Supplementary Material of "Finding Global Optima in Nonconvex Stochastic Semidefinite Optimization with Variance Reduction"

In this Supplementary Material, we present the main proofs.

Problem, Algorithms and Main Results

In this section, we first describe the considered problem, and then propose the stochastic variance reduced gradient (SVRG) method for the considered problem, and finally present the main convergence results.

Stochastic Semidefinite Optimization

The stochastic convex semidefinite optimization problem, arising in many applications like matrix sensing [12, 22], ordinal embedding [6, 1], is of the following form:

$$\min_{X\in\mathbb{R}^{p\times p}} \ f(X)=\frac{1}{n}\sum_{i=1}^{n} f_i(X) \quad \text{s.t.} \quad X\succeq 0, \tag{1}$$

where $f_i(X)$ is some convex, smooth cost function associated with the i-th sample, $X \succeq 0$ is the positive semidefinite (PSD) constraint.

Since the generic semidefinite optimization methods are generally not well-scalable, a nonconvex reformulation based on the low-rank factorization became very popular in recent years [8, 9, 5]. The main idea is to recast the original problem (1) into an unconstrained problem via introducing another rectangular matrix $U \in \mathbb{R}^{p\times r}$ with $r < p$. Specifically, let $X = UU^T$ and $g(U) := f(UU^T)$, then the following alternative problem is considered:

$$\underset{U\in\mathbb{R}^{p\times r}}{\text{minimize}} \ g(U) \quad \text{where} \quad r\leq p. \tag{2}$$

Notations: For any two matrices $X, Y \in \mathbb{R}^{p\times p}$, their inner product is defined as $\langle X, Y\rangle = \mathrm{tr}(X^TY)$. We denote $\mathbb{S}_+^p$ as the set of positive semi-definite matrices of size $p \times p$. For any matrix $X \in \mathbb{R}^{p\times p}$, $\|X\|_F$ and $\|X\|_2$ denote its Frobenius and spectral norms, respectively, and $\sigma_{\min}(X)$ and $\sigma_{\max}(X)$ denote the smallest and largest *strictly positive* singular values of X, denote $\tau(X) := \frac{\sigma_{\max}(X)}{\sigma_{\min}(X)}$, with a slight abuse of notation, we also use $\sigma_1(X) \equiv \sigma_{\max}(X) \equiv \|X\|_2$, and X_r denotes the rank-r approximation of X via its truncated singular value decomposition (SVD) for any $r \leq p$. $\mathbf{I}_p$ denotes the identity matrix with the size $p \times p$. We will omit the subscript p of $\mathbf{I}_p$ if there is no confusion in the context.

Algorithm 1 SVRG for Problem (1)

Parameters: update frequency m, step size (or learning rate) $\{\eta_k\}$, initial point $\tilde{U}^0 \in \mathbb{R}^{p\times r}$
for k=0,1,... **do**
  $\tilde{X}^k := \tilde{U}^k(\tilde{U}^k)^T$
  $g_k = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\tilde{X}^k)\tilde{U}^k$
  $U^0 = \tilde{U}^k$
  for $t = 0, \ldots, m-1$ **do**
    $X^t = U^t {U^t}^T$
    Randomly pick $i_t \in \{1, \ldots, n\}$
    $U^{t+1} = U^t - \eta_k(\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k + g_k)$
  end for
  $\tilde{U}^{k+1} = U^m$
end for

SVRG

The SVRG method for solving (2) can be described as in Algorithm 1. We consider the following three step-size strategies in Algorithm 1:

(a) Fixed step size [13]:

$$\eta_k \equiv \eta, \quad \text{for some } \eta > 0. \tag{5}$$

(b) Barzilai-Borwein (BB) step size [3, 21]: given an initial $\eta_0 > 0$, and for $k \geq 1$, let $\tilde{g}_k := \nabla f(\tilde{X}^k)$,

$$\eta_k = \frac{1}{m}\cdot\frac{\|\tilde{X}^k-\tilde{X}^{k-1}\|_F^2}{|\langle \tilde{X}^k-\tilde{X}^{k-1}, \tilde{g}_k-\tilde{g}_{k-1}\rangle|}. \tag{6}$$

(c) Stabilized BB step size: given an initial $\eta_0 > 0$ and an $\epsilon \geq 0$, for $k \geq 1$,

$$\eta_k = \frac{1}{m}\times \tag{7}$$
$$\frac{\|\tilde{X}^k-\tilde{X}^{k-1}\|_F^2}{|\langle \tilde{X}^k-\tilde{X}^{k-1}, \tilde{g}_k-\tilde{g}_{k-1}\rangle| + \epsilon\|\tilde{X}^k-\tilde{X}^{k-1}\|_F^2}.$$

Throughout the rest of supplementary material, with a slight abuse, we still name the original SVRG with a fixed step size as **SVRG**, and call the SVRG with stabilized BB step size (7) as **SVRG-SBB$_\epsilon$**, and particularly, we call SVRG with BB step size as **SVRG-SBB$_0$**.

Main Assumptions

To present our main convergence results, we need the following assumptions.

Assumption 1 *Each f_i $(i = 1, \dots, n)$ satisfies the following:*

(a) f_i is L-Lipschitz differentiable for some constant $L > 0$, i.e., f_i is smooth and ∇f_i is Lipschitz continuous satisfying

$$\|\nabla f_i(X) - \nabla f_i(Y)\|_F \le L\|X - Y\|_F, \ \forall X, Y \in \mathbb{S}^p_+.$$

(b) f_i is (μ, r)-restricted strongly convex for some constants $\mu > 0$ and $r \le p$, i.e., for any $X, Y \in \mathbb{S}^p_+$ with rank-r

$$f_i(Y) \ge f_i(X) + \langle \nabla f_i(X), Y - X\rangle + \frac{\mu}{2}\|Y - X\|_F^2.$$

Let X^* be a global optimum of problem (1) with rank $r^* := \mathrm{rank}(X^*)$, X_r^* be its rank-r $(r \le r^*)$ best approximation via truncated singular value decomposition (SVD), and U_r^* be a decomposition of X_r^* via $X_r^* = U_r^* U_r^{*T}$. Under Assumption 1, we define the following constants:

$$\kappa := \frac{L}{\mu}, \quad \gamma_0 := \frac{2(\sqrt{2}-1)}{3\kappa}, \tag{8}$$

$$\bar{\eta} := \min\left\{ \frac{(1-\sqrt{\gamma_0})^2}{\frac{\|\nabla f(X_r^*)\|_F}{L\sigma_r(X_r^*)} + (2\sqrt{\gamma_0} + \gamma_0)\tau(U_r^*)}, 1 \right\}, \tag{9}$$

$$\xi := \bar{\eta}(1 - \bar{\eta}/2), \tag{10}$$

where $\tau(X_r^*) := \frac{\sigma_1(X_r^*)}{\sigma_r(X_r^*)}$ and $\tau(U_r^*) := \frac{\sigma_1(U_r^*)}{\sigma_r(U_r^*)}$. $\kappa \ge 1$ is generally called the *condition number* of the objective function. Thus, $0 < \gamma_0 \le \frac{2(\sqrt{2}-1)}{3}$ and $0 < \xi \le 1/2$.

