– Supplementary Material – Learning from Few Positives: a Provably Accurate Metric Learning Algorithm to Deal with Imbalanced Data

Rémi Viola[1,2], Rémi Emonet[1], Amaury Habrard[1], Guillaume Metzler[1] and Marc Sebban[11]

[1]Laboratoire Hubert Curien UMR 5516, Univ Lyon, UJM F-42023, Saint-Etienne, France.
[2] Direction Générale des Finances Publiques, , Ministère de l'Économie et des Finances, France

1 Introduction and Notations

We will denote by $\mathbf{z} = (\mathbf{x}, y)$ the couple features-label where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{-1, 1\}$ and $S = \{\mathbf{z}_i\}_{i=1}^m$ a set of m training examples drawn from an unknown distribution $\mathcal{D}$. We denote by m_+ the number of positives and m_- the number of negatives. Thus the rate of positives α is equal to $\frac{m_+}{m}$. Suppose that $\mathbf{x}'$ is a test instance, we recall that:

- $d_{\mathbf{M}} = d_{\mathbf{M}}(\mathbf{x}', \mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{x}')^T M (\mathbf{x} - \mathbf{x}')}$ if $\mathbf{x}$ is a positive instance,
- $d = d_{\mathbf{I}}(\mathbf{x}', \mathbf{x}) = d_(\mathbf{x}', \mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')}$ otherwise.

We are considering the following optimization problem:

$$\min_{\mathbf{M} \in \mathbb{S}^+} \quad \frac{1}{m^3} \left((1-\alpha) \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \\ y_i = y_j \neq y_k = -1}} \ell_{\mathrm{FN}}(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) + \alpha \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \\ y_i = y_j \neq y_k = 1}} \ell_{\mathrm{FP}}(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) \right) + \mu \|\mathbf{M} - \mathbf{I}\|_{\mathcal{F}}^2 \quad (1)$$

Our loss function can thus be seen as :

$$\ell(\mathbf{M}, (z_1, z_2, z_3)) = \begin{cases} (1-\alpha) \times \ell_{\mathrm{FN}}(\mathbf{M}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) & \text{if } y_i = y_j = 1,\ y_k = -1, \\ \alpha \times \ell_{\mathrm{FP}}(\mathbf{M}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)] & \text{if } y_i = y_j = -1,\ y_k = 1, , \\ 0 & \text{otherwise,} \end{cases}$$

where ℓ_{FN} and ℓ_{FP} are defined by:

- $\ell_{\mathrm{FN}}(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) = [1 - c + d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)^2 - d(\mathbf{x}_i, \mathbf{x}_k)^2]_+$,

- $\ell_{\text{FP}}(\mathbf{M}, \mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k) = [1 - c + d(\mathbf{x}_i, \mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_k)^2]_+$.

In the following, we will also suppose that for all $\mathbf{x}$ we have: $\|\mathbf{x}\|_2 \leq K$. Furthermore, we will denote by $\mathcal{R}_S$ and $\mathcal{R}$ respectively the empirical risk of ℓ over the training sample S and the true risk. More precisely, the empirical risk $\mathcal{R}_S$ is evaluated using a training set of size m which is used to build all the triplets and the true risk $\mathcal{R}$ is its expectation over all the samples of size m, i.e. $\mathcal{R} = \underset{S \sim \mathcal{D}^m}{\mathbb{E}} [\mathcal{R}_S]$.

- $d = d_{\mathbf{M}}(\mathbf{x}', \mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{x}')^T M (\mathbf{x} - \mathbf{x}')}$ if $\mathbf{x}$ is a positive instance,
- $d = d_{\mathbf{I}}(\mathbf{x}', \mathbf{x}) = \sqrt{(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')}$.

In the following, we will also use the following constraint on $\mathbf{M}$:

$$\lambda_{\max}(\mathbf{M}) \leq 1, \text{ where } \lambda_{\max} \text{ is the largest eigenvalue of } \mathbf{M}.$$

Finally, due to the context of our study, i.e. imbalanced setting, $\alpha < 1/2$. Thus, $\alpha < 1 - \alpha$.

