Appendix: Posterior Refinement on Metric Matrix Improves Generalization Bound in Metric Learning

Mingda Wang[1], Canqian Yang[1], and Yi Xu[1] *

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{597924594,charles.young,xuyi}@sjtu.edu.cn

Appendix

A Proofs

A.1 Preliminary

In this section we give the detailed proof of Theorem 1 and Theorem 2. For the convenience of mathematical expression, let $X_{\Delta i} = [0, 0, \ldots, {x_i}' - x_i^{ref}, \ldots, 0] \in \mathcal{R}^{d \times n}$, where the ith column of $X_{\Delta i}$ is equal to ${x_i}' - x_i^{ref}$ and all the others are zero vectors. Then we have $X_i = X_{ref} + X_{\Delta i}$.

A.2 Proof of Theorem 1

First, we will introduce an important Theorem proposed in [3]. Let $R_{n,n}$ denote the space of all $n \times n$ square matrices and V_n denote the space of $n \times n$ positive semi-definite matrices. Let $\lambda_i(A)$ denote the ith eigenvalue of matrix A. Then we have the following theorem

Theorem 1. *For any $A \in R_{n,n}$ and $B \in V_n$, the following inequality holds*

$$tr(AB) \leq \lambda_1(\bar{A}) tr(B) \tag{1}$$

where $\bar{A} = (A + A^T)/2$. $\lambda_1(A)$ denotes the largest eigenvalue of matrix A

Proof. Please refer to [3]

Notice that we consist M' to be a symmetric positive semi-definite matrix, by introducing Theorem 1 into $tr(X_i^T M' X_i C_i)$, we have

$$tr(X_i^T M' X_i C_i) = tr(X_i C_i X_i^T M') \leq \lambda_1(\overline{X_i C_i X_i^T}) tr(M') \tag{2}$$

Let $A := X_i C_i X_i^T$, then we have $\overline{A} = (X_i C_i X_i^T + X_i C_i^T X_i^T)/2 = X_i \overline{C_i} X_i^T$, $\overline{A}$ is a $d \times d$ square matrix. We known that for any square matrix, the eigenvalues are all smaller or equal to is operator norm, denoted as $|\lambda(\overline{A})| \leq \|\overline{A}\|$, where

* Corresponding Author

$\|A\|$ is the operator norm of matrix A defined as $\|A\| = \sup_{|x|=1} |Ax|$ and $|x|$ is the norm of vector x. For simplicity, we choose induced L_2 norm as the operator norm, thus $\lambda_1(\overline{A}) \leq \|\overline{A}\|_2$. Since, $\|\overline{A}\|_2 \leq \|\overline{A}\|_F$, the induced L_2 norm is less or equal to the Frobenius norm for any square matrix, then the following inequality holds

$$tr({X_i}^T M' X_i C_i) \leq \|X_i \overline{C_i} X_i^T\|_F tr(M') \tag{3}$$

Notice that for any matrix A and B, such inequality always hold that $\|AB\|_F \leq \|A\|_F \|B\|_F$. Then we have $\|X_i \overline{C_i} X_i^T\|_F \leq \|X_i\|_F \|\overline{C_i}\|_F \|X_i^T\|_F = \|\overline{C_i}\|_F \|X_i\|_F^2$

For $\|\overline{C_i}\|_F$, recall that C is the summation of all sampling matrix of data triples multiplied by one $0, 1$ choosing matrix, $C = \sum_{t \in \mathcal{T}} C^t \Lambda^t$. Then, the elements of the C_i can be represented as

$$c_{ij} = \begin{cases} -1 & \times(\text{times}\{i,j\}\text{is selected as positive pair}), \{i,j\}\text{is a positive pair} \\ 1 & \times(\text{times}\{i,k\}\text{is selected as negative pair}), \{i,j\}\text{is a negative pair} \end{cases} \tag{4}$$

