Journal of Machine Learning Research 20 (2019) 1-30 Submitted 5/15; Revised 10/15; Published 1/19

Adaptation Based on Generalized Discrepancy

Corinna Cortes CORINNA@GOOGLE.COM
Google Research, 111 8th ave, New York, NY 10011

Mehryar Mohri MOHRI@CIMS.NYU.EDU
Andrés Muñoz Medina MUNOZ@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012

Editor: Multi Task Learning

Abstract

We present a new algorithm for domain adaptation improving upon a discrepancy minimization algorithm, (DM), previously shown to outperform a number of algorithms for this problem. Unlike many previously proposed solutions for domain adaptation, our algorithm does not consist of a fixed reweighting of the losses over the training sample. Instead, the reweighting depends on the hypothesis sought. The algorithm is derived from a less conservative notion of discrepancy than the DM algorithm called *generalized discrepancy*. We present a detailed description of our algorithm and show that it can be formulated as a convex optimization problem. We also give a detailed theoretical analysis of its learning guarantees which helps us select its parameters. Finally, we report the results of experiments demonstrating that it improves upon discrepancy minimization.

Keywords: domain adaptation, learning theory

1. Introduction

A standard assumption in statistical learning theory and PAC learning is that training and test samples are drawn from the same distribution (Vapnik, 1998; Valiant, 1984). In practice, however, this assumption often does not hold: the source and target distributions may somewhat differ. This problem is known as *domain adaptation* and arises in a variety of applications such as natural language processing and computer vision (Dredze et al., 2007; Blitzer et al., 2007; Jiang and Zhai, 2007; Leggetter and Woodland, 1995; Martínez, 2002; Hoffman et al., 2014). The domain adaptation problem may appear when the distributions over the instance space differ, the so-called *covariate shift* problem, or when the labeling functions associated with each domain disagree. In practice, a combination of both issues occurs and, for adaptation to succeed, the divergence between the two domains needs to be relatively small. This is clear for the labeling functions since, if the learner receives source labels that are vastly different from the target ones, no learning algorithm can generalize well to the target domain. The same holds when input distributions largely differ.

This intuition was formalized by Ben-David et al. (2010) and Ben-David and Urner (2012) who showed that even in the favorable scenario where the source and target distribution admit the same support, a sample of size in the order of that of the support is needed in order to solve the domain adaptation problem. As the authors point out, the domain adaptation problem becomes intractable when the labeling function for the training

data is vastly different from the labeling function used for testing. On the other hand, when some similarity between domains exist, it has been empirically and theoretically shown that adaptation algorithms can be beneficial and in fact a large number of algorithms for this task have been proposed over the past decade. The large majority of them fall in one of the following paradigms:

1. **Learning a new feature representation**. The core idea behind these algorithms is to map the source and target data into a new feature space where the difference between source and target distributions is reduced. Transfer Component Analysis (TCA) (Pan et al., 2011) and the work on Frustratingly Easy Domain Adaptation (FE) (Daumé III, 2007) belong to this family of algorithms. Whereas some empirical evidence of the effectiveness of these algorithms exists in the literature, to the best of our knowledge, no work has been done to provide learning guarantees for these algorithms.
2. **Reweighting**. Originated in the Statistics literature on sample bias correction, these techniques attempt to correct the difference between distributions by multiplying the loss at each training example by a positive weight. Most of the classical algorithms such as KMM (Huang et al., 2006), KLIEP (Sugiyama et al., 2007) and a two-step algorithm by Bickel et al. (2007) fall in this category.

The main focus of this work will be on the latter. A common trait shared by most algorithms in this category is that their reweighting schemes are based on the minimization of a divergence measure between the empirical source and target distributions. For instance, the KL-divergence in the case of KLIEP and the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) for KMM. The guarantees of these algorithms are therefore given as a function of the chosen divergence. The main drawback of these measures is that they do not take into account the hypothesis set or the loss function, both crucial components of any learning algorithm. In contrast, the *discrepancy* introduced by Mansour et al. (2009) and further studied by Cortes and Mohri (2011) is a measure of the divergence between distributions tailored to domain adaptation that precisely takes into account both the loss function and the hypothesis set. The $d_{\mathcal{A}}$-distance, introduced by Devroye et al. (1996)[pp. 271-272] under the name of *generalized Kolmogorov-Smirnov distance*, later by Ben-David et al. (2006), coincides with the discrepancy when the binary loss function is used. The discrepancy is a pivotal concept used in the analysis of several adaptation scenarios: the $\mathcal{Y}$-discrepancy or *integral probability metric* (Zhang et al., 2012) was successfully used by Mohri and Muñoz (2012) to provide tight learning guarantees for the related task of learning with drifting distributions, whereas a modified version of the discrepancy was used by Germain et al. (2013) to study the problem of domain adaptation in a PAC-Bayesian setting. The discrepancy-based generalization bounds given by Mansour et al. (2009) motivated a discrepancy minimization (DM) algorithm (Cortes and Mohri, 2013), which attempts to minimize said bounds. Besides its favorable theoretical guarantees, this algorithm was shown to perform well in a number of adaptation tasks and to match or outperform several other algorithms such as KMM, KLIEP and the aforementioned two stage algorithm by Bickel et al. (2007).

One shortcoming of the DM algorithm, however, is that it seeks to reweight the loss on the training samples to minimize a quantity defined as the maximum over *all* pairs of hypotheses, including hypotheses that the learning algorithm might not ever consider as candidates. Thus, the algorithm tends to be too conservative on its choice of weights. We

present an alternative theoretically well founded algorithm for domain adaptation that is based on minimizing a finer quantity, the *generalized discrepancy*, and that seeks to improve upon DM. Unlike the DM algorithm, our algorithm does not consist of a *fixed* reweighting of the losses over the training sample. Instead, the weights assigned to training sample losses vary as a function of the hypothesis h. This helps us ensure that for every hypothesis, h, the empirical loss on the source distribution is as close as possible to the empirical loss on the target distribution for that particular h.

We describe the learning scenario considered (Section 2), then present a detailed description of our algorithm and show that it can be formulated as a convex optimization problem (Section 3). Next, we analyze the theoretical properties of our algorithm, which guide us in choosing the surrogate hypothesis set defining our algorithm (Section 4). In Section 5, we further analyze the optimization problem defining our algorithm and derive an equivalent form that can be handled by a standard convex optimization solver. In Section 6, we report the results of experiments demonstrating that our algorithm improves upon the DM algorithm in several tasks.

2. Learning Scenario

This section defines the learning scenario of domain adaptation we consider, which coincides with that of Ben-David et al. (2006) or Mansour et al. (2009); Cortes and Mohri (2013). We first introduce the definitions and concepts needed for the following sections. For the most part, we follow the definitions and notation of Cortes and Mohri (2013).

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} \subseteq \mathbb{R}$ the output space. We define a *domain* as a pair formed by a distribution over $\mathcal{X}$ and a target labeling function mapping from $\mathcal{X}$ to $\mathcal{Y}$. Throughout the paper, (Q, f_Q) denotes the *source domain* and (P, f_P) the *target domain* with Q the source and P the target distribution over $\mathcal{X}$ and with $f_Q, f_P \colon \mathcal{X} \to \mathcal{Y}$ the source and target labeling functions, respectively.

In the scenario of *domain adaptation* we consider, the learner receives two samples: a labeled sample of m points from the source domain $\mathcal{S} = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$ with $x_1, \ldots, x_m$ drawn i.i.d. according to Q and $y_i = f_Q(x_i)$ for $i \in [1, m]$; and an unlabeled sample $\mathcal{T} = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$ of size n drawn i.i.d. according to the target distribution P. We denote by $\widehat{Q}$ the empirical distribution corresponding to the (unlabeled) sample $\mathcal{S}_{\mathcal{X}} = (x_1, \ldots, x_m)$ and by $\widehat{P}$ the empirical distribution corresponding to $\mathcal{T}$. We will be in fact more interested in the scenario commonly encountered in practice where, in addition to these two samples, the learner receives a small amount of labeled data $\mathcal{T}' = ((x''_1, y''_1), \ldots, (x''_s, y''_s)) \in (\mathcal{X} \times \mathcal{Y})^s$ from the target domain.

We consider a loss function $L \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ jointly convex in its two arguments. The L_p losses commonly used in regression and defined by $L_p(y, y') = |y' - y|^p$ for $p \geq 1$ are special instances of this definition. For any two functions $h, h' \colon \mathcal{X} \to \mathcal{Y}$ and any distribution D over $\mathcal{X}$, we denote by $\mathcal{L}_D(h, h')$ the expected loss of $h(x)$ and $h'(x)$: $\mathcal{L}_D(h, h') = \mathrm{E}_{x \sim D}[L(h(x), h'(x))]$. The learning problem consists of selecting a hypothesis h out of a hypothesis set H with a small expected loss $\mathcal{L}_P(h, f_P)$ with respect to the target domain. We further extend this notation to arbitrary functions $\mathsf{q} \colon \mathcal{X} \to \mathbb{R}$ with a finite support as follows: $\mathcal{L}_{\mathsf{q}}(h, h') = \sum_{x \in \mathcal{X}} q(x) L(h(x), h'(x))$.

$\mathcal{X}$	Input space	$\mathcal{Y}$	Output space
P	Target distribution	Q	Source distribution
$\widehat{P}$	Empirical target distribution	$\widehat{Q}$	Empirical source distribution
$\mathcal{T}$	Target unlabeled sample	$\mathcal{S}$	Labeled source sample
$\mathcal{T}'$	Small target labeled sample	$\mathcal{S}_{\mathcal{X}}$	Unlabeled source sample
f_P	Target labeling function	f_Q	Source labeling function
$\mathcal{L}_P(h,f_P)$	Expected target loss	$\mathcal{L}_Q(h,f_Q)$	Expected source loss
$\mathcal{L}_{\widehat{P}}(h,f_P)$	Empirical target loss	$\mathcal{L}_{\widehat{Q}}(h,f_Q)$	Empirical source loss
$\mathrm{disc}(P,Q)$	Discrepancy	$\mathrm{DISC}(\widehat{P},\mathsf{U})$	Generalized discrepancy
$\mathrm{disc}_{H''}(P,Q)$	Local Discrepancy	$\mathrm{disc}_{\mathcal{Y}}(P,Q)$	$\mathcal{Y}$-discrepancy
$\mathsf{q}_{\min}$	DM solution	Q_h	GDM solution

Table 1: Notation table.

3. Algorithm

In this section, we describe our new adaptation algorithm. We first review some related previous work. Next, we present the key idea behind our algorithm and derive its general form, and finally, formulate it as a convex optimization problem.

