Supplementary Material: Online Deep Metric Learning

Paper ID: 1535

1 Proof of Theorem 1

Theorem 1. *Suppose $\boldsymbol{M}_t$ is positive-definite, then $\boldsymbol{M}_{t+1}$ given by the MOML update, i.e., $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t - \gamma \boldsymbol{A}_{t+1}$ is positive definite by properly setting γ.*

Proof. As $\boldsymbol{A}_{t+1} = (\boldsymbol{x}_{t+1} - \boldsymbol{x}_p)(\boldsymbol{x}_{t+1} - \boldsymbol{x}_p)^\top - (\boldsymbol{x}_{t+1} - \boldsymbol{x}_q)(\boldsymbol{x}_{t+1} - \boldsymbol{x}_q)^\top$, whose rank is 1 or 2, it has at most 2 non-zero eigenvalues. That is to say, $\mathrm{Tr}(\boldsymbol{A}_{t+1}) = \lambda_1 + \lambda_2$. Specifically, we can also easily get that,

$$-\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_q\|_2^2 \leq \lambda(\boldsymbol{A}_{t+1}) \leq \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_p\|_2^2 \,, \quad (1)$$

where $\lambda(\boldsymbol{A}_{t+1})$ means the eigenvalue of $\boldsymbol{A}_{t+1}$ (*i.e.,* λ_1 or λ_2). For each sample $\boldsymbol{x}$ is ℓ_2 normalized, the ranges of $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_p\|_2^2$ and $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_q\|_2^2$ vary from $[0, 4]$. Thus,

$$\lambda_{\min}(\boldsymbol{M}_t) - 4\gamma \leq \lambda(\boldsymbol{M}_t - \gamma \boldsymbol{A}_{t+1}) \leq \lambda_{\max}(\boldsymbol{M}_t) + 4\gamma \,. \quad (2)$$

When $\gamma \leq \frac{1}{4}\lambda_{\min}(\boldsymbol{M}_t)$, it is guaranteed that the minimum eigenvalue of $\boldsymbol{M}_t - \gamma \boldsymbol{A}_{t+1}$ is greater than zero. As the initial matrix $\boldsymbol{M}_1 = \boldsymbol{I}$ is positive definite (*i.e.,* $\lambda_{\min}(\boldsymbol{M}_1) = 1$). By properly setting a small γ, the minimum eigenvalue of $\boldsymbol{M}_t - \gamma \boldsymbol{A}_{t+1}$ is generally large than zero. Thus, the positive definiteness of $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t - \gamma \boldsymbol{A}_{t+1}$ can be guaranteed. Same theoretical guarantee (*i.e.,* the small pertubations of positive definite matrix) can also be found in the chapter 9.6.12 of [Petersen *et al.*, 2008]. □

2 Proof of Theorem 2

Theorem 2. *Let $\langle \boldsymbol{x}_1, \boldsymbol{x}_p, \boldsymbol{x}_q \rangle, \cdots, \langle \boldsymbol{x}_T, \boldsymbol{x}_p, \boldsymbol{x}_q \rangle$ be a sequence of triplet constraints where each sample $\boldsymbol{x}_t|_{t=1}^T \in \mathbb{R}^d$ has $\|\boldsymbol{x}_t\|_2 = 1$ for all t. Let $\boldsymbol{M}_t \in \mathbb{R}^{d\times d}$ be the solution of MOML at the t-th time step, and $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ denotes an arbitrary parameter matrix. By setting $\gamma = \frac{1}{R\sqrt{T}}$ (where $R \in \mathbb{R}^+$), the regret bound is*

$$\begin{aligned} R(\boldsymbol{U}, T) &= \sum_{t=1}^{T} \ell(\boldsymbol{M}_t) - \sum_{t=1}^{T} \ell(\boldsymbol{U}) \\ &\leq \frac{1}{2}\|\boldsymbol{I} - \boldsymbol{U}\|_F^2 + \frac{32}{R^2} \,. \end{aligned} \quad (3)$$

Proof. According to the objective function of MOML, *i.e.,*

$$\Gamma = \arg\min_{\boldsymbol{M} \succcurlyeq 0} \frac{1}{2}\|\boldsymbol{M} - \boldsymbol{M}_{t-1}\|_F^2 + \gamma\Big[1 + \mathrm{Tr}(\boldsymbol{M}\boldsymbol{A}_t)\Big]_+ \,, \quad (4)$$

we denote ℓ_t as the instantaneous loss suffered by MOML at each t-time step with the learnt $\boldsymbol{M}_t \in \mathbb{R}^{d\times d}$, and denote by ℓ_t^* the loss suffered by an arbitrary parameter matrix $\boldsymbol{U} \in \mathbb{R}^{d\times d}$, which can be formalized as below:

$$\begin{aligned} \ell_t &= \ell(\boldsymbol{M}_t; \langle \boldsymbol{x}_t, \boldsymbol{x}_p, \boldsymbol{x}_q \rangle) = [1 + \mathrm{Tr}\,(\boldsymbol{M}_t \boldsymbol{A}_t)]_+ \\ \ell_t^* &= \ell(\boldsymbol{U}; \langle \boldsymbol{x}_t, \boldsymbol{x}_p, \boldsymbol{x}_q \rangle) = [1 + \mathrm{Tr}\,(\boldsymbol{U}\boldsymbol{A}_t)]_+ \,, \end{aligned} \quad (5)$$

where $\boldsymbol{A}_t = (\boldsymbol{x}_t - \boldsymbol{x}_p)(\boldsymbol{x}_t - \boldsymbol{x}_p)^\top - (\boldsymbol{x}_t - \boldsymbol{x}_q)(\boldsymbol{x}_t - \boldsymbol{x}_q)^\top$, Tr denotes trace and $[z]_+ = \max(0, z)$. As $\mathrm{Tr}(\boldsymbol{M}_t\boldsymbol{A}_t)$ is a linear function, it is convex *w.r.t* $\boldsymbol{M}_t$ by natural. Besides, the hinge loss function $[z]_+$ is a convex function (but not continuous at $z = 0$) *w.r.t* z. Hence, the resulting composite function $\ell_t(\boldsymbol{M}_t)$ is convex *w.r.t* $\boldsymbol{M}_t$. As ℓ is a convex function, we can introduce the first-order condition as follow:

$$\ell(\boldsymbol{Y}) \geq \ell(\boldsymbol{X}) + \mathrm{VEC}(\triangledown \ell(\boldsymbol{X}))^\top \,\mathrm{VEC}(\boldsymbol{Y} - \boldsymbol{X}) \,, \quad (6)$$

where $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times d}$, VEC denotes vectorization of a matrix, and $\triangledown \ell(\boldsymbol{X})$ is the gradient of function ℓ at $\boldsymbol{X}$.

Inspired by [Crammer *et al.*, 2006], we define Δ_t to be $\|\boldsymbol{M}_t - \boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_{t+1} - \boldsymbol{U}\|_F^2$. Then calculating the cumulative sum of Δ_t over all $t \in \{1, 2, \cdots, T\}$, we can easily obtain $\sum_t \Delta_t$,

$$\begin{aligned} \sum_{t=1}^{T} \Delta_t &= \sum_{t=1}^{T} (\|\boldsymbol{M}_t - \boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_{t+1} - \boldsymbol{U}\|_F^2) \\ &= \|\boldsymbol{M}_1 - \boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_{T+1} - \boldsymbol{U}\|_F^2 \\ &\leq \|\boldsymbol{M}_1 - \boldsymbol{U}\|_F^2 \,. \end{aligned} \quad (7)$$

For simplicity, we employ stochastic gradient descent (SGD) to update the parameter matrix $\boldsymbol{M}_t$. Hence, according to the definition of SGD, $\boldsymbol{M}_{t+1} = \boldsymbol{M}_t - \eta \triangledown \ell(\boldsymbol{M}_t)$, where η is the learning rate, and $\triangledown \ell(\boldsymbol{M}_t) = \gamma \boldsymbol{A}_{t+1}$. Then, we can rewrite the Δ_t as,

$$\begin{aligned} \Delta_t &= \|\boldsymbol{M}_t - \boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_{t+1} - \boldsymbol{U}\|_F^2 \\ &= \|\boldsymbol{M}_t - \boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_t - \eta \triangledown \ell(\boldsymbol{M}_t) - \boldsymbol{U}\|_F^2 \\ &= \|\boldsymbol{M}_t\|_F^2 - 2\langle \boldsymbol{M}_t, \boldsymbol{U}\rangle_F + \|\boldsymbol{U}\|_F^2 - \|\boldsymbol{M}_t - \boldsymbol{U}\|_F^2 \\ &\quad + 2\langle \boldsymbol{M}_t - \boldsymbol{U}, \eta \triangledown \ell(\boldsymbol{M}_t)\rangle_F - \eta^2 \|\triangledown \ell(\boldsymbol{M}_t)\|_F^2 \\ &= 2\eta \,\mathrm{VEC}(\boldsymbol{M}_t - \boldsymbol{U})^\top \,\mathrm{VEC}(\triangledown \ell(\boldsymbol{M}_t)) - \eta^2 \|\triangledown \ell(\boldsymbol{M}_t)\|_F^2 \\ &\quad \Big(\textit{employ the Eq. (6) i.e., } \ell(\boldsymbol{U}) \geq \ell(\boldsymbol{M}_t) + \mathrm{VEC}(\triangledown \ell(\boldsymbol{M}_t))^\top \,\mathrm{VEC}(\boldsymbol{U} - \boldsymbol{M}_t)\Big) \\ &\geq 2\eta(\ell_t - \ell_t^*) - \eta^2 \|\triangledown \ell(\boldsymbol{M}_t)\|_F^2 \,. \end{aligned} \quad (8)$$

We can easily get that,

$$\sum_{t=1}^{T}\left[2\eta(\ell_t-\ell_t^*)-\eta^2\|\bigtriangledown\ell(\boldsymbol{M}_t)\|_F^2\right]\le\|\boldsymbol{M}_1-\boldsymbol{U}\|_F^2\,. \quad (9)$$

