A Theoretical Analysis of Metric Hypothesis Transfer Learning

Michaël Perrot MICHAEL.PERROT@UNIV-ST-ETIENNE.FR
Amaury Habrard AMAURY.HABRARD@UNIV-ST-ETIENNE.FR
Université de Lyon, Université Jean Monnet de Saint-Etienne,
Laboratoire Hubert Curien, CNRS, UMR5516, F-42000, Saint-Etienne, France.

Abstract

We consider the problem of transferring some a priori knowledge in the context of supervised metric learning approaches. While this setting has been successfully applied in some empirical contexts, no theoretical evidence exists to justify this approach. In this paper, we provide a theoretical justification based on the notion of algorithmic stability adapted to the regularized metric learning setting. We propose an on-average-replace-two-stability model allowing us to prove fast generalization rates when an auxiliary source metric is used to bias the regularizer. Moreover, we prove a consistency result from which we show the interest of considering biased weighted regularized formulations and we provide a solution to estimate the associated weight. We also present some experiments illustrating the interest of the approach in standard metric learning tasks and in a transfer learning problem where few labelled data are available.

1. Introduction

A lot of machine learning problems, such as clustering, classification or ranking, require to accurately compare examples by means of distances or similarities. Designing a good metric for a task at hand is thus of crucial importance. Manually tuning a metric is in general difficult and tedious, a recent trend consists to learn the metrics directly from data. This has led to the emergence of *supervised metric learning*, see (Bellet et al., 2013; Kulis, 2013) for up-to-date surveys. The underlying idea is to infer automatically the parameters of a metric in order to capture the idiosyncrasies of the data. In a supervised classification perspective, this is generally done by trying to satisfy pair-based constraints aiming at assigning a small (resp. large) score to pairs of examples of the same class (resp. different class). Most of the existing work has notably focused on learning Mahalanobis-like distances of the form $d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')}$ where $\mathbf{M}$ is a positive semi-definite (PSD) matrix[1], the learned matrix being typically plugged in a k-Nearest Neighbor classifier allowing one to achieve a better accuracy than the standard Euclidean distance.

Recently, there is a growing interest for methods able to take into account some background knowledge (Parameswaran & Weinberger, 2010; Cao et al., 2013; Bohné et al., 2014) for learning $\mathbf{M}$. This is in particular the case for *supervised regularized metric learning approaches* where the regularizer is biased with respect to an auxiliary metric given under the form of a matrix. The main objective here is to make use of this a priori knowledge in a setting where *only few labelled data* are available to help learning. For example, in the context of learning a PSD matrix $\mathbf{M}$ plugged into a Mahalanobis-like distance as discussed above, let $\mathbf{I}$ be the identity matrix used as an auxiliary knowledge, $\|\mathbf{M} - \mathbf{I}\|$ is a biased regularizer often considered. This regularization can be interpreted as follows: learn $\mathbf{M}$ while trying to stay close to the Euclidean distance, or from another standpoint try to learn a matrix $\mathbf{M}$ which performs better than $\mathbf{I}$. Other standard matrices can be used such as Σ^{-1} the inverse of the variance-covariance matrix, note that if we take the $\mathbf{0}$ matrix, we retrieve the classical unbiased regularization term.

Another useful setting comes when $\mathbf{I}$ is replaced by any auxiliary matrix $\mathbf{M}_S$ learned from another task. This corresponds to a *transfer learning* approach where the biased regularization can be interpreted as transferring the knowledge brought by $\mathbf{M}_S$ for learning $\mathbf{M}$. This setting is appropriate when the distributions over training and testing domains are different but related. *Domain adaptation* strate-

[1]Note that this distance is a generalization of some well-known distances: when $\mathbf{M} = \mathbf{I}$, $\mathbf{I}$ being the identity matrix, we retrieve the Euclidean distance, when $\mathbf{M} = \Sigma^{-1}$ where Σ is the variance-covariance matrix of the data at hand, it actually corresponds to the original definition of a Mahalanobis distance.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

gies (Ben-David et al., 2010) propose to make use of the relationship between the training examples, called the *source domain*, and the testing instances, called the *target domain* to infer a model. However, it is sometimes not possible to have access to all the training examples, for example when some new domains are acquired incrementally. In this context, transferring the information directly from the model learned from the source domain without any other access to the source domain is of crucial importance. In the context of this paper, we call this setting *Metric Hypothesis Transfer Learning* in reference to the *Hypothesis Transfer Learning* model introduced in (Kuzborskij & Orabona, 2013) in the context of classical supervised learning.

Metric learning generally suffers from a lack of theoretical justifications, in particular *metric hypothesis transfer learning* has never been investigated from a theoretical standpoint. In this paper, we propose to bridge this gap by providing a theoretical analysis showing that *supervised regularized metric learning* approaches using a biased regularization are well-founded. Our theoretical analysis is based on *algorithmic stability* arguments allowing one to derive generalization guarantees when a learning algorithm does not suffer too much from a little change in the training sample. As a first contribution, we introduce a new notion of stability called *on-average-replace-two-stability* that is well-suited to regularized metric learning formulations. This notion allows us to prove a high probability generalization bound for metric hypothesis transfer learning achieving a fast converge rate in $\mathcal{O}(1/n)$ in the context of admissible, lipschitz and convex losses. In a second step, we provide a consistency result from which we justify the interest of *weighted biased regularization* of the form $\|\mathbf{M} - \beta\mathbf{M}_\mathcal{S}\|$ where β is a parameter to set. From this result, we derive an approach for assessing this parameter without resorting to a costly parameter tuning procedure. We also provide an experimental study showing the effectiveness of transfer metric learning with weighted biased regularization in the presence of few labeled data both on standard metric learning and transfer learning tasks.

This paper is organized as follows. Section 2 introduces some notations and definitions while Section 3 discusses some related work. Our theoretical analysis is presented in Section 4. We detail our experiments in Section 5 before concluding in Section 6.