2 Generalization Guarantees

The aim of this section is to provide some generalization guarantees on our loss function according to the used loss function. Note that the following results give guarantees on the learned metric $\mathbf{M}$ which aims to find a good compromise between achieving a low rate of False Negatives while keeping a reasonable rate of False Positives.

2.1 Uniform Stability

In this section, we briefly restate the definition of stability and the generalization bound based on this notion.
Roughly speaking, an algorithm is *stable* if its output, in terms of difference between losses, does not change significantly under a small modification of the training sample. This variation must be bounded in $O(1/m)$ in terms of infinite norm where m is the size of the training set S *i.i.d.* from an unknown distribution $\mathcal{D}$.

Definition 1. *[Definition 6 (Bousquet and Elisseeff, 2002)] A learning algorithm $\mathcal{A}$ has a uniform stability in $\frac{\kappa}{m}$ with respect to a loss function ℓ and parameter set θ, with κ a positive constant if:*

$$\forall S,\ \forall i,\ 1 \leq i \leq m,\ \sup_Z |\ell(\theta_S, Z) - \ell(\theta_{S^i}, Z)| \leq \frac{\kappa}{m},$$

where S is a learning sample of size m, $Z = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) = ((\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3))$ is a triplet of labeled examples, θ_S the model parameters learned from S, θ_{S^i} the model parameters learned from the sample S^i obtained by replacing the i^{th} example z_i from S by another example z_i' independent from S and drawn from $\mathcal{D}$. $\ell(\theta_S, \mathbf{x})$ is the loss suffered at $\mathbf{x}$.

In this definition, S^i represents the notion of small modification of the training sample. The following one aims to study the evolution of the loss function according to the label of the considered triplet.

Definition 2. *A loss function ℓ is said to be γ-admissible, with respect to the distance metric $\mathbf{M}$ if (i) it is convex with respect to its first argument and (ii) the following condition holds:*

$$\forall Z, Z' \quad |\ell(\mathbf{M}, Z) - \ell(\mathbf{M}, Z')| \leq \gamma,$$

where $Z = (z_i, z_j, z_k)$ and $Z' = (z_i', z_j', z_k')$ are two triplets of examples.

2.2 Preliminary Results

We now introduce the results we need to derive our generalization guarantees:

Proposition 1. *Let $X_1, ..., X_m$ be m independent random variables taking values in $\mathbb{R}$ and let $U = f(X_1, ..., X_m)$. If for each $1 \leq i \leq m$, there exists a constant c_i such that:*

$$\sup_{x_1,...x_m \in \mathbb{R}} |f(x_1, ..., x_m) - f(x_1, ...x_i', ..., x_m)| \leq c_i,$$

then for any positive constant B, we have:

$$\mathbb{P}[|U - \mathbb{E}[U]| \geq B] \leq 2 \exp\left(\frac{-2B^2}{\sum_{i=1}^m c_i^2}\right).$$

In the following, we set $D_S = \mathcal{R} - \mathcal{R}_S$. We then introduce the two following lemmas, for which the proof can be found in (Bellet et al., 2015) (see the proofs of Lemma 8.9 and 8.10 respectively). However, note that results have been adapted to our context, i.e. for triplet based loss function. But the proofs can be easily adapted.

Lemma 1. *For any learning method of estimation error D_S and satisfying a uniform stability in $\frac{\kappa}{m}$, we have $\mathbb{E}_S[D_S] \leq \frac{2\kappa}{m}$.*

Lemma 2. *For any parameter matrix $\mathbf{M}$ using m training examples, and any loss function ℓ satisfying the γ-admissibility, we have the following bound:*

$$\forall i,\ 1 \leq i \leq m,\ |D_S - D_{S^i}| \leq \frac{2\kappa}{m} + \frac{2\gamma}{m}.$$

Using the above Proposition and the two Lemmas, we are able to get the following generalization bound:

Theorem 1. *Let $\delta > 0$ and $m > 1$. Let S be a sample of m randomly selected training examples and let $\mathbf{M}$ be the learned parameter matrix from an algorithm with uniform stability $\frac{\kappa}{m}$. Assuming that the loss function ℓ is k-Lipschitz and γ-admissible and let us denote by $\mathcal{R}_S$ its empirical risk. With probability $1 - \delta$, we have the following bound on the true risk $\mathcal{R}$ of our loss function ℓ:*

$$\mathcal{R} \leq \mathcal{R}_S + 2\frac{\kappa}{m} + (2\kappa + 2\gamma)\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

2.3 Generalization Bound

We first prove that our function is k-Lipschitz according to the following definition.