Assumption 2 (rank-r approximation error) *Let X^* be a global optimum of problem (1), X_r^* be the rank-r approximation of X^* for a given positive integer $r \le r^* := \mathrm{rank}(X^*)$. The following holds*

$$\|X_r^* - X^*\|_F < \frac{\sqrt{2}-1}{\sqrt{3}} \xi^{1/2}\kappa^{-1} \cdot \sigma_r(X^*),$$

where κ is specified in (8), and $\sigma_r(X^*)$ is the r-th largest singular value of X^*.

Assumption 2 is a regular assumption used in literature (say, [5]). Under Assumption 2, we define several positive constants as follows: $\Delta := \frac{(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{3\kappa^2} - \xi\|X_r^* - X^*\|_F^2$, $\tilde{\Delta} := \frac{4(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{9\kappa^2} - \xi\|X_r^* - X^*\|_F^2$,

$$\gamma_l := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} - \sqrt{\Delta}, \tag{11}$$

$$\gamma_u := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} + \sqrt{\Delta}, \tag{12}$$

$$\tilde{\gamma}_l := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} - \sqrt{\tilde{\Delta}}, \tag{13}$$

$$\tilde{\gamma}_u := \frac{2(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} + \sqrt{\tilde{\Delta}}. \tag{14}$$

Note that the following relations hold

$$\gamma_l + \gamma_u = \frac{4(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}, \tag{15}$$

$$\tilde{\gamma}_l < \gamma_l < \gamma_u < \tilde{\gamma}_u \le \gamma_0\sigma_r(X_r^*), \tag{16}$$

where the last inequality of (16) holds for $0 < \xi \le 1/2$ and $\tilde{\gamma}_u \le 2\xi\gamma_0\sigma_r(X_r^*) \le \gamma_0\sigma_r(X_r^*)$.

We also need the following common assumption on the stochastic direction, which has been widely used in literature on stochastic algorithms (say, [7] and reference therein).

Assumption 3 (Unbiasedness) $\{\nabla f_{i_t}(X^t)U^t\}$ *satisfies* $\mathbb{E}_{i_t}[\nabla f_{i_t}(X^t)U^t] = \nabla f(X^t)U^t, \ \forall t \in \mathbb{N}$.

If i_t is uniformly sampled, then the above assumption can be satisfied. Under Assumptions 1-3, let $\mathcal{N}_{\gamma_0} := \{U : \|U - U_r^*\|_F^2 \le \gamma_0\sigma_r(X_r^*)\}$, and we define the following constants:

$$\mathcal{B} := \sup_{U \in \mathcal{N}_{\gamma_0}} \|UU^T\|_F, \tag{25}$$

$$B_0 := \sup_{U \in \mathcal{N}_{\gamma_0}} \left\{\mathbb{E}_{i_t}[\|\nabla f_{i_t}(UU^T)\|_F^2] - \|\nabla f(UU^T)\|_F^2\right\}, \tag{26}$$

$$B_1 := \sup_{U \in \mathcal{N}_{\gamma_0}} \|\nabla f(UU^T)\|_F^2, \tag{27}$$

$$B_2 := 4\left[2L^2\mathcal{B}(\mathcal{B} + \|X_r^*\|_F) + B_0 + B_1\right], \tag{17}$$

$$\theta := \frac{2\xi B_2}{L(\sqrt{\tilde{\Delta}} - \sqrt{\Delta})} = \frac{18 B_2\kappa\delta}{(\sqrt{2}-1)^2\xi\mu\sigma_r^2(X_r^*)}, \tag{18}$$

$$\eta_{\max} := \min\left\{\zeta_1, \zeta_2, \frac{1}{2\theta}\right\}, \tag{19}$$

where $\delta := \sqrt{\tilde{\Delta}} + \sqrt{\Delta}$, $\zeta_1 := \frac{1}{12\left[2L\cdot\kappa\mathcal{B} + \frac{B_0+B_1}{(\sqrt{2}-1)\mu\sigma_r(X_r^*)}\right]}$, $\zeta_2 := \frac{(\sqrt{2}-1)\mu\xi\sigma_r(X_r^*)}{12B_2}$.

Convergence Results

Let $\{\eta_k\}$ be a sequence satisfying $\eta_k \in (0, \eta_{\max})$ for any $k \in \mathbb{N}$. Given a positive integer m, define

$$\rho_k := 1 - \frac{\eta_k(\sqrt{2}-1)^2 \xi \mu \sigma_r^2(X_r^*)}{18\kappa\delta}, \tag{20}$$
$$\tilde{\rho}_k := \rho_k^m + (1-\rho_k^m)\eta_k\theta. \tag{21}$$

It is easy to check that $0 < \rho_k < 1$ and $0 < \tilde{\rho}_k < 1$. Based on the above defined constants, we present our main theorem as follows.

Theorem 1 (Linear convergence of SVRG) *Let $\{\tilde{U}^k\}$ be a sequence generated by Algorithm 1. Suppose that Assumptions 1-3 hold, and that $\eta_k \in (0, \eta_{\max})$. The following hold: (a) if $\gamma_l < \|\tilde{U}^0 - U_r^*\|_F^2 < \gamma_u$, there hold*

(a1) $\{\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2]\}$ is monotonically decreasing,

*(a2) (**Linear convergence**) for any $k \geq 1$,*

$$\begin{aligned} &\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \\ &\leq \left(\prod_{i=0}^{k-1}\tilde{\rho}_i\right)\cdot \|\tilde{U}^0 - U_r^*\|_F^2 + \tilde{\gamma}_l \times \\ &\left[\sum_{t=0}^{k-2}\left(\prod_{i=t+1}^{k-1}\tilde{\rho}_i \cdot (1-(\rho_t)^m)\right) + (1-(\rho_{k-1})^m)\right]. \end{aligned} \tag{22}$$

(b) In addition, if $\|\tilde{U}^0 - U_r^*\|_F^2 \leq \gamma_l$, then $\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \leq \gamma_l$ for any $k \in \mathbb{N}$.

The above theorem holds for a generic step size satisfying $\eta_k \in (0, \eta_{\max})$. In the following, we give a corollary to show the convergence of SVRG when adopting the considered three step-size strategies (5)-(7).

Corollary 1 (Convergence for different step sizes) *Under conditions of Theorem 1, all claims in Theorem 1 hold, if one of the following conditions holds:*

(1) $\eta \in (0, \eta_{\max})$ when a fixed step size is adopted;

(2) $m > \frac{1}{(\mu+\epsilon)\eta_{\max}}$ for any $\epsilon \geq 0$ when SBB step size is adopted.