As all samples are ℓ_2 normalized, the 2-norm of each sample is 1, namely $\|\boldsymbol{x}_t\|_2\equiv 1, t\in\{1,2,\cdots,T\}$. We can easily calculate the Frobenius norm of $\boldsymbol{A}_{t+1}$.

$$\begin{aligned}\|\boldsymbol{A}_{t+1}\|_F \le&\|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_p)(\boldsymbol{x}_{t+1}-\boldsymbol{x}_p)^\top\|_F+\|(\boldsymbol{x}_{t+1}-\boldsymbol{x}_q)(\boldsymbol{x}_{t+1}-\boldsymbol{x}_q)^\top\|_F\\ &\Big(\textit{employ } \|\boldsymbol{a}\boldsymbol{b}^\top\|_F^2=(\sum_{i=1}^{d}|\boldsymbol{a}_i|^2)(\sum_{j=1}^{d}|\boldsymbol{b}_j|^2)\textit{, where } \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d\Big)\\ =&\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_p\|_2\cdot\|\boldsymbol{x}_{t+1}^\top-\boldsymbol{x}_p^\top\|_2+\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_q\|_2\cdot\|\boldsymbol{x}_{t+1}^\top-\boldsymbol{x}_q^\top\|_2\\ =&\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_p\|_2^2+\|\boldsymbol{x}_{t+1}-\boldsymbol{x}_q\|_2^2\\ &\big(\textit{for } \|\boldsymbol{a}-\boldsymbol{b}\|_2^2\le(\|\boldsymbol{a}\|_2+\|\boldsymbol{b}\|_2)^2\big)\\ \le&8\,.\end{aligned} \quad (10)$$

Thus,

$$\begin{aligned}\sum_{t=1}^{T}(\ell_t-\ell_t^*)\le&\frac{1}{2\eta}\|\boldsymbol{M}_1-\boldsymbol{U}\|_F^2+\frac{\eta}{2}\sum_{t=1}^{T}\|\bigtriangledown\ell(\boldsymbol{M}_t)\|_F^2\\ =&\frac{1}{2\eta}\|\boldsymbol{M}_1-\boldsymbol{U}\|_F^2+\frac{\eta}{2}\sum_{t=1}^{T}\|\gamma\boldsymbol{A}_{t+1}\|_F^2\\ \le&\frac{1}{2\eta}\|\boldsymbol{M}_1-\boldsymbol{U}\|_F^2+32T\eta\gamma^2\\ &(\boldsymbol{M}_1 \textit{ is initialized to an identity matrix } \boldsymbol{I})\\ =&\frac{1}{2\eta}\|\boldsymbol{I}-\boldsymbol{U}\|_F^2+32T\eta\gamma^2\,.\end{aligned} \quad (11)$$

In particular, setting $\eta=\frac{1}{R\sqrt{T}}$ (where $R>0$ is a constant) yields the regret bound $R(\boldsymbol{U},T)\le\big(\frac{R}{2}\|\boldsymbol{I}-\boldsymbol{U}\|_F^2+\frac{32\gamma^2}{R}\big)\sqrt{T}$. In fact, in this study, as a closed-form solution is employed (*i.e.,* $\eta=1$), the regret bound is $R(\boldsymbol{U},T)\le\frac{1}{2}\|\boldsymbol{I}-\boldsymbol{U}\|_F^2+32T\gamma^2$. By setting γ in a decreasing way with the iteration number T, for example, $\gamma=\frac{1}{R\sqrt{T}}$, we can obtain a regret bound $R(\boldsymbol{U},T)\le\frac{1}{2}\|\boldsymbol{I}-\boldsymbol{U}\|_F^2+\frac{32}{R^2}$. Hence proved. □

3 Theoretical analysis of Proposition 1

Proposition 1. *Let $\boldsymbol{M}_1,\cdots,\boldsymbol{M}_n$ be the parameter matrixes learnt by each metric layer of ODML. The subsequent metric layer can learn a feature space that is at least as good as the one learnt by the former metric layer. That is, the composite feature space learnt by both $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ is better than the feature space learnt only by $\boldsymbol{M}_1$ in most cases (i.e., the feature space is more discriminative for classification).*

Proof. For simplicity, we just consider to analyze and prove this proposition of ODML-FP that only uses forward propagation strategy. In fact, as ODML-FP only has forward propagation, each metric layer is a relatively independent MOML algorithm. Thus, Theorem 2 is applicable to each metric layer. In other words, each metric layer (*i.e.,* a MOML algorithm) has its own tight regret bound. As the subsequent metric layer is learnt based on the output of the former metric layer, the metric space should not be worse according to the theoretical guarantee of regret bound. Moreover, ReLU activation function can introduce nonlinear and sparsity into the feature mapping, which is also beneficial to the exploration of feature space. In some cases, if the latter metric layer is in the wrong direction, backward propagation can be chosen to correct and adjust the direction to some extent. □

References

[Crammer *et al.*, 2006] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[Petersen *et al.*, 2008] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7:15, 2008.