A Theoretical Analysis of Metric Hypothesis Transfer Learning

Supplementary Material

Michaël Perrot and Amaury Habrard

1 Overview

This supplementary material is organised into three parts. In the first two parts we respectively state the proofs of the on-average and uniform stability analysis. In the last part, we show that the specific loss presented in the paper is k-lipschitz.

For the sake of readability we start by recalling our setting. Let T be a training set drawn from a distribution $\mathcal{D}_{\mathcal{T}}$ over $\mathcal{X} \times \mathcal{Y}$. We consider the following framework for biased regularization metric learning:

$$\mathbf{M}^* = \underset{\mathbf{M} \succeq 0}{\arg\min}\, L_T(\mathbf{M}) + \lambda \|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}} \tag{1}$$

where $L_T(\mathbf{M}) = \frac{1}{n^2} \sum_{\mathbf{z},\mathbf{z}' \in T} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$ stands for the empirical risk of hypothesis $\mathbf{M}$. Similarly we denote the true risk by $L_{\mathcal{D}_{\mathcal{T}}}(\mathbf{M}) = \mathbb{E}_{\mathbf{z},\mathbf{z}' \sim \mathcal{D}_{\mathcal{T}}} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$. We only consider convex, k-lipschitz and (σ, m)-admissible losses.

2 On-average-replace-two-stability analysis

In this part we show that our algorithm is on-average-replace-two-stable. For the sake of completeness, we also recall the proof of the bound already presented in the paper.

First we show in the following lemma that our algorithm is strongly convex. Before proving this result, we recall the definition of strong convexity.

Definition 1. *A function f is λ-strongly convex if for all $\mathbf{w}$, $\mathbf{u}$, and $\alpha \in [0, 1]$ we have:*

$$f(\alpha \mathbf{w} + (1-\alpha)\mathbf{u}) \leq \alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - \frac{\lambda}{2} \alpha (1-\alpha) \|\mathbf{w} - \mathbf{u}\|^2.$$

We can now state the lemma.

Lemma 1. *The algorithm presented in Eq.1 is 2λ-strongly convex.*

Proof. First we show that the regularization term $\lambda \|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2$ is 2λ-strongly convex in $\mathbf{M}$:

$$\begin{aligned}
&\lambda \|\alpha(\mathbf{M}) + (1-\alpha)(\mathbf{M}') - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \\
&\quad = \lambda \|\alpha(\mathbf{M} - \mathbf{M}_{\mathcal{S}}) + (1-\alpha)(\mathbf{M}' - \mathbf{M}_{\mathcal{S}})\|_{\mathcal{F}}^2 \\
&\quad \leq \lambda\alpha \|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 + \lambda(1-\alpha)\|\mathbf{M}' - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\alpha(1-\alpha)\|\mathbf{M} - \mathbf{M}_{\mathcal{S}} - \mathbf{M}' + \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \qquad (2) \\
&\quad \leq \lambda\alpha \|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 + \lambda(1-\alpha)\|\mathbf{M}' - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\alpha(1-\alpha)\|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}^2
\end{aligned}$$

Eq. 2 comes from the strong convexity of the squared Frobenius norm.

The regularization term is 2λ-strongly convex and $L_T(\mathbf{M})$ is convex since it a sum of convex functions. Thus our algorithm is 2λ-strongly convex because it is a sum of a 2λ-strongly convex and a convex function. □

We can now show the on-average-replace-two-stability of our algorithm.

Theorem 1 (On-average-replace-two-stability)**.** *Given a training sample T of n examples drawn i.i.d. from $\mathcal{D}_\mathcal{T}$, our algorithm is on-average-replace-two-stable with $\epsilon(n) = \frac{8k^2}{\lambda n}$.*