3.1. Previous Work

It was shown by Mansour et al. (2009) and Cortes and Mohri (2011) (see also the $d_{\mathcal{A}}$-*distance* (Ben-David et al., 2006) in the case of binary loss for classification) that a key measure of the difference of two distributions in the context of adaptation is the *discrepancy*. Given a hypothesis set H, the discrepancy, disc, between two distributions P and Q over $\mathcal{X}$ is defined by:

$$\mathrm{disc}(P,Q) = \max_{h,h' \in H} \left| \mathcal{L}_P(h',h) - \mathcal{L}_Q(h',h) \right|. \tag{1}$$

The discrepancy has several advantages over other common divergence measures such as the L_1 distance. We refer the reader to (Medina, 2015) for a detailed discussion on this subject. Several generalization bounds for adaptation in terms of the discrepancy have been given in the past (Ben-David et al., 2006; Mansour et al., 2009; Cortes and Mohri, 2011, 2013). including pointwise guarantees in the case of kernel-based regularization algorithms, which includes algorithms such as support vector machines (SVM), kernel ridge regression, or support vector regression (SVR). The bounds given in (Mansour et al., 2009) motivated a *discrepancy minimization* algorithm. Given a positive semi-definite (PSD) kernel K, the hypothesis returned by the algorithm is the solution of the following optimization problem

$$\min_{h \in \mathbb{H}} \quad \lambda \|h\|_K^2 + \mathcal{L}_{\mathsf{q}_{\min}}(h, f_Q), \tag{2}$$

where $\|\cdot\|_K$ is the norm in the reproducing Hilbert space $\mathbb{H}$ induced by the kernel K and $\mathsf{q}_{\min}$ is a distribution over the support of $\widehat{Q}$ such that $\mathsf{q}_{\min} = \mathrm{argmin}_{\mathsf{q} \in \mathcal{Q}} \mathrm{disc}(\mathsf{q}, \widehat{P})$, where $\mathcal{Q} = [0,1]^{\mathcal{S}_{\mathcal{X}}}$ is the set of all distributions defined over the support of $\widehat{Q}$. Besides its theoretical motivation, this algorithm has been shown to outperform several other algorithms in a series of experiments carried out by Cortes and Mohri (2013).

Observe that, by definition, the objective function optimized by $\mathsf{q}_{\min}$ corresponds to a maximum over all pairs of hypotheses. But, the maximizing pair of hypotheses may not be among the candidates ever considered by the learning algorithm. Thus, a learning algorithm based on discrepancy minimization tends to be too conservative.

3.2. Main Idea

From here on we assume the algorithm selected by the learner is an instance of a regularized risk minimization algorithm over the Hilbert space $\mathbb{H}$ induced by a PSD kernel K. With knowledge of the target labels, these algorithms return a hypothesis h^* solution of $\min_{h\in\mathbb{H}} F(h)$ where

$$F(h) = \lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P), \tag{3}$$

where $\lambda \geq 0$ is a regularization parameter. Thus, h^* can be viewed as the *ideal hypothesis*.

In view of that, we can formulate our objective, in the *presence* of a domain adaptation problem, as that of finding a hypothesis h whose loss $\mathcal{L}_P(h, f_P)$ with respect to the target domain is as close as possible to $\mathcal{L}_P(h^*, f_P)$. To do so, we will seek in fact a hypothesis h that is as close as possible to h^*, which would imply the closeness of the losses with respect to the target domains. We do not have access to f_P and can only access the labels of the training sample $\mathcal{S}$. Thus, we must resort to using in our objective function, instead of $\mathcal{L}_{\widehat{P}}(h, f_P)$, a reweighted empirical loss over the training sample $\mathcal{S}$. The main idea behind our algorithm is to define, for any $h \in \mathbb{H}$, a reweighting function $\mathsf{Q}_h \colon \mathcal{S}_{\mathcal{X}} = \{x_1, \ldots, x_m\} \to \mathbb{R}$ such that the objective function G defined for all $h \in \mathbb{H}$ by

$$G(h) = \lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{Q}_h}(h, f_Q) \tag{4}$$

is uniformly close to F, thereby resulting in close minimizers. Since the first term of (3) and (4) coincide, the idea consists equivalently of seeking Q_h such that $\mathcal{L}_{\mathsf{Q}_h}(h, f_Q)$ and $\mathcal{L}_{\widehat{P}}(h, f_P)$ be as close as possible. Observe that this departs from the standard reweighting methods: instead of reweighting the training sample with some fixed set of weights, we allow the weights to vary as a function of the hypothesis h. Note that we have further relaxed the condition commonly adopted by reweighting techniques that the weights must be non-negative and sum to one.

Of course, searching for Q_h to directly minimize $|\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)|$ is in general not possible since we do not have access to f_P, but it is instructive to consider the imaginary case where the average loss $\mathcal{L}_{\widehat{P}}(h, f_P)$ is known to us for any $h \in \mathbb{H}$. Q_h could then be determined via

$$\mathsf{Q}_h = \operatorname*{argmin}_{\mathsf{q}\in\mathcal{F}(\mathcal{S}_{\mathcal{X}},\mathbb{R})} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)|, \tag{5}$$

where $\mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})$ is the set of real-valued functions defined over $\mathcal{S}_{\mathcal{X}}$. For any h, we can in fact select Q_h such that $\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \mathcal{L}_{\widehat{P}}(h, f_P)$ since $\mathcal{L}_{\mathsf{q}}(h, f_Q)$ is a linear function of q. Thus, the optimization problem (5) reduces to solving a simple linear equation. With this choice of Q_h, the objective functions F and G coincide and by minimizing G we can recover the ideal solution h^*. Note that, in general, the DM algorithm could not recover that ideal solution. Even a finer discrepancy minimization algorithm exploiting the knowledge of $\mathcal{L}_{\widehat{P}}(h, f_P)$ for all h and seeking a distribution $\mathsf{q}'_{\min}$ minimizing $\max_{h\in H} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)|$ could not,

in general, recover the ideal solution since we could not have $\mathcal{L}_{\mathsf{q}'_{\min}}(h, f_Q) = \mathcal{L}_{\widehat{P}}(h, f_P)$ for all $h \in \mathbb{H}$.

Of course, $\mathcal{L}_{\widehat{P}}(h, f_P)$ is not accessible since the sample $\mathcal{T}$ is unlabeled. Instead, we will consider a non-empty convex set of candidate hypotheses $H'' \subseteq H$ that could contain a good approximation of f_P. Using H'' as a set of surrogate labeling functions leads to the following definition of Q_h instead of (5):

$$\mathsf{Q}_h = \operatorname*{argmin}_{\mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})} \max_{h'' \in H''} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, h'')|. \tag{6}$$

The choice of the subset H'' is of course key. Our choice will be based on the theoretical analysis of Section 4. Nevertheless, we now present the formulation of the optimization problem for an arbitrary choice of the convex subset H''.

Proposition 1 *For any $h \in \mathbb{H}$, let Q_h be defined by (6). Then, the following identity holds for any $h \in \mathbb{H}$:*

$$\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) = \frac{1}{2}\Big(\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big).$$

Proof For any $h \in \mathbb{H}$, the equation $\mathcal{L}_{\mathsf{q}}(h, f_Q) = l$ with $l \in \mathbb{R}$ admits a solution $\mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})$. Thus, $\{\mathcal{L}_{\mathsf{q}}(h, f_Q) : \mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})\} = \mathbb{R}$ and for any $h \in \mathbb{H}$, we can write

$$\begin{aligned}
\mathcal{L}_{\mathsf{Q}_h}(h, f_Q) &= \operatorname*{argmin}_{l \in \{\mathcal{L}_{\mathsf{q}}(h, f_Q) \colon \mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})\}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')| \\
&= \operatorname*{argmin}_{l \in \mathbb{R}} \max_{h'' \in H''} |l - \mathcal{L}_{\widehat{P}}(h, h'')| \\
&= \operatorname*{argmin}_{l \in \mathbb{R}} \max_{h'' \in H''} \max\Big\{\mathcal{L}_{\widehat{P}}(h, h'') - l, l - \mathcal{L}_{\widehat{P}}(h, h'')\Big\} \\
&= \operatorname*{argmin}_{l \in \mathbb{R}} \max\Big\{ \max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') - l, l - \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big\} \\
&= \frac{1}{2}\Big(\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big),
\end{aligned}$$

since the minimizing l is obtained for $\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') - l = l - \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')$. ■

In view of this proposition, with our choice of Q_h based on (6), the objective function G of our algorithm (4) can be equivalently written for all $h \in \mathbb{H}$ as follows:

$$G(h) = \lambda \|h\|_K^2 + \frac{1}{2}\Big(\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'') + \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')\Big). \tag{7}$$

Using the fact the $\mathcal{L}_{\widehat{P}}$ is a jointly convex function, it is easy to show (see for instance Boyd and Vandenberghe, 2004) that G is in fact a convex function too.

4. Learning Guarantees

Here, we present two different types of guarantees: a tight learning bound based on the Rademacher complexity and a pointwise bound derived from a stability analysis. We further show that our algorithm is in fact minimizing this pointwise bound. As in previous work, we assume that the loss function L is μ-*admissible*.

Definition 2 *A loss function L is μ-admissible if there exists $\mu > 0$ such that the inequality*

$$|L(h(x), y) - L(h'(x), y)| \leq \mu |h(x) - h'(x)| \tag{8}$$

holds for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $h', h \in H$.

The L_p losses commonly used in regression, $p \geq 1$, verify this condition (see Appendix C).

4.1. Rademacher Complexity Bounds

Definition 3 *Let $\mathcal{Z}$ be any set and G be a family of functions mapping $\mathcal{Z}$ to $\mathbb{R}$. Given a sample $\mathcal{S} = \{z_1, \ldots, z_n\} \subset \mathcal{Z}$, the empirical Rademacher complexity of G is denoted by $\widehat{\mathfrak{R}}_{\mathcal{S}}(G)$ and defined by*

$$\widehat{\mathfrak{R}}_{\mathcal{S}}(G) = \frac{1}{n} \mathop{\mathrm{E}}_{\sigma} \left[\sup_{g \in G} \sum_{i=1}^{n} \sigma_i g(z_i) \right],$$

where σ_is, called Rademacher variables, are independent random variables distributed according to the uniform distribution over $\{-1, 1\}$. The Rademacher complexity of G is defined as

$$\mathfrak{R}_n(G) = \mathop{\mathrm{E}}_{\mathcal{S}} \left[\widehat{\mathfrak{R}}_{\mathcal{S}}(G) \right].$$

Our first generalization bound is given in terms of the *$\mathcal{Y}$-discrepancy*, which is a generalization of the discrepancy distance. The $\mathcal{Y}$-discrepancy was first introduced by Mohri and Muñoz (2012) in the context of learning with drifting distributions.