To upper bound $\|C_i\|_F$, we might as well consider the extreme case of C_i, when all data triplet are selected and $\|C_i\|_F$ is the maximum one. In this case, C_i will be like

$$C_i = \begin{bmatrix} (m-n)\mathbf{1} & (m-1)\mathbf{1} & \dots & (m-1)\mathbf{1} \\ (m-1)\mathbf{1} & (m-n)\mathbf{1} & \dots & (m-1)\mathbf{1} \\ & \dots & & . \\ (m-1)\mathbf{1} & (m-1)\mathbf{1} & \dots & (m-n)\mathbf{1} \end{bmatrix} \tag{5}$$

where $\mathbf{1}$ is a $\frac{n}{m} \times \frac{n}{m}$ matrix whose elements are all 1. Therefore, we have $\|C_i\|_F \leq \sqrt{\frac{n}{m}(m-n)^2 + \frac{n^2-n}{m^2}(m-1)^2}$, where m and n are determined by the assumption of the dataset. For simplicity, let $\delta_2 := \sqrt{\frac{n}{m}(m-n)^2 + \frac{n^2-n}{m^2}(m-1)^2}$. The same analysis also applies for $\|C_i^T\|_F$, so similar conclusion can be derived that $\|C_i^T\|_F \leq \delta_2$ and $\|\overline{C_i}\|_F \leq \delta_2$

For $\|X_i\|_F^2$, it is easy to verify that

$$\|X_i\|_F^2 = \|X_{ref} + X_{\Delta i}\|_F^2 = \|X_{ref}\|_F^2 + 2({x_i}' - x_i^{ref})^T {x_i}^{ref} + \|X_{\Delta i}\|_F^2 \tag{6}$$

where $\|X_{ref}\|_F^2$ is regarded as a constant value. Assume that (1) the maximum distortion that random variable ${x_i}'$ shifts from ${x_i}^{ref}$ is bounded by $|{x_i}' - {x_i}^{ref}| \leq \epsilon_1$ (2) for any embedding vector in $x_i^{ref} \in X_{ref}$, its L_2 norm is bounded by $|x_i^{ref}| \leq \|X_{ref}\|_F$. Then, we have $({x_i}' - x_i^{ref})^T {x_i}^{ref} \leq |{x_i}' - x_i^{ref}||{x_i}^{ref}| \leq \epsilon_1 \|X_{ref}\|_F$, and $\|X_{\Delta i}\|_F^2 \leq \epsilon_1^2$. When we denote $\delta_1 := \|X_{ref}\|_F^2 + \epsilon_1^2 + 2\epsilon_1 \|X_{ref}\|_F = (\|X_{ref}\|_F + \epsilon_1)^2$, we have $\|X_i\|_F^2 \leq \delta_1$

Based on the analysis above, Eq. 3 can be rewritten as

$$tr({X_i}^T M' X_i C_i)/\|M'\|_F \leq \delta_1 \delta_2 tr(M')/\|M'\|_F \tag{7}$$

which is exactly the inequality proposed in Theorem 1.

Notice that during the proof of principle 1, when introducing Theorem 1 into Eq. 2, we assume that M' is one positive semi-definite matrix, which means principle 1 holds only if M' is positive semi-definite. Therefore, the precondition for principle 1 requests that M' is one symmetric positive semi-definite matrix. We attach one additional request of symmetry so that the similarity function parameterized by M' is symmetric.

A.3 Proof of Theorem 2

First, we will rewrite the objective formula as

$$\begin{aligned} & tr(X_i^T M' X_i C_i) = tr((X_{ref} + X_{\Delta i})^T M' (X_{ref} + X_{\Delta i}) C_i) \\ & = tr(X_{\Delta i}^T M' (X_{ref} + X_{\Delta i}) C_i) + tr(X_{\Delta i}^T M' X_{\Delta i} C_i) + tr(X_{ref}^T M' X_{ref} C_i) \end{aligned} \tag{8}$$