Proof. Let $\mathbf{M}^*$, respectively $\mathbf{M}^{i^*}$, be the optimal solution when learning with the training set T, respectively T^i. Let $\mathbf{z}_k, \mathbf{z}_k^i, \mathbf{z}_k^{i\,j}$ respectively be the k^{th} examples of training sets T, T^i, T^{ij}. We have:

$$
\begin{aligned}
&L_T(\mathbf{M}^{ij^*}) + \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - (L_T(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2) \\
&\quad = L_{T^i}(\mathbf{M}^{ij^*}) + \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - (L_{T^i}(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2) \\
&\quad + \frac{\sum_k l(\mathbf{M}^{ij^*}, \mathbf{z}_k, \mathbf{z}_i) - l(\mathbf{M}^*, \mathbf{z}_k, \mathbf{z}_i)}{n^2} + \frac{\sum_l l(\mathbf{M}^{ij^*}, \mathbf{z}_i, \mathbf{z}_l) - l(\mathbf{M}^*, \mathbf{z}_i, \mathbf{z}_l)}{n^2} \\
&\quad + \frac{\sum_k l(\mathbf{M}^*, \mathbf{z}_k^i, \mathbf{z}_i^i) - l(\mathbf{M}^{ij^*}, \mathbf{z}_k^i, \mathbf{z}_i^i)}{n^2} + \frac{\sum_l l(\mathbf{M}^*, \mathbf{z}_i^i, \mathbf{z}_l^i) - l(\mathbf{M}^{ij^*}, \mathbf{z}_i^i, \mathbf{z}_l^i)}{n^2} \qquad (3) \\
&\quad = L_{T^{ij}}(\mathbf{M}^{ij^*}) + \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - (L_{T^{ij}}(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2) \\
&\quad + \frac{\sum_k l(\mathbf{M}^{ij^*}, \mathbf{z}_k, \mathbf{z}_i) - l(\mathbf{M}^*, \mathbf{z}_k, \mathbf{z}_i)}{n^2} + \frac{\sum_l l(\mathbf{M}^{ij^*}, \mathbf{z}_i, \mathbf{z}_l) - l(\mathbf{M}^*, \mathbf{z}_i, \mathbf{z}_l)}{n^2} \\
&\quad + \frac{\sum_k l(\mathbf{M}^*, \mathbf{z}_k^i, \mathbf{z}_i^i) - l(\mathbf{M}^{ij^*}, \mathbf{z}_k^i, \mathbf{z}_i^i)}{n^2} + \frac{\sum_l l(\mathbf{M}^*, \mathbf{z}_i^i, \mathbf{z}_l^i) - l(\mathbf{M}^{ij^*}, \mathbf{z}_i^i, \mathbf{z}_l^i)}{n^2} \\
&\quad + \frac{\sum_k l(\mathbf{M}^{ij^*}, \mathbf{z}_k^i, \mathbf{z}_j^i) - l(\mathbf{M}^*, \mathbf{z}_k^i, \mathbf{z}_j^i)}{n^2} + \frac{\sum_l l(\mathbf{M}^{ij^*}, \mathbf{z}_j^i, \mathbf{z}_l^i) - l(\mathbf{M}^*, \mathbf{z}_j^i, \mathbf{z}_l^i)}{n^2} \\
&\quad + \frac{\sum_k l(\mathbf{M}^*, \mathbf{z}_k^{i\,j}, \mathbf{z}_j^{ij}) - l(\mathbf{M}^{ij^*}, \mathbf{z}_k^{i\,j}, \mathbf{z}_j^{ij})}{n^2} + \frac{\sum_l l(\mathbf{M}^*, \mathbf{z}_j^{ij}, \mathbf{z}_l^{ij}) - l(\mathbf{M}^{ij^*}, \mathbf{z}_j^{ij}, \mathbf{z}_l^{ij})}{n^2} \qquad (4) \\
&\quad \leq L_{T^{ij}}(\mathbf{M}^{ij^*}) + \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - (L_{T^{ij}}(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2) + \frac{8k\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F}}{n} \qquad (5) \\
&\quad \leq \frac{8k\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F}}{n} \qquad (6)
\end{aligned}
$$

Equalities (3) and (4) are obtained by successively adding and removing similar terms. Inequality (5) is due to the k-lipschitz property of the loss. Inequality (6) is obtained by noticing that $\mathbf{M}^{ij^*}$ is the minimizer of our algorithm when learning with training set T^{ij}.

