Supplemental Material for "Low-Rank Similarity Metric Learning in High Dimensions"

Wei Liu[†] **Cun Mu**[‡] **Rongrong Ji**[♮] **Shiqian Ma**[§] **John R. Smith**[†] **Shih-Fu Chang**[‡]
[†]IBM T. J. Watson Research Center, Yorktown Heights, NY, USA {weiliu,jsmith}@us.ibm.com
[‡]Columbia University, New York, NY, USA cm3052@columbia.edu sfchang@ee.columbia.edu
[♮]Xiamen University, Xiamen, China rrji@xmu.edu.cn
[§]The Chinese University of Hong Kong, Hong Kong SAR, China sqma@se.cuhk.edu.hk

A Proofs

A.1 Proof of Lemma 1

Proof. Let us select a matrix $\mathbf{U}' \in \mathbb{R}^{d\times(d-m)}$ such that the columns of $[\mathbf{U},\mathbf{U}']$ form an orthonormal basis in $\mathbb{R}^d$. Notice $\mathbf{U}^\top\mathbf{U}' = \mathbf{0}$ and define a matrix $\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix} = [\mathbf{U},\mathbf{U}']^\top\mathbf{M}^\star[\mathbf{U},\mathbf{U}']$, where $\mathbf{Q}_{11} = \mathbf{U}^\top\mathbf{M}^\star\mathbf{U}$, $\mathbf{Q}_{12} = \mathbf{Q}_{21}^\top = \mathbf{U}^\top\mathbf{M}^\star\mathbf{U}'$, and $\mathbf{Q}_{22} = {\mathbf{U}'}^\top\mathbf{M}^\star\mathbf{U}'$. Since $\mathbf{M}^\star \in \mathbb{S}_+^d$, we know $\mathbf{Q} \in \mathbb{S}_+^d$, $\mathbf{Q}_{11} \in \mathbb{S}_+^m$, and $\mathbf{Q}_{22} \in \mathbb{S}_+^{d-m}$. We also have $\mathbf{M}^\star = [\mathbf{U},\mathbf{U}']\mathbf{Q}[\mathbf{U},\mathbf{U}']^\top$.

We first prove that $\mathbf{Q}_{22}$ must be $\mathbf{0}$ for any optimal $\mathbf{M}^\star$. On the contrary, we assume $\mathbf{Q}_{22} \neq \mathbf{0}$. Denote by $\mathbf{M}' = \mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top$ another feasible solution. We now compare the values of $f(\mathbf{M}^\star)$ and $f(\mathbf{M}')$. Since $[\mathbf{U},\mathbf{U}']$ is orthonormal, we have $\mathrm{range}(\mathbf{U}') = \big(\mathrm{range}(\mathbf{U})\big)^\perp = \big(\mathrm{range}(\mathbf{X})\big)^\perp$, which implies $\boldsymbol{x}^\top\mathbf{U}' = \mathbf{0}$ for any $\boldsymbol{x} \in \mathrm{range}(\mathbf{X})$. Accordingly, for any $i,j \in [1:n]$ we can derive

$$\begin{aligned}\boldsymbol{x}_i^\top\mathbf{M}^\star\boldsymbol{x}_j &= \boldsymbol{x}_i^\top[\mathbf{U},\mathbf{U}']\mathbf{Q}[\mathbf{U},\mathbf{U}']^\top\boldsymbol{x}_j\\ &= [\boldsymbol{x}_i^\top\mathbf{U},\mathbf{0}]\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top\boldsymbol{x}_j\\ \mathbf{0}\end{bmatrix}\\ &= \boldsymbol{x}_i^\top\mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top\boldsymbol{x}_j = \boldsymbol{x}_i^\top\mathbf{M}'\boldsymbol{x}_j.\end{aligned}$$

Then for the first term in $f(\cdot)$, we have

$$\sum_{i,j}[\tilde{y}_{ij} - y_{ij}\mathcal{S}_{\mathbf{M}^\star}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_+ \equiv \sum_{i,j}[\tilde{y}_{ij} - y_{ij}\mathcal{S}_{\mathbf{M}'}(\boldsymbol{x}_i,\boldsymbol{x}_j)]_+$$

For the second term (the trace norm), we have

$$\begin{aligned}\mathrm{tr}(\mathbf{M}^\star) &= \mathrm{tr}\left([\mathbf{U},\mathbf{U}']\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}[\mathbf{U},\mathbf{U}']^\top\right)\\ &= \mathrm{tr}\left(\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}\right) = \mathrm{tr}(\mathbf{Q}_{11}) + \mathrm{tr}(\mathbf{Q}_{22})\\ &> \mathrm{tr}(\mathbf{Q}_{11}) = \mathrm{tr}\big(\mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top\big) = \mathrm{tr}(\mathbf{M}'),\end{aligned}$$

in which the strict inequality holds due to the assumption $\mathbf{Q}_{22} \neq \mathbf{0}$. Then we obtain $f(\mathbf{M}') < f(\mathbf{M}^\star)$, which contradicts with the fact that $\mathbf{M}^\star$ is an optimal solution. Thus, we have shown $\mathbf{Q}_{22} \equiv \mathbf{0}$ for any optimal $\mathbf{M}^\star$.

On the other hand, $\mathbf{Q}_{22} = \mathbf{0}$ automatically makes $\mathbf{Q}_{12} = \mathbf{Q}_{21}^\top = \mathbf{0}$ since $\mathbf{Q} \in \mathbb{S}_+^d$, which quickly leads to $\mathbf{M}^\star = [\mathbf{U},\mathbf{U}']\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}[\mathbf{U},\mathbf{U}']^\top = \mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top$. Therefore, we can say that any optimal solution $\mathbf{M}^\star$ must be in the set $\big\{\mathbf{U}\mathbf{W}\mathbf{U}^\top | \mathbf{W} \in \mathbb{S}_+^m\big\}$. □

A.2 Proof of Theorem 1

Proof. The proof is straightforward by showing that

$$\begin{aligned} f(\mathbf{M}^\star) &= \min_{\mathbf{M}\in\mathbb{S}_+^d} \sum_{i,j=1}^n \left[\tilde{y}_{ij} - y_{ij}\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right]_+ + \alpha\mathrm{tr}(\mathbf{M}) \\ &= \min_{\mathbf{W}\in\mathbb{S}_+^m} \sum_{i,j=1}^n \left[\tilde{y}_{ij} - y_{ij}\boldsymbol{x}_i^\top \mathbf{U}\mathbf{W}\mathbf{U}^\top \boldsymbol{x}_j\right]_+ + \alpha\mathrm{tr}\big(\mathbf{U}\mathbf{W}\mathbf{U}^\top\big) \\ &= \min_{\mathbf{W}\in\mathbb{S}_+^m} \sum_{i,j=1}^n \left[\tilde{y}_{ij} - y_{ij}\tilde{\boldsymbol{x}}_i^\top \mathbf{W}\tilde{\boldsymbol{x}}_j\right]_+ + \alpha\mathrm{tr}(\mathbf{W}) = \tilde{f}(\mathbf{W}^\star), \end{aligned}$$

where the second line is due to Lemma 1. □

A.3 Proof of Theorem 2

Our target optimization problem can be rewritten as

$$\begin{aligned} \min \quad & f_1(\mathbf{Z}) + f_2(\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{Z} = \tilde{\mathbf{Y}} - \mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big), \end{aligned} \tag{A.1}$$

where $f_1(\mathbf{Z}) = \sum_{i,j=1}^n [\boldsymbol{z}_{ij}]_+$, $f_2(\mathbf{W}) = \alpha\mathrm{tr}(\mathbf{W}) + \mathcal{I}(\mathbf{W}\succeq 0)$, and $\mathcal{I}(\mathbf{W}\succeq 0) = 0$ if $\mathbf{W}\succeq 0$ and $+\infty$ otherwise. Note that $\mathcal{I}(\mathbf{W}\succeq 0)$ is a convex function.