2. Notations and Definitions

We start by introducing several notations and definitions that will be used throughout the paper. Let $\mathcal{T}$ be a domain equipped with a probability distribution $\mathcal{D}_\mathcal{T}$ defined over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y}$ is the label set. We consider metrics corresponding to distance functions $\mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ parameterized by a $d \times d$ positive semi-definite (PSD) matrix $\mathbf{M}$ denoted $\mathbf{M} \succeq 0$. In the following, a metric will be represented by its matrix $\mathbf{M}$. We also consider that we have access to some additional information under the form of an auxiliary $d \times d$ matrix $\mathbf{M}_\mathcal{S}$, throughout this paper we call this additional information source metric or source hypothesis. We denote the Frobenius norm by $\|\cdot\|_\mathcal{F}$, $\mathbf{M}_{kl}$ represents the value of the entry at index (k, l) in matrix $\mathbf{M}$, $[a]_+ = max(a, 0)$ denotes the hinge loss and $[n]$ the set $\{1, \ldots, n\}$ for any $n \in \mathbb{N}$.

Let $T = \{\mathbf{z}_i = (\mathbf{x}_i, y_i)\}_{i=1}^n$ be a labeled training set drawn from $\mathcal{D}_\mathcal{T}$. We consider the following learning framework for *biased regularized metric learning*:

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

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 a metric hypothesis $\mathbf{M}$. Similarly we denote the true risk by $L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}) = \mathbb{E}_{\mathbf{z},\mathbf{z}' \sim \mathcal{D}_\mathcal{T}} l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$. In this work we only consider convex, k-lipschitz and (σ, m)-admissible losses for which we recall the definitions below.

Definition 1 (k-lipschitz continuity)**.** *A loss function $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$ is k-lipschitz w.r.t. its first argument if, for any matrices $\mathbf{M}$, $\mathbf{M}'$ and any pair of examples $\mathbf{z}, \mathbf{z}'$, there exists $k \geq 0$ such that:*

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

This property ensures that the loss deviation does not exceed the deviation between matrices $\mathbf{M}$ and $\mathbf{M}'$ with respect to a positive constant k.

Definition 2 ((σ, m)-admissibility)**.** *A loss function $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$ is (σ, m)-admissible, w.r.t. $\mathbf{M}$, if it is convex w.r.t. its first argument and if for any two pairs of examples $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3, \mathbf{z}_4$, we have:*

$$|l(\mathbf{M}, \mathbf{z}_1, \mathbf{z}_2) - l(\mathbf{M}, \mathbf{z}_3, \mathbf{z}_4)| \leq \sigma\,|y_1 y_2 - y_3 y_4| + m$$

where $y_i y_j = 1$ if $y_i = y_j$ and -1 otherwise. Thus $|y_1 y_2 - y_3 y_4| \in \{0, 2\}$.

This property bounds the difference between the losses of two pairs of examples by a value only related to the labels plus a constant independent from $\mathbf{M}$.

To derive our theoretical results, we make use of the notion of *algorithmic stability* which allows one to provide generalization guarantees. A learning algorithm is stable if a slight modification in its input does not change its output much. In our analysis we use two definitions of stability. On the one hand, we introduce in Section 4.1 the notion of *on-average-replace-two-stability* which is an adaptation to metric learning of the notion of on-average-replace-one-stability proposed in (Shalev-Shwartz & Ben-David, 2014) and recalled in Def. 3 below.

Definition 3 (On-average-replace-one-stability)**.** *Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be monotonically decreasing and $U(n)$ be the uniform distribution over $[n]$. An algorithm A is on-average-replace-one-stable with rate $\epsilon(n)$ if for any distribution $\mathcal{D}_\mathcal{T}$*

$$\mathbb{E}_{\substack{T\sim\mathcal{D}_\mathcal{T}{}^n \\ i\sim U(n) \\ \mathbf{z}'\sim\mathcal{D}_\mathcal{T}}} \left[l(A(T^i), \mathbf{z}^i) - l(A(T), \mathbf{z}^i) \right] \leq \epsilon(n)$$

where $A(T)$, respectively $A(T^i)$ is the optimal solution of algorithm A when learning with training set T, respectively T^i. T^i is obtained by replacing the i^{th} example of T by $\mathbf{z}'$.

This property ensures that, given an example, learning with or without it will not imply a big change in the hypothesis prediction. Note that the property is required to be true on average over all the possible training sets of size n.

On the other hand, we consider an adaptation of the framework of *uniform stability* for metric learning proposed in (Jin et al., 2009) and recalled in Def. 4.

Definition 4 (Uniform stability)**.** *A learning algorithm has a uniform stability in $\frac{\mathcal{K}}{n}$, with $\mathcal{K} \geq 0$ a constant, if $\forall i$,*

$$\sup_{\mathbf{z},\mathbf{z}'\sim\mathcal{D}_\mathcal{T}} \left| l(\mathbf{M}^*, \mathbf{z}, \mathbf{z}') - l(\mathbf{M}^{i*}, \mathbf{z}, \mathbf{z}') \right| \leq \frac{\mathcal{K}}{n}$$

where $\mathbf{M}^*$ is the matrix learned on the training set T, $\mathbf{M}^{i*}$ is the matrix learned on the training set T^i obtained by replacing the i^{th} example of T by a new independent one.

Uniform stability requires that a small change in the training set does not imply a significant variation in the learned models output. The constraint in $\mathcal{O}\left(\frac{1}{n}\right)$ over the supremum makes this property rather strong since it considers a worst case over the possible pairs of examples to compare, whatever the training set. It is actually one of the most general algorithmic stability setting (Bousquet & Elisseeff, 2002).

3. Related Work

3.1. Metric Learning

Based on the pioneering approach of (Xing et al., 2002), metric learning aims at finding the parameters of a distance function by maximizing the distance between dissimilar examples (*i.e.* examples of different class) while maintaining a small distance between similar ones (*i.e.* of similar class). Following this idea, one of the most famous approach, called LMNN (Weinberger et al., 2005), proposes to learn a PSD matrix dedicated to improve the k-nearest neighbours algorithm. To do so, the authors force the metric to respect triplet-based local constraints of the form $(\mathbf{z}_i, \mathbf{z}_j, \mathbf{z}_k)$ where $\mathbf{z}_j$ and $\mathbf{z}_k$ belong to the neighbourhood of $\mathbf{z}_i$, $\mathbf{z}_i$ and $\mathbf{z}_j$ being of the same class, and $\mathbf{z}_k$ being of opposite class. The constraints impose that $\mathbf{z}_i$ should be closer to $\mathbf{z}_j$ than to $\mathbf{z}_k$ with respect to a margin ε. In ITML, (Davis et al., 2007) propose to use a LogDet divergence as a regularizer allowing one to ensure an automatic enforcement of the PSD constraint. The idea is to force the learned matrix $\mathbf{M}$ to stay as close as possible to a good matrix $\mathbf{M}_\mathcal{S}$ defined a-priori (in general $\mathbf{M}_\mathcal{S}$ is chosen as the identity matrix). Indeed, if this divergence is finite, the authors show that $\mathbf{M}$ is guaranteed to be PSD. This constraint over $\mathbf{M}$ can be interpreted as a biased regularization w.r.t. $\mathbf{M}_\mathcal{S}$.