Furthermore, from the 2λ-strong convexity of our algorithm, proved in Lemma 1, we have:

$$
L_T(\mathbf{M}^{ij^*}) + \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - (L_T(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2) \geq \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F}^2.
$$

Thus we obtain:

$$
\begin{aligned}
& \lambda\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F}^2 \leq \frac{8k\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F}}{n} \\
\Rightarrow \quad & \|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F} \leq \frac{8k}{\lambda n} \\
\Rrightarrow \quad & \left|l(\mathbf{M}^{ij^*}, \mathbf{z}^i, \mathbf{z}^j) - l(\mathbf{M}^*, \mathbf{z}^i, \mathbf{z}^j)\right| \leq k\|\mathbf{M}^{ij^*} - \mathbf{M}^*\|_\mathcal{F} \leq \frac{8k^2}{\lambda n}.
\end{aligned} \qquad (7)
$$

The last inequality is obtained thanks to the k-lipschitz property of the loss. Taking the expectation of both sides gives the theorem. □

Using the on-average-replace-two-stability property of our algorithm, we derive our first bound.

Theorem 2 (On average bound)**.** *For any convex, k-lipschitz loss, we have:*

$$\mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*)\right] \leq L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}_\mathcal{S}) + \frac{8k^2}{\lambda n}$$

where the expected value is taken over training sets of size n.

Proof. We have:

$$\begin{aligned}
&\mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*)\right] \\
&\quad= \mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*)\right] + \mathbb{E}_T\left[L_T(\mathbf{M}^*)\right] - \mathbb{E}_T\left[L_T(\mathbf{M}^*)\right] \\
&\quad= \mathbb{E}_T\left[L_T(\mathbf{M}^*)\right] + \mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*)\right] \\
&\quad\leq \mathbb{E}_T\left[L_T(\mathbf{M}^*)\right] + \frac{8k^2}{\lambda n} \qquad (8) \\
&\quad\leq \mathbb{E}_T\left[L_T(\mathbf{M}_\mathcal{S})\right] + \frac{8k^2}{\lambda n} \qquad (9)
\end{aligned}$$

Inequality 8 is obtained by noting that from Th. 1 we have $\mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*)\right] \leq \frac{8k^2}{\lambda n}$. Inequality 9 comes from the convexity of our algorithm which gives $L_T(\mathbf{M}^*) \leq L_T(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 \leq L_T(\mathbf{M}_\mathcal{S}) + \lambda\|\mathbf{M}_\mathcal{S} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2$. Noting that $\mathbb{E}_T\left[L_T(\mathbf{M}_\mathcal{S})\right] = L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}_\mathcal{S})$ gives the theorem. □

3 Uniform stability analysis

In this second part, we show that our algorithm is uniformly stable before proving the generalization bound presented in the paper.

Theorem 3 (Uniform stability)**.** *Given a training sample T of n examples drawn i.i.d. from $\mathcal{D}_\mathcal{T}$, our algorithm has a uniform stability in $\frac{\mathcal{K}}{n}$ with $\mathcal{K} = \frac{4k^2}{\lambda}$.*

Proof. Let $\mathbf{\Delta M} = \mathbf{M} - \mathbf{M}^i$ where $\mathbf{M}$ is the optimal solution when learning with set T and $\mathbf{M}^i$ is the optimal solution when learning with set T^i. The empirical risk is convex by sum of convex functions, thus

$$\begin{aligned}
L_{T^i}(\mathbf{M} - t\mathbf{\Delta M}) - L_{T^i}(\mathbf{M}) &\leq t(L_{T^i}(\mathbf{M}^i) - L_{T^i}(\mathbf{M})) \\
L_{T^i}(\mathbf{M}^i + t\mathbf{\Delta M}) - L_{T^i}(\mathbf{M}^i) &\leq t(L_{T^i}(\mathbf{M}) - L_{T^i}(\mathbf{M}^i))
\end{aligned}$$

Summing up the two inequalities gives:

$$L_{T^i}(\mathbf{M} - t\mathbf{\Delta M}) - L_{T^i}(\mathbf{M}) + L_{T^i}(\mathbf{M}^i + t\mathbf{\Delta M}) - L_{T^i}(\mathbf{M}^i) \leq 0. \qquad (10)$$

