta\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 (\rho\sigma_r(U_r^*))^2 \\
&\leq \frac{\mu\sigma_r(X)}{40}D^2 + 16\cdot 43\eta\kappa\tau(X_r^*)\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}^2 \sigma_r(X_r^*)\rho^2 \\
&\overset{(ii)}{\leq} \frac{\mu\sigma_r(X)}{40}D^2 + 16\cdot 43\eta\kappa\tau(X_r^*)\|\nabla f(X)^{\#} U\|_{S_\infty}^2 \frac{\sigma_r(X_r^*)}{\sigma_r(X)}\rho^2 \\
&\overset{(iii)}{\leq} \frac{\mu\sigma_r(X)}{40}D^2 + 16\cdot 43\eta\kappa\tau(X_r^*)\|\nabla f(X)^{\#} U\|_{S_\infty}^2 (\frac{100}{99})^2\rho^2 \\
&\overset{(iv)}{\leq} \frac{\mu\sigma_r(X)}{40}D^2 + \frac{2\eta}{29}\|\nabla f(X)^{\#} U\|_{S_\infty}^2 \qquad \text{(C.16)}
\end{aligned}$$

(i) and (iii) is due to the assumption $D_F(U, U^*) \leq \rho\sigma_r(U_r^*)$ and lemma 6. (ii) is owing to

$$\|\nabla f(X)^{\#} U\|_{S_\infty} = \|U^T \nabla f(X)^{\#}\|_{S_\infty} = \|U^T Q_U Q_U^T \nabla f(X)^{\#}\|_{S_\infty} \geq \sigma_{min}(U)\|\nabla f(X)^{\#} Q_U Q_U^T\|_{S_\infty}$$

and $\sigma_{min}(U) = \sigma_r(U) = \sqrt{\sigma_r(X)}$. (iv) is obtained by plugging $\rho = \frac{1}{100\kappa\tau(X_r^*)}$.
We first note that $\nabla f(U^*(U^*)^T)U^* = 0$, since X^* is the optimum, and thus $\nabla f(X^*)Q_{U_r^*} = 0$. Now we start to bound $\|Q_{U_r^*} Q_{U_r^*}^T \nabla f(X)^{\#}\|_{S_\infty}$.

$$\begin{aligned}
\|Q_{U_r^*} Q_{U_r^*}^T \nabla f(X)^{\#}\|_{S_\infty} &\leq \|Q_{U_r^*} Q_{U_r^*}^T \nabla f(X)\|_{S_\infty} \\
&= \|Q_{U_r^*} Q_{U_r^*}^T (\nabla f(X) - \nabla f(X^*))\|_{S_\infty} \\
&\leq \|\nabla f(X) - \nabla f(X^*)\|_{S_\infty} \\
&\leq L\left(\|X - X_r^*\|_* + \|X_r^* - X^*\|_*\right) \qquad \text{(C.17)}
\end{aligned}$$

where the last inequality is owing to L-smoothness and the triangular inequality.

Plugging inequalities (C.16) and (C.17) into (C.13), we get

$$\left\langle \nabla f(X)^{\#}, \Delta\Delta^T \right\rangle \geq -\left[\frac{\mu\sigma_r(X)}{40}D^2 + \frac{2\eta}{29}\|\nabla f(X)^{\#} U\|_{S_\infty}^2 + L\left(\|X - X_r^*\|_* + \|X_r^* - X^*\|_*\right) D^2\right]. \qquad \text{(C.18)}$$

Now we plug two bounds (C.12) and (C.18) into (C.3) to get

$$\begin{aligned}
\left\langle \nabla f(X)^{\#} U, U - U_r^* R_U \right\rangle &\geq \frac{1}{2}\left[\frac{18\eta}{10}\|\nabla f(X)^{\#} U\|_F^2 + \frac{\mu}{2}\|X_r^* - X\|_*^2 - \frac{L}{2}\|X_r^* - X^*\|_*^2\right] \\
&\quad - \frac{1}{2}\left[\frac{\mu\sigma_r(X)}{40}D^2 + \frac{2\eta}{29}\|\nabla f(X)^{\#} U\|_{S_\infty}^2 + L\left(\|X - X_r^*\|_* + \|X_r^* - X^*\|_*\right) D^2\right] \\
&\geq 0.86\eta\|\nabla f(X)^{\#} U\|_F^2 - \frac{L}{4}\|X_r^* - X^*\|_*^2 \\
&\quad + \frac{\mu}{4}\left[\underline{|X_r^* - X\|_*^2 - \frac{\sigma_r(X^*)D^2}{20} - 2\kappa D^2\left(\|X - X_r^*\|_* + \|X_r^* - X^*\|_*\right)}\right] \qquad \text{(C.19)}
\end{aligned}$$

Now we present two lemmas to bound $|X_r^* - X\|_*$ and thus the underlined term in (C.19).

Lemma 4. *If $D_F(U,U^*) \le \rho\sigma_r(U_r^*)$ and $\rho \le \frac{1}{100}$, then for any unitary matrix R*

$$\begin{aligned}\|X - X_r^*\|_* &= \|UU^T - U_r^*(U_r^*)^T\|_* \\ &= \|UU^T - U_r^*RU^T + U_r^*RU^T - U_r^*R(U_r^*R)^T\|_* \\ &\le \|U - U_r^*R\|_*\|U\|_{S_\infty} + \|U - U_r^*R\|_*\|U_r^*\|_{S_\infty} \\ &\overset{(i)}{\le} \|U - U_r^*R\|_*(1+\rho)\|U_r^*\|_{S_\infty} + \|U - U_r^*R\|_*\|U_r^*\|_{S_\infty} \\ &\le (2+\rho)\|U - U_r^*R\|_*\|U_r^*\|_{S_\infty} \\ &\le (2.01)\|U - U_r^*R\|_*\|U_r^*\|_{S_\infty} \end{aligned} \tag{C.20}$$

where (i) is due to lemma 6.

Lemma 5. *Let $X = UU^T$ and $X_r^*U_r^*(U_r^*)^T$ then*

$$\|X - X_r^*\|_F^2 \ge 2(\sqrt{2}-1)\sigma_r(X_r^*)D_F(U,U_r^*)^2. \tag{C.21}$$

See reference [2].