Definition 4 *The $\mathcal{Y}$-discrepancy between two domains (P, f_P) and (Q, f_Q) is defined by*

$$\mathrm{disc}_{\mathcal{Y}}(P, Q) = \sup_{h \in H} \left| \mathcal{L}_Q(h, f_Q) - \mathcal{L}_P(h, f_P) \right|.$$

Note that the definition depends on the labeling functions f_P and f_Q. We do not explicitly indicate that dependency for the sake of simplicity of the notation.

We follow the analysis of (Mohri and Muñoz, 2012) to derive the following tight generalization bounds based on the notion of $\mathcal{Y}$-discrepancy.

Proposition 5 *Let $\mathcal{H}_Q$ and $\mathcal{H}_P$ be the families of functions defined as follows: $\mathcal{H}_Q := \{x \mapsto L(h(x), f_Q(x)) \colon h \in H\}$ and $\mathcal{H}_P := \{x \mapsto L(h(x), f_P(x)) \colon h \in H\}$. Define M_Q and M_P as $M_Q = \sup_{x \in \mathcal{X}, h \in H} L(h(x), f_Q(x))$ and $M_P = \sup_{x \in \mathcal{X}, h \in H} L(h(x), f_P(x))$. Then, for any $\delta > 0$,*

1. *with probability at least $1 - \delta$ over the choice of a labeled sample $\mathcal{S}$ of size m, the following inequality holds for all $h \in H$:*

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_{\widehat{Q}}(h, f_Q) + \mathrm{disc}_{\mathcal{Y}}(P, Q) + 2\mathfrak{R}_m(\mathcal{H}_Q) + M_Q \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}; \tag{9}$$

2. *with probability at least $1 - \delta$ over the choice of a sample $\mathcal{T}$ of size n, the following inequality holds for all $h \in H$ and any distribution q over a sample $\mathcal{S}_{\mathcal{X}}$:*

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_{\mathsf{q}}(h, f_Q) + \mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{10}$$

Proof Let $\Phi(\mathcal{S})$ denote $\sup_{h\in H} \mathcal{L}_{\widehat{Q}}(h, f_Q) - \mathcal{L}_P(h, f_P)$. Changing one point in $\mathcal{S}$ changes $\Phi(\mathcal{S})$ by at most $\frac{M_Q}{m}$. Thus, by McDiarmid's inequality, we have $\mathbb{P}\big(\Phi(\mathcal{S}) - \mathrm{E}[\Phi(\mathcal{S})] > \epsilon\big) \leq e^{-\frac{2m\epsilon^2}{M_Q^2}}$. Therefore, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in H$:

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_{\widehat{Q}}(h, f_Q) + \mathrm{E}[\Phi(\mathcal{S})] + M_Q \sqrt{\frac{\log(\frac{1}{\delta})}{2m}}.$$

Next, we can bound $\mathrm{E}[\Phi(\mathcal{S})]$ as follows:

$$\begin{aligned}
\mathrm{E}[\Phi(\mathcal{S})] &= \mathrm{E}\left[\sup_{h\in H} \mathcal{L}_{\widehat{Q}}(h, f_Q) - \mathcal{L}_P(h, f_P)\right] \\
&\leq \mathrm{E}\left[\sup_{h\in H} \mathcal{L}_{\widehat{Q}}(h, f_Q) - \mathcal{L}_Q(h, f_Q)\right] + \sup_{h\in H} \mathcal{L}_Q(h, f_Q) - \mathcal{L}_P(h, f_P) \\
&\leq 2\mathfrak{R}_m(\mathcal{H}_Q) + \mathrm{disc}_{\mathcal{Y}}(P, Q),
\end{aligned}$$

where the last inequality follows from a standard symmetrization inequality in terms of the Rademacher complexity and the definition of $\mathrm{disc}_{\mathcal{Y}}(P, Q)$.

For the second bound we have, starting with a standard Rademacher complexity bound for $\mathcal{H}_P$, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in H$:

$$\begin{aligned}
\mathcal{L}_P(h, f_P) &\leq \mathcal{L}_{\widehat{P}}(h, f_P) + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}} \\
&\leq \mathcal{L}_{\mathsf{q}}(h, f_Q) + \mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{q}}(h, f_Q) + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \qquad (11)
\end{aligned}$$

Moreover, by definition $\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{q}}(h, f_Q) \leq \mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q})$ for any q. Replacing this bound in (11) yields the result. ■

Observe that these bounds are tight as a function of the divergence measure (discrepancy) we use: in the absence of adaptation, the following standard Rademacher complexity learning bound holds:

$$\mathcal{L}_{\widehat{P}}(h, f_P) \leq \mathcal{L}_{\widehat{P}}(h, f_P) + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$

Our second adaptation bound differs from this inequality only by the fact that $\mathcal{L}_{\widehat{P}}(h, f_P)$ is replaced with $\mathcal{L}_{\mathsf{q}}(h, f_Q) + \mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q})$. But, by definition of $\mathcal{Y}$-discrepancy, there exists an $h \in H$ such that $|\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{q}}(h, f_Q)| = \mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q})$. A similar analysis shows that our first bound is also tight.

Given a labeled sample $\mathcal{S}$ from the source domain, Proposition 5 suggests choosing a distribution q with support $\mathcal{S}_{\mathcal{X}}$ that minimizes the right-hand side of (10). However, the quantity $\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q})$ depends, by definition, on the unknown labels from the target domain and therefore cannot be minimized. Thus, we will instead upper bound the $\mathcal{Y}$-discrepancy in terms of quantities that can be estimated.

Let $\mathcal{A}(H)$ denote the set of all functions $\mathsf{U}\colon h \mapsto \mathsf{U}_h$ mapping H to $\mathcal{F}(\mathcal{S}_{\mathcal{X}}, \mathbb{R})$ such that for all $h \in H$, $h \mapsto \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$ is a convex function. Thus, for any $h \in H$, U_h is a reweighting function defined over $\mathcal{S}_{\mathcal{X}}$. $\mathcal{A}(H)$ contains all constant functions U such that $\mathsf{U}_h = \mathsf{q}$ for all $h \in H$, where q is a distribution over $\mathcal{S}_{\mathcal{X}}$. We will abuse the notation and denote this functions also by q. By Proposition 1, $\mathcal{A}(H)$ also includes the function $\mathsf{Q}\colon h \to \mathsf{Q}_h$ used by our algorithm.

Definition 6 (Generalized discrepancy) *For any $\mathsf{U} \in \mathcal{A}(H)$, the* generalized discrepancy between $\widehat{P}$ and U *is denoted by* $\mathrm{DISC}(\widehat{P}, \mathsf{U})$ *and is defined by*

$$\mathrm{DISC}(\widehat{P}, \mathsf{U}) = \sup_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|. \tag{12}$$

We also denote by $d_1(f_P, H'')$ the L_1 distance of f_P to H'':

$$d_1(f_P, H'') = \min_{h_0 \in H''} \mathop{\mathrm{E}}_{\widehat{P}} |h_0(x) - f_P(x)|. \tag{13}$$

The following theorem gives an upper bound on the $\mathcal{Y}$-discrepancy in terms of the generalized discrepancy and $d_1(f_P, H'')$.

Proposition 7 *For any distribution q over $\mathcal{S}_{\mathcal{X}}$ and any set H'', the following inequality holds:*

$$\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) \leq \mathrm{DISC}(\widehat{P}, \mathsf{q}) + \mu\, d_1(f_P, H'').$$

Proof Let $h_0 \in H''$, by the triangle inequality, we can write

$$\begin{aligned}
\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) &= \sup_{h \in H} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)| \\
&\leq \sup_{h \in H} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, h_0)| + \sup_{h \in H} |\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{P}}(h, f_P)| \\
&\leq \sup_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, h'')| + \sup_{h \in H} |\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{P}}(h, f_P)|.
\end{aligned}$$

The hypothesis h_0 will later be chosen to minimize the distance of f_P to H''. By the μ-admissibility of the loss, the last term can be bounded as follows:

$$\sup_{h \in H} |\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{P}}(h, f_P)| \leq \mu \mathop{\mathrm{E}}_{\widehat{P}} |f_P(x) - h_0(x)|.$$

Using this inequality and minimizing over $h_0 \in H''$ yields:

$$\begin{aligned}
\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) &\leq \sup_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, h'')| + \mu\, d_1(f_P, H'') \\
&= \mathrm{DISC}(\widehat{P}, \mathsf{q}) + \mu\, d_1(f_P, H''),
\end{aligned}$$

which completes the proof. ∎

We can also bound the $\mathcal{Y}$-discrepancy in terms of the discrepancy measure and the following measure of the difference of the source and target labeling functions:

$$\eta_H(f_P, f_Q) = \min_{h_0 \in H} \Big(\max_{x \in \mathrm{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{x \in \mathrm{supp}(\widehat{Q})} |f_Q(x) - h_0(x)| \Big).$$

Proposition 8 *The following inequality holds for all distributions* q *over* $\mathcal{S}_{\mathcal{X}}$*:*

$$\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) \leq \mathrm{disc}(\widehat{P}, \mathsf{q}) + \mu\, \eta_H(f_P, f_Q).$$

Proof By the triangle inequality and the μ-admissibility of the loss, the following inequality holds for all $h_0 \in H$:

$$\begin{aligned}
&\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) \\
&= \sup_{h \in H} |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\widehat{P}}(h, f_P)| \\
&\leq \sup_{h \in H} \Big(|\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\widehat{P}}(h, f_P)| + |\mathcal{L}_{\mathsf{q}}(h, f_Q) - \mathcal{L}_{\mathsf{q}}(h, h_0)| \Big) + \sup_{h \in H} |\mathcal{L}_{\mathsf{q}}(h, h_0) - \mathcal{L}_{\widehat{P}}(h, h_0)| \\
&\leq \mu \Big(\sup_{x \in \mathrm{supp}(\widehat{P})} |h_0(x) - f_P(x)|] + \sup_{x \in \mathrm{supp}(\widehat{Q})} [|f_Q(x) - h_0(x)|] \Big) + \mathrm{disc}(\widehat{P}, \mathsf{q}).
\end{aligned}$$

Minimizing over all $h_0 \in H$ gives $\mathrm{disc}_{\mathcal{Y}}(\widehat{P}, \mathsf{q}) \leq \mu\, \eta_H(f_P, f_Q) + \mathrm{disc}(\widehat{P}, \mathsf{q})$ and completes the proof. ■

The following learning guarantees are immediate consequences of Propositions 5, 7 and 8.