Our linearized ADMM developed for solving Eq. (A.1) is

$$\mathbf{Z}^{k+1} := \arg\min_{\mathbf{Z}} \ f_1(\mathbf{Z}) + \frac{\rho}{2}\big\|\mathbf{Z} - \tilde{\mathbf{Y}} + \mathbf{Y}\circ\mathbf{S}^k + \mathbf{\Lambda}^k/\rho\big\|_{\mathrm{F}}^2, \tag{A.2}$$

$$\mathbf{W}^{k+1} := \arg\min_{\mathbf{W}} \ f_2(\mathbf{W}) + \frac{\rho}{2\tau}\left\|\mathbf{W} - \big(\mathbf{W}^k - \tau\tilde{\mathbf{X}}\mathbf{C}^k\tilde{\mathbf{X}}^\top - \tau\tilde{\mathbf{X}}\mathbf{S}^k\tilde{\mathbf{X}}^\top\big)\right\|_{\mathrm{F}}^2, \tag{A.3}$$

$$\mathbf{\Lambda}^{k+1} := \mathbf{\Lambda}^k + \rho\big(\mathbf{Z}^{k+1} - \tilde{\mathbf{Y}} + \mathbf{Y}\circ\mathbf{S}^{k+1}\big), \tag{A.4}$$

where $\mathbf{C}^k = \mathbf{Y}\circ\big(\mathbf{Z}^{k+1} + \mathbf{\Lambda}^k/\rho - \tilde{\mathbf{Y}}\big)$ and $\mathbf{S}^k = \tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}}$.

Before we prove the global convergence of the linearized ADMM, we need to prove the following lemma. We define an inner-product operator of matrices as $\langle\mathbf{A},\mathbf{B}\rangle = \mathrm{tr}\big(\mathbf{A}^\top\mathbf{B}\big)$.

Lemma 2. Suppose that $(\mathbf{Z}^\star, \mathbf{W}^\star)$ is an optimal solution of Eq. (A.1) and $\mathbf{\Lambda}^\star$ is the corresponding optimal dual variable. Let $\|\cdot\|_{op}$ denote the operator norm of matrices, and the initial $\mathbf{Z}^0$, $\mathbf{W}^0$ and $\mathbf{\Lambda}^0$ be any symmetric matrices. If the step size τ satisfies $0 < \tau < \frac{1}{\|\tilde{\mathbf{X}}\|_{op}^4}$, then there exists a $\xi > 0$ such that the sequence $\big\{\big(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\big)\big\}_k$ produced by the linearized ADMM in Eqs. (A.2)-(A.4) satisfies

$$\big\|\mathbf{R}^k - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 - \big\|\mathbf{R}^{k+1} - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 \geq \xi\big\|\mathbf{R}^k - \mathbf{R}^{k+1}\big\|_{\mathrm{H}}^2, \tag{A.5}$$

where $\mathbf{R}^\star = \begin{bmatrix}\mathbf{W}^\star & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}^\star\end{bmatrix}$, $\mathbf{R}^k = \begin{bmatrix}\mathbf{W}^k & \mathbf{0}\\ \mathbf{0} & \mathbf{\Lambda}^k\end{bmatrix}$, $\mathbf{H} = \begin{bmatrix}\frac{\rho}{\tau}\mathbf{I}_{m\times m} & \mathbf{0}\\ \mathbf{0} & \frac{1}{\rho}\mathbf{I}_{n\times n}\end{bmatrix}$, and the norm $\|\cdot\|_{\mathrm{H}}$ is defined by $\|\mathbf{R}\|_{\mathrm{H}} = \sqrt{\langle\mathbf{R},\mathbf{R}\rangle_{\mathrm{H}}}$ along with its corresponding inner-product $\langle\cdot,\cdot\rangle_{\mathrm{H}}$ being defined as $\langle\mathbf{A},\mathbf{B}\rangle_{\mathrm{H}} = \langle\mathbf{A},\mathbf{H}\mathbf{B}\rangle$.

Proof. Since $(\mathbf{Z}^\star, \mathbf{W}^\star, \mathbf{\Lambda}^\star)$ is optimal to Eq. (A.1), it follows the KKT conditions that lead to:

$$\mathbf{0} \in \partial f_1(\mathbf{Z}^\star) + \mathbf{\Lambda}^\star, \tag{A.6}$$

$$\mathbf{0} \in \partial f_2(\mathbf{W}^\star) + \tilde{\mathbf{X}}(\mathbf{\Lambda}^\star\circ\mathbf{Y})\tilde{\mathbf{X}}^\top, \tag{A.7}$$

and

$$\mathbf{0} = \mathbf{Z}^\star - \tilde{\mathbf{Y}} + \mathbf{Y}\circ\big(\tilde{\mathbf{X}}^\top\mathbf{W}^\star\tilde{\mathbf{X}}\big). \tag{A.8}$$

Note that the first-order optimality conditions for Eq. (A.2) are given by

$$\mathbf{0} \in \partial f_1\big(\mathbf{Z}^{k+1}\big) + \rho\big(\mathbf{Z}^{k+1} - \tilde{\mathbf{Y}} + \mathbf{Y} \circ \mathbf{S}^k + \mathbf{\Lambda}^k/\rho\big). \tag{A.9}$$

Using Eq. (A.4), Eq. (A.9) is reduced to

$$\mathbf{0} \in \partial f_1\big(\mathbf{Z}^{k+1}\big) + \mathbf{\Lambda}^{k+1} + \rho\mathbf{Y} \circ \big(\mathbf{S}^k - \mathbf{S}^{k+1}\big). \tag{A.10}$$

Combining Eqs. (A.6)(A.10) and utilizing the fact that $\partial f_1(\cdot)$ is a monotone operator, we have

$$\big\langle \mathbf{Z}^{k+1} - \mathbf{Z}^\star, \mathbf{\Lambda}^\star - \mathbf{\Lambda}^{k+1} - \rho\mathbf{Y} \circ \big(\mathbf{S}^k - \mathbf{S}^{k+1}\big)\big\rangle \geq 0. \tag{A.11}$$

The first-order optimality conditions for Eq. (A.3) are given by

$$\mathbf{0} \in \partial f_2\big(\mathbf{W}^{k+1}\big) + \frac{\rho}{\tau}\big(\mathbf{W}^{k+1} - \mathbf{W}^k + \tau\tilde{\mathbf{X}}\mathbf{C}^k\tilde{\mathbf{X}}^\top + \tau\tilde{\mathbf{X}}\mathbf{S}^k\tilde{\mathbf{X}}^\top\big). \tag{A.12}$$

Using Eq. (A.4), Eq. (A.12) can be reduced to

$$\mathbf{0} \in \partial f_2\big(\mathbf{W}^{k+1}\big) + \frac{\rho}{\tau}\left(\mathbf{W}^{k+1} - \mathbf{W}^k + \tau\tilde{\mathbf{X}}\big(\mathbf{Y} \circ \mathbf{\Lambda}^{k+1}/\rho\big)\tilde{\mathbf{X}}^\top + \tau\tilde{\mathbf{X}}\big(\mathbf{S}^k - \mathbf{S}^{k+1}\big)\tilde{\mathbf{X}}^\top\right). \tag{A.13}$$

Combining Eqs. (A.7)(A.13) and also utilizing the fact that $\partial f_2(\cdot)$ is a monotone operator, we have

$$\left\langle \mathbf{W}^{k+1} - \mathbf{W}^\star, \tilde{\mathbf{X}}\big((\mathbf{\Lambda}^\star - \mathbf{\Lambda}^{k+1}) \circ \mathbf{Y}\big)\tilde{\mathbf{X}}^\top - \frac{\rho}{\tau}\big(\mathbf{W}^{k+1} - \mathbf{W}^k\big) - \rho\tilde{\mathbf{X}}\big(\mathbf{S}^k - \mathbf{S}^{k+1}\big)\tilde{\mathbf{X}}^\top \right\rangle \geq 0. \tag{A.14}$$

Integrating Eqs. (A.11)(A.14)(A.4)(A.8), we obtain

$$\frac{\rho}{\tau}\big\langle \mathbf{W}^{k+1} - \mathbf{W}^\star, \mathbf{W}^k - \mathbf{W}^{k+1}\big\rangle + \frac{1}{\rho}\big\langle \mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^\star, \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}\big\rangle \geq -\left\langle \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}, \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\rangle. \tag{A.15}$$