Combining lemmas 4 and 5, we obtain a lower bound for the underlined term in (C.19).

$$\begin{aligned}&\|X_r^* - X\|_*^2 - \frac{\sigma_r(X^*)D^2}{20} - 2\kappa D^2\left(\|X - X_r^*\|_* + \|X_r^* - X^*\|_*\right) \\ \overset{(i)}{\ge}&\|X_r^* - X\|_F^2 - \frac{\sigma_r(X^*)D^2}{20} - 2\kappa D^2\left(\|X - X_r^*\|_* + \frac{\sigma_r(X^*)}{200\kappa^{1.5}\tau(X_r^*)}\right) \\ \overset{(ii)}{\ge}&2(\sqrt{2}-1)\sigma_r(X_r^*)D^2 - \frac{\sigma_r(X^*)D^2}{20} - 2\kappa D^2\|X - X_r^*\|_* - \frac{\sigma_r(X^*)D^2}{50} \\ \overset{(iii)}{\ge}&\left[2(\sqrt{2}-1) - \frac{1}{20} - \frac{1}{50}\right]\sigma_r(X_r^*)D^2 - 2\kappa D^2(2.01)\frac{1}{81\kappa}\frac{\sigma_r(X^*)}{\sigma_1(U^*)}\|U_r^*\|_{S_\infty} \\ \ge&\left[2(\sqrt{2}-1) - \frac{1}{20} - \frac{1}{50} - \frac{1}{20}\right]\sigma_r(X_r^*)D^2 \\ \ge&\, 0.7\,\sigma_r(X_r^*)\,D^2 \end{aligned} \tag{C.22}$$

where (i) is due to $\|\cdot\|_* \ge \|\cdot\|_F$, and (ii) is owing to lemma 5. (iii) is due to the assumption $\tilde{D}_* \le \frac{1}{81\kappa}\frac{\sigma_r(X^\star)}{\sigma_1(U^\star)}$. Combining (C.22) with (C.19), we get

$$\left\langle \nabla f(X)^{\#}U, U - U_r^*R_U\right\rangle \ge 0.86\eta\|\nabla f(X)^{\#}U\|_F^2 - \frac{L}{4}\|X_r^* - X^*\|_*^2 + \frac{0.7\,\mu}{4}\sigma_r(X_r^*)\,D^2, \tag{C.23}$$

the desired lemma 3.

$$\begin{aligned}D_F(U^+,U_r^*)^2 &= \min_{R\text{ is unitary}}\|U^+ - U_r^*\|_F^2 \le \|U^+ - U_r^*R_U\|_F^2 \\ &= \|U - U_r^*R_U\|_F^2 - 2\eta\left\langle\nabla f(X)^{\#}U, U - U_r^*R_U\right\rangle + \eta^2\|\nabla f(X)^{\#}U\|_F^2 \\ &\overset{(i)}{\le} D_F(U,U^*)^2 - 2\eta\left[-\frac{L}{4}\|X_r^* - X^*\|_*^2 + \frac{0.7\mu}{4}\sigma_r(X_r^*)D(U,U^*)^2\right] - \\ &(2(0.86)-1)\eta^2\|\nabla f(X)^{\#}U\|_F^2 \\ &\le \left[1 - \frac{0.7\,\mu\,\eta}{2}\sigma_r(X_r^*)\right]D(U,U^*)^2 + \frac{\eta\,L}{2}\|X_r^* - X^*\|_*^2 \end{aligned} \tag{C.24}$$

in which $R_U = \underset{R\text{ is unitary}}{\arg\min_R}\|U - U_r^*R\|_F$. (i) is by lemma 3.

Lemma 6. *Let U and U_r^* be two $n\times r$ matrices such that $D_F(U,U_r^*) \le \rho\sigma_r(U_r^*)$, for $\rho \le \frac{1}{100}$. Define $X \equiv UU^T$ and $X_r^* \equiv U_r^*(U_r^*)^T$. Then we have*

$$(1 - \frac{1}{100})\sigma_1(U_r^*) \le \sigma_1(U) \le (1 + \frac{1}{100})\sigma_1(U_r^*) \tag{C.25}$$

$$(1 - \frac{1}{100})\sigma_r(U_r^*) \le \sigma_r(U) \le (1 + \frac{1}{100})\sigma_r(U_r^*) \tag{C.26}$$

and thus

$$\tau(U) \le \frac{101}{99}\tau(U_r^*) \text{ and } \tau(X) \le (\frac{101}{99})^2\tau(X_r^*) \tag{C.27}$$

Proof. By $\|\cdot\|_{S_\infty} \leq \|\cdot\|_F$ and Weyl's inequality for perturbation of singular values, we obtain

$$|\sigma_i(U_r^*) - \sigma_i(U)| \leq \rho\sigma_r(U_r^*) \leq \frac{1}{100}\sigma_r(U_r^*). \tag{C.28}$$

□

Now we show $\min_{i\geq 0} \eta_i \geq \frac{1}{16}\overline{\eta}$ by verifying $\|U_iU_i^T\|_{S_\infty} \leq \left(\frac{1+\rho}{1-\rho}\right)^2 \|X_0\|_{S_\infty}$ and $\|\nabla f(U_iU_i^T)^{\#} Q_{U_i} Q_{U_i}^T\|_{S_\infty} \leq \frac{4L\sigma_1(U_0)\sigma_r(X^*)}{81\kappa\sigma_1(U^*)(1-\rho)} + \|\nabla f(X_0)\|_{S_\infty}$. The first one is an immediate result of lemma 6. Applying the same arguments of (A.20), (A.21) and (A.22), the second part is a direct consequence of assumption $\tilde{D}_* \leq \frac{1}{81\kappa}\frac{\sigma_r(X^\star)}{\sigma_1(U^\star)}$ and assumption $\tilde{D}_F \leq \rho\sigma_r(U_r^\star)$.

Appendix D
Proof of Sublinear Rate of Spectral Gradient Descent

We first use following lemma to prove the sublinear rate.

Lemma 7. *For the sequence of the iterates $\{U_i\}_{i=0}^k$, we have*

$$\begin{aligned} f(U_iU_i^T) - f(U_{i+1}{U_{i+1}}^T) &\geq \alpha_i \cdot \|[\nabla f(X_i)\cdot U_i]_\infty^{\#}\|_{S_\infty}^2 \\ &= \alpha_i \cdot \|\nabla f(X_i)\cdot U_i\|_*^2 \end{aligned} \tag{D.1}$$

and

$$f(U_iU_i^T) - f(U^*{U^*}^T) \leq \beta_i \cdot \|[\nabla f(X_i)\cdot U_i]_\infty^{\#}\|_{S_\infty} \tag{D.2}$$

where $\alpha_i = 1.117\ \eta_i$ and $\beta_i = (2 + \frac{19}{81})D_\infty(U_i, U^*)$.