Corollary 9 *Let* $H'' \subset H$ *be a convex set and* q *a distribution over* $\mathcal{S}_{\mathcal{X}}$*. Then, for any* $\delta > 0$*, each of the following inequalities holds with probability at least* $1 - \delta$ *for all* $h \in H$*:*

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_{\mathsf{q}}(h, f_Q) + \mathrm{DISC}(\widehat{P}, \mathsf{q}) + \mu\, d_1(f_P, H'') + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}, \tag{14}$$

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_{\mathsf{q}}(h, f_Q) + \mathrm{disc}(\widehat{P}, \mathsf{q}) + \mu\, \eta_H(f_P, f_Q) + 2\mathfrak{R}_n(\mathcal{H}_P) + M_P \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}. \tag{15}$$

In general, the bounds (14) and (15) are not comparable. However, when L is an L_P loss for some $p \geq 1$, we can show the existence of a set H'' for which (14) is a tighter bound than (15). The result is expressed in terms of the *local discrepancy* defined by:

$$\mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) = \sup_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, h'')|,$$

which is a finer measure than the standard discrepancy for which the supremum is defined over a pair of hypotheses *both* in $H \supseteq H''$.

Theorem 10 *Let* L *be the* L_P *loss for some* $p \geq 1$*. Let* $\mathcal{H} := \{B(r) \colon r \geq 0\}$ *be a set of all balls* $B(r) = \{h'' \in H | \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^p\}$*. Then, for any distribution* q *over* $\mathcal{S}_{\mathcal{X}}$*, there exists* $H'' \in \mathcal{H}$ *such that the following holds:*

$$\mathrm{DISC}(\widehat{P}, \mathsf{q}) + \mu\, d_1(f_P, H'') \leq \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) + \mu\, \eta_H(f_P, f_Q).$$

Proof Fix a distribution q over $\mathcal{S}_{\mathcal{X}}$. Let h_0^* be an element of $\operatorname{argmin}_{h_0 \in H} \left(\mathcal{L}_{\widehat{P}}(h_0, f_P)^{\frac{1}{p}} + \mathcal{L}_{\mathsf{q}}(h_0, f_Q)^{\frac{1}{p}}\right)$. Choose $H'' \in \mathcal{H}$ as $H'' = \{h'' \in H | \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^p\}$ with $r = \mathcal{L}_{\mathsf{q}}(h_0^*, f_Q)^{\frac{1}{p}}$. Then, by definition, h_0^* is in H''. For the L_p loss, it is not hard to show that for all $h, h'' \in H$, $|\mathcal{L}_{\mathsf{q}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \leq \mu[\mathcal{L}_{\mathsf{q}}(h'', f_Q)]^{\frac{1}{p}}$ (see Appendix C). In view of this inequality, we can write:

$$\begin{aligned}
\mathrm{DISC}(\widehat{P}, \mathsf{q}) &= \sup_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \\
&\leq \sup_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, h'')| + \sup_{h \in H, h'' \in H''} |\mathcal{L}_{\mathsf{q}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \\
&\leq \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) + \max_{h'' \in H''} \mu[\mathcal{L}_{\mathsf{q}}(h'', f_Q)]^{\frac{1}{p}} \\
&= \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) + \mu r = \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) + \mu \mathcal{L}_{\mathsf{q}}(h_0^*, f_Q)^{\frac{1}{p}}.
\end{aligned}$$

Using this inequality, Jensen's inequality, and the fact that h_0^* is in H'', we can write

$$\begin{aligned}
&\mu\, d_1(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{q}) \\
&\leq \mu \min_{h_0 \in H''} \mathop{\mathrm{E}}_{x \in \widehat{P}}[|f_P(x) - h_0(x)|] + \mu \mathcal{L}_{\mathsf{q}}(h_0^*, f_Q)^{\frac{1}{p}} + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) \\
&\leq \mu \min_{h_0 \in H''} \mathop{\mathrm{E}}_{x \in \widehat{P}}[|f_P(x) - h_0(x)|^p]^{\frac{1}{p}} + \mu \mathcal{L}_{\mathsf{q}}(h_0^*, f_Q)^{\frac{1}{p}} + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) \\
&\leq \mu\, \mathcal{L}_{\widehat{P}}(h_0^*, f_P)^{\frac{1}{p}} + \mu \mathcal{L}_{\mathsf{q}}(h_0^*, f_Q)^{\frac{1}{p}} + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}).
\end{aligned}$$

Moreover, by definition of h_0^* the last expression is equal to

$$\begin{aligned}
&\mu \min_{h_0 \in H} \left(\mathcal{L}_{\widehat{P}}(h_0, f_P)^{\frac{1}{p}} + \mathcal{L}_{\mathsf{q}}(h_0, f_Q)^{\frac{1}{p}}\right) + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) \\
&\qquad \leq \mu \min_{h_0 \in H} \big(\max_{x \in \mathrm{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{x \in \mathrm{supp}(\widehat{Q})} |f_Q(x) - h_0(x)| \big) + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) \\
&\qquad = \mu\, \eta_H(f_P, f_Q) + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}).
\end{aligned}$$

which concludes the proof. ■

Theorem 10 shows that the generalized discrepancy can provide a finer measure of the difference between two domains for some choices of H''. Therefore, for a good choice of H'', an algorithm minimizing the right-hand side of (14) would benefit from better theoretical guarantees than the DM algorithm. However, the optimization problem defined by (14) is not jointly convex in q and h. Instead, we propose to first minimize the generalized discrepancy and then use this reweighting function as input to our learning algorithm. Further motivation for this two-stage algorithm is given in the following section.

4.2. Pointwise Guarantees

Similar to the guarantee presented by Cortes and Mohri (2013), we will seek to bound the difference between an *ideal solution* h^* and the solution obtained by our algorithm. We begin by stating the following bound motivating the DM algorithm.

Theorem 11 (Cortes and Mohri, 2013) *Let* q *be an arbitrary distribution over* $\mathcal{S}_{\mathcal{X}}$ *and let* h^* *and* h_{q} *be the hypotheses minimizing* $\lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ *and* $\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{q}}(h, f_Q)$ *respectively. Then, the following inequality holds:*

$$\lambda\|h^* - h_{\mathsf{q}}\|_K^2 \leq \mu\, \eta_H(f_P, f_Q) + \mathrm{disc}(\widehat{P}, \mathsf{q}). \tag{16}$$

Notice that the solution of DM minimizes the right-hand side of (16), that is $\mathrm{disc}(\widehat{P}, \mathsf{q})$. The following theorem provides an analogous bound for our algorithm.

Theorem 12 *Let* U *be an arbitrary element of* $\mathcal{A}(H)$ *and let* h^* *and* h_{U} *be the hypotheses minimizing* $\lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ *and* $\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$ *respectively. Then, the following inequality holds for any convex set* $H'' \subseteq H$*:*

$$\lambda\|h^* - h_{\mathsf{U}}\|_K^2 \leq \mu\, d_1(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{U}). \tag{17}$$

Proof Fix $\mathsf{U} \in \mathcal{A}(H)$ and let $G_{\widehat{P}}$ denote $h \mapsto \mathcal{L}_{\widehat{P}}(h, f_P)$ and G_{U} the function $h \mapsto \mathcal{L}_{\mathsf{U}_h}(h, f_Q)$. Since $h \mapsto \lambda\|h\|_K^2 + G_{\widehat{P}}(h)$ is convex and differentiable and since h^* is its minimizer, the gradient is zero at h^*, that is $2\lambda h^* = -\nabla G_{\widehat{P}}(h^*)$. Similarly, since $h \mapsto \lambda\|h\|_K^2 + G_{\mathsf{U}}(h)$ is convex, it admits a sub-differential at any $h \in \mathbb{H}$. Since h_{U} is a minimizer, its sub-differential at h_{U} must contain 0. Thus, there exists a sub-gradient $g_0 \in \partial G_{\mathsf{U}}(h_{\mathsf{U}})$ such that $2\lambda h_{\mathsf{U}} = -g_0$, where $\partial G_{\mathsf{U}}(h_{\mathsf{U}})$ denotes the sub-differential of G_{U} at h_{U}. Using these two equalities we can write

$$\begin{aligned}
2\lambda\|h^* - h_{\mathsf{U}}\|_K^2 &= \langle h^* - h_{\mathsf{U}}, g_0 - \nabla G_{\widehat{P}}(h^*)\rangle = \langle g_0, h^* - h_{\mathsf{U}}\rangle - \langle \nabla G_{\widehat{P}}(h^*), h^* - h_{\mathsf{U}}\rangle \\
&\leq G_{\mathsf{U}}(h^*) - G_{\mathsf{U}}(h_{\mathsf{U}}) + G_{\widehat{P}}(h_{\mathsf{U}}) - G_{\widehat{P}}(h^*) \\
&= \mathcal{L}_{\widehat{P}}(h_{\mathsf{U}}, f_P) - \mathcal{L}_{\mathsf{U}_h}(h_{\mathsf{U}}, f_Q) + \mathcal{L}_{\mathsf{U}_h}(h^*, f_Q) - \mathcal{L}_{\widehat{P}}(h^*, f_P) \\
&\leq 2 \sup_{h \in H} |\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|,
\end{aligned}$$

where we used for the first inequality the convexity of G_{U} combined with the sub-gradient property of $g_0 \in \partial G_{\mathsf{U}}(h_{\mathsf{U}})$, and the convexity of $G_{\widehat{P}}$. For any $h \in H$, using the μ-admissibility of the loss, we can upper bound the operand of the max operator as follows:

$$\begin{aligned}
|\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| &\leq |\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{P}}(h, h_0)| + |\mathcal{L}_{\widehat{P}}(h, h_0) - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| \\
&\leq \mu \mathop{\mathrm{E}}_{x \sim \widehat{P}} |f_P(x) - h_0(x)| + \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|,
\end{aligned}$$

where h_0 is an arbitrary element of H''. Since this bound holds for all $h_0 \in H''$, it follows immediately that

$$\lambda\|h^* - h_{\mathsf{U}}\|_K^2 \leq \mu \min_{h_0 \in H''} \mathop{\mathrm{E}}_{\widehat{P}} |f_P(x) - h_0(x)| + \sup_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|,$$

which concludes the proof. ∎

Note that our choice of $\mathsf{Q}\colon h \mapsto \mathsf{Q}_h$ minimizes the right-hand side of (17) among all functions $\mathsf{U} \in \mathcal{A}(H)$ since, for any U, we can write

$$\begin{aligned}
\mathrm{DISC}(\widehat{P}, \mathsf{U}) &= \sup_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)| \geq \sup_{h \in H} \min_{\mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathcal{X}})} \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \\
&= \sup_{h \in H} \max_{h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{Q}_h}(h, f_Q)| = \mathrm{DISC}(\widehat{P}, \mathsf{Q}).
\end{aligned}$$

Thus, in view of Theorem 10, for any constant function $\mathsf{U} \in \mathcal{A}(H)$ with $\mathsf{U}_h = \mathsf{q}$ for some fixed distribution q over $\mathcal{S}_{\mathcal{X}}$, the right-hand side of the bound of Theorem 11 is lower bounded by the right-hand side of the bound of Theorem 12, since the local discrepancy is a finer quantity than the discrepancy: $\mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) \leq \mathrm{disc}(\widehat{P}, \mathsf{q})$. Thus, as expected from the discussion after Theorem 10, our algorithm benefits from a more favorable guarantee than the DM algorithm for some particular choices of H'', especially since, our choice of Q is based on the minimization over all elements in $\mathcal{A}(H)$ and not just the subset of constant functions mapping to a distribution. The following pointwise guarantee follows directly from Theorem 12.