From (11)-(13), if $r = r^*$ then $\|X_r^* - X^*\|_F = 0$, and thus $\tilde{\gamma}_l = 0$ and $\gamma_u = \frac{(2+\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}$. However, in this case, $\gamma_l = \frac{(2-\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} > 0$, and thus, we cannot claim the exact recovery of a global optimum directly from Theorem 1 even if $\|\tilde{U}^0 - U_r^*\|_F^2 \leq \gamma_l$. To circumvent this problem, we use a more consecutive step size, and get the following corollary showing the exact recovery of SVRG. Let

$$\bar{\eta}_{\max} := \min\left\{\frac{L\gamma_u}{2B_2\xi}, \eta_{\max}\right\}. \tag{23}$$

Corollary 2 (Exact recovery when $r = r^*$) *Let $\{\tilde{U}^k\}$ be a sequence generated by Algorithm 1. Let Assumptions 1 and 3 hold. If the following conditions hold: (a) $r = r^*$, (b) $\eta_k \in (0, \bar{\eta}_{\max})$, and (c) $\|\tilde{U}^0 - U_r^*\|_F^2 < \frac{(2+\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}$, then SVRG exactly recover the global optimum X^* in expectation at a linear rate.*

Proofs

For any matrix $U \in \mathbb{R}^{p\times r}$, let Q_U be a basis of the column space of U. Denote $\mathcal{P}_U := Q_U Q_U^T$. Then $\mathcal{P}_U \cdot U = U$. For any matrix $Y \in \mathbb{R}^{p\times p}$, $\mathcal{P}_U Y$ is a projection of Y onto the subspace spanned by $X := UU^T$.

Proof of Lemma 1

In the following, we describe a key lemma for the convergence of SVRG.

Lemma 1 (A key lemma) *Let $\{U^t\}_{t=0}^m$ be the sequence at the k-th inner loop. Let Assumptions 1, 2 and 3 hold. Let $\eta_k \in (0, \eta_{\max})$. If $\gamma_l < \mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] < \gamma_u$, then the sequence $\{\mathbb{E}[\|U^t - U_r^*\|_F^2]\}$ is monotonically decreasing for $t = 0, \ldots, m-1$, and*

$$\begin{aligned} &\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] \\ &\leq \|U^t - U_r^*\|_F^2 - \frac{2(\sqrt{2}-1)}{3}\eta_k\mu\sigma_r(X_r^*)\|U^t - U_r^*\|_F^2 \\ &+ \frac{\eta_k L}{2\xi}\|U^t - U_r^*\|_F^4 + \eta_k^2 B_2 \cdot \|\tilde{U}^k - U_r^*\|_F^2 \\ &+ \frac{\eta_k L}{2}\|X^* - X_r^*\|_F^2, \end{aligned} \tag{24}$$

where B_2 is specified in (17)*; while if $\mathbb{E}[\|\tilde{U}^k - U_r^*\|_F^2] \leq \gamma_l$, then $\mathbb{E}[\|U^t - U_r^*\|_F^2] \leq \gamma_l$ for any $t = 0, \ldots, m-1$.*

The sketch proof of Lemma 1 is show as follows. We prove this lemma by induction. Specifically, we first show that if $\gamma_l < \mathbb{E}[\|U^t - U_r^*\|_F^2] < \gamma_u$, then $\mathbb{E}[\|U^{t+1} - U_r^*\|_F^2] \leq \mathbb{E}[\|U^t - U_r^*\|_F^2] < \gamma_u$ for $t = 0, \ldots, m-1$. Furthermore, $\mathbb{E}[\|U^{t+1} - U_r^*\|_F^2]$ can be estimated via noting that

$$\begin{aligned} &\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] \\ &= \|U^t - U_r^*\|_F^2 + \eta_k^2\mathbb{E}_{i_t}[\|v_k^t\|_F^2] \\ &- 2\eta_k\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle], \end{aligned}$$

where $v_k^t = \nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k + \nabla f(\tilde{X}^k)\tilde{U}^k$, and establishing the bounds of both $\mathbb{E}_{i_t}[\|v_k^t\|_F^2]$ and $\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle]$ shown as the following two lemmas, respectively.

Lemma 2 (Bound of $2\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle]$) *Let Assumptions 1 and 3 hold. Let $\{U^t\}_{t=0}^{m-1}$ be a sequence* *generated by SVRG in Algorithm 1 at the k-th inner loop. Let $X^t = U^t {U^t}^T$ and $v_k^t = \nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k + \nabla f(\tilde{X}^k)\tilde{U}^k$. If $\|U^t - U_r^*\|_F^2 < \gamma_0 \sigma_r(X_r^*)$, then there holds*

$$\begin{aligned}
&2\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle] \\
&\geq \frac{\mu}{2}\|X^t - X_r^*\|_F^2 + \frac{\xi}{2L}\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2 \\
&- \frac{L}{2}\|X^* - X_r^*\|_F^2 - \frac{L}{2\xi}\|U^t - U_r^*\|_F^4.
\end{aligned}$$

where ξ is specified in (10).

Proof. By Assumption 3,

$$\begin{aligned}
&2\mathbb{E}_{i_t}[\langle v_k^t, U^t - U_r^*\rangle] \\
&= 2\langle \nabla f(X^t)U^t, U^t - U_r^*\rangle \\
&= 2\langle \nabla f(X^t), X^t - U_r^*{U^t}^T\rangle \\
&= \langle \nabla f(X^t), X^t - X_r^*\rangle \\
&+ \langle \nabla f(X^t), X^t + X_r^* - 2U_r^*{U^t}^T\rangle. \qquad (28)
\end{aligned}$$

To bound the first term of (28), we utilize the following three inequalities mainly by the Lipschitz differentiability and restricted strong convexity of f, that is,

(i) $f(X_r^*) \geq$
$f(X^t) + \langle \nabla f(X^t), X_r^* - X^t\rangle + \frac{\mu}{2}\|X_r^* - X^t\|_F^2$,

(ii) $f(X^t) \geq f(X^*) + (1-\bar{\eta}/2)\bar{\eta}L^{-1}\cdot\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2$,

(iii) $f(X^*) \geq f(X_r^*) - \frac{L}{2}\|X^* - X_r^*\|_F^2$,

where (i) holds for the (μ, r)-restricted strong convexity of f, (ii) holds the following inequality induced by the L-Lipschitz differentiability of f, i.e.,

$$\begin{aligned}
f(X^t) &\geq f(\bar{X}) + \langle \nabla f(X^t), X^t - \bar{X}\rangle - \frac{L}{2}\|X^t - \bar{X}\|_F^2 \\
&\quad (\text{where } \bar{X} := X^t - \frac{\bar{\eta}}{L}\mathcal{P}_{U^t}\nabla f(X^t)\mathcal{P}_{U^t}) \\
&= f(\bar{X}) + (1-\bar{\eta}/2)\bar{\eta}L^{-1}\cdot\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2,
\end{aligned}$$

and $f(\bar{X}) \geq f(X^*)$ since X^* is an optimum and $\bar{X}$ is a feasible point by Lemma 8(b), and (iii) holds for the L-Lipschitz differentiability of f and the optimality condition $\nabla f(X^*)U^* = 0$, i.e.,

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

where the equality holds for $\nabla f(X^*)U^* = 0$, which directly implies the following facts: $\nabla f(X^*)U_r^* = 0$, $\nabla f(X^*)X^* = 0$ and $\nabla f(X^*)X_r^* = 0$ due to $X^* = U^*{U^*}^T$ and $X_r^* = U_r^*{U_r^*}^T$. Summing the inequalities (i)-(iii) yields

$$\begin{aligned}
&\langle \nabla f(X^t), X^t - X_r^*\rangle \geq \frac{\mu}{2}\|X^t - X_r^*\|_F^2 \qquad (29) \\
&+ (1-\bar{\eta}/2)\bar{\eta}L^{-1}\cdot\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2 - \frac{L}{2}\|X^* - X_r^*\|_F^2.
\end{aligned}$$