Define $\delta_i = f(U_iU_i^T) - f(U^*{U^*}^T)$ and follow the previous lemma. We know $\{\delta_i\}$ is an positive decreasing sequence and

$$\begin{aligned} \delta_{i+1} &\leq \delta_i - \alpha_i \cdot \|[\nabla f(X)\cdot U]_\infty^{\#}\|_{S_\infty}^2 \\ &\leq \delta_i - \frac{\alpha_i}{\beta_i^2}\cdot \delta_i^2 \end{aligned}$$

Dividing both sides with $(\delta_i \cdot \delta_{i+1})$, we obtain, by assumption (IV.5),

$$\frac{1}{\delta_{i+1}} - \frac{1}{\delta_i} \geq \frac{\alpha_i}{\beta_i^2}\cdot\frac{\delta_i}{\delta_{i+1}} \geq \frac{\alpha_i}{\beta_i^2} \geq \frac{\alpha_i}{\tilde{D}_{S_\infty}^2}.$$

Telescoping the inequality we get the desired result.
Now we prove (D.1) of lemma 7. The smoothness gives

$$\begin{aligned} f(U_iU_i^T) - f(U_{i+1}{U_{i+1}}^T) &\geq \langle \nabla f(X_i), X_i - X_{i+1}\rangle - \frac{L}{2}\|X_i - X_{i+1}\|_{S_\infty}^2 \\ &= \underbrace{\left\langle \nabla f(X_i), (U_i - U_{i+1})U_i^T + U_i(U_i - U_{i+1})^T\right\rangle}_{①} \\ &\quad - \underbrace{\left\langle \nabla f(X_i), (U_i - U_{i+1})(U_i - U_{i+1})^T\right\rangle}_{②} - \frac{L}{2}\underbrace{\|X_i - X_{i+1}\|_{S_\infty}^2}_{③} \end{aligned} \tag{D.3}$$

For ① we have

$$\begin{aligned} &\left\langle \nabla f(X_i), (U_i - U_{i+1})U_i^T + U_i(U_i - U_{i+1})^T\right\rangle \\ =&2\langle \nabla f(X_i)U_i, U_i - U_{i+1}\rangle \\ =&2\eta_i\left\langle \nabla f(X_i)U_i, [\nabla f(X_i)U_i]_\infty^{\#}\right\rangle \\ =&2\eta_i\|\nabla f(X_i)U_i\|_*^2. \end{aligned} \tag{D.4}$$

To upper bound ②, we use

$$\begin{aligned} &\left\langle \nabla f(X_i), (U_i - U_{i+1})(U_i - U_{i+1})^T\right\rangle \\ =&\eta_i^2\|\nabla f(X_i)\cdot U_i\|_*^2 \cdot Trace(A^T\nabla f(X_i)A\begin{bmatrix}\mathbf{I_r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0_{(n-r)\times(n-r)}}\end{bmatrix}) \\ \leq&\eta_i^2\|\nabla f(X_i)\cdot U_i\|_*^2 \cdot \|\nabla f(X_i)_r\|_* \\ \overset{(*)}{\leq}&\frac{1}{4}\eta_i\|\nabla f(X_i)\cdot U_i\|_*^2 \end{aligned} \tag{D.5}$$

in which A is the left-singular vectors of $\nabla f(X_i) \cdot U_i$ and $\|\nabla f(X_i)_r\|_*$ equals to the sum of the top r singular values of $\nabla f(X_i)$. $(*)$ is by $\eta_i \leq \frac{1}{4\|\nabla f(X_i)_r\|_*}$.
To upper bound ③, we use

$$\begin{aligned}
&\|U_i U_i^T - U_{i+1} U_{i+1}^T\|_{S_\infty} \\
=&\|U_i(U_i - U_{i+1})^T + (U_i - U_{i+1})U_i^T - (U_i - U_{i+1})(U_i - U_{i+1})^T\|_{S_\infty} \\
\leq&2\|U_i\|_{S_\infty}\|U_i - U_{i+1}\|_{S_\infty} + \|U_i - U_{i+1}\|_{S_\infty}^2 \\
=&2\eta_i\|U_i\|_{S_\infty}\|\nabla f(X_i)\cdot U_i\|_* + \eta_i^2\|\nabla f(X_i)\cdot U_i\|_*^2 \\
=&\eta_i\ \|\nabla f(X_i)\cdot U_i\|_*\ [2\|U_i\|_{S_\infty} + \eta_i\|\nabla f(X_i)\cdot U_i\|_*] \\
\overset{(1)}{\leq}&\eta_i\ \|\nabla f(X_i)\cdot U_i\|_*\ [2\|U_i\|_{S_\infty} + \eta_i\|\nabla f(X_i)_r\|_*\|U_i\|_{S_\infty}] \\
\overset{(2)}{\leq}&\eta_i\ \|\nabla f(X_i)\cdot U_i\|_*\frac{9}{4}\|U_i\|_{S_\infty}
\end{aligned} \tag{D.6}$$

where (1) is due to the rank of U_i is less than r and (2) is by $\eta_i \leq \frac{1}{4\|\nabla f(X_i)_r\|_*}$. Plugging above inequalities into (D.3), we obtain

$$\begin{aligned}
&f(U_i U_i^T) - f(U_{i+1}{U_{i+1}}^T) \\
\geq&2\eta_i\|\nabla f(X_i)U_i\|_*^2 - \frac{1}{4}\eta_i\|\nabla f(X_i)\cdot U_i\|_*^2 \\
&\qquad - \frac{L}{2}(\frac{9}{4}\eta_i\ \|U_i\|_{S_\infty}\|\nabla f(X_i)\cdot U_i\|_*)^2 \\
\geq&\eta_i\|\nabla f(X_i)\cdot U_i\|_*^2\left[\frac{7}{4} - \frac{L}{2}\left(\frac{9}{4}\right)^2\eta_i\|U_i\|_{S_\infty}^2\right] \\
\overset{(*)}{\geq}&1.117\ \eta_i\ \|\nabla f(X_i)\cdot U_i\|_*^2
\end{aligned} \tag{D.7}$$

$(*)$ is by $\eta_i \leq \frac{1}{4L\|X_i\|_{S_\infty}} = \frac{1}{4L\|U_i\|_{S_\infty}^2}$. We have thus finished the first part of lemma 7.
Now we give the proof of (D.2) of lemma 7.