Definition 3. *A loss function ℓ is k-Lipschitz with respect to its first argument if for any parameters matrices $\mathbf{M}$ and $\mathbf{M}'$, and for any triplets of labeled examples $Z = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$, we have:*

$$|\ell(\mathbf{M}, Z) - \ell(\mathbf{M}', Z)| \leq k\|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}.$$

Lemma 3. *We now show that our loss function ℓ is k-Lipschitz with $k = 4(1 - \alpha)K^2$*

Proof. We need to study two cases, according to the label of the triplets.

Case 1: $y_i = y_j = 1,\ y_k = -1$

$$\begin{aligned}
|\ell(\mathbf{M}, Z) - \ell(\mathbf{M}', Z)| &= (1-\alpha)|[1 - c + d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)^2 - d(\mathbf{x}_i, \mathbf{x}_k)^2]_+ - [1 - c + d_{\mathbf{M}'}(\mathbf{x}_i, \mathbf{x}_j)^2 - d(\mathbf{x}_i, \mathbf{x}_k)^2]_+|, \\
&\leq (1-\alpha)|d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j)^2 - d_{\mathbf{M}'}(\mathbf{x}_i, \mathbf{x}_j)^2|, \\
&= (1-\alpha)|(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{M} - \mathbf{M}')(\mathbf{x}_i - \mathbf{x}_j)|, \\
&= (1-\alpha)|\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}, \\
|\ell(\mathbf{M}, Z) - \ell(\mathbf{M}', Z)| &\leq 4(1-\alpha)K^2\|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}
\end{aligned}$$

where the second line uses the fact that the hinge loss is 1-Lipschitz, the third line uses the linearity of the difference with respect to $\mathbf{M}, \mathbf{M}'$, the fourth line uses usual properties on norms and the last line the fact that $\|\mathbf{x}\| \leq K$.

Case 2: $y_i = y_j = -1,\ y_k = 1$
The proof is similar to the proof given in the previous case and leads to the following result:

$$|\ell(\mathbf{M}, Z) - \ell(\mathbf{M}', Z)| \leq 4\alpha K^2\|\mathbf{M} - \mathbf{M}'\|_{\mathcal{F}}.$$

We conclude by taking the maximum of the three previous values. Thus $k = 4(1-\alpha)K^2$ □

Now, we have to prove that our loss function is γ-admissible according to the definition 2.

Lemma 4. *The loss function ℓ defined by (1) is γ-admissible with respect to the distance metric $\mathbf{M}$, with $\gamma = (1-\alpha)(1 - c + 4K^2)$.*

Proof. Needless to say that the loss function ℓ is convex with respect to $\mathbf{M}$ as the sum of two convex functions. Indeed, both of them are linear w.r.t. $\mathbf{M}$ and the maximum of two convex functions remains convex.
Furthermore, because our loss function can be equal to zero for some labels of our triplets, we are looking for the greatest value than our loss function ℓ can achieve.

Using our previous result, we can bound the first part ell_{FN} by: $(1-\alpha)(1 - c + 4K^2)$ and the last term ell_{FP} by: $\alpha(1 - c + 4K^2)$.

Finally:

$$\forall\ Z, Z'\ |\ell(\mathbf{M}, Z) - \ell(\mathbf{M}, Z')| \leq \max((1-\alpha)(1 - c + 4K^2), \alpha(1 - c + 4K^2)).$$

Thus, $\gamma = (1-\alpha)(1 - c + 4K^2)$.