On the other hand, we observe that

$$\begin{aligned}
&\langle \nabla f(X^t), X^t + X_r^* - 2U_r^*{U^t}^T\rangle \\
&= \langle \mathcal{P}_{U^t}\nabla f(X^t) + (\mathbf{I} - \mathcal{P}_{U^t})\nabla f(X^t), X^t + X_r^* - 2U_r^*{U^t}^T\rangle \\
&= \langle \mathcal{P}_{U^t}\nabla f(X^t), X^t + X_r^* - 2U_r^*{U^t}^T\rangle \\
&= \langle \mathcal{P}_{U^t}\nabla f(X^t), (U^t - U_r^*)(U - U_r^*)^T\rangle \\
&\geq -\frac{(1-\bar{\eta}/2)\bar{\eta}}{2L}\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2 \\
&- \frac{L}{2\bar{\eta}(1-\bar{\eta}/2)}\cdot\|U^t - U_r^*\|_F^4. \qquad (30)
\end{aligned}$$

where the second equality is due to $\langle(\mathbf{I} - \mathcal{P}_{U^t})\nabla f(X^t), X^t\rangle = 0$, $\langle(\mathbf{I} - \mathcal{P}_{U^t})\nabla f(X^t), U_r^*{U^t}^T\rangle = 0$ and $\langle(\mathbf{I} - \mathcal{P}_{U^t})\nabla f(X^t), X_r^*\rangle = 0$ by Lemma 8(c), the last equality holds for $X_r^* = U_r^*{U_r^*}^T$, and the inequality holds for the basic inequality: $\langle Y, Z\rangle \geq -\frac{c}{2}\|Y\|_F^2 - \frac{1}{2c}\|Z\|_F^2$ for any $Y, Z \in \mathbb{R}^{p\times p}$ and $c > 0$. Substituting (29) and (30) into (28) concludes this lemma. $\square$

Lemma 3 (Bound of $\mathbb{E}_{i_t}[\|v_k^t\|_F^2]$) *Let Assumptions 1, 2 and 3 hold. Assume that $\|U^t - U_r^*\|_F^2 < \gamma_u$ and $\|\tilde{U}^k - U_r^*\|_F^2 < \gamma_u$, then*

$$\begin{aligned}
\mathbb{E}_{i_t}[\|v_k^t\|_F^2] &\leq 4(B_0 + B_1)(\|U^t - U_r^*\|_F^2 + \|\tilde{U}^k - U_r^*\|_F^2) \\
&+ 4L^2\mathcal{B}(\|X^t - X_r^*\|_F^2 + \|\tilde{X}^k - X_r^*\|_F^2) \\
&+ \|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2\cdot\|X^t\|_F.
\end{aligned}$$

Proof. Note that

$$\begin{aligned}
\|v_k^t\|_F^2 &= \|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2 \\
&+ \|\nabla f(\tilde{X}^k)\tilde{U}^k\|_F^2 \\
&+ 2\langle \nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k, \nabla f(\tilde{X}^k)\tilde{U}^k\rangle.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{i_t}[\|v_k^t\|_F^2] \\
&= \mathbb{E}_{i_t}[\|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2] \\
&+ \|\nabla f(\tilde{X}^k)\tilde{U}^k\|_F^2 \\
&+ 2\langle \nabla f(X^t)U^t - \nabla f(\tilde{X}^k)\tilde{U}^k, \nabla f(\tilde{X}^k)\tilde{U}^k\rangle \\
&= \mathbb{E}_{i_t}[\|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2] \\
&- \|\nabla f(X^t)U^t - \nabla f(\tilde{X}^k)\tilde{U}^k\|_F^2 + \|\nabla f(X^t)U^t\|_F^2 \\
&\leq \mathbb{E}_{i_t}[\|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2] \\
&+ \|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2\cdot\|X^t\|_F, \qquad (31)
\end{aligned}$$

where the last inequality holds for $\|\nabla f(X^t)U^t - \nabla f(\tilde{X}^k)\tilde{U}^k\|_F^2 \geq 0$ and

$$\begin{aligned}
&\|\nabla f(X^t)U^t\|_F^2 \\
&= \|\mathcal{P}_{U^t}\nabla f(X^t)U^t + (\mathbf{I} - \mathcal{P}_{U^t})\nabla f(X^t)U^t\|_F^2 \\
&= \|\mathcal{P}_{U^t}\nabla f(X^t)U^t\|^2 \\
&\leq \|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2 \cdot \|X^t\|_F.
\end{aligned}$$

In the following, we bound the first term of (31). Note that

$$\begin{aligned}
&\|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2 \\
&= \|\nabla f_{i_t}(X^t)(U^t - \tilde{U}^k) + (\nabla f_{i_t}(X^t) - \nabla f_{i_t}(\tilde{X}^k))\tilde{U}^k\|_F^2 \\
&\leq 2\|\nabla f_{i_t}(X^t)(U^t - \tilde{U}^k)\|_F^2 \\
&+ 2\|(\nabla f_{i_t}(X^t) - \nabla f_{i_t}(\tilde{X}^k))\tilde{U}^k\|_F^2 \\
&\leq 2\|\nabla f_{i_t}(X^t)\|_F^2\|U^t - \tilde{U}^k\|_F^2 \\
&+ 2L^2\|\tilde{X}^k\|_F\|X^t - \tilde{X}^k\|_F^2,
\end{aligned}$$

which follows

$$\begin{aligned}
&\mathbb{E}_{i_t}[\|\nabla f_{i_t}(X^t)U^t - \nabla f_{i_t}(\tilde{X}^k)\tilde{U}^k\|_F^2] \\
&\leq 2(\mathbb{E}_{i_t}[\|\nabla f_{i_t}(X^t)\|_F^2] - \|\nabla f(X^t)\|_F^2)\|U^t - \tilde{U}^k\|_F^2 \\
&+ 2\|\nabla f(X^t)\|_F^2 \cdot \|U^t - \tilde{U}^k\|_F^2 + 2L^2\|\tilde{X}^k\|_F\|X^t - \tilde{X}^k\|_F^2 \\
&\leq 2(B_0 + B_1) \cdot \|U^t - \tilde{U}^k\|_F^2 + 2L^2\mathcal{B}\|X^t - \tilde{X}^k\|_F^2 \\
&\leq 4(B_0 + B_1) \cdot (\|U^t - U_r^*\|_F^2 + \|\tilde{U}^k - U_r^*\|_F^2) \\
&+ 4L^2\mathcal{B}(\|X^t - X_r^*\|_F^2 + \|\tilde{X}^k - X_r^*\|_F^2),
\end{aligned}$$

where B_0, B_1 and $\mathcal{B}$ are specified in (26), (27) and (25), respectively. Substituting the above inequality into (31), we can conclude this lemma. □

Based on the above two lemmas, we give the proof of Lemma 1.