Using the notations of $\mathbf{R}^k$, $\mathbf{R}^\star$ and $\mathbf{H}$, Eq. (A.15) can be rewritten as

$$\big\langle \mathbf{R}^{k+1} - \mathbf{R}^\star, \mathbf{R}^k - \mathbf{R}^{k+1}\big\rangle_{\mathrm{H}} \geq -\left\langle \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}, \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\rangle, \tag{A.16}$$

which can further be written as

$$\big\langle \mathbf{R}^k - \mathbf{R}^\star, \mathbf{R}^k - \mathbf{R}^{k+1}\big\rangle_{\mathrm{H}} \geq \big\|\mathbf{R}^k - \mathbf{R}^{k+1}\big\|_{\mathrm{H}} - \left\langle \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}, \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\rangle. \tag{A.17}$$

Combining Eq. (A.17) with the following equation

$$\big\|\mathbf{R}^{k+1} - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 = \big\|\mathbf{R}^{k+1} - \mathbf{R}^k\big\|_{\mathrm{H}}^2 - 2\big\langle \mathbf{R}^k - \mathbf{R}^{k+1}, \mathbf{R}^k - \mathbf{R}^\star\big\rangle_{\mathrm{H}} + \big\|\mathbf{R}^k - \mathbf{R}^\star\big\|_{\mathrm{H}}^2,$$

we derive

$$\begin{aligned}
&\big\|\mathbf{R}^k - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 - \big\|\mathbf{R}^{k+1} - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 \\
&= 2\big\langle \mathbf{R}^k - \mathbf{R}^{k+1}, \mathbf{R}^k - \mathbf{R}^\star\big\rangle_{\mathrm{H}} - \big\|\mathbf{R}^{k+1} - \mathbf{R}^k\big\|_{\mathrm{H}}^2 \\
&\geq \big\|\mathbf{R}^{k+1} - \mathbf{R}^k\big\|_{\mathrm{H}}^2 - 2\left\langle \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}, \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\rangle.
\end{aligned} \tag{A.18}$$

Let $\mathcal{A}$ represent the linear operator $\mathcal{A}[\mathbf{W}] = \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)$ with $\mathcal{A}^\star$ being its adjoint operator. The operator norm of $\mathcal{A}$ is defined as $\|\mathcal{A}\|_{op} := \sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \big\|\mathcal{A}[\mathbf{W}]\big\|_{\mathrm{F}}$. Then we have

$$\begin{aligned}
\|\mathcal{A}\|_{op}^2 &= \sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \big\|\mathcal{A}[\mathbf{W}]\big\|_{\mathrm{F}}^2 = \sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \left\|\mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big)\right\|_{\mathrm{F}}^2 = \sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \big\|\tilde{\mathbf{X}}^\top\mathbf{W}\tilde{\mathbf{X}}\big\|_{\mathrm{F}}^2 \\
&\leq \sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \Big(\|\tilde{\mathbf{X}}\|_{op}\|\mathbf{W}\|_{\mathrm{F}}\|\tilde{\mathbf{X}}\|_{op}\Big)^2 = \|\tilde{\mathbf{X}}\|_{op}^2 \left(\sup_{\|\mathbf{W}\|_{\mathrm{F}}=1} \|\mathbf{W}\|_{\mathrm{F}}^2\right) \|\tilde{\mathbf{X}}\|_{op}^2 = \|\tilde{\mathbf{X}}\|_{op}^4 = \|\mathbf{X}\|_{op}^4,
\end{aligned} \tag{A.19}$$

where we have used the fact that for any matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ of compatible size, $\|\mathbf{AB}\|_{\mathrm{F}} \leq \|\mathbf{A}\|_{op}\|\mathbf{B}\|_{\mathrm{F}}$ and $\|\mathbf{BC}\|_{\mathrm{F}} \leq \|\mathbf{B}\|_{\mathrm{F}}\|\mathbf{A}\|_{op}$. Let $\eta = \frac{\tau\|\mathcal{A}\|_{op}^2+1}{2\rho}$, then $\frac{\tau\|\mathcal{A}\|_{op}^2}{\rho} < \eta < \frac{1}{\rho}$ holds because $0 < \tau < \frac{1}{\|\tilde{\mathbf{X}}\|_{op}^4} \leq \frac{1}{\|\mathcal{A}\|_{op}^2}$. By taking advantage of the Cauchy-Schwartz inequality, we can derive

$$\begin{aligned} & -2\left\langle \mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}, \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\rangle \\ \geq\ & -\eta\big\|\mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}\big\|_{\mathrm{F}}^2 - \frac{1}{\eta}\left\|\mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top(\mathbf{W}^k - \mathbf{W}^{k+1})\tilde{\mathbf{X}}\big)\right\|_{\mathrm{F}}^2 \\ \geq\ & -\eta\big\|\mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}\big\|_{\mathrm{F}}^2 - \frac{\|\mathcal{A}\|_{op}^2}{\eta}\big\|\mathbf{W}^k - \mathbf{W}^{k+1}\big\|_{\mathrm{F}}^2. \end{aligned} \tag{A.20}$$

Combining Eqs. (A.18)(A.20), eventually we obtain

$$\begin{aligned} & \big\|\mathbf{R}^k - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 - \big\|\mathbf{R}^{k+1} - \mathbf{R}^\star\big\|_{\mathrm{H}}^2 \\ \geq\ & \left(\frac{\rho}{\tau} - \frac{\|\mathcal{A}\|_{op}^2}{\eta}\right)\big\|\mathbf{W}^k - \mathbf{W}^{k+1}\big\|_{\mathrm{F}}^2 + \left(\frac{1}{\rho} - \eta\right)\big\|\mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1}\big\|_{\mathrm{F}}^2 \\ \geq\ & \xi\big\|\mathbf{R}^k - \mathbf{R}^{k+1}\big\|_{\mathrm{H}}^2, \end{aligned} \tag{A.21}$$

where $\xi = \min\big\{\frac{\rho}{\tau} - \frac{\|\mathcal{A}\|_{op}^2}{\eta}, \frac{1}{\rho} - \eta\big\} > 0$. This completes the proof. □

We are now ready to give the main convergence result of the linearized ADMM. Recall our Theorem 2 in the main paper.

Theorem 2. *Given* $0 < \tau < \frac{1}{\|\tilde{\mathbf{X}}\|_{op}^4} = \frac{1}{\|\mathbf{X}\|_{op}^4}$, *the sequence* $\big\{\big(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\big)\big\}_k$ *generated by the linearized ADMM in Eqs.* (A.2)-(A.4) *starting with any symmetric* $\big(\mathbf{Z}^0, \mathbf{W}^0, \mathbf{\Lambda}^0\big)$ *converges to an optimal solution of the original problem in Eq.* (A.1).

Proof. Due to Lemma 2, we can easily achieve that

- (i) $\big\|\mathbf{R}^k - \mathbf{R}^{k+1}\big\|_{\mathrm{H}} \to 0$;
- (ii) $\big\{\mathbf{R}^k\big\}_k$ lies in a compact region;
- (iii) $\big\|\mathbf{R}^k - \mathbf{R}^\star\big\|_{\mathrm{H}}^2$ is monotonically non-increasing and thus converges.