Corollary 13 *Let h^* be a minimizer of $\lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ and h_{Q} a minimizer of $\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{Q}_h}(h, f_Q)$. Then, the following holds for any convex set $H'' \subseteq H$ and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$:*

$$|L(h_{\mathsf{Q}}(x), y) - L(h^*(x), y)| \leq \mu R \sqrt{\frac{\mu\, d_1(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{Q})}{\lambda}}, \tag{18}$$

where $R^2 = \sup_{x \in \mathcal{X}} K(x, x)$.

Proof By the μ-admissibility of the loss, the reproducing property of $\mathbb{H}$, and the Cauchy-Schwarz inequality, the following holds for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$:

$$\begin{aligned}
|L(h_{\mathsf{Q}}(x), y) - L(h^*(x), y)| &\leq \mu |h_{\mathsf{Q}}(x) - h^*(x)| \\
&= \mu |\langle h_{\mathsf{Q}} - h^*, K(x, \cdot)\rangle_K| \\
&\leq \mu \|h_{\mathsf{Q}} - h^*\|_K \sqrt{K(x, x)} \leq R \|h_{\mathsf{Q}} - h^*\|_K.
\end{aligned}$$

Upper bounding $\|h_{\mathsf{Q}} - h^*\|_K$ using Theorem 12 and using the fact that $\mathsf{Q} \colon h \to \mathsf{Q}_h$ is a minimizer of the bound over all choices of $\mathsf{U} \in \mathcal{A}(H)$ yields the desired result. ■

The pointwise loss guarantee just presented can be directly used to bound the difference of the expected loss of h^* and h_{Q} in terms of the same upper bounds, e.g.,

$$\mathcal{L}_P(h_{\mathsf{Q}}, f_P) \leq \mathcal{L}_P(h^*, f_P) + \mu R \sqrt{\frac{\mu\, d_1(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{Q})}{\lambda}}. \tag{19}$$

Similarly, Theorem 10 directly implies the following Corollary.

Corollary 14 *Let h^* be a minimizer of $\lambda\|h\|_K^2 + \mathcal{L}_{\widehat{P}}(h, f_P)$ and h_{Q} a minimizer of $\lambda\|h\|_K^2 + \mathcal{L}_{\mathsf{Q}_h}(h, f_Q)$. Let $\sup_{x \in \mathcal{X}} K(x, x) = R^2$. Then, there exists a choice of $H'' \in \mathcal{H}$ for which the following inequality holds uniformly over over $(x, y) \in \mathcal{X} \times \mathcal{Y}$:*

$$|L(h_{\mathsf{Q}}(x), y) - L(h^*(x), y)| \leq \mu R \sqrt{\frac{\mu \eta_H(f_P, f_Q) + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}_{min})}{\lambda}},$$

where q_{min} is the solution of the DM algorithm.

The choice of the set H'' defining our algorithm is strongly motivated by the theoretical results of this section. In view of Theorem 10, we restrict our choice of H'' to the family $\mathcal{H}$, parametrized only by the radius r. Since the generalized discrepancy DISC is a function of the set H'' which in turn depends only on r, the radius r is chosen to minimize (19). This can be done by using as a validation set a small amount of labeled data from the target domain which is typically available in practice. In particular, as the size of the unlabeled sample $\mathcal{T}'$ increases, our estimate of the optimal radius r becomes more accurate. We provide a detailed description of our algorithm's implementation in Section 5.

4.3. Comparison against Other Learning Bounds

We now compare the learning bounds just derived for our algorithm with those of some common reweighting techniques. In particular, we compare our bounds with those of Cortes et al. (2008) for the KMM algorithm. A similar comparison however can be derived for other algorithms based on importance weighting such as KLIEP or uLSIF.

Assume P and Q admit densities p and q respectively. For every $x \in \mathcal{X}$ we denote by $\beta(x) = \frac{p(x)}{q(x)}$ the importance ratio and by $\overline{\beta} = \beta\big|_{\mathcal{S}_\mathcal{X}}$ its restriction to $\mathcal{S}_\mathcal{X}$. We also let $\widehat{\beta}$ be the solution to the optimization problem solved by the KMM algorithm. Let $h_{\overline{\beta}}$ denote the solution to

$$\min_{h \in \mathbb{H}} \lambda \|h\|^2 + \mathcal{L}_{\overline{\beta}}(h, f_Q), \tag{20}$$

and $h_{\widehat{\beta}}$ be the solution to

$$\min_{h \in \mathbb{H}} \lambda \|h\|^2 + \mathcal{L}_{\widehat{\beta}}(h, f_Q). \tag{21}$$

The following proposition due to Cortes et al. (2008) relates the error of these hypotheses. The proposition requires the kernel K to be a strictly positive definite universal kernel, with Gram matrix $\mathbf{K}$ given by $\mathbf{K}_{ij} = K(x_i, x_j)$.

Proposition 15 *Assume $L(h(x), y) \leq 1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}, h \in H$. For any $\delta > 0$, with probability at least $1 - \delta$ we have:*

$$|\mathcal{L}_P(h_{\overline{\beta}}, f_P) - \mathcal{L}_P(h_{\widehat{\beta}}, f_P)| \leq \frac{\mu^2 R^2 \lambda_{\max}^{\frac{1}{2}}(\mathbf{K})}{\lambda} \Big(\frac{\epsilon B'}{\sqrt{m}} + \frac{\kappa^{1/2}}{\lambda_{\min}^{1/2}(\mathbf{K})} \sqrt{\frac{B'^2}{m} + \frac{1}{n}} \Big(1 + \sqrt{2 \log \frac{2}{\delta}} \Big) \Big), \tag{22}$$

where ϵ and B' are the hyperparameters defining the KMM algorithm and $\lambda_{\max}(\mathbf{K}), \lambda_{\min}(\mathbf{K})$ denote the largest and smallest eigenvalues of $\mathbf{K}$ respectively.

This bound and the one obtained in (19) are of course not comparable since the dependence on μ, R and λ is different. In some cases this dependency can be more favorable in (22) whereas for other values of these parameters (19) provides a better bound. Moreover, (22) depends on the condition number of $\mathbf{K}$ which can become really large in practice. However, the most important difference between these bounds is that (19) is given in terms of the ideal hypothesis h^* while (22) is given in terms of $h_{\overline{\beta}}$, which, in view of the results of Cortes et al. (2010) is not guaranteed to have a good performance on the target distribution. Therefore (22) does not, in general, provide an informative bound.

4.4. Scenario of Additional Labeled Data

Here, we consider a rather common scenario in practice where, in addition to the labeled sample $\mathcal{S}$ drawn from the source domain and the unlabeled sample $\mathcal{T}$ from the target domain, the learner receives a small amount of labeled data from the target domain $\mathcal{T}' = ((x'_1, y'_1), \ldots, (x'_s, y'_s)) \in (\mathcal{X} \times \mathcal{Y})^s$. This sample is typically too small to be used solely to train an algorithm and achieve a good performance. However, it can be useful in at least two ways that we discuss here.

One important benefit of $\mathcal{T}'$ is to serve as a validation set to determine the parameter r that defines the convex set H'' used by our algorithm. The sample $\mathcal{T}'$ can also be used to enhance the discrepancy minimization algorithm as we now show. Let $\widehat{P}'$ denote the empirical distribution associated with $\mathcal{T}'$. To take advantage of $\mathcal{T}'$, the DM algorithm can be trained on the sample of size $(m+s)$ obtained by combining $\mathcal{S}$ and $\mathcal{T}'$, which corresponds to the new empirical distribution $\widehat{Q}' = \frac{m}{m+s}\widehat{Q} + \frac{s}{m+s}\widehat{P}'$. Note that for a fixed value m and large values of s, $\widehat{Q}'$ essentially ignores the points from the source distribution Q, which corresponds to the standard supervised learning scenario in the absence of adaptation. Let $\mathsf{q}'_{\min}$ denote the discrepancy minimization solution when using $\widehat{Q}'$. Since $\mathrm{supp}(\widehat{Q}') \supseteq \mathrm{supp}(\widehat{Q})$, the discrepancy using $\mathsf{q}'_{\min}$ is a lower bound on the discrepancy using $\mathsf{q}_{\min}$:

$$\mathrm{disc}(\mathsf{q}'_{\min}, \widehat{P}) = \min_{\mathrm{supp}(\mathsf{q}) \subseteq \mathrm{supp}(\widehat{Q}')} \mathrm{disc}(\widehat{P}, \mathsf{q}) \leq \min_{\mathrm{supp}(\mathsf{q}) \subseteq \mathrm{supp}(\widehat{Q})} \mathrm{disc}(\widehat{P}, \mathsf{q}) = \mathrm{disc}(\mathsf{q}_{\min}, \widehat{P}).$$

5. Optimization Solution

As shown in Section 3.2, the function G defining our algorithm is convex and the problem of minimizing the expression (7) is a convex optimization problem. Nevertheless, the problem is not straightforward to solve, in particular because evaluating a term like $\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')$ that it contains requires solving a non-convex optimization problem. Here, we present an exact solution in the case of the L_2 loss by solving a semi-definite programming (SDP) problem.