Proof of Lemma 1: By Lemma 2 and Lemma 3,

$$\begin{aligned}
&\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] \leq \|U^t - U_r^*\|_F^2 \\
&- \eta_k\left[\frac{\mu}{2}\|X^t - X_r^*\|_F^2 + \frac{\xi}{2L}\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2\right] \\
&+ \eta_k\left[\frac{L}{2}\|X^* - X_r^*\|_F^2 + \frac{L}{2\xi}\|U^t - U_r^*\|_F^4\right] \\
&+ 4\eta_k^2(B_0 + B_1)(\|U^t - U_r^*\|_F^2 + \|\tilde{U}^k - U_r^*\|_F^2) \\
&+ 4\eta_k^2L^2\mathcal{B}\left(\|X^t - X_r^*\|_F^2 + \|\tilde{X}^k - X_r^*\|_F^2\right) \\
&+ \eta_k^2\|\mathcal{P}_{U^t}\nabla f(X^t)\|_F^2 \cdot \|X^t\|_F \\
&\leq \|U^t - U_r^*\|_F^2 - \left(\frac{\eta_k\mu}{2} - 4\eta_k^2L^2\mathcal{B}\right)\|X^t - X_r^*\|_F^2 \\
&+ 4\eta_k^2(B_0 + B_1)\|U^t - U_r^*\|_F^2 + \frac{\eta_kL}{2\xi}\|U^t - U_r^*\|_F^4 \\
&+ 4\eta_k^2\left[L^2\mathcal{B}\|\tilde{X}^k - X_r^*\|_F^2 + (B_0 + B_1)\|\tilde{U}^k - U_r^*\|_F^2\right] \\
&+ \frac{\eta_kL}{2}\|X^* - X_r^*\|_F^2 \\
&\leq \|U^t - U_r^*\|_F^2 - \eta_k\|U^t - U_r^*\|_F^2 \times \\
&\left[2(\sqrt{2}-1)\sigma_r(X_r^*)\left(\frac{\mu}{2} - 4\eta_kL^2\mathcal{B}\right) - 4\eta_k(B_0 + B_1)\right] \\
&+ \frac{\eta_kL}{2\xi}\|U^t - U_r^*\|_F^4 \\
&+ 4\eta_k^2\left[L^2\mathcal{B}\|\tilde{X}^k - X_r^*\|_F^2 + (B_0 + B_1)\|\tilde{U}^k - U_r^*\|_F^2\right] \\
&+ \frac{\eta_kL}{2}\|X^* - X_r^*\|_F^2, \\
&\leq \|U^t - U_r^*\|_F^2 + \frac{\eta_kL}{2\xi}\|U^t - U_r^*\|_F^4 - \eta_k\|U^t - U_r^*\|_F^2 \times \\
&\left[2(\sqrt{2}-1)\sigma_r(X_r^*)\left(\frac{\mu}{2} - 4\eta_kL^2\mathcal{B}\right) - 4\eta_k(B_0 + B_1)\right] \\
&+ 4\eta_k^2\left[2L^2\mathcal{B}(\mathcal{B} + \|X_r^*\|_F) + B_0 + B_1\right] \cdot \|\tilde{U}^k - U_r^*\|_F^2 \\
&+ \frac{\eta_kL}{2}\|X^* - X_r^*\|_F^2,
\end{aligned}$$

where the first inequality is due to Lemma 2 and Lemma 3, the second inequality holds for $\eta_k < \eta_{\max} \leq \frac{\xi}{2\mathcal{B}L}$, the third inequality holds for $\eta_k < \eta_{\max} \leq \frac{1}{8\mathcal{B}\kappa L}$ and Lemma 5(b), the final inequality holds for Lemma 5(a). Since $\eta_k < \eta_{\max} \leq \frac{(\sqrt{2}-1)\mu\sigma_r(X_r^*)}{12[2(\sqrt{2}-1)\sigma_r(X_r^*)L^2\mathcal{B}+B_0+B_1]}$, then

$$\begin{aligned}
&2(\sqrt{2}-1)\sigma_r(X_r^*)\left(\frac{\mu}{2} - 4\eta_kL^2\mathcal{B}\right) - 4\eta_k(B_0 + B_1) \\
&\geq \frac{2(\sqrt{2}-1)\mu\sigma_r(X_r^*)}{3}.
\end{aligned}$$

Thus, substituting the above inequality into the first main inequality yields (24).

Furthermore, by the assumption of this lemma and $\gamma_u \leq \gamma_0\sigma_r(X_r^*)$, we have

$$\|\tilde{U}^k - U_r^*\|_F^2 < \gamma_0\sigma_r(X_r^*).$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{i_t}[\|U^{t+1} - U_r^*\|_F^2] \\
&\leq \|U^t - U_r^*\|_F^2 + \frac{\eta_kL}{2\xi}\times \\
&\left[\|U^t - U_r^*\|_F^4 + \xi\|X^* - X_r^*\|_F^2 + \frac{2\xi\eta_kB_2}{L}\cdot\gamma_0\sigma_r(X_r^*)\right] \\
&- \frac{\eta_kL}{2\xi}\cdot\frac{4(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}\|U^t - U_r^*\|_F^2 \qquad (32) \\
&\leq \|U^t - U_r^*\|_F^2 + \frac{\eta_kL}{2\xi}\times \\
&\left[\|U^t - U_r^*\|_F^4 - \frac{4(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}\|U^t - U_r^*\|_F^2\right] \\
&+ \frac{\eta_kL}{2\xi}\cdot\left[\xi\|X^* - X_r^*\|_F^2 + \frac{(\sqrt{2}-1)^2\xi^2\sigma_r^2(X_r^*)}{9\kappa^2}\right],
\end{aligned}$$

where the second inequality holds for $\eta_k < \eta_{\max} \leq \frac{(\sqrt{2}-1)\mu\xi\sigma_r(X_r^*)}{12B_2}$, and $\eta_k < \eta_{\max} \leq \frac{(\sqrt{2}-1)^2\xi\mu\sigma_r(X_r^*)}{18\kappa B_2\gamma_0}$. By the definitions of γ_l (11) and γ_u (12), the above inequality implies

$$\mathbb{E}_{i_t}[\|U^{t+1}-U_r^*\|_F^2] \leq \|U^t-U_r^*\|_F^2 - \frac{\eta_k L}{2\xi}(\gamma_u - \|U^t-U_r^*\|_F^2)(\|U^t-U_r^*\|_F^2-\gamma_l),$$

which implies $\mathbb{E}[\|U^{t+1}-U_r^*\|_F^2] \leq \mathbb{E}[\|U^t-U_r^*\|_F^2]$ if $\gamma_l < \mathbb{E}[\|U^t-U_r^*\|_F^2] < \gamma_u$. Inductively, we can claim the first part of this lemma.