For the first component of Eq. 8, recall that $X_{\Delta i} = [0, 0, \dots, {x_i}' - x_i^{ref}, \dots, 0]$, therefore for any matrix A, $tr(X_{\Delta i}^T A) = ({x_i}' - x_i^{ref})^T a_i \le |{x_i}' - x_i^{ref}||a_i|$, where a_i is the ith column of matrix A. Since $|{x_i}' - x_i^{ref}| \le \epsilon_1$ by assumption and $|a_i| \le \|A\|_F$ holds for all matrix, we have

$$\begin{aligned} & tr(X_{\Delta i}^T M' (X_{ref} + X_{\Delta i}) C_i) \\ \le & \epsilon_1 \|M' (X_{ref} + X_{\Delta i}) C_i\|_F \\ \le & \epsilon_1 \|X_{ref} + X_{\Delta i}\|_F \|C_i\|_F \|M'\|_F \\ \le & \epsilon_1 \sqrt{\delta_1} \delta_2 \|M'\|_F \end{aligned} \tag{9}$$

where $\delta_1 := (\|X_{ref}\|_F^2 + \epsilon_1)^2$, $\delta_2 := \sqrt{\frac{n}{m}(m-n)^2 + \frac{n^2-n}{m^2}(m-1)^2}$ is defined in Section A.2. Similarly, the second component of Eq. 8 is bounded by

$$tr(X_{\Delta i}^T M' X_{\Delta i} C_i) \le \epsilon_1^2 \delta_2 \|M'\|_F \tag{10}$$

For the third component, we directly apply the inequality that for any matrix A and B, we have $tr(AB) \le \|A\|_F \|B\|_F$. Then

$$\begin{aligned} & tr(X_{ref}^T M' X_{ref} C_i) \\ \le & \|X_{ref}\|_F^2 \|M'\|_F \|C_i\|_F \\ \le & \delta_2 \|X_{ref}\|_F^2 \|M'\|_F \end{aligned} \tag{11}$$

Summarizing Eq. 9, Eq. 10 and Eq. 11, we can derive the second formula of the upper bound as

$$tr(X_i^T M' X_i C_i)/tr(M') \le \left(\epsilon_1 \sqrt{\delta_1} \delta_2 + \epsilon_1^2 \delta_2 + \delta_2 \|X_{ref}\|_F^2\right) \|M'\|_F / tr(M') \tag{12}$$

which is exactly the formula in Theorem 2.

It is worth pointing out that Eq. 12 holds for any matrix M', not as strict as the precondition of principle 2. Thus the precondition of principle 2 can be released that M' is symmetric.

B Experimental Results

B.1 Problem Observation

In table 1, we compare the performance of the trained metric, the identity metric matrix and the best performance of diagonal or random restraint metric matrix cross all DML methods mentioned in [6] on three benchmark datasets simultaneously. As can be clearly observed, for each DML methods, the performance of identity metric matrix has an remarkable margin advanced to the trained metric, and the evaluated two restraint methods are comparable to the identical metric. Such experimental results imply that there exists a certain data-free pluggable posterior refinement operation on the trained metric matrix which can significantly improve the generalization ability of DML methods.

B.2 Discussion

One may concern that the posterior refinement operations in Section 4.2 do not satisfy the prerequisite in two principles, which acquires that the refined metric matrix M' is still a symmetric positive semi-definite matrix. Here we give a brief explanation that even though in some cases these operations may not strictly satisfy such a prerequisite, they can still be used to verify the correctness of two principles.

For **Identity Refinement**, obviously an identity matrix is a symmetric positive definite matrix. Besides, during the restraint procedure we never restrain any diagonal elements, thus $tr(M')$ is fixed to $tr(M^*)$.

For **Random Restraint**, M' is still a symmetric matrix because we always restrain $m_{ij} = 0$ and $m_{ji} = 0$ at the same time. Thus random restraint satisfy the precondition for principle 2. It is worth pointing out that even though the positive semi-definiteness of M' can not be mathematically guaranteed, M' can still serve as one metric matrix. As pointed out in [2], we do not have to require the metric matrix to be positive semi-definite in metric learning. It can be tolerated if there exists a d dimensional vector x so that $x^T M' x < 0$, since we concern more about the ranking of the similarity between data pairs instead of the similarity itself. We only need to require the metric matrix to be symmetric so that the similarity function is symmetric.