The idea behind biased regularization has been successfully used in many metric learning approaches. For example, (Zha et al., 2009) have proposed to replace the identity matrix ($\mathbf{M}_\mathcal{S} = \mathbf{I}$) originally used in ITML by matrices previously learned on so called auxiliary data sets. Similarly, in (Parameswaran & Weinberger, 2010) the authors are interested in Multi-Task metric learning. They propose to learn one metric for each task and a global metric common to all the tasks. For this global metric, they consider a biased regularization of the form $\|\mathbf{M} - \mathbf{I}\|_\mathcal{F}^2$ where $\mathbf{I}$ is the identity matrix but they do not study any other kind of source information. In (Cao et al., 2013), the authors use a similar biased regularization to learn a metric learning model for face recognition. As a last example, (Bohné et al., 2014) introduce a regularization of the form $\|\mathbf{M} - \beta\mathbf{I}\|_\mathcal{F}$ where they learn $\mathbf{M}$ and β. In our work, instead of optimizing these two parameters, we derive a theoretically founded algorithm to choose beforehand the optimal value of β.

3.2. Theoretical Frameworks in Metric Learning

Theoretically speaking, there is not a lot of frameworks for metric learning. The goal of generalization guarantees is to show that the empirical estimation of the error of an algorithm does not deviate much from the true error. One of the main difficulty in deriving bounds for metric learning is the fact that instead of considering examples drawn i.i.d. from a distribution, we consider pairs of examples which might not be independent. Building upon the framework of stability proposed in (Bousquet & Elisseeff, 2002), (Jin et al., 2009) propose one of the first study of the generalization ability of a metric learning algorithm. Building upon this work, (Perrot et al., 2014) give theoretical guarantees for a local metric learning algorithm and (Bellet et al., 2012) derive generalization guarantees for a similarity learning algorithm. Other ways to derive generalization guarantees are to use the Rademacher complexity as in (Cao et al., 2012; Guo & Ying, 2014) or to use the notion of algorithmic robustness (Bellet & Habrard, 2015).

3.3. Biased Regularization in Supervised Learning

Biased regularization has already been studied in non metric learning settings. For example in (Kienzle & Chellapilla, 2006), the authors propose to use biased regular-

ization to learn SVM classifiers. A first theoretical study of biased regularization in the context of regularized least squares has been proposed in (Kuzborskij & Orabona, 2013). Their study is based on a notion of *hypothesis stability* less general than the *uniform stability* used in our approach. In (Kuzborskij & Orabona, 2014), the authors derive generalization bounds based on the Rademacher complexity for regularized empirical risk minimization methods in a supervised learning setting. Their results show that if the true risk of the source hypothesis on the target domain is low, then the generalization rate can be improved. However computing the true risk of the source hypothesis is not possible in practice. In our analysis, we derive a generalization bound which depends on the empirical risk and the complexity (w.r.t. the regularization term) of the source metric. It allows us to derive an algorithm to minimize the generalization bound taking into account the performance and the complexity of the source metric.

4. Contribution

We divide our contribution consisting of a theoretical analysis of Alg. 1 given convex, k-lipschitz and (σ, m)-admissible losses into three parts. First, we provide an on average analysis for $\mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*)\right]$ where $\mathbf{M}^*$ represents the metric learned with Alg. 1 using training set T. This analysis allows us to bound the expected loss over distribution $\mathcal{D}_\mathcal{T}$ with respect to the loss of the auxiliary metric $\mathbf{M}_\mathcal{S}$ over $\mathcal{D}_\mathcal{T}$. It shows that on average the learned metric tends to be better than the given source $\mathbf{M}_\mathcal{S}$, with a fast convergence rate in $\mathcal{O}(1/n)$. Second, we provide a consistency analysis of our framework leading to a standard convergence rate of $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ w.r.t the empirical loss over T optimized in Alg. 1. In a third part, we specialize the previous consistency result to a specific loss and show that it is possible to refine our generalization bound in order to depend both on the complexity of our source metric $\mathbf{M}_\mathcal{S}$ and its empirical performance on the training set T. We then deduce an approach to weight the importance of the source hypothesis for optimizing the generalization bound.

4.1. On average analysis

Def. 3 allows one to perform an average analysis over the expected loss, however its formulation is not tailored to metric learning approaches that work with pair of examples. Thus we propose an adaptation of it that we call *on-average-replace-two-stability* allowing one to derive sharp bounds for metric learning.

Definition 5 (On-average-replace-two-stability)**.** *Let $\epsilon : \mathbb{N} \to \mathbb{R}$ be monotonically decreasing and let $U(n)$ be the uniform distribution over $[n]$. A metric learning algorithm is on-average-replace-two-stable with rate $\epsilon(n)$ if for every distribution $\mathcal{D}_\mathcal{T}$:*

$$\mathbb{E}_{\substack{T\sim\mathcal{D}_\mathcal{T}^n \\ i,j\sim U(n) \\ \mathbf{z}_1,\mathbf{z}_2\sim\mathcal{D}_\mathcal{T}}}\left[l(\mathbf{M}^{ij^*}, \mathbf{z}^i, \mathbf{z}^j) - l(\mathbf{M}^*, \mathbf{z}^i, \mathbf{z}^j)\right] \leq \epsilon(n)$$

where $\mathbf{M}^*$, respectively $\mathbf{M}^{ij^*}$, is the optimal solution when learning with the training set T, respectively T^{ij}. T^{ij} is obtained by replacing $\mathbf{z}^i$, the i^{th} example of T, by $\mathbf{z}_1$ to get a training set T^i and then by replacing $\mathbf{z}^j$, the j^{th} example of T^i, by $\mathbf{z}_2$.