5.1. SDP Formulation

As discussed in Section 4, the choice of H'' is a key component of our algorithm. In view of Corollary 14, we will consider the set $H'' = \{h'' \,|\, \mathcal{L}_{\mathsf{q}_{\min}}(h'', f_Q) \leq r^2\}$. Equivalently, as a result of the reproducing property of $\mathbb{H}$ and the representer theorem, H'' may be defined as $\{\mathbf{a} \in \mathbb{R}^m | \sum_{j=1}^m \mathsf{q}_{\min}(x_j)(\sum_{i=1}^m a_i \mathsf{q}_{\min}(x_i)^{1/2} K(x_i, x_j) - y_j)^2 \leq r^2\}$. Also by the representer theorem, the solution to (7) will be of the form $h = n^{-1/2} \sum_{i=1}^n b_i K(x'_i, \cdot)$. Therefore, given *normalized* kernel matrices $\mathbf{K}_t$, $\mathbf{K}_s$, $\mathbf{K}_{st}$ defined respectively as $\mathbf{K}_t^{ij} = n^{-1} K(x'_i, x'_j)$, $\mathbf{K}_s^{ij} = \mathsf{q}_{\min}(x_i)^{1/2} \mathsf{q}_{\min}(x_j)^{1/2} K(x_i, x_j)$ and $\mathbf{K}_{st}^{ij} = n^{-1/2} \mathsf{q}_{\min}(x_j)^{1/2} K(x'_i, x_j)$, problem (7) is equivalent to

$$\min_{\mathbf{b} \in \mathbb{R}^n} \lambda \mathbf{b}^\top \mathbf{K}_t \mathbf{b} + \frac{1}{2} \Bigg(\max_{\substack{\mathbf{a} \in \mathbb{R}^m \\ \|\mathbf{K}_s \mathbf{a} - \mathbf{y}\|^2 \leq r^2}} \|\mathbf{K}_{st}\mathbf{a} - \mathbf{K}_t \mathbf{b}\|^2 + \min_{\substack{\mathbf{a} \in \mathbb{R}^m \\ \|\mathbf{K}_s \mathbf{a} - \mathbf{y}\|^2 \leq r^2}} \|\mathbf{K}_{st}\mathbf{a} - \mathbf{K}_t \mathbf{b}\|^2 \Bigg), \quad (23)$$

where $\mathbf{y} = (\mathsf{q}_{\min}(x_1)^{1/2} y_1, \ldots, \mathsf{q}_{\min}(x_m)^{1/2} y_m)$ is the vector of normalized labels.

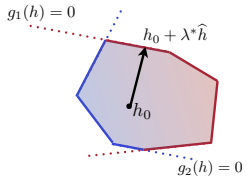


Figure 1: Illustration of the sampling process on the set H''.

Lemma 16 *The Lagrangian dual of the problem*

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^m \\ \|\mathbf{K}_s\mathbf{a}-\mathbf{y}\|^2\leq r^2}} \frac{1}{2}\|\mathbf{K}_{st}\mathbf{a}\|^2 - \mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st}\mathbf{a},$$

is given by

$$\min_{\eta\geq 0,\gamma} \gamma$$
$$s.t. \begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} + \eta\mathbf{K}_s^2 & \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} - \eta\mathbf{K}_s\mathbf{y} \\ \frac{1}{2}\mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st} - \eta\mathbf{y}^\top\mathbf{K}_s & \eta(\|\mathbf{y}\|^2 - r^2) + \gamma \end{pmatrix} \succeq 0.$$

Furthermore, the duality gap for these problems is zero.

The proof of the lemma is given in Appendix A. The lemma helps us derive the following equivalent SDP formulation for our original optimization problem. Its solution can be found in polynomial time using standard convex optimization solvers.

Proposition 17 *The optimization problem* (23) *is equivalent to the following SDP:*

$$\max_{\alpha,\beta,\nu,\mathbf{Z},\mathbf{z}} \frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{Z}) - \beta - \alpha$$
$$s.\ t \begin{pmatrix} \nu\mathbf{K}_s^2 + \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} - \frac{1}{4}\widetilde{\mathbf{K}} & \nu\mathbf{K}_s\mathbf{y} + \frac{1}{4}\widetilde{\mathbf{K}}\mathbf{z} \\ \nu\mathbf{y}^\top\mathbf{K}_s + \frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}} & \alpha + \nu(\|\mathbf{y}\|^2 - r^2) \end{pmatrix} \succeq 0 \quad \begin{pmatrix} \lambda\mathbf{K}_t + \mathbf{K}_t^2 & \frac{1}{2}\mathbf{K}_t\mathbf{K}_{st}\mathbf{z} \\ \frac{1}{2}\mathbf{z}^\top\mathbf{K}_{st}^\top\mathbf{K}_t & \beta \end{pmatrix} \succeq 0$$
$$\begin{pmatrix} \mathbf{Z} & \mathbf{z} \\ \mathbf{z}^\top & 1 \end{pmatrix} \succeq 0 \wedge \nu \geq 0 \wedge \operatorname{Tr}(\mathbf{K}_s^2\mathbf{Z}) - 2\mathbf{y}^\top\mathbf{K}_s\mathbf{z} + \|\mathbf{y}\|^2 \leq r^2,$$

where $\widetilde{\mathbf{K}} = \mathbf{K}_{st}^\top\mathbf{K}_t(\lambda\mathbf{K}_t + \mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}$*, and* $\mathbf{A}^\dagger$ *denotes the pseudo-inverse of matrix* A*.*

In the following section we derive a more efficient approximate solution to the optimization problem using sampling, which helps reducing the problem to a simple QP.

5.2. QP Formulation

The SDP formulation described in the previous section is applicable for a specific choice of H''. In this section, we present an analysis that holds for an arbitrary compact, convex set H''. First, notice that the problem of minimizing G (expression (7)) is related to the

minimum enclosing ball (MEB) problem. For a set $D \subseteq \mathbb{R}^d$, the MEB problem is defined as follows:

$$\min_{\mathbf{u}\in\mathbb{R}^d} \max_{\mathbf{v}\in D} \|\mathbf{u}-\mathbf{v}\|^2.$$

Omitting the regularization and the min term from (7) leads to a problem similar to the MEB. Thus, we could benefit from the extensive literature and algorithmic study available for this problem (Kumar et al., 2003; Schőnherr, 2002; Yildirim, 2008). However, to the best of our knowledge, there is currently no solution available to this problem in the case of an infinite set D, as in the case of our problem. Instead, we present a solution for solving an approximation of (7) based on sampling.

Let $\{h_1, \ldots, h_k\}$ be a set of hypotheses on the boundary of H'', $\partial H''$ and let $\mathcal{C} = \mathcal{C}(h_1, \ldots, h_k)$ denote their convex hull. The following is the sampling-based approximation of (7) that we consider:

$$\min_{h\in\mathbb{H}} \lambda\|h\|_K^2 + \frac{1}{2}\max_{i=1,\ldots,k} \mathcal{L}_{\widehat{P}}(h, h_i) + \frac{1}{2}\min_{h'\in\mathcal{C}} \mathcal{L}_{\widehat{P}}(h, h'). \tag{24}$$

Proposition 18 *Let $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{n\times k}$ be the matrix defined by $Y_{ij} = n^{-1/2}h_j(x'_i)$ and $\mathbf{y}' = (y'_1, \ldots, y'_k)^\top \in \mathbb{R}^k$ the vector defined by $y'_i = n^{-1}\sum_{j=1}^n h_i(x'_j)^2$. Then, the dual problem of (24) is given by*

$$\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\beta} -\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big)^\top \mathbf{K}_t\Big(\lambda\mathbf{I} + \frac{1}{2}\mathbf{K}_t\Big)^{-1}\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big) - \frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{K}_t\mathbf{K}_t^\dagger\boldsymbol{\gamma} + \boldsymbol{\alpha}^\top\mathbf{y}' - \beta \tag{25}$$
$$s.t.\ \mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2}, \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top\boldsymbol{\gamma}, \qquad \boldsymbol{\alpha} \geq 0,$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^k$ with all components equal to 1. Furthermore, the solution h of (24) can be recovered from a solution $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta)$ of (25) by $\forall x, h(x) = \sum_{i=1}^n a_i K(x_i, x)$, where $\boldsymbol{a} = \big(\lambda\mathbf{I} + \frac{1}{2}\mathbf{K}_t\big)^{-1}(\mathbf{Y}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\gamma})$.

The proof of the proposition is given in Appendix B. The result shows that, given a finite sample $h_1, \ldots, h_k$ on the boundary of H'', (24) is in fact equivalent to a standard QP. Hence, a solution can be found efficiently with one of the many off-the-shelf algorithms for quadratic programming.

We now describe the process of sampling from the boundary of the set H'', which is a necessary step for defining problem (24). We consider compact sets of the form $H'' := \{h'' \in \mathbb{H} \mid g_i(h'') \leq 0\}$, where the functions g_i are continuous and convex. For instance, we could consider the set H'' defined in the previous section. More generally, we can consider a family of sets $H''_p = \{h'' \in H| \mid \sum_{i=1}^m \mathfrak{q}_{\min}(x_i)|h(x_i) - y_i|^p \leq r^p\}$.

Assume that there exists h_0 satisfying $g_i(h_0) < 0$. Our sampling process is illustrated by Figure 1 and works as follows: pick a random direction $\widehat{h}$ and define λ_i to be the minimal solution to the system

$$(\lambda \geq 0) \wedge (g_i(h_0 + \lambda\widehat{h}) = 0).$$

Set $\lambda_i = \infty$ if no solution is found and define $\lambda^* = \min_i \lambda_i$. By the convexity and compactness of H'' we can guarantee that $\lambda^* < \infty$. The hypothesis $h = h_0 + \lambda^*\widehat{h}$ satisfies $h \in H''$ and $g_j(h) = 0$ for j such that $\lambda_j = \lambda^*$. The latter is straightforward. To verify the former,

assume that $g_i(h_0 + \lambda^* \widehat{h}) > 0$ for some i. The continuity of g_i would imply the existence of λ_i' with $0 < \lambda_i' < \lambda^* \leq \lambda_i$ such that $g_i(h_0 + \lambda_i' \widehat{h}) = 0$. This would contradict the choice of λ_i, thus, the inequality $g_i(h_0 + \lambda^* \widehat{h}) \leq 0$ must hold for all i.

Since a point h_0 with $g_i(h_0) < 0$ can be obtained by solving a convex program and solving the equations defining λ_i is, in general, simple, the process described provides an efficient way of sampling points from the convex set H''.

5.3. Implementation for the L_2 Loss

We now describe how to fully implement our sampling-based algorithm for the case where L is equal to the L_2 loss. In view of the results of Section 4, we let $H'' = \{h'' | \|h''\|_K \leq \Lambda \ \wedge \ \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^2\}$. We first describe the steps needed to find a point $h_0 \in H''$. Let h_Λ be such that $\|h_\Lambda\|_K = \Lambda$ and $\lambda_r \in \mathbb{R}_+$ be such that the solution h_r to the optimization problem

$$\min_{h \in \mathbb{H}} \lambda_r \|h\|^2 + \mathcal{L}_{\mathsf{q}}(h, f_Q),$$

satisfies $\mathcal{L}_{\mathsf{q}}(h_r, f_Q) = r^2$. It is easy to verify that the existence of λ_r is guaranteed for $\min_{h \in H} \mathcal{L}_{\mathsf{q}}(h, f_Q) \leq r^2 \leq \sum_{i=1}^m \mathsf{q}(x_i) y_i^2$. It is clear that the point $h_0 = \frac{1}{2}(h_r + h_\Lambda)$ is in the interior of H''. Of course, finding λ_r with the desired properties is not straightforward. However, since r is chosen via validation, we do not need to find λ_r as a function of r. Instead, we can simply select λ_r through validation too.