It follows from (i) that $\mathbf{\Lambda}^k - \mathbf{\Lambda}^{k+1} \to \mathbf{0}$ and $\mathbf{W}^k - \mathbf{W}^{k+1} \to \mathbf{0}$. Then Eq. (A.4) implies $\mathbf{Z}^k - \mathbf{Z}^{k+1} \to \mathbf{0}$ and $\mathbf{Z}^k - \tilde{\mathbf{Y}} + \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}}\big) \to \mathbf{0}$. From (ii) we know that $\big\{\mathbf{R}^k\big\}_k$ must have a subsequence $\big\{\mathbf{R}^{k_j}\big\}_j$ converging to $\hat{\mathbf{R}} = \big(\hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}\big)$, *i.e.*, $\mathbf{W}^{k_j} \to \hat{\mathbf{W}}$ and $\mathbf{\Lambda}^{k_j} \to \hat{\mathbf{\Lambda}}$. $\mathbf{Z}^k - \tilde{\mathbf{Y}} + \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\mathbf{W}^k\tilde{\mathbf{X}}\big) \to \mathbf{0}$ leads to $\mathbf{Z}^{k_j} \to \hat{\mathbf{Z}} = \tilde{\mathbf{Y}} - \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\hat{\mathbf{W}}\tilde{\mathbf{X}}\big)$. Therefore, $\big(\hat{\mathbf{Z}}, \hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}\big)$ is a limit point of the sequence $\big\{\big(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\big)\big\}_k$. Note that Eq. (A.10) implies

$$\mathbf{0} \in \partial f_1\big(\hat{\mathbf{Z}}\big) + \hat{\mathbf{\Lambda}}, \tag{A.22}$$

and Eq. (A.13) implies

$$\mathbf{0} \in \partial f_2\big(\hat{\mathbf{W}}\big) + \tilde{\mathbf{X}}\big(\hat{\mathbf{\Lambda}} \circ \mathbf{Y}\big)\tilde{\mathbf{X}}^\top. \tag{A.23}$$

Eqs. (A.22)(A.23) and $\hat{\mathbf{Z}} - \tilde{\mathbf{Y}} + \mathbf{Y} \circ \big(\tilde{\mathbf{X}}^\top\hat{\mathbf{W}}\tilde{\mathbf{X}}\big) = \mathbf{0}$ imply that $\big(\hat{\mathbf{Z}}, \hat{\mathbf{W}}, \hat{\mathbf{\Lambda}}\big)$ satisfies the KKT conditions for Eq. (A.1) and is thus an optimal solution to Eq. (A.1). Therefore, we have shown that any limit point of $\big\{\big(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\big)\big\}_k$ is an optimal solution to Eq. (A.1).

To complete the proof, we need to further show that such a limit point is unique. Let $\big(\hat{\mathbf{Z}}_1, \hat{\mathbf{W}}_1, \hat{\mathbf{\Lambda}}_1\big)$ and $\big(\hat{\mathbf{Z}}_2, \hat{\mathbf{W}}_2, \hat{\mathbf{\Lambda}}_2\big)$ be any two limit points of $\big\{\big(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\big)\big\}_k$. As we have shown, both $\big(\hat{\mathbf{Z}}_1, \hat{\mathbf{W}}_1, \hat{\mathbf{\Lambda}}_1\big)$ and $\big(\hat{\mathbf{Z}}_2, \hat{\mathbf{W}}_2, \hat{\mathbf{\Lambda}}_2\big)$ are optimal solutions to Eq. (A.1). Thus, $\mathbf{R}^\star$ in Eq. (A.21) can be replaced by $\hat{\mathbf{R}}_1 = \big(\hat{\mathbf{W}}_1, \hat{\mathbf{\Lambda}}_1\big)$ or $\hat{\mathbf{R}}_2 = \big(\hat{\mathbf{W}}_2, \hat{\mathbf{\Lambda}}_2\big)$, which results in

$$\big\|\mathbf{R}^{k+1} - \hat{\mathbf{R}}_i\big\|_{\mathrm{H}}^2 \leq \big\|\mathbf{R}^k - \hat{\mathbf{R}}_i\big\|_{\mathrm{H}}^2, \quad i = 1, 2.$$

Thus, we know the existence of the limits

$$\lim_{k\to\infty} \left\|\mathbf{R}^k - \hat{\mathbf{R}}_i\right\|_{\mathrm{H}} = \eta_i < +\infty, \quad i = 1, 2.$$

Now using the equation

$$\left\|\mathbf{R}^k - \hat{\mathbf{R}}_1\right\|_{\mathrm{H}}^2 - \left\|\mathbf{R}^k - \hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2 = -2\left\langle \mathbf{R}^k, \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2 \right\rangle_{\mathrm{H}} + \left\|\hat{\mathbf{R}}_1\right\|_{\mathrm{H}}^2 - \left\|\hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2$$

and passing the limits, we arrive at

$$\eta_1^2 - \eta_2^2 = -2\left\langle \hat{\mathbf{R}}_1, \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2 \right\rangle_{\mathrm{H}} + \left\|\hat{\mathbf{R}}_1\right\|_{\mathrm{H}}^2 - \left\|\hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2 = -\left\|\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2,$$

and

$$\eta_1^2 - \eta_2^2 = -2\left\langle \hat{\mathbf{R}}_2, \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2 \right\rangle_{\mathrm{H}} + \left\|\hat{\mathbf{R}}_1\right\|_{\mathrm{H}}^2 - \left\|\hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2 = \left\|\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2.$$

Then, we must have $\left\|\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2\right\|_{\mathrm{H}}^2 = 0$, indicating that the limit point of $\left\{\left(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\right)\right\}_k$ is unique.

Consequently, we can conclude that starting from any symmetric $\left(\mathbf{Z}^0, \mathbf{W}^0, \mathbf{\Lambda}^0\right)$ the sequence $\left\{\left(\mathbf{Z}^k, \mathbf{W}^k, \mathbf{\Lambda}^k\right)\right\}_k$ produced by the linearized ADMM in Eqs. (A.2)-(A.4) converges to an optimal solution of the original problem in Eq. (A.1). □

B Theoretic Analysis

A primary advantage of our analysis in Lemma 1 and Theorem 1 is avoiding the expensive projection onto the high-dimensional PSD cone $\mathbb{S}_+^d$, which was required by the previous trace norm regularized metric learning methods such as ([McFee and Lanckriet, 2010, Lim et al., 2013]). In this section, we further provide in-depth theoretic analysis (see Lemma 3 and Theorem 3) to comprehensively justify the low-rank solution structure $\mathbf{M}^\star = \mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$ for any convex loss function in terms of $\boldsymbol{x}_i^\top\mathbf{M}\boldsymbol{x}_j$ regularized by trace norm $\mathrm{tr}(\mathbf{M})$ or squared Frobenius norm $\|\mathbf{M}\|_{\mathrm{F}}^2$.

As a result, our analysis would directly lead to scalable $O(d)$ algorithms for a task of low-rank distance or similarity metric learning supervised by instance-level, pairwise, or listwise label information. For example, our analysis would give an $O(d)$-time algorithm for optimizing the low-rank distance metric learning objective (a hinge loss based on listwise supervision plus a trace norm regularizer) in ([McFee and Lanckriet, 2010]) through following our proposed two-step scheme, SVD projection + lower dimensional metric learning.

Suppose that we have a collection of data examples $\left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^n$ for training a metric. We formulate a more general convex objective as follows

$$\min_{\mathbf{M}\in\mathbb{S}_+^d} \; h(\mathbf{M}) := \ell\left(\left\{\boldsymbol{x}_i^\top\mathbf{M}\boldsymbol{x}_j\right\}_{i,j\in[1:n]}\right) + \alpha_1\mathrm{tr}(\mathbf{M}) + \alpha_2\|\mathbf{M}\|_{\mathrm{F}}^2, \tag{B.1}$$

where $\ell(\cdot)$ denotes any convex loss function, and $\alpha_1, \alpha_2 \geq 0$ are two regularization parameters. This extended learning framework in Eq. (B.1) is capable of adapting to a much larger class of loss functions (squared loss, ℓ_1 loss, hinge loss, logistic loss, *etc.*), supervision types (instance-level label, pairwise label, and triplet-level rank), and regularizers (trace norm, squared Frobenius norm, and mixed norm), which will also encompass many distance and similarity metric learning approaches including our proposed Low-Rank Similarity Metric Learning (LRSML), Online Regularized Distance Metric Learning (Online-Reg) ([Jin et al., 2009]), and Metric Learning to Rank (MLR) ([McFee and Lanckriet, 2010]).

It turns out that the optimality characterization for the objective of LRSML (hinge loss based on pairwise labels + trace norm regularizer), shown in Lemma 1 and Theorem 1 in our main paper, can still be applied to the generic objective in Eq. (B.1) with minor modifications. The techniques for the proofs are very similar but we still provide the proofs for completeness.