Our algorithm is convex as stated in Lemma 1, thus:

$$\begin{aligned}
&L_T(\mathbf{M}) + \lambda\|\mathbf{M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - L_T(\mathbf{M} - t\mathbf{\Delta M}) - \lambda\|\mathbf{M} - t\mathbf{\Delta M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 \\
&+ L_{T^i}(\mathbf{M}^i) + \lambda\|\mathbf{M}^i - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - L_{T^i}(\mathbf{M}^i + t\mathbf{\Delta M}) - \lambda\|\mathbf{M}^i + t\mathbf{\Delta M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 \leq 0. \qquad (11)
\end{aligned}$$

And thus summing inequalities 10 and 11 gives:

$$\begin{aligned}
&L_T(\mathbf{M}) - L_{T^i}(\mathbf{M}) + L_{T^i}(\mathbf{M} - t\mathbf{\Delta M}) - L_T(\mathbf{M} - t\mathbf{\Delta M}) \\
&+ \lambda\|\mathbf{M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - \lambda\|\mathbf{M} - t\mathbf{\Delta M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 + \lambda\|\mathbf{M}^i - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 - \lambda\|\mathbf{M}^i + t\mathbf{\Delta M} - \mathbf{M}_\mathcal{S}\|_\mathcal{F}^2 \leq 0. \qquad (12)
\end{aligned}$$

Let $B = \lambda\|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{M}^i - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M}^i + t\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2$, we have:

$$
\begin{aligned}
B \leq & |-L_T(\mathbf{M}) + L_{T^i}(\mathbf{M}) - L_{T^i}(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}) + L_T(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M})| \\
\leq & \frac{1}{n^2}|\sum_j \sum_k l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j, \mathbf{z}_k) - l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_k^i) + l(\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_k^i) - l(\mathbf{M}, \mathbf{z}_j, \mathbf{z}_k)| \\
\leq & \frac{1}{n^2}|\sum_j l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j, \mathbf{z}_i) - l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_i^i) + l(\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_i^i) - l(\mathbf{M}, \mathbf{z}_j, \mathbf{z}_i) \\
& + \sum_k l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_i, \mathbf{z}_k) - l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_i^i, \mathbf{z}_k^i) + l(\mathbf{M}, \mathbf{z}_i^i, \mathbf{z}_k^i) - l(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_k)| \\
\leq & \frac{1}{n^2}\left(\sum_j |l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j, \mathbf{z}_i) - l(\mathbf{M}, \mathbf{z}_j, \mathbf{z}_i)| + \sum_j |l(\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_i^i) - l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_j^i, \mathbf{z}_i^i)|\right. \\
& \left. + \sum_k |l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_i, \mathbf{z}_k) - l(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_k)| + \sum_k |l(\mathbf{M}, \mathbf{z}_i^i, \mathbf{z}_k^i) - l(\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M}, \mathbf{z}_i^i, \mathbf{z}_k^i)|\right) \\
\leq & \frac{nk}{n^2}\left(\|\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}\|_{\mathcal{F}} + \|\mathbf{M} - \mathbf{M} + t\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}} + \|\mathbf{M} - t\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}\|_{\mathcal{F}} + \|\mathbf{M} - \mathbf{M} + t\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}}\right) \\
\leq & \frac{4kt}{n}\|\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}}
\end{aligned}
$$