Define a univariate function $h(z) = z - \frac{\eta_k L}{2\xi}\cdot\big((\gamma_l+\gamma_u)z - z^2 - \gamma_l\cdot\gamma_u\big)$ for any $z\in\mathbb{R}_+$. Then its derivative is

$$\begin{aligned} h'(z) &= 1 - \frac{\eta_k L}{2\xi}\cdot(\gamma_l+\gamma_u) + \frac{\eta_k L}{\xi}\cdot z \\ &= 1 - (\sqrt{2}-1)\eta_k\mu\sigma_r(X_r^*) + \frac{\eta_k L}{\xi}\cdot z > 0, \end{aligned}$$

for $0 < z \leq \gamma_l$, where the second equality holds for (15), and the inequality is due to $1-(\sqrt{2}-1)\eta_k\mu\sigma_r(X_r^*) > 0$ for any $\eta_k\in(0,\eta_{\max})$. Thus, for any $0<z\leq\gamma_l$,

$$h(z) \leq h(\gamma_l) = \gamma_l,$$

which shows that the last statement of this lemma holds. Therefore, we end the proof of this lemma. □

Proof of Corollary 2

Proof. Note that $\bar{\eta}_{\max}\leq\eta_{\max}$. By Theorem 1, if

$$\begin{aligned} &\gamma_l := \frac{(2-\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} \\ &< \|\tilde{U}^0-U_r^*\|_F^2 < \frac{(2+\sqrt{3})\cdot(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa} := \gamma_u, \end{aligned}$$

then it is obvious that SVRG converges to the optimum X^* at a linear rate. As a consequence, we only need to prove the exact recovery of SVRG when $\|\tilde{U}^0-U_r^*\|_F^2 \leq \gamma_l$. By Theorem 1, in this case, $\mathbb{E}[\|\tilde{U}^k-U_r^*\|_F^2]\leq\gamma_l$ for all $k\in\mathbb{N}$. Actually, by the proof of Theorem 1, at any k-th inner loop,

$$\mathbb{E}[\|\tilde{U}^t-U_r^*\|_F^2] \leq \gamma_l \tag{33}$$

for any $t=1,\ldots,m$.

In this case, it is obvious that Lemmas 2 and 3 still hold, and (32) in the proof of Lemma 1 should be revised as

$$\begin{aligned} &\mathbb{E}_{i_t}[\|U^{t+1}-U_r^*\|_F^2] \\ &\leq \|U^t-U^*\|_F^2 + \frac{\eta_k L}{2\xi}\times \\ &\left[\|U^t-U^*\|_F^4 - \frac{4(\sqrt{2}-1)\xi\sigma_r(X_r^*)}{3\kappa}\|U^t-U^*\|_F^2\right] \\ &+ \eta_k^2 B_2\cdot\|\tilde{U}^k-U^*\|_F^2 \\ &\leq \left(1-\frac{\eta_k L}{2\xi}\cdot\gamma_u\right)\cdot\|U^t-U^*\|_F^2 \\ &+ \eta_k^2 B_2\cdot\|\tilde{U}^k-U^*\|_F^2, \end{aligned} \tag{34}$$

where the second inequality holds for (33) and (15). By (34), recursively, after some simplifications we have

$$\begin{aligned} &\mathbb{E}[\|\tilde{U}^{k+1}-U^*\|_F^2] \\ &\leq \left(1-\frac{\eta_k L}{2\xi}\cdot\gamma_u\right)^m \|\tilde{U}^k-U^*\|_F^2 \\ &+ \frac{2B_2\eta_k\xi}{L\gamma_u}\cdot\left[1-\left(1-\frac{\eta_k L}{2\xi}\cdot\gamma_u\right)^m\right]\|\tilde{U}^k-U^*\|_F^2 \end{aligned} \tag{35}$$

Since $\eta_k\in(0,\bar{\eta}_{\max})$, then

$$0 < \frac{2B_2\eta_k\xi}{L\gamma_u} < 1,$$

and thus,

$$\begin{aligned} 0 &< \left(1-\frac{\eta_k L}{2\xi}\cdot\gamma_u\right)^m \\ &+ \frac{2B_2\eta_k\xi}{L\gamma_u}\cdot\left[1-\left(1-\frac{\eta_k L}{2\xi}\cdot\gamma_u\right)^m\right] < 1, \end{aligned}$$

which implies that SVRG converges to X^* at a linear rate. Therefore, we finish the proof of this corollary. □

Appendix

In the appendix, we first present several lemmas, which are frequently used in this paper, and then provide the embedding results of *eurodist* dataset .

A. Several Important Lemmas

Lemma 4 ([1]) *Let A and B be two positive semi-definite matrices with the size $p\times p$. Assume that A is full rank, then*

$$\mathrm{tr}(AB) \geq \sigma_{\min}(A)\mathrm{tr}(B).$$

Lemma 5 *For any $U\in\mathbb{R}^{p\times r}$, let $X=UU^T$, $X_r^* = U_r^*{U_r^*}^T$, the following hold:*

(a) ***(Upper bound)*** $\|X-X_r^*\|_F^2 \le 2(\|X\|_F+\|X_r^*\|_F)\cdot \|U-U_r^*\|_F^2$, *and*

(b) ***(Lower bound)*** *if* $\|U-U_r^*\|_F \le \gamma\sigma_r(U_r^*)$ *for some* $0<\gamma<1$, *then*

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

Proof. (a) Note that

$$\begin{aligned}\|UU^T - U_r^*U_r^{*T}\|_F &= \|U(U-U_r^*)^T + (U-U_r^*)U_r^{*T}\|_F \\ &\le (\|U\|_F + \|U_r^*\|_F)\|U-U_r^*\|_F.\end{aligned}$$

Thus, $\|X-X_r^*\|_F^2 \le (\|U\|_F+\|U_r^*\|_F)^2\|U-U_r^*\|_F^2 \le 2(\|U\|_F^2+\|U_r^*\|_F^2)\|U-U_r^*\|_F^2$.

(b) For any $x\in\mathbb{R}^r$, note that

$$\begin{aligned}2x^TU^TU_r^*x &= \|Ux\|_2^2 + \|U_r^*x\|_2^2 - \|(U-U_r^*)x\|_2^2 \\ &\ge \|U_r^*x\|_2^2 - \|U-U_r^*\|_2^2\cdot\|x\|_2^2 \\ &\ge (1-\gamma^2)\sigma_r(X_r^*)\|x\|_2^2 \\ &\ge 0 \quad (\because 0<\gamma<1), \qquad (36)\end{aligned}$$

where the first inequality is due to $\|Ux\|_2^2 \ge 0$ and $\|(U-U_r^*)x\|_2 \le \|U-U_r^*\|_2\cdot\|x\|_2$, and the second inequality holds for $\|U_r^*x\|_2^2 \ge \sigma_r(X_r^*)\|x\|_2^2$ and $\|U-U_r^*\|_2^2 \le \|U-U_r^*\|_F^2 \le \gamma^2\sigma_r(X_r^*)$ by the assumption of this lemma. Thus, (36) implies

$$U^TU_r^* \succ 0, \qquad (37)$$

and $U^TU_r^*$ is full rank. Based on (37), we prove part (b). Let $H=U-U_r^*$. Then

$$\begin{aligned}\|X-X_r^*\|_F^2 &= \operatorname{tr}\left((H^TH)^2 + 4H^THH^TU_r^*\right) \\ &\quad + \operatorname{tr}\left(2(H^TU_r^*)^2 + 2U_r^{*T}U_r^*H^TH\right).\end{aligned}$$