Write the singular value decomposition (SVD) of the data matrix $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ as

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^m \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top,$$

where m ($\leq n \ll d$) is the rank of $\mathbf{X}$, $\sigma_1, \cdots, \sigma_m$ are the positive singular values, and $\mathbf{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m] \in \mathbb{R}^{d\times m}$, $\mathbf{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_m] \in \mathbb{R}^{n\times m}$ are the matrices containing left- and right-singular vectors, respectively. Then the optimal solutions to the problem in Eq. (B.1) can be characterized in what follows.

Lemma 3. If $\alpha_1 + \alpha_2 > 0$, then any optimal solution $\mathbf{M}^\star$ to Eq. (B.1) must be in the set $\{\mathbf{U}\mathbf{W}\mathbf{U}^\top | \mathbf{W} \in \mathbb{S}_+^m\}$.

Proof. Let us select a matrix $\mathbf{U}' \in \mathbb{R}^{d\times(d-m)}$ such that $[\mathbf{U}, \mathbf{U}']$ forms an orthonormal matrix. Notice $\mathbf{U}^\top\mathbf{U}' = \mathbf{0}$ and define a matrix $\mathbf{Q} = \begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix} = [\mathbf{U}, \mathbf{U}']^\top\mathbf{M}^\star[\mathbf{U}, \mathbf{U}']$, where $\mathbf{Q}_{11} = \mathbf{U}^\top\mathbf{M}^\star\mathbf{U}$, $\mathbf{Q}_{12} = \mathbf{Q}_{21}^\top = \mathbf{U}^\top\mathbf{M}^\star\mathbf{U}'$, and $\mathbf{Q}_{22} = {\mathbf{U}'}^\top\mathbf{M}^\star\mathbf{U}'$. Since $\mathbf{M}^\star \in \mathbb{S}_+^d$, we know $\mathbf{Q} \in \mathbb{S}_+^d$, $\mathbf{Q}_{11} \in \mathbb{S}_+^m$, and $\mathbf{Q}_{22} \in \mathbb{S}_+^{d-m}$. We also have $\mathbf{M}^\star = [\mathbf{U}, \mathbf{U}']\mathbf{Q}[\mathbf{U}, \mathbf{U}']^\top$ as $[\mathbf{U}, \mathbf{U}']$ is an orthonormal matrix.

We first prove that $\mathbf{Q}_{22}$ must be $\mathbf{0}$ for any optimal $\mathbf{M}^\star$. On the contrary, we assume $\mathbf{Q}_{22} \neq \mathbf{0}$. Denote by $\mathbf{M}' = \mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top$ another feasible solution. Let us compare the values of $h(\mathbf{M}^\star)$ and $h(\mathbf{M}')$. Since $[\mathbf{U}, \mathbf{U}']$ is orthonormal, we have $\mathrm{range}(\mathbf{U}') = \big(\mathrm{range}(\mathbf{U})\big)^\perp = \big(\mathrm{range}(\mathbf{X})\big)^\perp$, which implies $\boldsymbol{x}^\top\mathbf{U}' = \mathbf{0}$ for any $\boldsymbol{x} \in \mathrm{range}(\mathbf{X})$. Accordingly, for any $i, j \in [1:n]$ we can derive

$$\begin{aligned}\boldsymbol{x}_i^\top\mathbf{M}^\star\boldsymbol{x}_j &= \boldsymbol{x}_i^\top[\mathbf{U}, \mathbf{U}']\mathbf{Q}[\mathbf{U}, \mathbf{U}']^\top\boldsymbol{x}_j\\ &= [\boldsymbol{x}_i^\top\mathbf{U}, \mathbf{0}]\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{U}^\top\boldsymbol{x}_j\\ \mathbf{0}\end{bmatrix}\\ &= \boldsymbol{x}_i^\top\mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top\boldsymbol{x}_j = \boldsymbol{x}_i^\top\mathbf{M}'\boldsymbol{x}_j.\end{aligned}$$

Hence, for the first term in $h(\cdot)$ we obtain $\ell\left(\{\boldsymbol{x}_i^\top\mathbf{M}^\star\boldsymbol{x}_j\}_{i,j\in[1:n]}\right) \equiv \ell\left(\{\boldsymbol{x}_i^\top\mathbf{M}'\boldsymbol{x}_j\}_{i,j\in[1:n]}\right)$.

For the second term (the trace norm), we have

$$\begin{aligned}\mathrm{tr}(\mathbf{M}^\star) &= \mathrm{tr}\left([\mathbf{U}, \mathbf{U}']\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}[\mathbf{U}, \mathbf{U}']^\top\right)\\ &= \mathrm{tr}\left(\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}\right) = \mathrm{tr}(\mathbf{Q}_{11}) + \mathrm{tr}(\mathbf{Q}_{22})\\ &> \mathrm{tr}(\mathbf{Q}_{11}) = \mathrm{tr}\big(\mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top\big) = \mathrm{tr}(\mathbf{M}'),\end{aligned}$$

in which the strict inequality holds due to the assumption $\mathbf{Q}_{22} \neq \mathbf{0}$ that implies $\mathrm{tr}(\mathbf{Q}_{22}) > 0$ (notice $\mathbf{Q}_{22} \in \mathbb{S}_+^{d-m}$).

For the third term (the squared Frobenius norm), we have

$$\begin{aligned}\big\|\mathbf{M}^\star\big\|_{\mathrm{F}}^2 &= \left\|[\mathbf{U}, \mathbf{U}']\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}[\mathbf{U}, \mathbf{U}']^\top\right\|_{\mathrm{F}}^2\\ &= \left\|\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{Q}_{12}\\ \mathbf{Q}_{21} & \mathbf{Q}_{22}\end{bmatrix}\right\|_{\mathrm{F}}^2 \geq \big\|\mathbf{Q}_{11}\big\|_{\mathrm{F}}^2 + \big\|\mathbf{Q}_{22}\big\|_{\mathrm{F}}^2\\ &> \big\|\mathbf{Q}_{11}\big\|_{\mathrm{F}}^2 = \big\|\mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top\big\|_{\mathrm{F}}^2 = \big\|\mathbf{M}'\big\|_{\mathrm{F}}^2,\end{aligned}$$

where the strict inequality holds due to the assumption $\mathbf{Q}_{22} \neq \mathbf{0}$ that implies $\big\|\mathbf{Q}_{22}\big\|_{\mathrm{F}}^2 > 0$.

Because at least one of α_1, α_2 is positive, we arrive at $h(\mathbf{M}') < h(\mathbf{M}^\star)$, which contradicts with the fact that $\mathbf{M}^\star$ is an optimal solution. So far, we have proven $\mathbf{Q}_{22} \equiv \mathbf{0}$ for any optimal $\mathbf{M}^\star$.

On the other hand, $\mathbf{Q}_{22} = \mathbf{0}$ automatically makes $\mathbf{Q}_{12} = \mathbf{Q}_{21}^\top = \mathbf{0}$ since $\mathbf{Q} \in \mathbb{S}_+^d$, which quickly leads to

$$\mathbf{M}^\star = [\mathbf{U}, \mathbf{U}']\begin{bmatrix}\mathbf{Q}_{11} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}[\mathbf{U}, \mathbf{U}']^\top = \mathbf{U}\mathbf{Q}_{11}\mathbf{U}^\top.$$

Therefore, we can say that any optimal solution $\mathbf{M}^\star$ must be in the set $\{\mathbf{U}\mathbf{W}\mathbf{U}^\top | \mathbf{W} \in \mathbb{S}_+^m\}$. □

Theorem 3. Project each data point $\boldsymbol{x}_i \in \mathbb{R}^d$ ($i \in [1:n]$) onto the subspace $\mathrm{range}(\mathbf{U}) = \mathrm{range}(\mathbf{X})$, obtaining a new data point $\tilde{\boldsymbol{x}}_i = \mathbf{U}^\top\boldsymbol{x}_i \in \mathbb{R}^m$. If $\alpha_1 + \alpha_2 > 0$, then $\mathbf{M}^\star$ is an optimal solution to the raw problem in