In order to complete the sampling process, we must have an efficient way of selecting a random direction $\widehat{h}$. If $H \subset \mathbb{R}^d$ is a set of linear hypotheses, a direction $\widehat{h}$ can be sampled uniformly by letting $\widehat{h} = \frac{\xi}{\|\xi\|}$, where ξ is a standard Gaussian random variable in $\mathbb{R}^d$. If H is a subset of a RKHS, by the representer theorem, we only need to consider hypotheses of the form $h = \sum_{i=1}^m \alpha_i K(x_i, \cdot)$. Therefore, we can sample a direction $\widehat{h} = \sum_{i=1}^m \alpha_i' K(x_i, \cdot)$, where the vector $\boldsymbol{\alpha'} = (\alpha_1', \ldots, \alpha_m')$ is drawn uniformly from the unit sphere in $\mathbb{R}^m$. A full implementation of our algorithm thus consists of the following steps:

- find the distribution $\mathsf{q}_{\min} = \operatorname{argmin}_{\mathsf{q} \in \mathcal{Q}} \operatorname{disc}(\mathsf{q}, \widehat{P})$. This can be done by using the smooth approximation algorithm of Cortes and Mohri (2013);
- sample points from the set H'' using the sampling process described above;
- solve the QP introduced in Section 5.2.

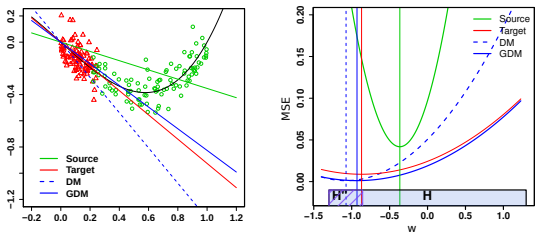
Notice that our algorithm only requires solving a simple QP and therefore its complexity is the same as other adaptation algorithms such as KMM, KLIEP and DM.

6. Experiments

Here, we report the results of extensive comparisons between GDM and several other adaptation algorithms which demonstrate the benefits of our algorithm. We use the implementation described in the previous section. The source code for our algorithm as well as all other baselines described in this section can be found at `http://cims.nyu.edu/~munoz`.

6.1. Synthetic Data Set

To compare the performances of the GDM and DM algorithms, we considered the following synthetic one-dimensional task, which is similar to the one considered by Huang et al.



(a) (b)

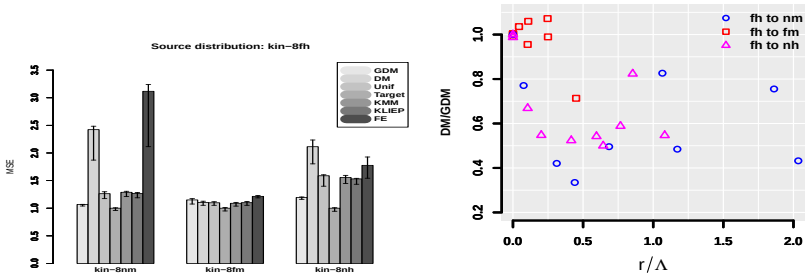
Figure 2: (a) Hypotheses obtained by training on source (green circles), target (red triangles) and using DM (dashed blue) and GDM algorithms (solid blue). (b) Objective functions for source and target distribution as well as GDM and DM algorithms. Sets H and surrogate hypothesis set $H'' \subseteq H$ are shown at the bottom. The vertical lines represent the minimizing hypothesis for each loss.

(2006): the source domain examples were sampled from the uniform distribution over the interval $[.2, 1]$ and target ones sampled uniformly over $[0, .25]$. The labels were given by the map $x \mapsto -x + x^3 + \xi$, where ξ is a Gaussian random variable with mean 0 and standard deviation 0.1. Our hypothesis set was defined by the family of linear functions without an offset. Figure 2(a) shows the regression hypotheses obtained by training the DM and GDM algorithm as well as those obtained by training on the source and target distributions. The ideal hypothesis is shown in red. Notice how the GDM solution gives a closer approximation than DM to the ideal solution. In order to better understand the difference between the solutions of these algorithms, Figure 2(b) depicts the objective function minimized by each algorithm as a function of the slope w of the linear function, the only variable of the hypothesis. The vertical lines show the value of the minimizing hypothesis for each loss. Keeping in mind that the regularization parameter λ used in ridge regression corresponds to a Lagrange multiplier for the constraint $w^2 \leq \Lambda^2$ for some Λ (Cortes and Mohri, 2013) [Lemma 1], the hypothesis set $H = \{w | w^2 \leq \Lambda^2\}$ is depicted at the bottom of this plot. The shaded region represents the set $H'' = H \cap \{h'' | \mathcal{L}_{\mathbf{q}_{\min}}(h'') \leq r\}$. It is clear from this plot that DM helps approximate the target loss function. Nevertheless, only GDM seems to uniformly approach it. This should come as no surprise since our algorithm was precisely designed to achieve that.

6.2. Adaptation Data Sets

We now present the results of evaluating our algorithm against several other adaptation algorithms. GDM is compared against DM and training on the uniform distribution. The following baselines were also used:

1. The KMM algorithm (Huang et al., 2006), which reweights examples from the source distribution in order to match the mean of the source and target data in a feature space induced by a universal kernel. The hyper-parameters of this algorithm were set to the recommended values of $B = 1000$ and $\epsilon = \frac{\sqrt{m}}{\sqrt{m}-1}$.
2. KLIEP (Sugiyama et al., 2007). This algorithm estimates the importance ratio of the source and target distribution by modeling this ratio as a mixture of basis functions and



(a) (b)

Figure 3: (a)MSE performance for different adaptation algorithms when adapting from `kin-8fh` to the three other `kin-8xy` domains. (b) Relative error of DM over GDM as a function of the ratio $\frac{r}{\Lambda}$.

learning the mixture coefficients from the data. Gaussian kernels were used as basis functions for this algorithm and KMM. The bandwidth for the kernel was selected from the set $\{\sigma d\colon \sigma = 2^{-5}, \ldots, 2^{5}\}$ via validation on the *test* set, where d is the mean distance between points sampled from the source domain.

3. FE (Daumé III, 2007). This algorithm maps source and target data into a common high-dimensional feature space where the difference of the distributions is expected to reduce.

Since the two-stage algorithm of Bickel et al. (2007) was already shown to perform similarly to KMM and KLIEP (Cortes and Mohri, 2013), for the sake of readability of our results, we omitted the results of comparison with this algorithm. Finally, we compare our algorithm with the ideal hypothesis h^* returned by training on the target sample $\mathcal{T}$, which we denote by `Tar`. Notice that in practice, this is impossible as $\mathcal{T}$ is unlabeled so we use this result only to show the best attainable performance.

We selected the set of linear functions as our hypothesis set. The learning algorithm used for all tasks was ridge regression and the performance evaluated by the mean squared error. We follow the setup of Cortes and Mohri (2011) and for all adaptation algorithms we selected the parameter λ via 10-fold cross validation over the training data by using a grid search over the set of values $\lambda \in \{2^{-10}, \ldots, 2^{10}\}$. The results of training on the target distribution are presented for a parameter λ tuned via 10-fold cross validation over the target data. We used the QP implementation of our algorithm with the sampling set H'' and the sampling mechanism defined at the end of Section 5.2, where the parameter λ_r was chosen from the same set as λ via validation on a small amount of data from the target distribution. Whereas there are other methods such as transfer cross validation (Zhong et al., 2010) to select the parameters for our algorithm, these methods require the use of importance weighting which as shown in (Cortes et al., 2010) is not theoretically justified.

In order to achieve a fair comparison, all other algorithms were allowed to use the small amount of labeled data too. Since, with the exception of FE, all other baselines do not propose a way of dealing with labeled data from the target distribution, we simply added this data to the training set and ran the algorithms on the extended source data as discussed in Section 4.4.

Task: Sentiment								
S	T	GDM	DM	Unif	Tar	KMM	KLIEP	F
B	K	**0.763±(0.222)**	1.056±(0.289)	1.00	0.517±(0.152)	3.328±(0.845)	3.494±(1.144)	0.942±(0.093)
	E	**0.574±(0.211)**	1.018±(0.206)	1.00	0.367±(0.124)	3.018±(0.319)	3.022±(0.318)	0.857±(0.135)
	D	0.936±(0.256)	1.215±(0.255)	1.00	0.623±(0.152)	2.842±(0.492)	2.764±(0.446)	**0.936±(0.110)**
K	B	**0.854±(0.119)**	1.258±(0.117)	1.00	0.665±(0.085)	2.784±(0.244)	2.642±(0.218)	1.047±(0.047)
	E	0.975±(0.131)	1.460±(0.633)	1.00	0.653±(0.201)	2.408±(0.582)	2.157±(0.255)	**0.969±(0.131)**
	D	**0.884±(0.101)**	1.174±(0.140)	1.00	0.665±(0.071)	2.771±(0.157)	2.620±(0.210)	1.111±(0.059)
E	B	**0.723±(0.138)**	1.016±(0.187)	1.00	0.551±(0.109)	3.433±(0.694)	3.290±(0.583)	1.035±(0.059)
	K	1.030±(0.312)	1.277±(0.283)	1.00	0.636±(0.176)	2.173±(0.249)	2.223±(0.293)	**0.955±(0.199)**
	D	**0.731±(0.171)**	1.005±(0.166)	1.00	0.518±(0.117)	3.363±(0.402)	3.231±(0.483)	0.974±(0.102)
D	B	0.992±(0.191)	1.026±(0.090)	1.00	0.740±(0.138)	2.571±(0.616)	2.475±(0.400)	**0.986±(0.041)**
	K	**0.870±(0.212)**	1.062±(0.318)	1.00	0.557±(0.137)	2.755±(0.375)	2.741±(0.347)	0.940±(0.087)
	E	**0.674±(0.135)**	0.994±(0.171)	1.00	0.478±(0.098)	2.939±(0.501)	2.878±(0.418)	0.907±(0.081)

Table 2: Adaptation from `books` (B), `kitchen` (K), `electronics` (E) and `dvd` (D) to all other domains. Normalized results: MSE of training on the unweighted source data is equal to 1. Results in bold represent the algorithm with the lowest MSE.