□

Definition 4. *A learning algorithm has a uniform stability in $\frac{\kappa}{m}$ where κ is a positive constant, if given any training set S we have:*

$$\forall i,\ \sup_Z |\ell(\mathbf{M}, Z) - \ell(\mathbf{M}^i, Z)| \leq \frac{\kappa}{m},$$

where M^i is the matrix learned with a training set S^i which differs from S of only one example $(\mathbf{x}_i \to \mathbf{x}_i')$.

For the sake of clarity for the following development, let us denote by F_S the objective function to optimize over the training set S, i.e. $F_S = \frac{1}{m^3}\sum_{\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k} \ell(\mathbf{M}, Z) + \mu\|\mathbf{M}-\mathbf{I}\|^2_{\mathcal{F}}$.
To compute the constant of uniform stability, we first need the following technical lemma:

Lemma 5. *Let S be a learning sample, let F_S and F_{S^i} be two objective functions with respect to two samples S and S^i and let $\mathbf{M}$ and $\mathbf{M}^i$ be their respective minimizers. We also define $\Delta\mathbf{M} = \mathbf{M}^i - \mathbf{M}$ and recall that $N(\mathbf{M}) = \mu\|\mathbf{M}-\mathbf{I}\|^2_{\mathcal{F}}$. For all $t \in [0,1]$, we have:*

$$N(\mathbf{M}) - N(\mathbf{M}+t\Delta\mathbf{M}) + N(\mathbf{M}^i) - N(\mathbf{M}^i - t\Delta\mathbf{M}) \\ \leq \frac{2t}{\mu m^3}[3m(m-1)+1] \times \left(4(1-\alpha)K^2\right) \times \|\Delta\mathbf{M}\|_{\mathcal{F}}.$$

Proof. Since ℓ (the hinge loss) is convex, so is the empirical risk and thus for all $t \in [0,1]$ we have the two following inequalities:

$$\mathcal{R}_{S^i}(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M},R) \leq t\mathcal{R}_{S^i}(\mathbf{M}^i) - t\mathcal{R}_{S^i}(\mathbf{M}).$$

and

$$\mathcal{R}_{S^i}(\mathbf{M}^i - t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}^i) \leq t\mathcal{R}_{S^i}(\mathbf{M}) - t\mathcal{R}_{S^i}(\mathbf{M}^i).$$

We get the second inequality by swapping the role of $\mathbf{M}$ and $\mathbf{M}^i$. If we sum these two inequalities, the right hand side vanishes and we obtain:

$$\mathcal{R}_{S^i}(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}) + \mathcal{R}_{S^i}(\mathbf{M}^i - t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}^i) \leq 0. \quad (2)$$

By assumption on $\mathbf{M}$ and $\mathbf{M}^i$ we have:

$$\begin{aligned} F_S(\mathbf{M},R) - F_S(\mathbf{M}+t\Delta\mathbf{M}) &\leq 0, \\ F_{S^i}(\mathbf{M}^i) - F_{S^i}(\mathbf{M}^i - t\Delta\mathbf{M}) &\leq 0, \end{aligned}$$

then, summing the two previous inequalities and using (2), we get:

$$\mathcal{R}_{S^i}(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_S(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}) + \mathcal{R}_S(\mathbf{M}) \\ + \mu[\|\mathbf{M}-\mathbf{I}\|^2_F + \|\mathbf{M}^i-\mathbf{I}\|^2_F - \|\mathbf{M}+t\Delta\mathbf{M}-\mathbf{I}\|^2_F - \|\mathbf{M}^i - t\Delta\mathbf{M}-\mathbf{I}\|^2_F] \leq 0. \quad (3)$$

We now focus on the first part of the previous inequality. For the sake of simplicity, let us set:

$$H = \mathcal{R}_S(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}+t\Delta\mathbf{M}) + \mathcal{R}_{S^i}(\mathbf{M}) - \mathcal{R}_S(\mathbf{M}).$$