Thus, establishing Lemma 5(b) is equivalent to show that

$$\begin{aligned}&\operatorname{tr}\left((H^TH)^2 + 4H^THH^TU_r^* + 2(H^TU_r^*)^2\right) \\ &+ \operatorname{tr}\left(2U_r^{*T}U_r^*H^TH - cH^TH\right) \\ &\ge 0,\end{aligned}$$

where $c:=2(\sqrt{2}-1)\sigma_r^2(U_r^*)$. By some simple derivations, we can observe that

$$\begin{aligned}&\operatorname{tr}\left((H^TH)^2 + 4H^THH^TU_r^* + 2(H^TU_r^*)^2\right) \\ &+ \operatorname{tr}\left(2U_r^{*T}U_r^*H^TH - cH^TH\right) \\ &= \operatorname{tr}\left((H^TH+\sqrt{2}H^TU_r^*)^2 + (4-2\sqrt{2})H^THH^TU_r^*\right) \\ &+ \operatorname{tr}\left(2U_r^{*T}U_r^*H^TH - cH^TH\right) \\ &\ge \operatorname{tr}\left((4-2\sqrt{2})H^THH^TU_r^* + 2U_r^{*T}U_r^*H^TH - cH^TH\right) \\ &= \operatorname{tr}\left(\left((4-2\sqrt{2})H^TU_r^* + 2U_r^{*T}U_r^* - c\mathbf{I}\right)\cdot H^TH\right).\end{aligned}$$

Recalling $H^TU_r^* = U^TU_r^* - U_r^{*T}U_r^*$, we have

$$\begin{aligned}&\operatorname{tr}\left(\left((4-2\sqrt{2})H^TU_r^* + 2U_r^{*T}U_r^* - c\mathbf{I}\right)\cdot H^TH\right) \\ &= \operatorname{tr}\left((4-2\sqrt{2})U^TU_r^*\cdot H^TH\right) \\ &+ \operatorname{tr}\left(\left(2(\sqrt{2}-1)U_r^{*T}U_r^* - c\mathbf{I}\right)\cdot H^TH\right) \\ &= \operatorname{tr}\left((2-\sqrt{2})(U^TU_r^*+U_r^{*T}U)\cdot H^TH\right) \\ &+ \operatorname{tr}\left(\left(2(\sqrt{2}-1)U_r^{*T}U_r^* - c\mathbf{I}\right)\cdot H^TH\right) \\ &\ge \operatorname{tr}\left(\left(2(\sqrt{2}-1)U_r^{*T}U_r^* - c\mathbf{I}\right)\cdot H^TH\right) \\ &\ge 0, \quad (\because c:=2(\sqrt{2}-1)\sigma_r^2(U_r^*),\ \text{Lemma 4})\end{aligned}$$

where the second equality is due to $\operatorname{tr}(U^TU_r^*H^TH) = \operatorname{tr}(H^THU_r^{*T}U) = \operatorname{tr}(U_r^{*T}UH^TH)$, and the first inequality holds for (37) and Lemma 4. Therefore, the above inequality implies

$$\begin{aligned}&\operatorname{tr}\left((H^TH)^2 + 4H^THH^TU_r^* + 2(H^TU_r^*)^2\right) \\ &+ \operatorname{tr}\left(2U_r^{*T}U_r^*H^TH - cH^TH\right) \ge 0,\end{aligned}$$

which concludes part (b) of this lemma. □

The following lemma is similar to [5, Lemma 18].

Lemma 6 *Let* $X=UU^T$ *and* $X_r^* = U_r^*U_r^{*T}$ *be two* $p\times p$ *rank-r positive semidefinite matrices. Let* $\|U-U_r^*\|_F \le \gamma\sigma_r(U_r^*)$ *for some constant* $0<\gamma<1$. *Then*

$$\|X-X_r^*\|_F \le (2\gamma+\gamma^2)\cdot\tau(U_r^*)\cdot\sigma_r(X_r^*),$$

where $\tau(U_r^*) := \frac{\sigma_1(U_r^*)}{\sigma_r(U_r^*)}$.

Proof. Note that

$$\begin{aligned}\|X-X_r^*\|_F &= \|U(U-U_r^*)^T + (U-U_r^*)U_r^{*T}\|_F \\ &\le \|U-U_r^*\|_F(\|U\|_2 + \|U_r^*\|_2) \\ &\le (2\|U_r^*\|_2 + \gamma\sigma_r(U_r^*))\|U-U_r^*\|_F \\ &\le (2+\gamma)\gamma\cdot\|U_r^*\|_2\cdot\sigma_r(U_r^*),\end{aligned}$$

where the first inequality holds for the triangle inequality, Cauchy-Schwartz inequality and the fact that the spectral norm is invariant with respect to the orthogonal transformation, the second inequality is due to the following sequence of inequalities, based on the hypothesis of the lemma:

$$\|U\|_2 - \|U_r^*\|_2 \le \|U-U_r^*\|_2 \le \|U-U_r^*\|_F \le \gamma\sigma_r(U_r^*),$$

and the last inequality holds for the fact $\sigma_r(U_r^*) \le \|U_r^*\|_2$ and the assumption of this lemma. The above inequality directly implies the claim of this lemma by the definition of $\tau(U_r^*)$. □

Moreover, we need modify [5, Lemma 19] as follows.

Lemma 7 *Let $X = UU^T$ and $X_r^* = U_r^* U_r^{*T}$ be two $p \times p$ rank-r positive semidefinite matrices. Let $\|U - U_r^*\|_F \leq \gamma \sigma_r(U_r^*)$ for some constant $0 < \gamma < 1$. Then*

$$\sigma_r(U) \geq (1-\gamma)\sigma_r(U_r^*).$$

Proof. Using the norm ordering $\|\cdot\|_2 \leq \|\cdot\|_F$ and the Weyl's inequality for perturbation of singular values (see, [2, Theorem 3.3.16]), we get

$$|\sigma_i(U) - \sigma_i(U_r^*)| \leq \gamma \sigma_r(U_r^*), \ 1 \leq i \leq r,$$

which implies that

$$\sigma_r(U) \geq (1-\gamma)\sigma_r(U_r^*).$$

□

Lemma 8 *Let Assumption 1 hold. Let $X = UU^T$ and $X_r^* = U_r^* U_r^{*T}$ be two $p \times p$ rank-r ($r < p$) positive semidefinite matrices. Suppse that $\|U - U_r^*\|_F^2 < \gamma_0 \sigma_r(X_r^*)$, where γ_0 is specified in (8). Then the following hold:*

(a) ***(Bounded gradient)***

$$\begin{aligned}&\|\nabla f(X)\|_F \\ &\leq \|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0) L \tau(U_r^*) \sigma_r(X_r^*),\end{aligned}$$

(b) ***(Feasibility of** $\bar{X}$**)*** *Let $\bar{X} := X - \frac{\bar{\eta}}{L}\mathcal{P}_U \nabla f(X) \mathcal{P}_U$, where $\bar{\eta}$ is specified in (9), then $\bar{X}$ is a feasible point, i.e., $\bar{X}$ is symmetric and positive semidefinite.*

(c) $(\mathbf{I} - \mathcal{P}_U) X_r^* = 0$.