Note that when this definition holds, it implies $\mathbb{E}_T\left[L_{\mathcal{D}_\mathcal{T}}(\mathbf{M}^*) - L_T(\mathbf{M}^*)\right] \leq \epsilon(n)$. The next theorem shows that our algorithm is on-average-replace-two-stable.

Theorem 1 (On-average-replace-two-stability)**.** *Given a training sample T of size n 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. The proof of Th. 1 can be found in the supplementary material. □

We can now bound the expected true risk of our algorithm.

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

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

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

Proof. We have:

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

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

This bound shows that with a sufficient number of examples w.r.t. a fast convergence rate in $\mathcal{O}(1/n)$, we will on average obtain a metric which is at least as good as the source hypothesis. Thus choosing a good source metric is key to learn well.

4.2. Consistency analysis

We now provide a consistency analysis taking into account the empirical risk optimized in Alg. 1. We begin by showing that our algorithm is uniformly stable w.r.t. Def. 4 in the next theorem.

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. The beginning of the proof follows closely the one proposed in (Bousquet & Elisseeff, 2002) and is postponed to the supplementary material for the sake of readability. We consider the end of the proof here. We have

$$B \leq \frac{4kt}{n}\|\mathbf{\Delta M}\|_{\mathcal{F}}$$

where $B = \lambda\|\mathbf{M}-\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M}-t\mathbf{\Delta M}-\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 + \lambda\|\mathbf{M}^i - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 - \lambda\|\mathbf{M}^i + t\mathbf{\Delta M} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2$.
Setting $t = \frac{1}{2}$ we have:

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

Then we obtain

$$\frac{\lambda}{2}\|\mathbf{\Delta M}\|_{\mathcal{F}}^2 \leq \frac{4k}{2n}\|\mathbf{\Delta M}\|_{\mathcal{F}} \Leftrightarrow \|\mathbf{\Delta M}\|_{\mathcal{F}} \leq \frac{4k}{\lambda n}.$$

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

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

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

Using the fact that our algorithm is uniformly stable, we can derive generalization guarantees as stated in Th. 4.

Theorem 4 (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:*

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

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

Proof. The proof is available in the supplementary. □

This bound shows that with a convergence rate in $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ the true risk of our algorithm is bounded above by the empirical risk justifying the consistency of the approach. In the next section, we propose an extension of this analysis to include the performance of the source metric. This extension allows us to introduce a natural weighting of the source metric in order to improve the proposed bound.

4.3. Refinement with weighted source hypothesis

In this part we study a specific loss, namely $l(\mathbf{M}, \mathbf{z}, \mathbf{z}') = \left[yy'((\mathbf{x}-\mathbf{x}')^T\mathbf{M}(\mathbf{x}-\mathbf{x}') - \gamma_{yy'})\right]_+$ where $yy' = 1$ if $y = y'$ and -1 otherwise. The convexity follows from the use of the hinge loss. In the next two lemmas, we show that this loss is k-lipschitz continuous and (σ, m)-admissible. The (σ, m)-admissibility result is of high importance because it allows us to introduce some information coming from the source matrix $\mathbf{M}_{\mathcal{S}}$.

Lemma 1 (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. The proof is available in the supplementary. □

Lemma 2 ((σ, m)-admissibility). *Let $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4$ be four examples and $\mathbf{M}^*$ be the optimal solution of Problem 1. The convex and k-lipschitz loss function $l(\mathbf{M}, \mathbf{z}, \mathbf{z}')$ is (σ, m)-admissible with $\sigma = \max(\gamma_{y_3y_4}, \gamma_{y_1y_2})$ and $m = 2\max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x} - \mathbf{x}'\|^2 \left(\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}\right)$.*

Proof. Let $\boldsymbol{\varepsilon}^* = \mathbf{M}^* - \mathbf{M}_{\mathcal{S}}$ be the difference between the learned and the source metric. We first bound the frobenius norm of $\boldsymbol{\varepsilon}^*$ w.r.t. the performance of the source metric.

$$\begin{aligned}
& L_T(\mathbf{M}^*) + \lambda\|\mathbf{M}^* - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \leq L_T(\mathbf{M}_{\mathcal{S}}) + \lambda\|\mathbf{M}_{\mathcal{S}} - \mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}^2 \\
\Rightarrow \quad & \lambda\|\boldsymbol{\varepsilon}^*\|_{\mathcal{F}}^2 \leq L_T(\mathbf{M}_{\mathcal{S}}) \Leftrightarrow \|\boldsymbol{\varepsilon}^*\|_{\mathcal{F}} \leq \sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}}
\end{aligned}$$

Now we can prove the (σ, m)-admissibility of our loss.

$$\begin{aligned}
&|l(\mathbf{M}^*, \mathbf{z}_1, \mathbf{z}_2) - l(\mathbf{M}^*, \mathbf{z}_3, \mathbf{z}_4)| \\
&\quad = |\left[y_1y_2((\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{M}^*(\mathbf{x}_1 - \mathbf{x}_2) - \gamma_{y_1y_2})\right]_+ \\
&\qquad - \left[y_3y_4((\mathbf{x}_3 - \mathbf{x}_4)^T\mathbf{M}^*(x_3 - x_4) - \gamma_{y_3y_4})\right]_+| \\
&\quad \leq |y_1y_2((\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{M}^*(\mathbf{x}_1 - \mathbf{x}_2) - \gamma_{y_1y_2}) \\
&\qquad - y_3y_4((\mathbf{x}_3 - \mathbf{x}_4)^T\mathbf{M}^*(\mathbf{x}_3 - \mathbf{x}_4) - \gamma_{y_3y_4})| \qquad (3) \\
&\quad \leq |y_1y_2(\mathbf{x}_1 - \mathbf{x}_2)^T\mathbf{M}^*(\mathbf{x}_1 - \mathbf{x}_2) \\
&\qquad - y_3y_4(\mathbf{x}_3 - \mathbf{x}_4)^T\mathbf{M}^*(\mathbf{x}_3 - \mathbf{x}_4)| \\
&\qquad + |y_3y_4\gamma_{y_3y_4} - y_1y_2\gamma_{y_1y_2}|
\end{aligned}$$