$$\begin{aligned} H &\leq \left|\mathcal{R}_S(\mathbf{M}+t\Delta\mathbf{M}) - \mathcal{R}_{S^i}(\mathbf{M}+t\Delta\mathbf{M}) + \mathcal{R}_{S^i}(\mathbf{M}) - \mathcal{R}_S(\mathbf{M})\right|, \\ &\leq \frac{1}{m^3}\left| \sum_{z_i,z_j,z_k \in S^l} \ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \sum_{z_i,z_j,z_k \in S} \ell(\mathbf{M}, z_i, z_j, z_k) \right. \\ &\qquad \left. \sum_{z_i,z_j,z_k \in S} \ell(\mathbf{M}+t\Delta\mathbf{M}, z_i, z_j, z_k) - \sum_{z_i,z_j,z_k \in S^l} \ell(\mathbf{M}+t\Delta\mathbf{M}, z_i^l, z_j^l, z_k^l)\right|, \end{aligned}$$

where S and S^l differ from the l-th example, i.e. $\forall i,j,k \neq l,\ z_i = z_i^l,\ z_j = z_j^l$ and $z_k = z_k^l$.

We will now focus on the first difference in the previous expression, i.e. on:

$$\sum_{z_i,z_j,z_k \in S^l} \ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \sum_{z_i,z_j,z_k \in S} \ell(\mathbf{M}, z_i, z_j, z_k).$$

This difference can be decomposed into two parts according to the value of the index i: when $i = l$ and when $i \neq l$:

$$\begin{aligned}
&\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_l^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_l, z_j, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_i, z_j, z_k)\Big)
\end{aligned}$$

The first part of the decomposition is composed of m^2 terms that are at least not equal to zero. We, thus have to work on the second part of the decomposition has it contains some terms that are equal to zero. We will have to do this process two times as follows:

$$\begin{aligned}
&\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_l^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_l, z_j, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_i, z_j, z_k)\Big), \\
=\ &\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_l^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_l, z_j, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_l^l, z_k^l) - \ell(\mathbf{M}, z_i, z_l, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{j\neq l}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_i, z_j, z_k)\Big), \\
=\ &\sum_{j=1}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_l^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_l, z_j, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{k=1}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_l^l, z_k^l) - \ell(\mathbf{M}, z_i, z_l, z_k)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{j\neq l}^{m}\Big(\ell(\mathbf{M}, z_i^l, z_j^l, z_l^l) - \ell(\mathbf{M}, z_i, z_j, z_l)\Big) \\
&+\sum_{i\neq l}^{m}\sum_{j\neq l}^{m}\sum_{k\neq l}^{m}\underbrace{\Big(\ell(\mathbf{M}, z_i^l, z_j^l, z_k^l) - \ell(\mathbf{M}, z_i, z_j, z_k)\Big)}_{=0}.
\end{aligned}$$

All these sums are respectively composed of m^2, $m(m-1)$ and $(m-1)^2$ terms and the last $(m-1)^3$ terms are all equal to zero. Furthermore: $m^2 + m(m-1) + (m-1)^2 = 3m(m-1) + 1$, so that we have to find a bound on the supremum of the difference:

$$[3m(m-1)+1] \sup_{Z,Z'} |\ell(\mathbf{M}, Z) - \ell(\mathbf{M}, Z') + \ell(\mathbf{M} + t\Delta\mathbf{M}, Z) - \ell(\mathbf{M} + t\Delta\mathbf{M}, Z')|.$$

Thus, H can be upper-bounded by:

$$H \leq \frac{1}{m^3}(m-1)+1] \, (\sup_{Z,Z'} |\ell(\mathbf{M},Z) - \ell(\mathbf{M}+,Z') + \ell(\mathbf{M}+t\Delta\mathbf{M},Z) - \ell(\mathbf{M}+t\Delta\mathbf{M},Z')|).$$

We can then write:

$$\begin{aligned} H &\leq \frac{1}{m^3}(m-1)+1] \, (\sup_{Z} |\ell(\mathbf{M}+t\Delta\mathbf{M},Z) - \ell(\mathbf{M},Z)| + (\sup_{Z'} |\ell(\mathbf{M}+t\Delta\mathbf{M},Z') - \ell(\mathbf{M}+,Z')|), \\ &\leq \frac{2t}{m^3}[3m(m-1)+1] \times \|\Delta\mathbf{M}\|_{\mathcal{F}} \times \left(4(1-\alpha)K^2\right), \end{aligned}$$

where the last lines uses Lemma 3 and properties on norms. Finally, we have :

$$N(\mathbf{M}) - N(\mathbf{M}+t\Delta\mathbf{M}) + N(\mathbf{M}^i) - N(\mathbf{M}^i - t\Delta\mathbf{M}) \leq \frac{2t}{\mu m^3}[3m(m-1)+1] \times \left(4(1-\alpha)K^2\right) \times \|\Delta\mathbf{M}\|_{\mathcal{F}} \tag{4}$$

□

We are now able to prove the uniform stability of our algorithm.