For **Diagonal Restraint**, notice that Theorem 1 holds only if M' is a positive semi-definite matrix. First, restraining the diagonal elements can still remain the symmetry of the matrix, thus the matrices generated by diagonal restraint are all symmetry matrices. A symmetric matrix is a positive semi-definite matrix if and only if all its eigenvalues are non negative. Therefore, in Fig. 1 we statistics the eigenvalues of the refined matrix M' under different restraint degree r. It can be observed that the majority of the eigenvalues are far larger that zero, and only a small set of eigenvalues are around zero. Considering the inevitable numerical error in the computation of eigenvalues, these eigenvalues are negligible, thus the matrices generated by diagonal restraint can be regarded as positive semi-definite matrices. Thus, we claim that diagonal restraint satisfies the precondition of Principle 1.

Table 1. Comparison between the trained metric, identity matrix (marked as *) and the best refined matrix under random restraint or diagonal restraint (marked as †). The superscript $D/R, r$ represents such best refinement method is **D**iagonal or **R**andom restraint with restraint degree r. The best and the second best metric for each DML method is highlighted in red and blue, respectively.

Methods	CUB			Cars			SOP		
	MAP	Pre@1	RP	MAP	Pre@1	RP	MAP	Pre@1	RP
ArcFace[9]	21.51	60.12	32.39	17.48	72.88	27.66	42.23	71.98	45.15
ArcFace*	24.20	65.06	35.05	18.10	78.79	27.78	42.18	72.28	44.93
ArcFace†	24.18	64.90	$35.03^{R,1}$	18.66	78.80	$28.55^{D,1}$	43.08	73.02	$45.88^{D,1}$
Contrastive [4]	21.39	60.07	32.28	17.69	70.70	28.24	40.80	69.68	43.84
Contrastive*	23.28	64.13	34.05	18.74	77.54	28.64	41.41	70.86	44.29
Contrastive†	23.30	64.12	$34.06^{R,1}$	18.74	77.53	$28.64^{R,1}$	41.69	70.84	$44.63^{R,1}$
CosFace [10]	20.69	58.24	31.45	18.60	74.56	28.81	40.78	70.77	43.67
CosFace*	23.58	64.13	34.40	19.20	80.78	28.77	41.44	71.54	44.21
CosFace†	23.57	64.06	$34.38^{D,1}$	19.21	80.76	$28.77^{D,1}$	41.59	71.71	$44.34^{D,1}$
FastAP [1]	18.99	55.47	29.62	15.99	65.55	26.68	38.89	68.17	42.07
FastAP*	21.74	61.51	32.46	17.12	75.13	27.35	40.27	70.18	43.25
FastAP†	21.71	61.59	$32.43^{R,1}$	17.12	75.04	$27.34^{R,1}$	40.53	70.16	$43.58^{D,1}$
Margin [13]	17.77	53.84	28.46	16.65	68.31	27.48	36.38	65.67	39.62
Margin*	20.72	60.34	31.55	17.69	76.99	27.89	38.75	68.76	41.78
Margin†	20.70	60.32	$31.53^{R,1}$	18.37	75.59	$28.94^{D,1}$	38.63	68.41	$41.74^{D,1}$
Margin/class [13]	18.37	55.13	29.21	15.41	67.12	26.18	37.51	66.97	40.73
Margin/class*	21.43	61.69	32.33	15.90	74.53	26.12	39.54	69.56	42.56
Margin/class†	21.42	61.70	$32.32^{R,1}$	16.67	73.70	$27.24^{D,1}$	39.60	69.62	$42.61^{R,1}$
MS+Miner [11]	20.96	58.79	31.85	19.36	71.89	29.94	41.88	70.99	44.99
MS+Miner*	23.68	64.89	34.51	21.29	80.84	31.15	42.25	71.90	45.15
MS+Miner†	23.68	64.89	$34.50^{R,1}$	21.30	80.90	$31.16^{R,1}$	42.94	72.29	$45.90^{D,1}$
MS [11]	20.06	57.25	30.77	18.89	73.54	29.50	40.87	70.24	43.96
MS*	22.35	62.71	33.04	19.72	80.42	29.49	41.95	71.78	44.86
MS†	22.35	62.67	$33.04^{R,1}$	19.73	80.51	$29.50^{R,1}$	42.42	71.99	$45.37^{D,1}$
NTXent [8]	19.61	58.05	30.56	16.92	68.19	27.81	40.14	69.67	43.35
NTXent*	22.67	64.60	33.58	17.90	76.91	28.15	41.03	71.02	44.06
NTXent†	22.66	64.65	$33.57^{R,1}$	17.91	76.86	$28.16^{R,1}$	41.65	71.34	$44.75^{D,1}$
ProxyNCA[5]	19.39	56.99	30.18	18.85	73.91	29.43	41.90	71.51	44.86
ProxyNCA*	22.16	63.64	33.03	18.43	80.10	28.27	41.96	71.79	44.75
ProxyNCA†	22.16	63.55	$33.03^{R,1}$	19.58	79.82	$29.76^{D,1}$	42.80	72.48	$45.65^{D,1}$
SNR [14]	20.18	57.98	30.96	17.04	69.54	27.49	40.34	69.45	43.38
SNR*	22.63	62.87	33.45	17.48	76.17	27.33	41.14	70.78	44.00
SNR†	22.62	62.96	$33.45^{R,1}$	17.47	76.14	$27.33^{R,1}$	41.43	70.84	$44.35^{D,1}$
SoftTriple [7]	21.54	60.02	32.36	19.02	73.97	29.42	40.82	70.72	43.75
SoftTriple*	24.38	65.43	35.19	19.93	80.81	29.78	39.38	69.72	42.11
SoftTriple†	23.85	64.35	$34.75^{R,1}$	19.92	80.86	$29.78^{R,1}$	40.55	70.89	$43.31^{D,1}$
Triplet[12]	18.29	54.95	29.09	15.35	64.92	26.14	38.50	67.96	41.75
Triplet*	21.16	61.52	32.07	16.90	74.86	27.21	40.33	70.49	43.33
Triplet†	21.16	61.61	$32.08^{R,1}$	17.16	72.53	$27.82^{D,1}$	40.45	70.28	$43.55^{D,1}$