Furthermore, setting $t = \frac{1}{2}$, we have

$$
\begin{aligned}
B = & \lambda\|\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M} - \frac{1}{2}\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{M}^i - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M}^i + \frac{1}{2}\boldsymbol{\Delta}\mathbf{M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \\
= & \lambda\sum_k\sum_l\left[(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl})^2 - (\mathbf{M}_{kl} - \frac{1}{2}(\mathbf{M}_{kl} - \mathbf{M}_{kl}^i) - \mathbf{M}_{\mathcal{S}kl})^2 + (\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2\right. \\
& \left. - (\mathbf{M}_{kl}^i + \frac{1}{2}(\mathbf{M}_{kl} - \mathbf{M}_{kl}^i) - \mathbf{M}_{\mathcal{S}kl})^2\right] \\
= & \lambda\sum_k\sum_l\left[(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl})^2 - (\frac{1}{2}\mathbf{M}_{kl} + \frac{1}{2}\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2 + (\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2 - (\frac{1}{2}\mathbf{M}_{kl} + \frac{1}{2}\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2\right] \\
= & \lambda\sum_i\sum_j\left[(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl})^2 - (\frac{1}{2}(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl}) + \frac{1}{2}(\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl}))^2 + (\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2\right. \\
& \left. - (\frac{1}{2}(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl}) + \frac{1}{2}(\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl}))^2\right] \\
= & \lambda\sum_i\sum_j\left[\frac{1}{2}((\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl})^2 + (\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2 - 2(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl})(\mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl}))\right] \\
= & \lambda\sum_i\sum_j\left[\frac{1}{2}(\mathbf{M}_{kl} - \mathbf{M}_{\mathcal{S}kl} - \mathbf{M}_{kl}^i - \mathbf{M}_{\mathcal{S}kl})^2\right] \\
= & \frac{\lambda}{2}\|\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}}^2.
\end{aligned}
$$

Then we obtain:

$$
\begin{aligned}
& \frac{\lambda}{2}\|\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}}^2 \leq \frac{4k}{2n}\|\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}} \\
\Leftrightarrow \quad & \|\boldsymbol{\Delta}\mathbf{M}\|_{\mathcal{F}} \leq \frac{4k}{\lambda n}.
\end{aligned}
$$

Using the k-lipschitz continuity of the loss, we have:

$$\sup_{\mathbf{z},\mathbf{z}'} |l(\mathbf{M},\mathbf{z},\mathbf{z}') - l(\mathbf{M}^i,\mathbf{z},\mathbf{z}')| \leq \|\mathbf{\Delta M}\|_{\mathcal{F}} \leq \frac{4k^2}{\lambda n}.$$

Setting $\mathcal{K} = \frac{4k^2}{\lambda}$ concludes the proof. □

We now recall the McDiarmid inequality McDiarmid (1989), used to prove our main theorem.

Theorem 4 (McDiarmid inequality)**.** *Let $X_1, ..., X_n$ be n independent random variables taking values in X and let $Z = f(X_1, ..., X_n)$. If for each $1 \leq i \leq n$, there exists a constant c_i such that*

$$\sup_{x_1,...,x_n,x_i' \in \mathcal{X}} |f(x_1, ..., x_n) - f(x_1, ..., x_i', ..., x_n)| \leq c_i, \forall 1 \leq i \leq n,$$

$$\textit{then for any } \epsilon > 0, Pr\left[|Z - \mathbb{E}\left[Z\right]| \geq \epsilon\right] \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Using Th. 3 which state the uniform stability of our algorithm and the McDiarmid inequality we can derive our generalization bound. For this purpose, we replace Z by $R_T = L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*)$ in Theorem 4 and we need to bound $\mathbb{E}_T\left[R_T\right]$ and $|R_T - R_{T^i}|$, which is done in the following two lemmas.

Lemma 2. *For any learning method of estimation error R_T and satisfying a uniform stability in $\frac{\mathcal{K}}{n}$, we have*

$$\mathbb{E}_T\left[R_T\right] \leq \frac{2\mathcal{K}}{n}.$$

Proof.