Proof. **(a)** Note that

$$\begin{aligned}\|\nabla f(X)\|_F &\leq \|\nabla f(X_r^*)\|_F + L\|X - X_r^*\|_F \\ &\leq \|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0) L \tau(U_r^*) \sigma_r(X_r^*),\end{aligned}$$

where the first inequality holds for the L-Lipschitz differentiability of f, and the second inequality holds for Lemma 6.

(b) Since $\mathcal{P}_U X \mathcal{P}_U = X$ and X is rank-r, then

$$X - \frac{\bar{\eta}}{L} \cdot \mathcal{P}_U \nabla f(X) \mathcal{P}_U = \mathcal{P}_U (X - \frac{\bar{\eta}}{L} \cdot \nabla f(X)) \mathcal{P}_U,$$

which implies that $\bar{X}$ is symmetric and that the last $p-r$ eigenvalues of the matrix $X - \frac{\bar{\eta}}{L} \cdot \mathcal{P}_U \nabla f(X) \mathcal{P}_U$ are zero, that is, $\lambda_i(X - \frac{\bar{\eta}}{L} \cdot \mathcal{P}_U \nabla f(X) \mathcal{P}_U) = 0$ for $i = r+1, \ldots, p$. While for any $i = 1, \ldots, r$,

$$\begin{aligned}&\lambda_i \left(X - \frac{\bar{\eta}}{L} \cdot \mathcal{P}_U \nabla f(X) \mathcal{P}_U \right) \\ &\geq \lambda_i(X) - \frac{\bar{\eta}}{L} \cdot \lambda_{\max}(\mathcal{P}_U \nabla f(X) \mathcal{P}_U) \\ &\geq \sigma_r(X) - \frac{\bar{\eta}}{L} \cdot \sigma_{\max}(\nabla f(X)) \\ &\geq (1 - \sqrt{\gamma_0})^2 \sigma_r(X_r^*) \\ &- \frac{\bar{\eta}}{L} \cdot [\|\nabla f(X_r^*)\|_F + (2\sqrt{\gamma_0} + \gamma_0) L \tau(U_r^*) \sigma_r(X_r^*)] \\ &\geq 0,\end{aligned}$$

where the third inequality holds for Lemma 7 and **(a)** of this lemma, and the final inequality holds for the definition of $\bar{\eta}$. Therefore, $\bar{X}$ is positive semidefinite.

(c) By $\|U - U_r^*\|_F < \sqrt{\gamma_0} \sigma_r(U_r^*)$ and $0 < \sqrt{\gamma_0} < 1$, we have

$$\sigma_i(U) \cdot \sigma_i(U_r^*) > 0, \quad i \in \{1, \ldots, r\},$$

and

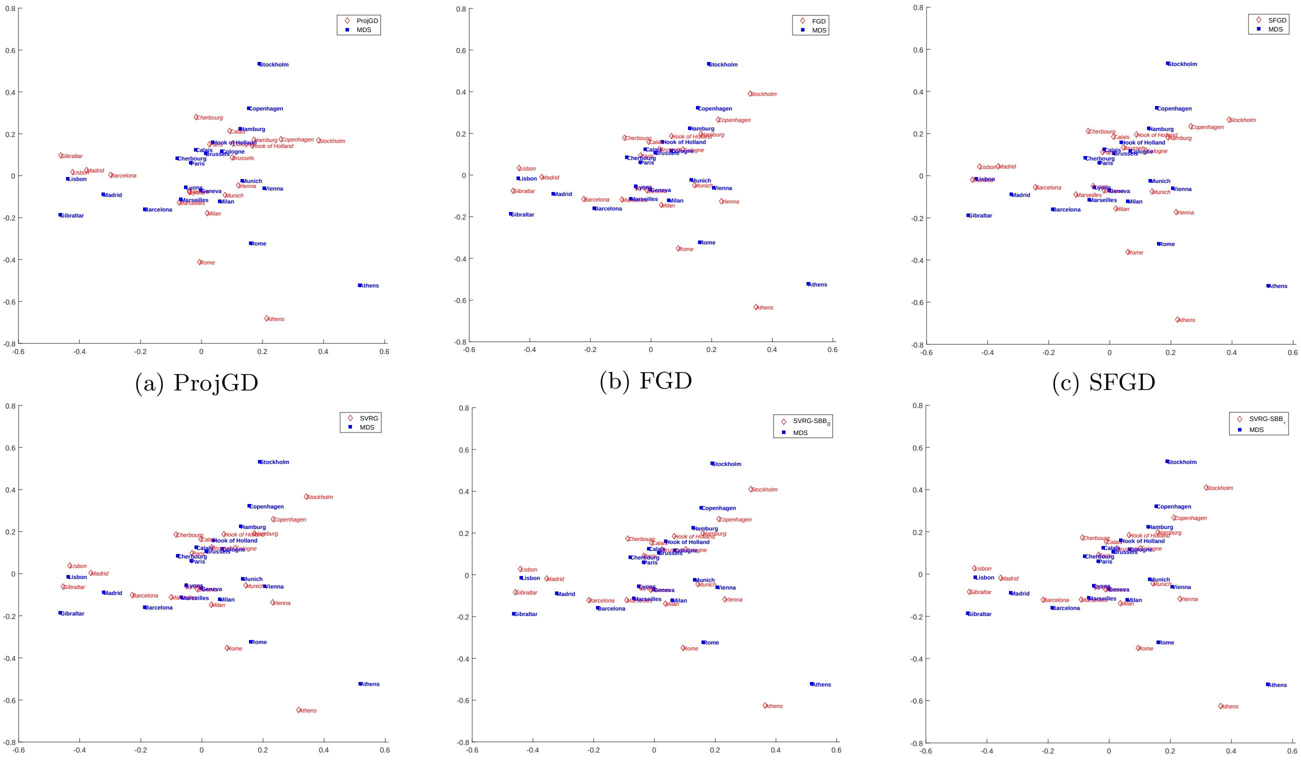
$$\sigma_i(U_r^*) = 0, \ \sigma_i(U) = 0, \quad i \in \{r+1, \ldots, p\},$$

which implies that U_r^* lies in the subspace spanned by U. In other words, U_r^* does not lie in the orthogonal subspace of the subspace spanned by U, that is, the following holds

$$(\mathbf{I} - \mathcal{P}_U) U_r^* = 0.$$

Thus, $(\mathbf{I} - \mathcal{P}_U) X_r^* = 0$. □

B. Embedding results of eurodist dataset



(a) ProjGD (b) FGD (c) SFGD

(d) SVRG (e) SVRG-SBB$_0$ (f) SVRG-SBB$_{0.1}$

Figure 5: Embedding results of eurodist dataset.

References

[1] R. Bhatia (1987). *Perturbation bounds for matrix eigenvalues*, volume 53, SIAM.

[2] R. A Horn, and C. R Johnson (1991). *Topics in matrix analysis*, Cambridge University Press, Cambridge.