Eq. (B.1) if and only if $\mathbf{M}^\star = \mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$ in which

$$\mathbf{W}^\star \in \arg\min_{\mathbf{W}\in\mathbb{S}_+^m} \tilde{h}(\mathbf{W}) := \ell\left(\left\{\tilde{\boldsymbol{x}}_i^\top\mathbf{W}\tilde{\boldsymbol{x}}_j\right\}_{i,j\in[1:n]}\right) + \alpha_1\mathrm{tr}(\mathbf{W}) + \alpha_2\left\|\mathbf{W}\right\|_{\mathrm{F}}^2. \tag{B.2}$$

Proof. The proof is straightforward by showing that

$$\begin{aligned}
h(\mathbf{M}^\star) &= \min_{\mathbf{M}\in\mathbb{S}_+^d} \ell\left(\left\{\boldsymbol{x}_i^\top\mathbf{M}\boldsymbol{x}_j\right\}_{i,j\in[1:n]}\right) + \alpha_1\mathrm{tr}(\mathbf{M}) + \alpha_2\left\|\mathbf{M}\right\|_{\mathrm{F}}^2 \\
&= \min_{\mathbf{W}\in\mathbb{S}_+^m} \ell\left(\left\{\boldsymbol{x}_i^\top\mathbf{U}\mathbf{W}\mathbf{U}^\top\boldsymbol{x}_j\right\}_{i,j\in[1:n]}\right) + \alpha_1\mathrm{tr}\big(\mathbf{U}\mathbf{W}\mathbf{U}^\top\big) + \alpha_2\left\|\mathbf{U}\mathbf{W}\mathbf{U}^\top\right\|_{\mathrm{F}}^2 \\
&= \min_{\mathbf{W}\in\mathbb{S}_+^m} \ell\left(\left\{\tilde{\boldsymbol{x}}_i^\top\mathbf{W}\tilde{\boldsymbol{x}}_j\right\}_{i,j\in[1:n]}\right) + \alpha_1\mathrm{tr}(\mathbf{W}) + \alpha_2\left\|\mathbf{W}\right\|_{\mathrm{F}}^2 \\
&= \tilde{h}(\mathbf{W}^\star),
\end{aligned}$$

where the second line is due to Lemma 3. □

Lemma 3 and Theorem 3 immediately indicate that under the small sample setting $n \ll d$, low-rank distance or similarity metric learning with trace norm or squared Frobenius norm regularization guarantees to yield an optimal low-rank solution as $\mathbf{M}^\star = \mathbf{U}\mathbf{W}^\star\mathbf{U}^\top$, and can therefore be implemented efficiently by specialized scalable $O(d)$ algorithms. Such algorithms may follow the two-step scheme, SVD projection + lower dimensional metric learning, which has been employed by our LRSML algorithm.

Finally, we would like to point out that our generic objective in Eq. (B.1) for learning low-rank metrics does not accommodate ℓ_1 norm $\|\mathbf{M}\|_1$ which has been exploited for encouraging sparse metrics ([Qi et al., 2009, Liu et al., 2010]), nor $\ell_{2,1}$ norm $\|\mathbf{M}\|_{2,1}$ which has been used for inducing group sparsity in the learned metrics ([Rosales and Fung, 2006, Ying et al., 2009, Lim et al., 2013]). When being imposed on the target metric matrix $\mathbf{M}$, the ℓ_1 and $\ell_{2,1}$ norms may not result in the justified low-rank solution structure $\left\{\mathbf{U}\mathbf{W}\mathbf{U}^\top|\mathbf{W}\in\mathbb{S}_+^m\right\}$.

C Extended Experiments

Under a single training/validation/testing trial of the experiments, the eight compared metric/similarity learning methods leverage the validation data subset to acquire the optimal parameters associated with their models or algorithms. For linear SVM, MLR and LRSML, the regularization (or trade-off) parameters need to be tuned; for LSA, FLDA, LSA-ITML and MLNCM which use the low-rank basis $\mathbf{L}^0 \in \mathbb{R}^{d\times r^0}$ produced by LSA in a preprocessing step or as a heuristic, the initial rank r^0 (or the output dimension of the corresponding linear transformation $(\mathbf{L}^0)^\top$) requires a tuning as well. Since any of LSA, FLDA, LSA-ITML, CM-ITML, MLR, MLNCM and LRSML yields a low-rank basis $\mathbf{L} \in \mathbb{R}^{d\times r}$ eventually, we record the final rank r (*i.e.*, the output dimension of the resulting linear transformation $\mathbf{L}^\top$). Note that r differs in various metric learning methods. FLDA always gives rise to $r \equiv C-1$, while CM-ITML always results in $r \equiv C$.

To thoroughly evaluate our proposed approach LRSML which applies the linearized ADMM optimization algorithm to seek the low-rank similarity metric, we need to know the convergence property of the linearized ADMM. Across all datasets we have tried, we find out that in almost all cases, the linearized ADMM converges within $T = 1,000$ iterations under the setting of $\rho = 1, \tau = 0.01$. Figure 1 shows the convergence curves of the linearized ADMM working with four different groups of training samples. From Figure 1, we also observe that the linearized ADMM decreases the objective function value abruptly during the earliest 20 iterations, thereby achieving a fast convergence rate. To point out, we plot the following scaled objective function values

$$\mathcal{Q}(\mathbf{M}) = \left(\sum_{i,j=1}^n \left[\tilde{y}_{ij} - y_{ij}\mathcal{S}_{\mathbf{M}}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]_+ + \alpha\mathrm{tr}(\mathbf{M})\right)/n^2$$

in Figure 1.

Further, we present more experimental results in Tables 1 and 2, and Figures 2, 3, 4 and 5, where the classification error rates (or recognition rates) as well as the output dimensions achieved by various metric/similarity learning

Table 1: **UIUC-Sports** dataset: classification accuracy and training time of eight competing methods.

	10×8 training samples			70×8 training samples		
Method	**Measure**	**Accuracy** (%)	**Train Time** (sec)	**Measure**	**Accuracy** (%)	**Train Time** (sec)
Original	–	58.89±2.64	–	–	74.81±1.44	–
Linear SVM	–	69.64±1.69	0.78	–	83.87±1.13	5.9
LSA	distance	58.89±2.64	0.22	distance	74.81±1.44	8.9
	cosine	58.89±2.64		cosine	74.81±1.44	
FLDA	distance	48.21±5.52	0.19	distance	80.89±1.32	8.1
	cosine	46.27±5.91		cosine	78.12±2.26	
LSA-ITML	distance	69.33±1.82	20.2	distance	57.87±6.79	1611.7
	cosine	67.21±1.85		cosine	77.06±1.80	
CM-ITML	distance	62.44±2.28	2.0	distance	74.75±1.64	47.9
	cosine	62.94±2.38		cosine	74.50±1.53	
MLR	distance	65.44±2.89	17.4	distance	74.18±2.16	1317.8
	cosine	65.48±2.36		cosine	78.33±1.67	
MLNCM	distance	43.16±3.87	0.80	distance	72.13±1.97	105.1
	cosine	42.29±3.19		cosine	71.20±1.62	
AROMA	similarity	31.09±3.55	169.8	similarity	37.87±2.16	2773.8
LRSML	inner-product	**70.75±1.58**	2.5	inner-product	**84.85±1.40**	278.3
	cosine	**70.92±1.51**		cosine	**84.96±1.36**	

Table 2: **UIUC-Scene** dataset: classification accuracy and training time of eight competing methods.