We denote

$$R_{U_i} \equiv \underset{\substack{R \\ R \text{ is unitary}}}{\arg\min}\|U_i - U^* R\|_{S_\infty}. \tag{D.8}$$

and define $\Delta_{U_i} \equiv U_i - U^* R_{U_i}$. Then we have

$$\begin{aligned}
&f(U_i U_i^T) - f(U^* {U^*}^T) \\
\leq&\langle\nabla f(X_i), X_i - X^*\rangle \\
=&\left\langle\nabla f(X_i), \Delta_{U_i}U_i^T\right\rangle + \left\langle\nabla f(X_i), U_i\Delta_{U_i}^T\right\rangle - \left\langle\nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T\right\rangle \\
=&2\langle\nabla f(X_i)U_i, \Delta_{U_i}\rangle - \left\langle\nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T\right\rangle \\
\leq&2\|\nabla f(X_i)\cdot U_i\|_*\|\Delta_{U_i}\|_{S_\infty} + \underbrace{\left|\left\langle\nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T\right\rangle\right|}_{①}
\end{aligned} \tag{D.9}$$

To upper bound ①, we use

$$\begin{aligned}
\left\langle\nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T\right\rangle &= \langle\nabla f(X_i)\Delta_{U_i}, \Delta_{U_i}\rangle \\
&\leq \|\nabla f(X_i)\Delta_{U_i}\|_*\|\Delta_{U_i}\|_{S_\infty} \\
&= \|\nabla f(X_i)P_{\Delta_{U_i}}\Delta_{U_i}\|_*\|\Delta_{U_i}\|_{S_\infty} \\
&\leq \|\nabla f(X_i)P_{\Delta_{U_i}}\|_*\|\Delta_{U_i}\|_{S_\infty}^2 \\
&\overset{(*)}{\leq} \Big(\|\nabla f(X_i)P_{U_i}\|_* \\
&\quad + \|\nabla f(X_i)P_{U^*}\|_*\Big)\|\Delta_{U_i}\|_{S_\infty}^2
\end{aligned} \tag{D.10}$$

in which P_U denotes the projection onto $\mathbf{Col}(U)$, and $(*)$ is due to $\mathbf{Span}(\mathbf{Col}(\Delta_{U_i})) \subseteq \mathbf{Span}(\mathbf{Col}(U_i) \cup \mathbf{Col}(U_r^*))$. Continuing, we compute

$$
\begin{aligned}
\|\nabla f(X_i)P_{U_i}\|_* &= \|\nabla f(X_i)U_iU_i^\dagger\|_* \\
&\overset{(1)}{\leq} \|\nabla f(X_i)U_i\|_* \frac{1}{\sigma_r(U_i)} \\
&\overset{(2)}{\leq} \|\nabla f(X_i)U_i\|_* \frac{10}{9\sigma_r(U^*)}
\end{aligned} \tag{D.11}
$$

in which $U_i^\dagger$ denotes the pseudoinverse of U_i. Here (1) is due to $\sigma_1(U_i^\dagger) = \sigma_r(U_i)^{-1}$, and (2) is by assumption (IV.5), Weyl's inequality and $\sigma_r(U^* R_{U_i}) = \sigma_r(U^*)$. Similarly, we have

$$
\begin{aligned}
\|\nabla f(X_i)P_{U^*}\|_* &= \|\nabla f(X_i)U^*(U^*)^\dagger\|_* \\
&\leq \underbrace{\|\nabla f(X_i)U^*\|_*}_{\text{Ⓐ}} \frac{1}{\sigma_r(U^*)}.
\end{aligned} \tag{D.12}
$$

To upper bound Ⓐ, we use the following inequality.

$$
\begin{aligned}
\|\nabla f(X_i)U^*\|_* &= \|\nabla f(X_i)U^*R_{U_i}\|_* \\
&\leq \|\nabla f(X_i)U_i\|_* + \|\nabla f(X_i)\Delta_{U_i}\|_* \\
&= \|\nabla f(X_i)U_i\|_* + \|\nabla f(X_i)P_{\Delta_{U_i}}\Delta_{U_i}\|_* \\
&\leq \|\nabla f(X_i)U_i\|_* + \|\nabla f(X_i)P_{\Delta_{U_i}}\|_* \|\Delta_{U_i}\|_{S_\infty} \\
&\overset{(1)}{\leq} \|\nabla f(X_i)U_i\|_* + (\|\nabla f(X_i)P_{U_i}\|_* \\
&\qquad + \|\nabla f(X_i)P_{U^*}\|_*)\|\Delta_{U_i}\|_{S_\infty} \\
&\overset{(2)}{\leq} \|\nabla f(X_i)U_i\|_* + \Big(\frac{10}{9}\|\nabla f(X_i)U_i\|_* \\
&\qquad + \|\nabla f(X_i)U^*\|_*\Big)\frac{\|\Delta_{U_i}\|_{S_\infty}}{\sigma_r(U^*)} \\
&\overset{(3)}{\leq} \|\nabla f(X_i)U_i\|_* + \frac{1}{10}\Big(\frac{10}{9}\|\nabla f(X_i)U_i\|_* \\
&\qquad + \|\nabla f(X_i)U^*\|_*\Big) \\
&= \frac{10}{9}\|\nabla f(X_i)U_i\|_* + \frac{1}{10}\|\nabla f(X_i)U^*\|_*.
\end{aligned} \tag{D.13}
$$

Here, (1) is owing to the similar reason of (D.10), (2) is obtained by plugging in (D.11) and (D.12) and (3) is by assumption (IV.5). Thus we arrive at

$$
\|\nabla f(X_i)U^*\|_* \leq \left(\frac{10}{9}\right)^2 \|\nabla f(X_i)U_i\|_*. \tag{D.14}
$$

Plugging this into (D.12), we get

$$
\|\nabla f(X_i)P_{U^*}\|_* \leq \left(\frac{10}{9}\right)^2 \|\nabla f(X_i)U_i\|_* \frac{1}{\sigma_r(U^*)}. \tag{D.15}
$$

Combining (D.11) and (D.15) with (D.10), we obtain

$$
\begin{aligned}
\left\langle \nabla f(X_i), \Delta_{U_i}\Delta_{U_i}^T \right\rangle &\leq \Big(\|\nabla f(X_i)U_i\|_* \frac{10}{9\sigma_r(U^*)} \\
&\qquad + \left(\frac{10}{9}\right)^2 \|\nabla f(X_i)U_i\|_* \frac{1}{\sigma_r(U^*)}\Big)\|\Delta_{U_i}\|_{S_\infty}^2 \\
&= \|\nabla f(X_i)U_i\|_* \frac{190}{81}\frac{\|\Delta_{U_i}\|_{S_\infty}}{\sigma_r(U^*)}\|\Delta_{U_i}\|_{S_\infty} \\
&\overset{(*)}{\leq} \frac{19}{81}\|\nabla f(X_i)U_i\|_* \|\Delta_{U_i}\|_{S_\infty}
\end{aligned} \tag{D.16}
$$

where $(*)$ is by assumption (IV.5). Now we plug (D.16) into (D.9) and obtain

$$
f(U_iU_i^T) - f(U^*U^{*T}) \leq \left(2 + \frac{19}{81}\right)\|\Delta_{U_i}\|_{S_\infty}\|\nabla f(X_i)\cdot U_i\|_*. \tag{D.17}
$$