Theorem 2. *Let S be a learning sample of size m, the algorithm (1) has a uniform stability in $\frac{\kappa}{m}$ with $\kappa = \frac{6}{\mu} \times \left(4(1-\alpha)K^2\right)^2$.*

Proof. Let us set $t = \frac{1}{2}$ in the result of Lemma 5 and we focus on the left hand side of this result. We have:

$$\begin{aligned} f(\mathbf{M}) &= \|\mathbf{M}-\mathbf{I}\|_F^2 + \|\mathbf{M}^i-\mathbf{I}\|_F^2 - \|\frac{1}{2}(\mathbf{M}+\mathbf{M}^i)-\mathbf{I}\|_F^2 - \|\frac{1}{2}(\mathbf{M}+\mathbf{M}^i)-\mathbf{I}\|_F^2, \\ &= \|\mathbf{M}-\mathbf{I}\|_F^2 + \|\mathbf{M}^i-\mathbf{I}\|_F^2 - \frac{1}{2}\|\mathbf{M}+\mathbf{M}^i-\mathbf{I}\|_F^2, \\ f(\mathbf{M}) &= \frac{1}{2}\|\mathbf{M}-\mathbf{M}^i\|_F^2. \end{aligned}$$

Then, using Lemma 5, we get the following bound on $\|\Delta\mathbf{M}\|_{\mathcal{F}}$.

$$\begin{aligned} \|\Delta\mathbf{M}\|_{\mathcal{F}}^2 &\leq \frac{8}{\mu m^3}[3m(m-1)+1] \times \left((1-\alpha)K^2\right) \times \|\Delta\mathbf{M}\|_{\mathcal{F}}, \\ \|\Delta\mathbf{M}\|_{\mathcal{F}} &\leq \frac{8}{\mu m^3}[3m(m-1)+1] \times \left((1-\alpha)K^2\right). \end{aligned}$$

To prove the uniform stability of our algorithm, it remains to find the value κ such that:

$$\forall S, \ \forall i, \ 1 \leq i \leq m, \ \sup_Z |\ell(\mathbf{M},Z) - \ell(\mathbf{M}^i,Z)| \leq \frac{\kappa}{m}.$$

To do this, we use the fact that our loss function ℓ is k-Lipsichtz with $k = \left(4(1-\alpha)K^2\right)$ and our upper-bound on $\|\Delta\mathbf{M}\|_{\mathcal{F}}$. It gives:

$$\begin{aligned} |\ell(\mathbf{M},Z) - \ell(\mathbf{M}^i,Z)| &\leq k\|\Delta\mathbf{M}\|_{\mathcal{F}}, \\ &\leq \frac{2k^2(3m^2-3m+1)}{\mu m^3}. \end{aligned}$$

Finally:

$$\forall S,\ \forall i,\ 1 \leq i \leq m,\ \sup_{Z} |\ell(\mathbf{M}, Z) - \ell(\mathbf{M}^i, Z)| \leq \frac{\kappa}{m^3},$$

with $\kappa = \dfrac{4(3m^2 - 3m + 1)}{\mu} \times \left((1-\alpha)K^2\right)^2$. □

For the sake of simplicity, we will simplify this result in the following. Note that for all $m \geq 1$, $\dfrac{3m^2 - 3m + 1}{m^3} \leq \dfrac{3}{m}$. Thus, our algorithm has a uniform stability in $\dfrac{\kappa}{m}$ with $\kappa = \dfrac{12}{\mu} \times \left((1-\alpha)K^2\right)^2$. We can now apply Theorem 1 to our algorithm and get the following result:

Theorem 3. *Let $\delta > 0$ and $m > 1$. With probability $1 - \delta$, we have the following bound on the true risk $\mathcal{R}$ of our loss function ℓ:*

$$\mathcal{R} \leq \mathcal{R}_S + 2\frac{\kappa}{m} + (2\kappa + 2\gamma)\sqrt{\frac{\ln(2/\delta)}{2m}},$$

with:

$$\kappa = \frac{12}{\mu} \times \left((1-\alpha)K^2\right).$$

and

$$\gamma = (1-\alpha)(1 - c + 4K^2).$$

Proof. The proof is consequence of Theorem 1 and Lemma 4. □

3 Classification Guarantees - Proof

We now give a proof of the Theorem 3 provided in the paper.

Proof. We first begin with the FP rate. We can note the hinge loss can be a surrogate for the indicator function as follows:

$$\mathbb{1}_{\{d_{\mathbf{M}}(\mathbf{x},\mathbf{x}_{p_i}) \leq d(\mathbf{x},\mathbf{x}_n)\}} = \mathbb{1}_{\{d_{\mathbf{M}}(\mathbf{x},\mathbf{x}_{p_i})^2 \leq d(\mathbf{x},\mathbf{x}_n)^2\}} \leq \left[1 + d(\mathbf{x},\mathbf{x}_n)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_p)^2\right]_+,$$

We can recognize one of the term of our optimization Problem (1) with the hyper-parameter $c = 0$. We recall that each labeled example is denoted as $\mathbf{z} = (\mathbf{x}, y)$. Then, we have:

$$\begin{aligned}
\mathcal{R}_{FP} &\leq \mathbb{E}_{S\sim D^m}\mathbb{E}_{\mathbf{z}\sim D}\left[1 + d(\mathbf{x},\mathbf{x}_n)^2 - d_{\mathbf{M}}(\mathbf{x},\mathbf{x}_p)^2\right]_+ \times \mathbb{1}_{\{y=-1\}} \\
&\leq \mathbb{E}_{S'\sim D^{m+1}}\mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k\in S'}\left[1 + d(\mathbf{x}_i,\mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+, \times \mathbb{1}_{\{y_i=y_j=-1\neq y_k\}} \\
&\leq \mathbb{E}_{S'\sim D^{m+1}}\mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k\in S'}\left[\frac{\alpha}{\alpha}\left[1 + d(\mathbf{x}_i,\mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+ \times \mathbb{1}_{\{y_i=y_j=-1\neq y_k\}} + \right. \\
&\qquad \left. \frac{1-\alpha}{\alpha}\left(\left[1 + d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+ \times \mathbb{1}_{\{y_i=y_j=1\neq y_k\}}\right)\right], \\
&\leq \mathbb{E}_{S'\sim D^{m+1}}\mathbb{E}_{\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_k\in S'}\left[\frac{1}{\alpha}\left(\alpha\left[1 + d(\mathbf{x}_i,\mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+ \times \mathbb{1}_{\{y_i=y_j=-1\neq y_k\}} + \right.\right. \\
&\qquad \left.\left. (1-\alpha)\left[1 + d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+ \times \mathbb{1}_{\{y_i=y_j=1\neq y_k\}}\right)\right], \\
&\leq \frac{1}{\alpha}\mathcal{R}.
\end{aligned}$$

The second inequality is obtained by the i.i.d. aspect of the expectation. The third inequality is due to the fact that the second term in the sum is positive. Finally, one can note that the right-hand side of the last inequality corresponds to a weighted version of the true risk with respect to the loss used in Problem (1) with $c = 0$ and where we take an expectation over all the samples of size $m + 1$. The result is obtained by combining the results of Theorems 3 and 1 over the true risk defined above.

The bound for the false negative can be obtained in a similar way. Using the same arguments, one can show that:

$$\mathcal{R}_{FN} \leq \frac{1}{1-\alpha}\mathcal{R}.$$

Applying Theorems 3 and 1 to the above risk leads to the result. □

References

Bellet, A., Habrard, A., and Sebban, M. (2015). *Metric Learning*. Morgan & Claypool Publishers.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526.