	10×15 training samples			100×15 training samples		
Method	**Measure**	**Accuracy** (%)	**Train Time** (sec)	**Measure**	**Accuracy** (%)	**Train Time** (sec)
Original	–	54.39±0.93	–	–	66.29±1.11	–
Linear SVM	–	65.31±1.45	1.1	–	81.09±0.78	12.8
LSA	distance	54.39±0.93	0.37	distance	66.29±1.11	45.1
	cosine	54.39±0.93		cosine	66.29±1.11	
FLDA	distance	53.54±2.77	0.44	distance	79.50±0.93	32.3
	cosine	53.41±3.09		cosine	78.38±0.75	
LSA-ITML	distance	65.34±1.35	38.2	distance	77.98±0.86	6519.2
	cosine	63.84±1.49		cosine	76.70±1.12	
CM-ITML	distance	61.11±1.76	4.7	distance	70.39±1.02	2139.6
	cosine	58.73±2.24		cosine	69.90±1.81	
MLR	distance	55.13±2.32	23.8	distance	63.17±1.60	3925.7
	cosine	60.88±1.86		cosine	74.33±1.28	
MLNCM	distance	48.94±1.66	3.3	distance	67.72±0.92	1010.1
	cosine	49.33±1.52		cosine	67.19±0.34	
AROMA	similarity	45.65±3.63	42.2	similarity	60.67±2.24	8565.0
LRSML	inner-product	**67.63±1.51**	12.4	inner-product	**81.15±0.86**	3534.9
	cosine	**68.12±1.30**		cosine	**82.38±0.87**	

methods working with an increasing amount of supervision are plotted. Note that for each method in comparison, we choose a proper measure among distance, inner-product, and cosine, such that higher classification accuracy is obtained. These results again corroborate that the proposed low-rank similarity metric learning method LRSML is superior to the state-of-the-art metric/similarity learning methods in terms of 1NN classification accuracy on high-dimensional datasets. Compared against the competing low-rank distance metric learning methods LSA-ITML, MLR and MLNCM, LRSML achieves faster metric training than LSA-ITML and MLR (MLNCM is fastest because of its nonconvex and stochastic metric optimization), and yields the basis $\mathbf{L}$ with a much lower rank in most cases. Figures 2(b), 3(b) and 5(b) disclose that the output dimension of $\mathbf{L}$ produced by LRSML is usually very close to the class number C.

Last but not least, we would like to discuss the scalability in the training sample size n of our proposed LRSML. It is true that LRSML scales cubically with n. Nonetheless, our extensive experiments have confirmed that LRSML is able to accomplish good classification performance with just a relatively small number of training samples (*e.g.*, $80 \leq n \leq 1500$), which implies that our approach LRSML trained on a relatively small training set could well adapt to a much larger test set. Please note that the scalability to a large data size claimed by the method MLNCM in ([Mensink et al., 2013]) is achieved by a Stochastic Gradient Descent (SGD) based inexact

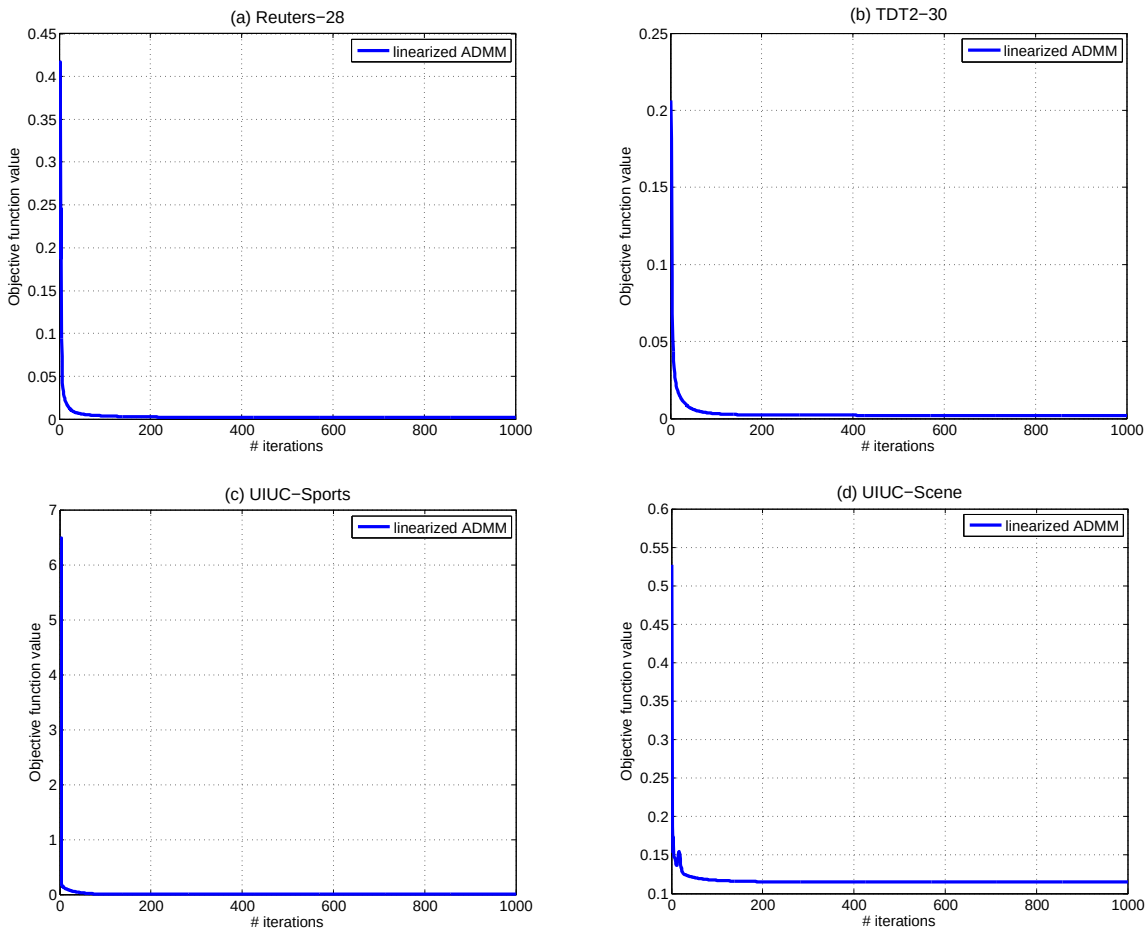

Figure 1: Convergence test of the linearized ADMM optimization algorithm. (a) On the **Reuters-28** dataset with $30 \times C$ training samples, (b) on the **TDT2-30** dataset with $30 \times C$ training samples, (c) on the **UIUC-Sports** dataset with $60 \times C$ training samples, and (d) on the **UIUC-Scene** dataset with $60 \times C$ training samples.

optimization algorithm over a nonconvex objective. In our experiments, we have found that MLNCM performs worse than linear SVM, but our LRSML outperforms linear SVM consistently. For a more fair comparison, we would make LRSML scalable to the training data size n by developing a stochastic version of the linearized ADMM algorithm. We will pursue it in future work.

References

R. Jin, S. Wang, and Y. Zhou. Regularized distance metric learning: Theory and algorithm. In *NIPS 22*, 2009.

D. K. H. Lim, B. McFee, and G. Lanckriet. Robust structural metric learning. In *Proc. ICML*, 2013.

W. Liu, S. Ma, D. Tao, J. Liu, and P. Liu. Semi-supervised sparse metric learning using alternating linearization optimization. In *Proc. KDD*, 2010.

B. McFee and G. Lanckriet. Metric learning to rank. In *Proc. ICML*, 2010.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near zero cost. *TPAMI*, 35(11):2624–2637, 2013.

G.-J. Qi, J. Tang, Z.-J. Zha, T.-S. Chua, and H.-J. Zhang. An efficient sparse metric learning in high-dimensional space via ℓ_1-penalized log-determinant regularization. In *Proc. ICML*, 2009.