$$\begin{aligned}
\mathbb{E}_T\left[R_T\right] &\leq \mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*)\right] \\
&\leq \mathbb{E}_{T,\mathbf{z},\mathbf{z}'\sim\mathcal{D}_\mathcal{T}}\left[\left|l(\mathbf{M}^*,\mathbf{z},\mathbf{z}') - \frac{1}{n^2}\sum_i\sum_j l(\mathbf{M}^*,\mathbf{z}_i,\mathbf{z}_j)\right|\right] \\
&\leq \mathbb{E}_{T,\mathbf{z},\mathbf{z}'\sim\mathcal{D}_\mathcal{T}}\left[\left|\frac{1}{n^2}\sum_i\sum_j l(\mathbf{M}^*,\mathbf{z},\mathbf{z}') - l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j) + l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j) - l(\mathbf{M}^*,\mathbf{z}_i,\mathbf{z}_j)\right|\right] \\
&\leq \mathbb{E}_{T,\mathbf{z},\mathbf{z}'\sim\mathcal{D}_\mathcal{T}}\left[\left|\frac{1}{n^2}\sum_i\sum_j l(\mathbf{M}^{ij^*},\mathbf{z}_i,\mathbf{z}_j) - l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j) + l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j) - l(\mathbf{M}^*,\mathbf{z}_i,\mathbf{z}_j)\right|\right] \quad (13) \\
&\leq \mathbb{E}_{T,\mathbf{z},\mathbf{z}'\sim\mathcal{D}_\mathcal{T}}\left[\frac{1}{n^2}\sum_i\sum_j\left|l(\mathbf{M}^{ij^*},\mathbf{z}_i,\mathbf{z}_j) - l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j)\right| + \frac{1}{n^2}\sum_i\sum_j\left|l(\mathbf{M}^{i^*},\mathbf{z}_i,\mathbf{z}_j) - l(\mathbf{M}^*,\mathbf{z}_i,\mathbf{z}_j)\right|\right] \quad (14) \\
&\leq \frac{2\mathcal{K}}{n} \quad (15)
\end{aligned}$$

Inequality (13) comes from the fact that $T, \mathbf{z}, \mathbf{z}'$ are drawn i.i.d. from the distribution $\mathcal{D}_\mathcal{T}$ and thus we do not change the expected value by replacing one example with another. Inequality (14) is obtained by applying triangle inequality. The lemma comes from applying the property of uniform stability twice (Th. 3). □

Lemma 3. *For any matrix $\mathbf{M}^*$ learned by our algorithm using n training examples, and any loss function l satisfying the (σ, m)-admissibility, we have*

$$|R_T - R_{T^i}| \leq \frac{2\mathcal{K} + (4\sigma + 2m)}{n}.$$

Proof.

$$
\begin{aligned}
|R_T - R_{T^i}| &= \left| L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*) - (L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^{i^*}) - L_{T^i}(\mathbf{M}^{i^*})) \right| \\
&= \left| L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^{i^*}) + L_{T^i}(\mathbf{M}^{i^*}) - L_{T^i}(\mathbf{M}^*) + L_{T^i}(\mathbf{M}^*) - L_T(\mathbf{M}^*) \right| \\
&\leq \left| L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^{i^*}) \right| + \left| L_{T^i}(\mathbf{M}^{i^*}) - L_{T^i}(\mathbf{M}^*) \right| + |L_{T^i}(\mathbf{M}^*) - L_T(\mathbf{M}^*)| \qquad (16) \\
&\leq \frac{2\mathcal{K}}{n} + \left| \frac{1}{n^2} \sum_j \sum_k l(\mathbf{M}^*, \mathbf{z}_j^i, \mathbf{z}_k^i) - l(\mathbf{M}^*, \mathbf{z}_j, \mathbf{z}_k) \right| \qquad (17) \\
&\leq \frac{2\mathcal{K}}{n} + \left| \frac{1}{n^2} \sum_j l(\mathbf{M}^*, \mathbf{z}_j^i, \mathbf{z}_i^i) - l(\mathbf{M}^*, \mathbf{z}_j, \mathbf{z}_i) + \frac{1}{n^2} \sum_j l(\mathbf{M}^*, \mathbf{z}_i^i, \mathbf{z}_k^i) - l(\mathbf{M}^*, \mathbf{z}_i, \mathbf{z}_k) \right| \qquad (18) \\
&\leq \frac{2\mathcal{K}}{n} + \frac{1}{n^2} \sum_j \left| l(\mathbf{M}^*, \mathbf{z}_j^i, \mathbf{z}_i^i) - l(\mathbf{M}^*, \mathbf{z}_j, \mathbf{z}_i) \right| + \frac{1}{n^2} \sum_k \left| l(\mathbf{M}^*, \mathbf{z}_i^i, \mathbf{z}_k^i) - l(\mathbf{M}^*, \mathbf{z}_i, \mathbf{z}_k) \right| \qquad (19) \\
&\leq \frac{2\mathcal{K}}{n} + \frac{2(2\sigma + m)}{n} \qquad (20) \\
& \qquad (21)
\end{aligned}
$$

Inequalities (16) and (19) are due to the triangle inequality. (17) comes from the application of uniform stability (Th. 3). (18) comes from the fact that T and T^i only differ by their i^{th} example. (20) comes from the (σ, m)-admissibility of the loss and the fact that $|y_1 y_2 - y_3 y_4| \leq 2$. □

We are now ready to prove our generalization bound.