$$
\begin{aligned}
&\leq 2 \max_{\mathbf{x},\mathbf{x}'} ((\mathbf{x}-\mathbf{x}')^T \mathbf{M}^* (\mathbf{x}-\mathbf{x}')) \\
&\quad + |y_3 y_4 - y_1 y_2| \max(\gamma_{y_3 y_4}, \gamma_{y_1 y_2}) \\
&\leq 2 \max_{\mathbf{x},\mathbf{x}'} ((\mathbf{x}-\mathbf{x}')^T (\boldsymbol{\varepsilon}^* + \mathbf{M}_{\mathcal{S}}) (\mathbf{x}-\mathbf{x}')) \\
&\quad + |y_3 y_4 - y_1 y_2| \max(\gamma_{y_3 y_4}, \gamma_{y_1 y_2}) \\
&\leq 2 \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x}-\mathbf{x}'\|^2 (\|\boldsymbol{\varepsilon}^*\|_{\mathcal{F}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}) \\
&\quad + |y_3 y_4 - y_1 y_2| \max(\gamma_{y_3 y_4}, \gamma_{y_1 y_2}) \qquad (4) \\
&\leq 2 \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x}-\mathbf{x}'\|^2 (\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}) \\
&\quad + |y_3 y_4 - y_1 y_2| \max(\gamma_{y_3 y_4}, \gamma_{y_1 y_2}).
\end{aligned}
$$

Inequality 3 comes from the 1-lipschitz property of the hinge loss. We obtain inequality 4 by applying the Cauchy-Schwarz inequality and some classical norm properties. Setting $m = 2 \max_{\mathbf{x},\mathbf{x}'} \|\mathbf{x}-\mathbf{x}'\|^2 (\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}})$ and $\sigma = \max(\gamma_{y_3 y_4}, \gamma_{y_1 y_2})$ gives the lemma. □

Using Lemmas 1 and 2 we can now derive, in Th. 5, a generalization bound associated with our specific loss.

Theorem 5 (Generalization bound)**.** *With probability* $1-\delta$ *for any matrix* $\mathbf{M}$ *learned with Alg. 1, we have:*

$$
\begin{aligned}
L_{\mathcal{D}_T}(\mathbf{M}) \leq & L_T(\mathbf{M}) + \mathcal{O}\left(\frac{1}{n}\right) \\
& + \left(\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}} + c_\gamma\right) \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}
\end{aligned}
$$

where c_γ *is a constant linked to the* k*-lipschitz property of the loss and the chosen margins.*

Proof. The proof is the same as for Th. 4 replacing k, σ and m by their values. □

As for Th. 4, the convergence rate is in $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. The term $C(\mathbf{M}_{\mathcal{S}}) \stackrel{\text{def}}{=} \left(\sqrt{\frac{L_T(\mathbf{M}_{\mathcal{S}})}{\lambda}} + \|\mathbf{M}_{\mathcal{S}}\|_{\mathcal{F}}\right)$ mainly depends on the quality of the source hypothesis $\mathbf{M}_{\mathcal{S}}$. The product $C(\mathbf{M}_{\mathcal{S}})\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ means that as the number of examples available for learning increases, the quality of the source metric is of decreasing importance. A similar result has already been stated in domain adaptation or transfer learning in (Ben-David et al., 2010; Kuzborskij & Orabona, 2013) where they show that as the number of target examples increases, the necessity of having source examples decreases.

Given a source hypothesis $\mathbf{M}_{\mathcal{S}}$, it is possible to optimize it w.r.t. the bound derived in Th. 5. Indeed, note that $C(\mathbf{M}_{\mathcal{S}})$ corresponds to a trade-off between the complexity of the source metric and its performance on the training set. The lower the value of this term, the tighter the bound. Hence, we propose a way to minimize the generalization bound and more specifically $C(\mathbf{M}_{\mathcal{S}})$ by adding a weighting parameter $\beta \geq 0$ on the source metric $\mathbf{M}_{\mathcal{S}}$. This parameter is a way to control the trade-off between complexity and performance of the source metric. It can be assessed by means of the following optimization problem:

$$
\beta^* = \arg\min_{\beta} C(\beta \mathbf{M}_{\mathcal{S}}) \qquad (5)
$$

Note that the bound derived in Th. 5 holds whatever the value of $\mathbf{M}_{\mathcal{S}}$. Thus replacing it with $\beta^* \mathbf{M}_{\mathcal{S}}$ does not impact the theoretical study proposed in this section.

Interpretation of the value of β^* We can distinguish three main cases. First if the source hypothesis performs poorly on the training set at hand we expect β^* to be as small as possible to reduce the importance of $\mathbf{M}_{\mathcal{S}}$. In a sense, we tend to go back to the classical case were $\mathbf{M}_{\mathcal{S}} = \mathbf{0}$. Second if the source hypothesis is complex and performs well, we expect β^* to be rather small to reduce the complexity of the hypothesis while keeping a good performance on the training set. Third if the source hypothesis is simple and performs well, we expect β^* to be closer to one since $\mathbf{M}_{\mathcal{S}}$ is already a good choice.

Learning β^* Problem 5 is highly non differentiable[2] and non convex. However, it remains simple in the sense that we have only one parameter to assess and we used a classical subgradient descent to solve it. Even if it is not convex, our empirical study shows no need to perform many restarts to output a good solution: we always found almost the same solution. As a consequence, we applied only one optimization procedure in our experiments.

In this section we presented a new framework for metric learning where one can use a source hypothesis to add some side information during the learning process. We have shown that our approach is consistent with a convergence rate in $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Furthermore, given a specific loss, we have shown that the use of a weighting parameter to control the importance of the source metric is theoretically founded. In the next part we empirically demonstrate that we can obtain competitive results both in a classical metric learning setting and in a domain adaptation setting.

5. Experiments

We propose an empirical study according to two directions depending on the choice of the source metric. First, using some well-known distances as a source metric, we show that our framework performs well on classical supervised metric learning tasks of the UCI database and we empirically demonstrate the interest of learning the β parameter.