The last part is to prove $\min_i \gamma_i \geq \frac{1}{4}\overline{\eta}$ by showing $\|U_i\|_{S_\infty} \leq \frac{11}{9}\|U_0\|_{S_\infty}$ and

$$\|\nabla f(X_i)_r\|_* \leq \frac{40L}{81}\sigma_r(U_0)\sigma_1(U_0) + \|\nabla f(X_0)_r\|_*. \tag{D.18}$$

By assumption (IV.5) and Weyl's inequality, we have for every $i \geq 0$

$$\begin{aligned}&(1-\frac{1}{10})\sigma_1(U^*) \leq \sigma_1(U_i) \leq (1+\frac{1}{10})\sigma_1(U^*)\,,\ \text{and thus}\\ &\frac{1+\frac{1}{10}}{1-\frac{1}{10}}\sigma_1(U_0) \geq \sigma_1(U_i).\end{aligned} \tag{D.19}$$

Since $\|\nabla f(X_i)_r\|_*$ is the Ky Fan r-norm of $\nabla f(X_i)$, we have

$$\begin{aligned}\|\nabla f(X_i)_r\|_* &\leq \|((\nabla f(X_i) - \nabla f(X_0))_r\|_* + \|\nabla f(X_0)_r\|_*\\ &\leq \|\nabla f(X_i) - \nabla f(X_0)\|_* + \|\nabla f(X_0)_r\|_*\\ &\leq L_{S_\infty \to S_1}\|X_i - X_0\|_{S_\infty} + \|\nabla f(X_0)_r\|_*\\ &\leq L_{S_\infty \to S_1}\Big(\|X_i - X^*\|_{S_\infty} + \|X_0 - X^*\|_{S_\infty}\Big)\\ &\qquad + \|\nabla f(X_0)_r\|_*.\end{aligned}$$

Since

$$\begin{aligned}\|X_i - X^*\|_{S_\infty} &= \|U_i(U_i - U^*R_{U_i})^T\\ &\qquad + (U_i - U^*R_{U_i})(U^*R_{U_i})^T\|_{S_\infty}\\ &\leq \|U_i - U^*R_{U_i}\|_{S_\infty}(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty}),\end{aligned}$$

we have

$$\begin{aligned}&\|X_i - X^*\|_{S_\infty} + \|X_0 - X^*\|_{S_\infty}\\ \leq&\|U_i - U^*R_{U_i}\|_{S_\infty}(\|U_i\|_{S_\infty} + \|U^*\|_{S_\infty})\\ &\qquad + \|U_0 - U^*R_{U_0}\|_{S_\infty}(\|U_0\|_{S_\infty} + \|U^*\|_{S_\infty})\\ \leq&\frac{\sigma_r(U^*)}{10}\sigma_1(U_0)(\frac{11}{9} + \frac{10}{9} + 1 + \frac{10}{9})\\ \leq&\frac{1}{1-\frac{1}{10}}\frac{\sigma_r(U_0)}{10}\sigma_1(U_0)\frac{40}{9}\\ =&\frac{40}{81}\sigma_r(U_0)\sigma_1(U_0)\end{aligned}$$

by applying inequality (D.19).

Appendix E
Proof of Lemma 1

Using chain rules, we see that

$$\nabla f(A) = \nabla \mathrm{lse}(Ax)x^\top. \tag{E.1}$$

To prove the convexity of f, we compute

$$\begin{aligned}\langle \nabla f(A) - \nabla f(A'), A - A'\rangle &= \big\langle \nabla \mathrm{lse}(Ax)x^\top - \nabla \mathrm{lse}(A'x)x^\top, A - A'\big\rangle\\ &= \langle \nabla \mathrm{lse}(Ax) - \nabla \mathrm{lse}(A'x), Ax - A'x\rangle\\ &\geq 0\end{aligned}$$

since the lse function is convex.

We now turn to the smoothness parameters. Since transposing a matrix does not alter the Schatten-p norm, we have

$$\begin{aligned}\|\nabla f(A) - \nabla f(A')\|_F &= \Big\|\Big(\nabla \mathrm{lse}(Ax) - \nabla \mathrm{lse}(A'x)\Big)x^\top\Big\|_F\\ &\leq \|x\|_2\|\nabla \mathrm{lse}(Ax) - \nabla \mathrm{lse}(A'x)\|_2\\ &\leq \|x\|_2\|(A - A')x\|_2, \qquad \text{by (IV.2)}\\ &\leq \|x\|_2^2\|A - A'\|_F\end{aligned}$$

which proves (IV.8).

Using similar arguments, we compute

$$\begin{aligned}\|\nabla f(A)-\nabla f(A')\|_{S_1} &= \left\|\Big(\nabla \text{lse}(Ax)-\nabla \text{lse}(A'x)\Big)x^\top\right\|_{S_1} \\ &\le \|x\|_2\|\nabla \text{lse}(Ax)-\nabla \text{lse}(A'x)\|_2 \\ &\le \|x\|_2\|\nabla \text{lse}(Ax)-\nabla \text{lse}(A'x)\|_1 \\ &\le \|x\|_2\|(A-A')x\|_\infty, && \text{by (IV.3)} \\ &\le \|x\|_2^2\|A-A'\|_{S_\infty}\end{aligned}$$

which establishes (IV.9).

Appendix F
Proof of Lemma 2

The convexity of $\hat{f}$ follows immediately from the convexity of f.

For any two positive semi-definite matrices $Z_1 = \begin{bmatrix} A_1 & B_1 \\ B_1^\top & D_1\end{bmatrix}$ and $Z_2 = \begin{bmatrix} A_2 & B_2 \\ B_2^\top & D_2\end{bmatrix}$, we have

$$\begin{aligned}\|\nabla \hat{f}(Z_1)-\nabla \hat{f}(Z_2)\|_{S_q}^q &= \frac{1}{2}\left\|\begin{bmatrix} 0 & \nabla f(B_1)-\nabla f(B_2) \\ \nabla f(B_1)-\nabla f(B_2) & 0\end{bmatrix}\right\|_{S_q}^q \\ &\overset{(1)}{=} \frac{1}{2}\left\|\begin{bmatrix} \nabla f(B_1)-\nabla f(B_2) & 0 \\ 0 & \nabla f(B_1)-\nabla f(B_2)\end{bmatrix}\right\|_{S_q}^q \\ &\overset{(2)}{=} \frac{1}{2}\left(\|\nabla f(B_1)-\nabla f(B_2)\|_{S_q}^q + \|\nabla f(B_1)-\nabla f(B_2)\|_{S_q}^q\right) \\ &\overset{(3)}{\le} L^q\|B_1-B_2\|_{S_p}^q, \end{aligned} \tag{F.1}$$

where

(1) is because $||\cdot||_{S_q}$ is permutation invariant,
(2) is because of the block-diagonal structure, and
(3) uses the smoothness of f.