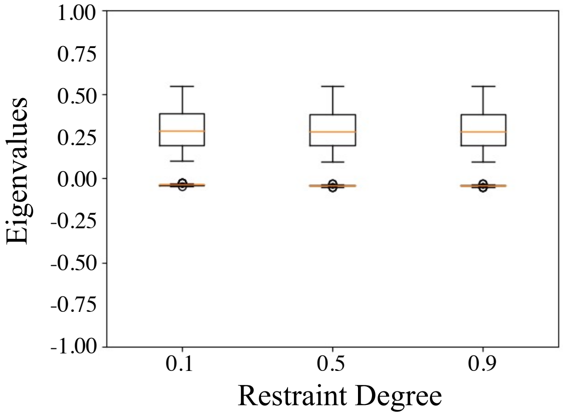

Fig. 1. Comparison between the eigenvalues of the matrices generated by the diagonal restraint under different restrain degree. The metric matrix is trained by ArcFace on CUB.

References

1. Cakir, F., He, K., Xia, X., Kulis, B., Sclaroff, S.: Deep metric learning to rank. In: CVPR. pp. 1861–1870 (2019)
2. Cao, Q., Guo, Z.C., Ying, Y.: Generalization bounds for metric and similarity learning. Machine Learning **102**(1), 115–132 (2016)
3. Fang, Y., Loparo, K.A., Feng, X.: Inequalities for the trace of matrix product. IEEE Transactions on Automatic Control **39**(12), 2489–2490 (1994)
4. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR. vol. 2, pp. 1735–1742. IEEE (2006)
5. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: ICCV. pp. 360–368 (2017)
6. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. arXiv preprint arXiv:2003.08505 (2020)
7. Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R.: Softtriple loss: Deep metric learning without triplet sampling. In: ICCV. pp. 6450–6458 (2019)
8. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. NIPS **29**, 1857–1865 (2016)
9. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
10. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR. pp. 5265–5274 (2018)
11. Wang, X., Han, X., Huang, W., Dong, D., Scott, M.R.: Multi-similarity loss with general pair weighting for deep metric learning. In: CVPR. pp. 5022–5030 (2019)
12. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS. pp. 1473–1480 (2006)
13. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: ICCV. pp. 2840–2848 (2017)
14. Yuan, T., Deng, W., Tang, J., Tang, Y., Chen, B.: Signal-to-noise ratio: A robust distance metric for deep metric learning. In: CVPR. pp. 4815–4824 (2019)