[2] To avoid this problem, we can use the classical relaxation with slack variables.

	Baselines			Our approach			
Dataset	1-NN	ITML	LMNN	IDENTITY	IDENTITY-B1	MAHALANOBIS	MAHALANOBIS-B1
Breast	95.31 ± 1.11	95.40 ± 1.37	95.60 ± 0.92	$\mathbf{96.06 \pm 0.77}$	95.75 ± 0.87	95.71 ± 0.84	94.76 ± 1.38
Pima	67.92 ± 1.95	68.13 ± 1.86	67.90 ± 2.05	67.87 ± 1.57	67.54 ± 1.99	$\mathbf{68.37 \pm 2.00}$	66.31 ± 2.37
Scale	78.73 ± 1.69	$\mathbf{87.31 \pm 2.35}$	86.20 ± 2.83	80.98 ± 1.51	80.82 ± 1.27	81.35 ± 1.17	80.88 ± 1.43
Wine	93.40 ± 2.70	93.82 ± 2.63	93.47 ± 1.80	$\mathbf{95.42 \pm 1.71}$	95.07 ± 1.68	94.31 ± 2.01	80.56 ± 5.75

Table 1. Results of the experiments conducted on the UCI datasets. Each value corresponds to the mean and standard deviation over 10 runs. For each dataset we highlight the best result using a bold font. Approaches with the suffix -B1 do not learn β, it is fixed to 1.

Second, we apply our framework in a semi-supervised Domain Adaptation task. We show that, using only source information through a learned metric, our method is able to compete with state of the art algorithms.

Setup In all our experiments we use limited training dataset, making it difficult to apply any kind of cross-validation to set the parameters. Thus we propose to fix them as follows. First the positive and negative margin are respectively set to the 5^{th} and 95^{th} percentile of the training set possible distances computed with the source metric as proposed in (Davis et al., 2007). Next we set λ such that the two terms of Eq. 5 are equals, i.e. we balance the complexity and performance importance with respect to the source metric. The β parameter is then learned using Algorithm 5. In all the experiments we plug our metric in a 1-nearest neighbour classifier to classify the examples of the test set. Furthermore, the significance of the results is assessed with a paired samples t-test considering that an approach is significantly better when the p-value is lower than 0.05.

5.1. Classical Supervised Metric Learning

First we start by conducting experiments on several UCI datasets (Lichman, 2013), namely breast, pima, scale and wine. We propose to consider three source metrics: (i) **Zero**: No source hypothesis, (ii) **Identity**: Euclidean distance, (iii) **Mahalanobis**: Inverse of the variance-covariance matrix computed on the training set.

For the last two hypothesis we propose two experiments, one where we set $\beta = 1$ and one where we learn β using Algorithm 5. The goal of this experiment is to show the interest of automatically setting β. We consider a 1-nearest neighbour (1-NN) classifier using the Euclidean Distance as the baseline and also report the results of two well known metric learning algorithms, namely ITML, (Davis et al., 2007) and LMNN (Weinberger et al., 2005). The results averaged over 10 runs are reported in Table 1. For each run we randomly draw a training set containing 20% of the data available for each class and we test the metric on the remaining 80% of data.

These experiments highlight the interest of learning the β parameter. When we consider the performance of our approach with and without learning β, we mainly notice the following facts. First, learning β always leads to an improvement on all the datasets and the final result is better than the $1NN$ classifier. Second, learning β when considering the identity matrix as the source metric seems to be of limited interest as the differences in accuracy are only significant for the wine dataset. This can be justified by the fact that, in this case, it only consists of a rescaling of the diagonal of the matrix and it does not change much the behaviour of the distance. Finally, learning β when considering the variance-covariance matrix as the source metric leads to a significant improvement of the performance of the metric except on the breast dataset. This is particularly true for the wine dataset with a gain of nearly 14% in accuracy. It can be explained by the fact that, for this dataset, we are learning with less than 40 examples. Thus the original Mahalanobis distance does not carry as much information as in the other datasets and is thus of a lower quality. Learning β allows us to compensate this drawback and to obtain results which are even better than ITML or LMNN.

5.2. Metric learning for Semi-supervised Domain Adaptation

In this section we consider a Semi-supervised Domain Adaptation (DA) task with the Office-Caltech dataset. This dataset consists of four domains: Amazon (A), Caltech (C), DSLR (D) and Webcam (W) for which we consider 10 classes. This leads to consider 12 different adaptation problems when we alternatively take each domain as the source or the target dataset. In these experiments we use the same splits as the ones considered in (Hoffman et al., 2013) since they are freely available from the authors website and follow their experimental setup. The results averaged over 20 runs and for each run 8 labelled source examples (20 if the source is Amazon) and 3 labelled target examples are selected. The data is normalized thanks to the zscore and the dimensionality is reduced to 20 thanks to a simple PCA. The results are presented in Table 2 where we compare the performance of our algorithm to 6 baselines: (i) 1-NN$_{\mathcal{S}}$: a 1-NN using the source examples, (ii) 1-NN$_{\mathcal{T}}$: a 1-NN using the target examples, (iii) LMNN$_{\mathcal{T}}$: a 1-NN on the target examples using the metric learned by LMNN on the source examples, (iv) ITML$_{\mathcal{T}}$: a 1-NN on the target examples using the metric learned by ITML on the source examples, (v) MMDT: a DA method (Hoffman et al., 2013), (vi) GFK:

	Baselines						Our approach		
Task	1-NN$_{\mathcal{S}}$	1-NN$_{\mathcal{T}}$	LMNN$_{\mathcal{T}}$	ITML$_{\mathcal{T}}$	MMDT	GFK	MAHALANOBIS	ITML	LMNN
A $\rightarrow$ C	35.95 $\pm$ 1.30	31.92 $\pm$ 3.24	32.42 $\pm$ 3.03	32.56 $\pm$ 4.17	**39.76 $\pm$ 2.25**	37.81 $\pm$ 1.85	32.65 $\pm$ 3.76	32.93 $\pm$ 4.60	34.66 $\pm$ 3.66
A $\rightarrow$ D	33.58 $\pm$ 4.37	53.31 $\pm$ 4.31	49.96 $\pm$ 3.53	44.33 $\pm$ 8.18	54.25 $\pm$ 4.32	51.54 $\pm$ 3.55	54.69 $\pm$ 3.96	51.54 $\pm$ 4.03	**54.72 $\pm$ 5.00**
A $\rightarrow$ W	33.68 $\pm$ 3.60	66.25 $\pm$ 3.87	62.62 $\pm$ 4.49	58.17 $\pm$ 10.63	64.91 $\pm$ 5.71	59.36 $\pm$ 4.30	67.11 $\pm$ 5.11	64.09 $\pm$ 5.20	**67.62 $\pm$ 5.18**
C $\rightarrow$ A	37.37 $\pm$ 2.95	47.28 $\pm$ 4.15	42.97 $\pm$ 3.76	45.16 $\pm$ 7.60	**51.05 $\pm$ 3.38**	46.36 $\pm$ 2.94	50.15 $\pm$ 4.87	49.89 $\pm$ 5.25	50.36 $\pm$ 4.67
C $\rightarrow$ D	31.89 $\pm$ 5.77	54.17 $\pm$ 4.76	46.02 $\pm$ 6.54	48.07 $\pm$ 8.98	52.80 $\pm$ 4.84	**58.07 $\pm$ 3.90**	56.77 $\pm$ 4.63	53.78 $\pm$ 7.23	57.44 $\pm$ 4.48
C $\rightarrow$ W	28.60 $\pm$ 6.13	65.06 $\pm$ 6.27	55.79 $\pm$ 5.09	59.21 $\pm$ 9.71	62.75 $\pm$ 5.19	63.26 $\pm$ 5.89	64.64 $\pm$ 6.44	64.00 $\pm$ 6.08	**65.11 $\pm$ 5.25**
D $\rightarrow$ A	33.59 $\pm$ 1.77	47.81 $\pm$ 3.56	40.57 $\pm$ 3.79	45.06 $\pm$ 6.78	**50.39 $\pm$ 3.40**	40.77 $\pm$ 2.55	49.48 $\pm$ 4.41	49.11 $\pm$ 4.09	49.67 $\pm$ 4.00
D $\rightarrow$ C	31.16 $\pm$ 1.19	32.22 $\pm$ 2.98	27.96 $\pm$ 3.03	29.93 $\pm$ 4.84	**35.70 $\pm$ 3.25**	30.64 $\pm$ 1.98	32.90 $\pm$ 3.14	32.99 $\pm$ 3.58	33.84 $\pm$ 2.99
D $\rightarrow$ W	**76.92 $\pm$ 2.18**	66.19 $\pm$ 4.60	65.36 $\pm$ 3.82	66.74 $\pm$ 7.16	74.43 $\pm$ 3.10	74.98 $\pm$ 2.89	65.57 $\pm$ 4.52	66.38 $\pm$ 6.04	69.72 $\pm$ 3.78
W $\rightarrow$ A	32.19 $\pm$ 3.04	48.25 $\pm$ 3.52	41.69 $\pm$ 3.71	45.11 $\pm$ 5.72	50.56 $\pm$ 3.66	43.26 $\pm$ 2.34	50.80 $\pm$ 3.63	50.16 $\pm$ 4.32	**50.92 $\pm$ 4.00**
W $\rightarrow$ C	27.67 $\pm$ 2.58	30.74 $\pm$ 3.92	28.60 $\pm$ 3.41	28.99 $\pm$ 4.31	**34.86 $\pm$ 3.62**	29.95 $\pm$ 3.05	31.54 $\pm$ 3.60	31.40 $\pm$ 4.29	32.64 $\pm$ 3.52
W $\rightarrow$ D	64.61 $\pm$ 4.30	54.84 $\pm$ 5.17	56.89 $\pm$ 5.06	57.76 $\pm$ 7.03	62.52 $\pm$ 4.40	**71.93 $\pm$ 4.07**	57.17 $\pm$ 6.50	56.85 $\pm$ 5.51	61.14 $\pm$ 5.78
Mean	38.93 $\pm$ 3.26	49.84 $\pm$ 4.20	45.90 $\pm$ 4.11	46.76 $\pm$ 7.09	**52.83 $\pm$ 3.93**	50.66 $\pm$ 3.28	51.12 $\pm$ 4.55	50.26 $\pm$ 5.02	52.32 $\pm$ 4.36

Table 2. Metric Learning for Semi-Supervised Domain Adaptation. For the sake of readability we design the considered domains by their initials. $\mathcal{S} \rightarrow \mathcal{T}$ stands for adaptation from the source domain to the target domain. Each time we consider the mean and standard deviation over 20 runs. For each task, the best result is highlighted with a bold font.

another DA approach (Gong et al., 2012).

The last two methods need the source sample while in our case we only use a source metric learned from the source instances. For our biased regularization framework we consider 3 possible metrics learned on the sources examples, namely (i) Mahalanobis, (ii) ITML and (iii) LMNN.

These results show that metric hypothesis transfer learning can perform well in a Semi-supervised Domain Adaptation setting. Indeed, we perform better than directly plugging the metrics learned by LMNN and ITML in a 1-nearest neighbour classifier. Moreover, we obtain accuracies which are competitive with state of the art approaches like MMDT or GFK while using less information. If we compare our approach using LMNN as the source metric with MMDT, we note that MMDT is significantly better than our approach on 4 out of 12 tasks while we are significantly better on 3 and 5 ends as a draw. Hence we can conclude that our method presents a similar level of performance than MMDT. Similarly, if we compare our approach using LMNN as the source metric with GFK, we obtain that GFK is significantly better than our approach on 3 tasks, we are significantly better on 7 and 2 lead to a draw. Hence, we can conclude that our approach performs better than GFK.

If we compare the performances of both ITML and LMNN as metrics used directly in a nearest neighbour classifier one can intuitively expect ITML to be a better source hypothesis than LMNN. However, in practice using the metric learned by LMNN as the source hypothesis yields better results. This suggests that using a learned source model that tends to overfit reasonably the learning source sample can be of potential interest in a transfer learning context. Indeed LMNN does not use a regularization term in its formulation and it is well know that LMNN is prone to overfitting. Since, the parameter β penalizes the source metric w.r.t. its complexity it may limit the impact of the source metric to what is needed for the transfer. Nevertheless, this point deserves further investigation.