It remains to see that $\|B_1-B_2\|_{S_p} \le \|Z_1-Z_2\|_{S_p}$. In order to prove this, we use the permutation invariance of the $||\cdot||_{S_p}$, and the *Pinching inequality [1]*:

$$\begin{aligned}\|Z_1-Z_2\|_{S_p}^p &= \left\|\begin{bmatrix} B_1-B_2 & A_1-A_2 \\ D_1-D_2 & B_1^\top - B_2^\top\end{bmatrix}\right\|_{S_p}^p \\ &\ge \|B_1-B_2\|_{S_p}^p.\end{aligned}$$

Appendix G
Synthetic Data for Phase Retrieval

Two synthetic datasets are further presented in Figure 1 and 2. The results are in accordance with Section VI in the main text.

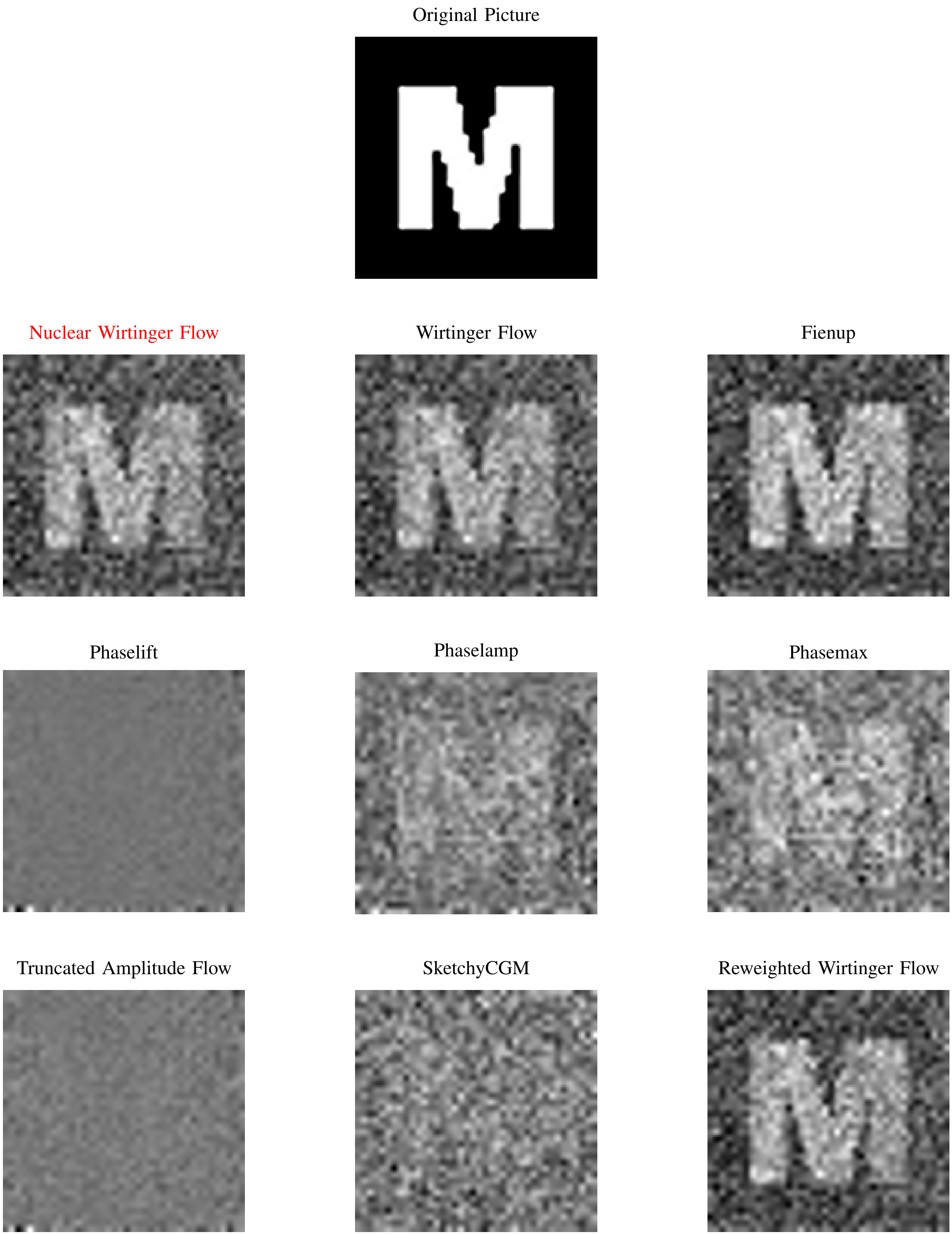


Fig. 1: Comparison of phase retrieval algorithms, synthetic dataset 2.