R. Rosales and G. Fung. Learning sparse metrics via linear programming. In *Proc. KDD*, 2006.

Y. Ying, K. Huang, and C. Campbell. Sparse metric learning via smooth optimization. In *NIPS 22*, 2009.

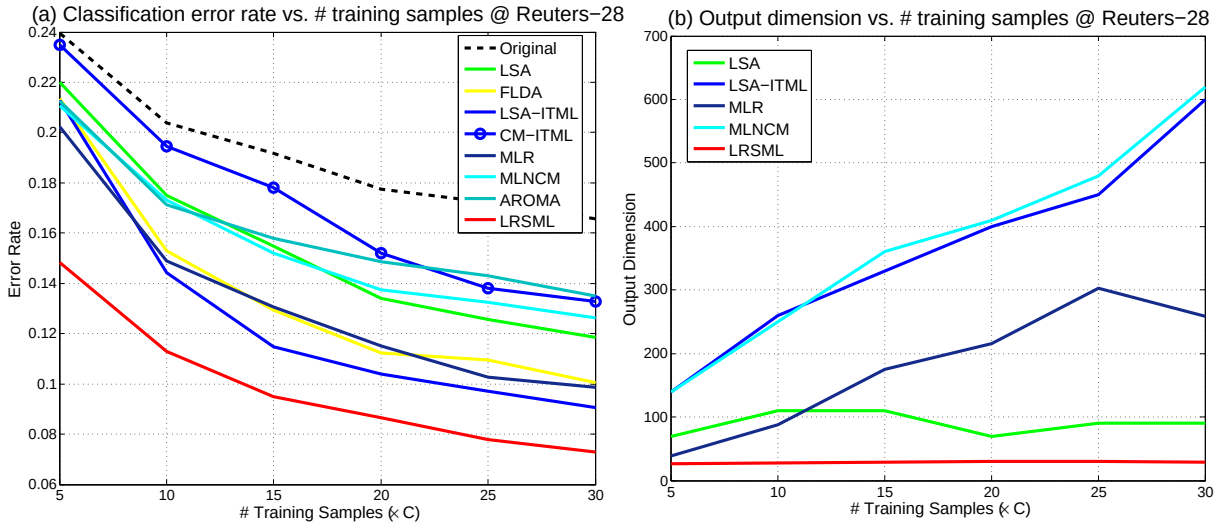

Figure 2: The results with a varying number of training samples on the **Reuters-28** dataset.

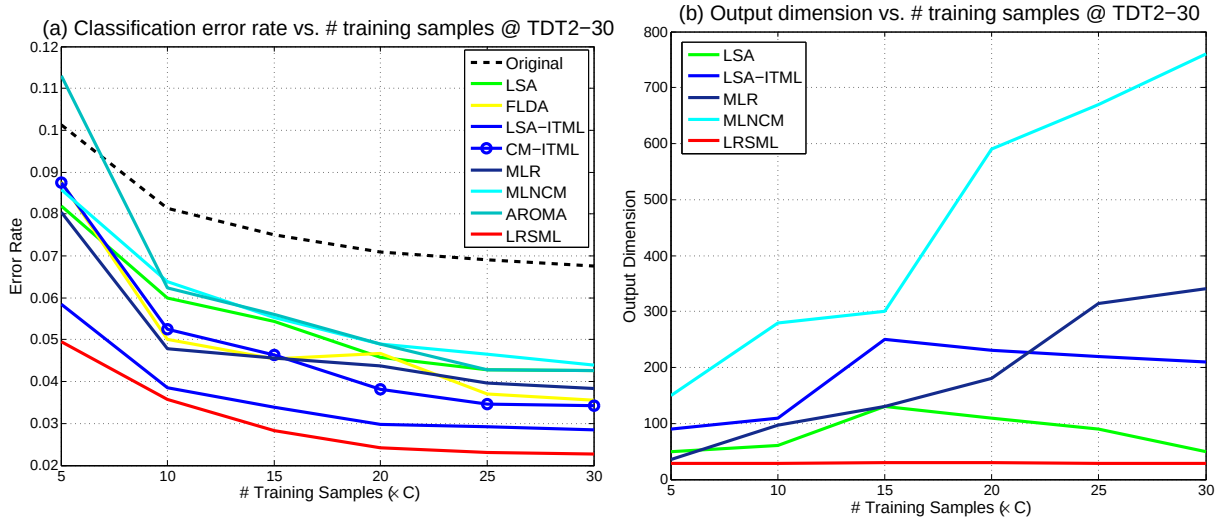

Figure 3: The results with a varying number of training samples on the **TDT2-30** dataset.

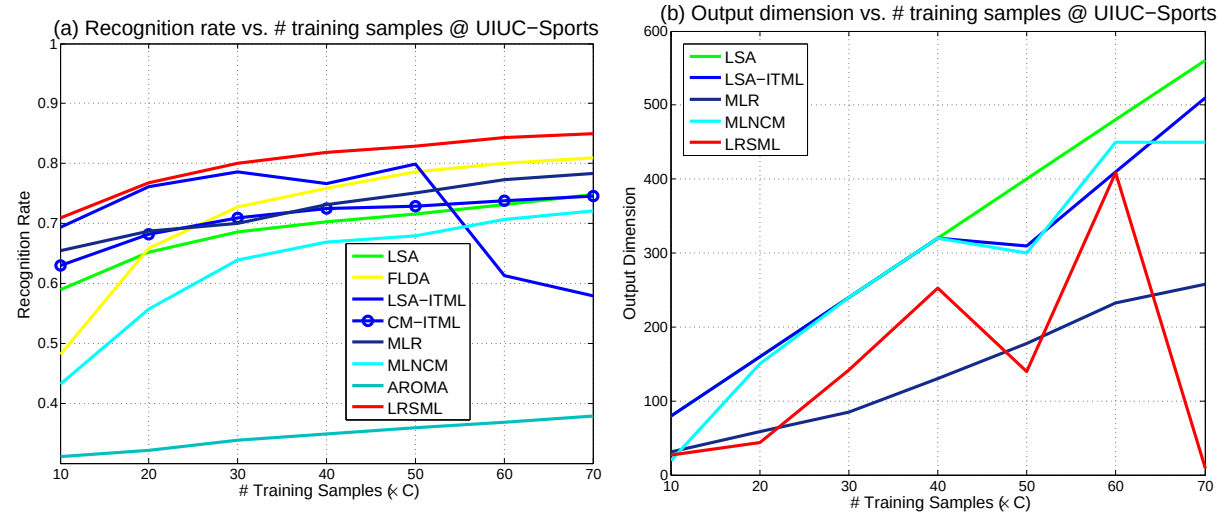

Figure 4: The results with a varying number of training samples on the **UIUC-Sports** dataset.

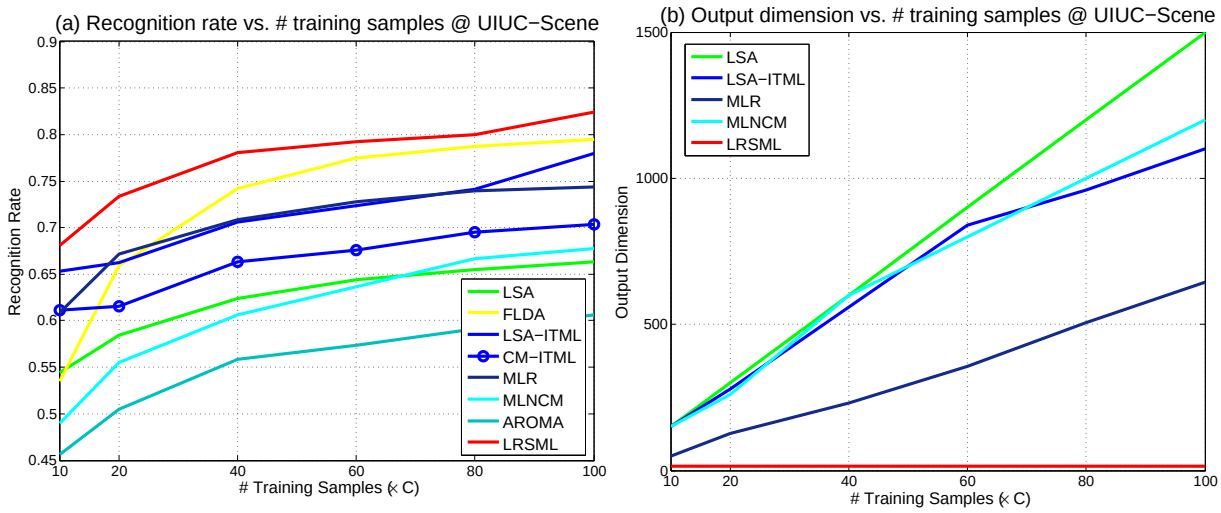

Figure 5: The results with a varying number of training samples on the **UIUC-Scene** dataset.