Theorem 5 (Generalization bound)**.** *With probability $1 - \delta$, for any matrix $\mathbf{M}$ learned with our $\mathcal{K}$ uniformly stable algorithm and for any convex, k-lipschitz and (σ, m)-admissible loss, we have:*

$$
L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}) \leq L_T(\mathbf{M}) + (4\sigma + 2m + c)\, \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}\left(\frac{1}{n}\right)
$$

where c is a constant linked to the k-lipschitz property of the loss.

Proof. Using the McDiarmid inequality (Th. 4) and Lemma 3 we have:

$$
\begin{aligned}
\Pr\left[|R_T - \mathbb{E}_T\left[R_T\right]| \geq \epsilon\right] &\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n \left(\frac{2\mathcal{K} + 4\sigma + 2m}{n}\right)^2}\right) \\
&\leq 2\exp\left(-\frac{2\epsilon^2}{\frac{1}{n}\left(2\mathcal{K} + 4\sigma + 2m\right)^2}\right).
\end{aligned}
$$

Then, by setting:

$$
\delta = 2\exp\left(-\frac{2\epsilon^2}{\frac{1}{n}\left(2\mathcal{K} + 4\sigma + 2m\right)^2}\right)
$$

we obtain:

$$
\epsilon = (2\mathcal{K} + 4\sigma + 2m)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}
$$

and thus:

$$
\Pr\left[|R_T - \mathbb{E}_T\left[R_T\right]| < \epsilon\right] > 1 - \delta.
$$

Then, with probability $1-\delta$:

$$\begin{aligned} & & R_T &< \mathbb{E}_T\left[R_T\right] + \epsilon \\ \Leftrightarrow & & L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*) &< \mathbb{E}_T\left[R_T\right] + \epsilon \\ \Rightarrow & & L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) &< L_T(\mathbf{M}^*) + \frac{2\mathcal{K}}{n} + (2\mathcal{K} + 4\sigma + 2m)\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}. \end{aligned}$$

The last inequality is obtained using Lem. 2 and replacing ϵ by its value. □

4 Specific loss

We show the k-lipschitz property of our loss.

Lemma 4 (k-lipschitz continuity)**.** *Let $\mathbf{M}$ and $\mathbf{M}'$ be two matrices and $\mathbf{z}, \mathbf{z}'$ be two examples. Our loss $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$ is k-lipschitz continuous with $k = \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|^2$.*

Proof.

$$\begin{aligned} |l(\mathbf{M}, \mathbf{z}, \mathbf{z}') - l(\mathbf{M}', \mathbf{z}, \mathbf{z}')| &= |\left[yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})\right]_+ - \left[yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}'(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})\right]_+| \\ &\leq |yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - \gamma_{yy'}) - yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}'(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})| \\ &\leq |yy'(\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - yy'(\mathbf{x}-\mathbf{x}')^T\mathbf{M}'(\mathbf{x}-\mathbf{x}')| \\ &\leq |(\mathbf{x}-\mathbf{x}')^T(\mathbf{M}-\mathbf{M}')(\mathbf{x}-\mathbf{x}')| \\ &\leq \|\mathbf{x}-\mathbf{x}'\|^2\|\mathbf{M}-\mathbf{M}'\|_\mathcal{F} \\ &\leq \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x}-\mathbf{x}'\|^2\|\mathbf{M}-\mathbf{M}'\|_\mathcal{F} \end{aligned}$$

Setting $k = \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|^2$ concludes the proof. □

References

McDiarmid, Colin. *Surveys in Combinatorics*, chapter On the method of bounded differences, pp. 148–188. Cambridge University Press, 1989.