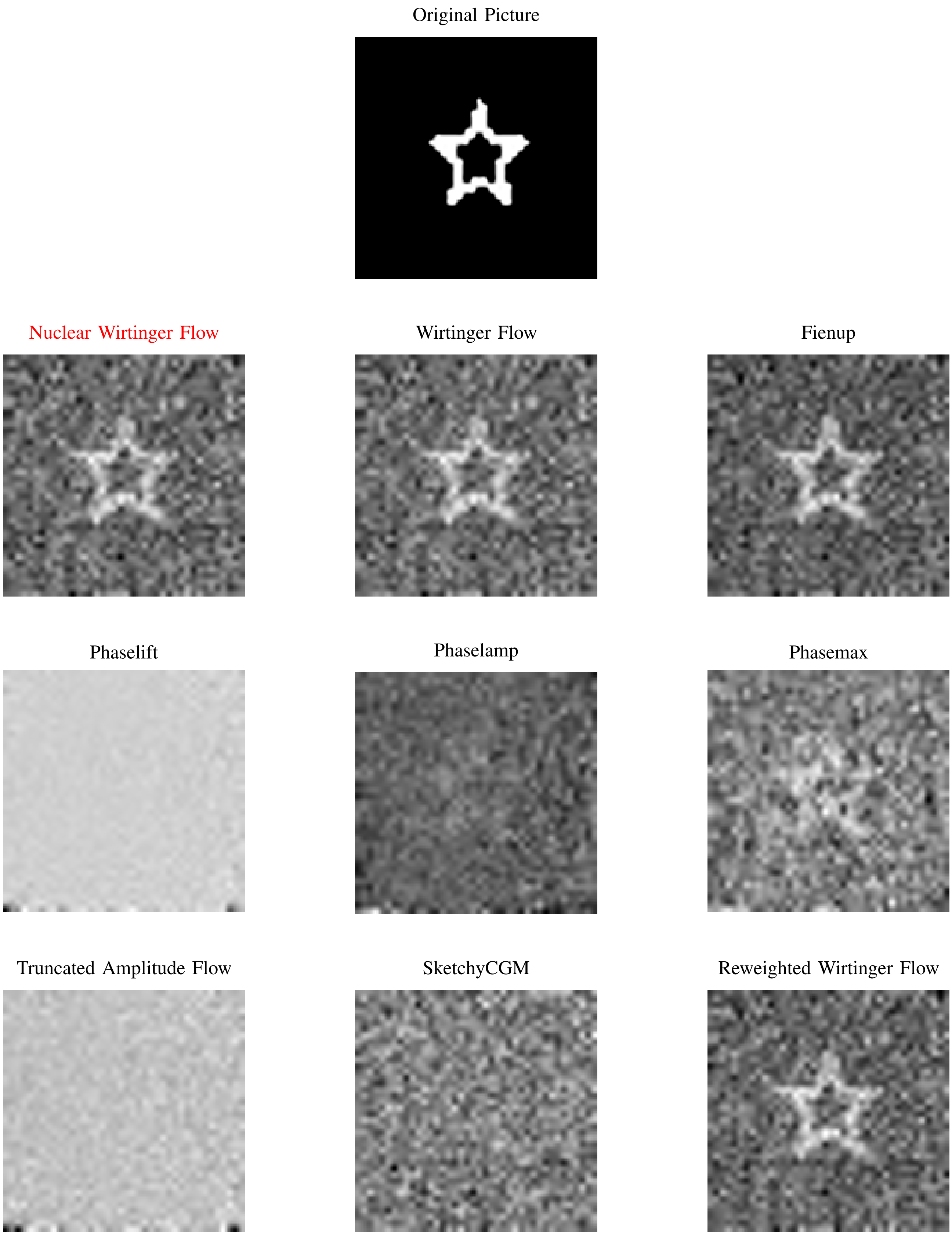


Fig. 2: Comparison of phase retrieval algorithms, synthetic dataset 3.

APPENDIX H
WIRTINGER FLOW V.S. NUCLEAR WIRTINGER FLOW

In Figure 4 we present the images recovered from nuclear Wirtinger flow and Wirtinger flow, indexed by time. Our experiments show that the nuclear Wirtinger flow quickly finds area (at $t = 4$s) where meaningful image characteristics start to emerge. At $t = 8$s, a fully visible image is recovered, and the reconstruction stays at the solution for a short period. However, the nuclear Wirtinger flow eventually overfits and returns a noisy figure; see Figure 3. This phenomenon is possibly due to the mismatch of the mathematical model and real Fourier Ptychographic reconstructions.

In contrast, the Wirtinger flow recovers only partial image at $t = 8$s, and exhibits oscillating behaviors. Eventually the Wirtinger flow overfits, and return solutions like random noise.

We stress that the Wirtinger flow fails to recover the image for *all* the initializations we have tried, whereas the nuclear Wirtinger flow is quite robust to the choice of initial point.

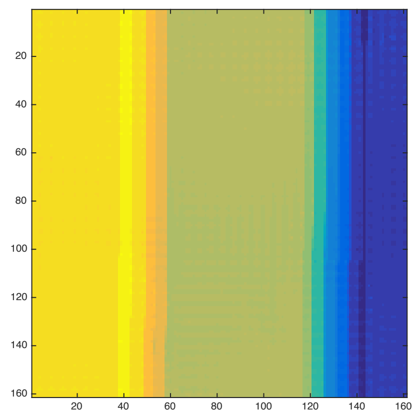


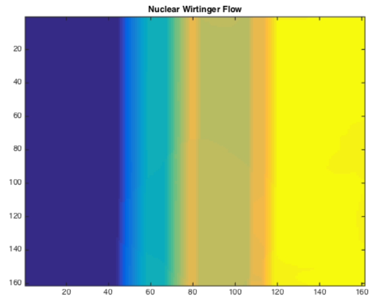
Fig. 3: Final solution of the nuclear Wirtinger flow, $t = 37$s.

APPENDIX I
SPECTRAL GRADIENT METHODS FOR FASTTEXT

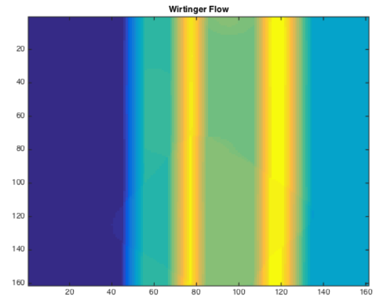
Four more datasets are presented in Figure 5. The results are in accordance with our observations in Section VI-B: The heuristic version of (V.4) is the best optimization algorithm, in that it solves the training problem most efficiently, but is prone to overfitting. On the other hand, the theoretical iterates (V.4) is either the best or comparable to the other methods in terms of prediction accuracy.

References

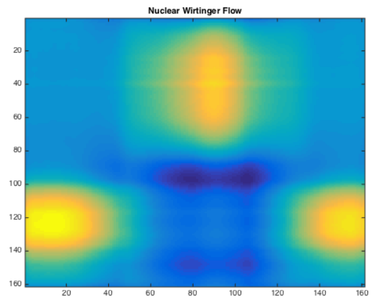
[1] R. Bhatia, *Matrix analysis*. Springer Science & Business Media, 2013, vol. 169.
[2] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 964–973.



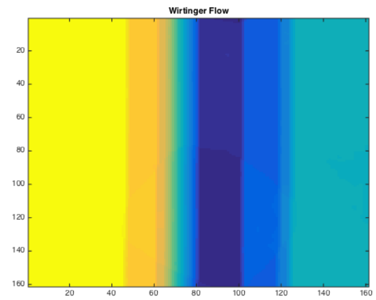
(a) Nuclear Wirtinger flow at $t = 2$s.



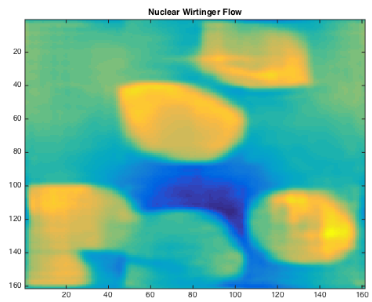
(b) Wirtinger flow at $t = 2$s.