6. Conclusion

In this paper we presented a new theoretical analysis for metric hypothesis transfer learning. This framework takes into account a source hypothesis information to help learning by means of a biased regularization. This biased regularization can be interpreted into two ways: (i) when the source metric is an a priori known metric such as the identity matrix, the objective is to infer a new metric that performs better than the source metric, (ii) when the source metric has been learned from another domain, the formulation allows one to transfer the knowledge from the source metric to the new domain. This last interpretation refers to a transfer learning setting where the learner does not have access to source examples and can only make use of the source model in the presence of few labelled data.

Our analysis has shown that this framework is theoretically well founded and that a good source hypothesis can facilitate fast generalization in $\mathcal{O}(1/n)$. Moreover, we have provided a consistency analysis from which we have developed a generalization bound able to consider both the performance and the complexity of the source hypothesis. This has led to the use of weighted source hypothesis to optimize the bound in a theoretically sound way.

As stated in (Kuzborskij & Orabona, 2014) in another context, our results stress the importance of choosing good source hypothesis. However, choosing the best source metric from few labelled data is a difficult problem of crucial importance. One perspective could be to consider notions of reverse validations as used in some transfer learning/domain adaptation tasks (Bruzzone & Marconcini, 2010; Zhong et al., 2010). Another perspective would be to extend our framework to other settings and other kind of regularizers.

References

Bellet, Aurélien and Habrard, Amaury. Robustness and Generalization for Metric Learning. *Neurocomputing*, 151(1):259–267, 2015.

Bellet, Aurélien, Habrard, Amaury, and Sebban, Marc. Similarity learning for provably accurate sparse linear classification. In *Proc. of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

Bellet, Aurélien, Habrard, Amaury, and Sebban, Marc. A survey on metric learning for feature vectors and structured data. *CoRR*, abs/1306.6709, 2013.

Ben-David, Shai, Blitzer, John, Crammer, Koby, Kulesza, Alex, Pereira, Fernando, and Vaughan, Jennifer Wortman. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Bohné, Julien, Ying, Yiming, Gentric, Stéphane, and Pontil, Massimiliano. Large margin local metric learning. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proc., Part II*, pp. 679–694, 2014.

Bousquet, Olivier and Elisseeff, André. Stability and generalization. *Journal of Machine Learning Research*, 2: 499–526, 2002.

Bruzzone, Lorenzo and Marconcini, Mattia. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *Transaction Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010.

Cao, Qiong, Guo, Zheng-Chu, and Ying, Yiming. Generalization bounds for metric and similarity learning. *CoRR*, abs/1207.5437, 2012.

Cao, Qiong, Ying, Yiming, and Li, Peng. Similarity metric learning for face recognition. In *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2408–2415, 2013.

Davis, Jason V., Kulis, Brian, Jain, Prateek, Sra, Suvrit, and Dhillon, Inderjit S. Information-theoretic metric learning. In *Machine Learning, Proc. of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pp. 209–216, 2007.

Gong, Boqing, Shi, Yuan, Sha, Fei, and Grauman, Kristen. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 2066–2073, 2012.

Guo, Zheng-Chu and Ying, Yiming. Guaranteed classification via regularized similarity learning. *Neural Computation*, 26(3):497–522, 2014. doi: 10.1162/NECO_a_00556. URL http://dx.doi.org/10.1162/NECO_a_00556.

Hoffman, Judy, Rodner, Erik, Donahue, Jeff, Saenko, Kate, and Darrell, Trevor. Efficient learning of domain-invariant image representations. *CoRR*, abs/1301.3224, 2013.

Jin, Rong, Wang, Shijun, and Zhou, Yang. Regularized distance metric learning: Theory and algorithm. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proc. of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, pp. 862–870, 2009.

Kienzle, Wolf and Chellapilla, Kumar. Personalized handwriting recognition via biased regularization. In *Machine Learning, Proc. of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, pp. 457–464, 2006.

Kulis, Brian. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013.

Kuzborskij, Ilja and Orabona, Francesco. Stability and hypothesis transfer learning. In *Proc. of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 942–950, 2013.

Kuzborskij, Ilja and Orabona, Francesco. Learning by transferring from auxiliary hypotheses. *CoRR*, abs/1412.1619, 2014.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Parameswaran, Shibin and Weinberger, Kilian Q. Large margin multi-task metric learning. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proc. of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pp. 1867–1875, 2010.

Perrot, Michaël, Habrard, Amaury, Muselet, Damien, and Sebban, Marc. Modeling perceptual color differences by local metric learning. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proc., Part V*, pp. 96–111, 2014.

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning - From Theory to Algorithms*, chapter Regularization and Stability, pp. 137–149. Cambridge University Press, 2014.

Weinberger, Kilian Q., Blitzer, John, and Saul, Lawrence K. Distance metric learning for large margin nearest neighbor classification. In *Advances in Neural Information Processing Systems 18 [Neural Information Processing Systems, NIPS 2005, December 5-8, 2005, Vancouver, British Columbia, Canada]*, pp. 1473–1480, 2005.

Xing, Eric P., Ng, Andrew Y., Jordan, Michael I., and Russell, Stuart J. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pp. 505–512, 2002.

Zha, Zheng-Jun, Mei, Tao, Wang, Meng, Wang, Zengfu, and Hua, Xian-Sheng. Robust distance metric learning with auxiliary knowledge. In *IJCAI 2009, Proc. of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pp. 1327–1332, 2009.

Zhong, ErHeng, Fan, Wei, Yang, Qiang, Verscheure, Olivier, and Ren, Jiangtao. Cross validation framework to choose amongst models and datasets for transfer learning. In *Proc. of European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, volume 6323 of *LNCS*, pp. 547–562. Springer, 2010.

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 $\mathbf{M}^{ij^*}$, be the optimal solution when learning with the training set T, respectively T^{ij}. Let $\mathbf{z}_k, \mathbf{z}_k^i, \mathbf{z}_k^{ij}$ 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^{ij}, \mathbf{z}_j^{ij}) - l(\mathbf{M}^{ij^*}, \mathbf{z}_k^{ij}, \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:

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

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.

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

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.