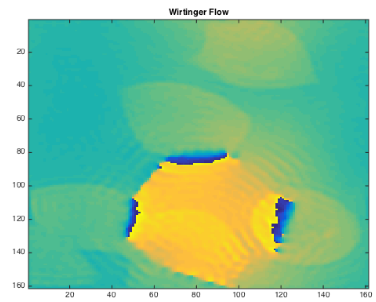
(c) Nuclear Wirtinger flow at $t = 4$s.



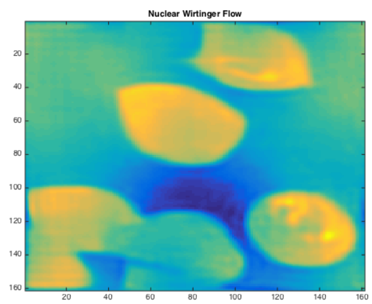
(d) Wirtinger flow at $t = 4$s.



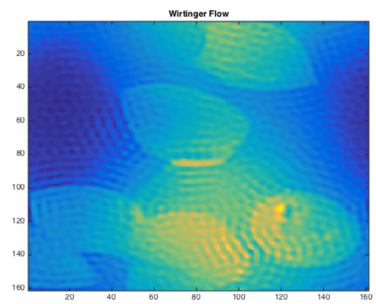
(e) Nuclear Wirtinger flow at $t = 8$s.



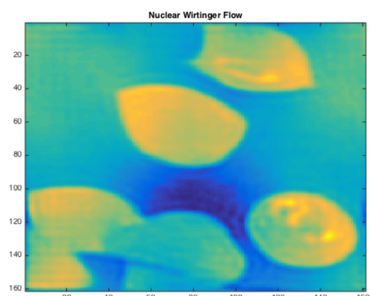
(f) Wirtinger flow at $t = 8$s.



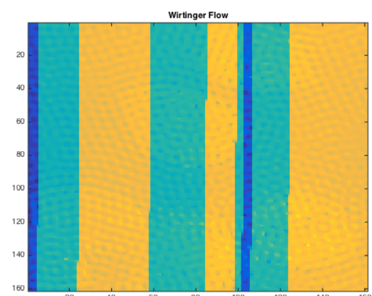
(g) Nuclear Wirtinger flow at $t = 14$s.



(h) Wirtinger flow at $t = 14$s.



(i) Nuclear Wirtinger flow at $t = 16$s.



(j) Wirtinger flow at $t = 16$s.

Fig. 4: Nuclear Wirtinger Flow v.s. Wirtinger Flow

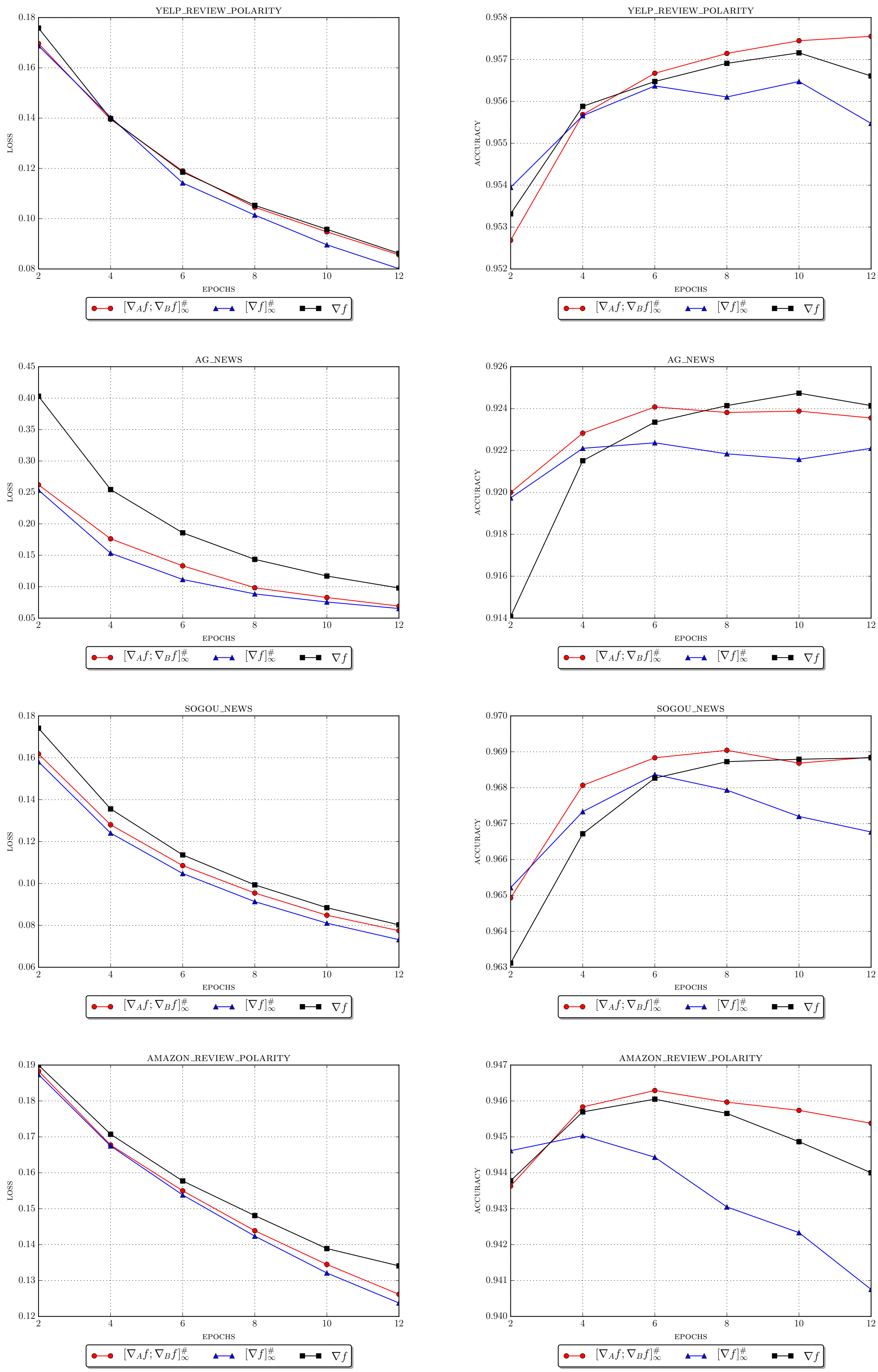


Fig. 5: From left to right, training loss and test accuracy. From top to bottom, results on `Yelp Review Polarity`, `AG News`, `Sogou News`, and `Amazon Review Polarity`