The first task we considered is given by the 4 `kin-8xy` Delve data sets (Rasmussen et al., 1996). These data sets are variations of the same model: a realistic simulation of the forward dynamics of an 8 link all-revolute robot arm. The task in all data sets consists of predicting the distance of the end-effector from a target. The data sets differ by the degree of non-linearity (fairly linear, `x=f`, or non-linear, `x=n`) and the amount of noise in the output (moderate, `y=m`, or high, `y=h`). The data set defines 4 different domains, that is 12 pairs of different distributions and labeling functions. A sample of 200 points from each domain was used for training and 10 labeled points from the target distribution were used to select H''. The experiment was carried out 10 times and the results of testing on a sample of 400 points from the target domain are reported in Figure 3(a). The bars represent the median performance of each algorithm. The error bars are the low and high 25% quartiles respectively. All results were normalized in such a way that the median performance of training on the source is equal to 1. Notice that the performance of all algorithms is comparable when adapting to `kin8-fm` since both labeling functions are fairly linear, yet only GDM is able to reasonably adapt to the two data sets with different labeling functions. In order to better understand the advantages of GDM over DM we plot the relative error of DM against GDM as a function of the ratio r/Λ in Figure 3(b), where r is the radius defining H'' and is selected through cross validation. Notice that when the ratio r/Λ is small then both algorithms behave similarly which is most of the times for the adaptation task `fh` to `fm`. On the other hand, a better performance of GDM can be obtained when the ratio is larger. This is due to the fact that r/Λ measures the effective size of the set H''. A small ratio means that the size of H'' is small and therefore the hypothesis returned by GDM will be close to that of DM where as if H'' is large then GDM has the possibility of finding a better hypothesis.

For our next experiment we considered the cross-domain sentiment analysis data set of Blitzer et al. (2007). This data set consists of consumer reviews from 4 different domains: `books, kitchen, electronics` and `dvds`. We used the top 1000 uni-grams and bi-grams

Task: Images								
S	T	GDM	DM	Unif	Tar	KMM	KLIEP	F
C	I	**0.927±(0.051)**	1.005±(0.010)	1.00	0.879±(0.048)	2.752±(3.820)	0.936±(0.016)	0.959±(0.035)
	S	0.938±(0.064)	0.993±(0.018)	1.00	0.840±(0.057)	**0.827±(0.017)**	0.835±(0.020)	0.947±(0.025)
	B	**0.909±(0.040)**	1.003±(0.013)	1.00	0.886±(0.052)	0.945±(0.022)	0.942±(0.017)	0.947±(0.019)
I	C	1.011±(0.015)	**0.951±(0.011)**	1.00	0.802±(0.040)	0.989±(0.036)	1.009±(0.042)	0.971±(0.024)
	S	1.006±(0.030)	0.992±(0.016)	1.00	0.871±(0.030)	**0.930±(0.018)**	0.936±(0.016)	0.973±(0.017)
	B	**0.987±(0.022)**	1.009±(0.010)	1.00	0.986±(0.028)	1.011±(0.028)	1.011±(0.028)	0.994±(0.018)
S	C	1.022±(0.037)	0.982±(0.035)	1.00	0.759±(0.033)	1.172±(0.043)	1.201±(0.038)	**0.938±(0.036)**
	I	**0.924±(0.049)**	0.998±(0.030)	1.00	0.831±(0.047)	3.868±(4.231)	1.227±(0.039)	0.947±(0.028)
	B	**0.898±(0.072)**	1.003±(0.044)	1.00	0.821±(0.053)	1.240±(0.039)	1.248±(0.041)	0.945±(0.021)
B	C	1.010±(0.014)	**0.956±(0.017)**	1.00	0.777±(0.031)	1.028±(0.033)	1.032±(0.031)	0.980±(0.019)
	I	1.012±(0.010)	1.004±(0.007)	1.00	0.966±(0.009)	2.785±(3.803)	**0.981±(0.018)**	1.000±(0.004)
	S	1.009±(0.018)	0.988±(0.010)	1.00	0.850±(0.035)	**0.930±(0.022)**	0.934±(0.024)	0.983±(0.013)

Table 3: Adaptation from `caltech256` (C), `imagenet` (I), `sun` (S) and `bing` (B).

as the features for this task. For each pair of adaptation tasks we sampled 700 points from the source distribution and 700 unlabeled points from the target. Only 50 labeled points from the target distribution were used to tune the parameter r of our algorithm. The final evaluation is done on a test set of 1000 points. The mean results and standard deviations of this task are shown in Table 2 where the MSE values have been normalized in such a way that the performance of training on the source without reweighting is always 1.

Finally, we considered a novel domain adaptation task (Tommasi et al., 2014) of paramount importance in the computer vision community. The domains correspond to 4 well known collections of images: `bing, caltech256, sun` and `imagenet`. These data sets have been standardized so that they all share the same feature representation and labeling function (Tommasi et al., 2014). We sampled 800 labeled points from the source distribution and 800 unlabeled points from the target distribution as well as 50 labeled target points to be used for validation of r. The results of testing on 1000 points from the target domain are presented in Table 3 where, again, the results were normalized in such a way that the performance of training on the source data is always 1.

After analyzing the results of this section we notice that the GDM algorithm consistently outperforms DM and achieves similar or better performance than all other common adaptation algorithms. It is worth noticing that in some cases, other algorithms perform even worse than training on the unweighted sample. This deficiency of the KLIEP algorithm had already been pointed out by Sugiyama et al. (2007) but here we observe that this problem can also affect the KMM algorithm. Finally, let us point out that even though the FE algorithm also achieved performances similar to GDM on the sentiment and image adaptation, its performance was far from optimal adapting on the `kin-8xy` task. Since there is a lack of theoretical understanding for this algorithm, it is hard to characterize the scenarios where FE would perform better than GDM.

7. Conclusion

We presented a new theoretically well-founded domain adaptation algorithm seeking to minimize a less conservative quantity than the DM algorithm. We presented an SDP so-

lution for the particular case of the L_2 loss which can be solved in polynomial time. Our empirical results show that our new algorithm always performs better than or is on par with the otherwise state-of-the-art DM algorithm. We also provided tight generalization bounds for the domain adaptation problem based on the $\mathcal{Y}$-discrepancy. As pointed out in Section 4, an algorithm that minimizes the $\mathcal{Y}$-discrepancy would benefit from the best possible guarantees. However, the lack of labeled data from the target distribution makes this algorithm not viable. This suggests analyzing a richer scenario where the learner is allowed to ask for a limited number of labels from the target distribution. This setup, which is related to active learning, seems to be in fact the closest one to real-life applications and has started to receive attention from the research community (Berlind and Urner, 2015). We believe that the discrepancy disc will play a central role in the analysis of that scenario as well.

Acknowledgments

We would like to thank the reviewers of KDD and JMLR for their suggestions to improve this paper. This work was partly funded by the NSF awards IIS-1117591 and CCF-1535987.

Appendix A. SDP Formulation

Lemma 19 *The Lagrangian dual of the problem*

$$\max_{\substack{\mathbf{a}\in\mathbb{R}^m \\ \|\mathbf{K}_s\mathbf{a}-\mathbf{y}\|^2\leq r^2}} \frac{1}{2}\|\mathbf{K}_{st}\mathbf{a}\|^2 - \mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st}\mathbf{a}, \tag{26}$$

is given by

$$\begin{aligned} &\min_{\eta\geq 0,\gamma} \ \gamma \\ &\text{s. t.} \begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} + \eta\mathbf{K}_s^2 & \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} - \eta\mathbf{K}_s\mathbf{y} \\ \frac{1}{2}\mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st} - \eta\mathbf{y}^\top\mathbf{K}_s & \eta(\|\mathbf{y}\|^2 - r^2) + \gamma \end{pmatrix} \succeq 0. \end{aligned}$$

Furthermore, the duality gap for these problems is zero.

Proof For $\eta \geq 0$ the Lagrangian of (26) is given by

$$\begin{aligned} L(\mathbf{a},\eta) &= \frac{1}{2}\|\mathbf{K}_{st}\mathbf{a}\|^2 - \mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st}\mathbf{a} - \eta(\|\mathbf{K}_s\mathbf{a}-\mathbf{y}\|^2 - r^2) \\ &= \mathbf{a}^\top\Big(\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} - \eta\mathbf{K}_s^2\Big)\mathbf{a} + (2\eta\mathbf{K}_s\mathbf{y} - \mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b})^\top\mathbf{a} - \eta(\|\mathbf{y}\|^2 - r^2). \end{aligned}$$

Since the Lagrangian is a quadratic function of $\mathbf{a}$ and that the conjugate function of a quadratic can be expressed in terms of the pseudo-inverse, the dual is given by

$$\begin{aligned} &\min_{\eta\geq 0} \ \frac{1}{4}(2\eta\mathbf{K}_s\mathbf{y} - \mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b})^\top\Big(\eta\mathbf{K}_s^2 - \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}\Big)^\dagger(2\eta\mathbf{K}_s\mathbf{y} - \mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b}) - \eta(\|\mathbf{y}\|^2 - r^2) \\ &\text{s. t.} \ \ \eta\mathbf{K}_s^2 - \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} \succeq 0. \end{aligned}$$

Introducing the variable γ to replace the objective function yields the equivalent problem

$$\begin{aligned}
&\min_{\eta\geq 0,\gamma} \gamma \\
&\text{s. t.} \quad \eta\mathbf{K}_s^2 - \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} \succeq 0 \\
&\qquad \gamma - \frac{1}{4}(2\eta\mathbf{K}_s\mathbf{y} - \mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b})^\top\Big(\eta\mathbf{K}_s^2 - \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}\Big)^\dagger(2\eta\mathbf{K}_s\mathbf{y} - \mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b}) + \eta(\|\mathbf{y}\|^2 - r^2) \geq 0
\end{aligned}$$

Finally, by the properties of the Schur complement (Boyd and Vandenberghe, 2004), the two constraints above are equivalent to

$$\begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} + \eta\mathbf{K}_s^2 & \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} - \eta\mathbf{K}_s\mathbf{y} \\ \Big(\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} - \eta\mathbf{K}_s\mathbf{y}\Big)^\top & \eta(\|\mathbf{y}\|^2 - r) + \gamma \end{pmatrix} \succeq 0.$$

Since duality holds for a general QCQP with only one constraint (Boyd and Vandenberghe, 2004)[Appendix B], the duality gap between these problems is 0. ■

Proposition 20 *The optimization problem* (23) *is equivalent to the following SDP:*

$$\begin{aligned}
&\max_{\alpha,\beta,\nu,\mathbf{Z},\mathbf{z}} \frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{Z}) - \beta - \alpha \\
&\quad s.\ t \begin{pmatrix} \nu\mathbf{K}_s^2 + \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} - \frac{1}{4}\widetilde{\mathbf{K}} & \nu\mathbf{K}_s\mathbf{y} + \frac{1}{4}\widetilde{\mathbf{K}}\mathbf{z} \\ \nu\mathbf{y}^\top\mathbf{K}_s + \frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}} & \alpha + \nu(\|\mathbf{y}\|^2 - r^2) \end{pmatrix} \succeq 0 \quad \wedge \quad \begin{pmatrix} \mathbf{Z} & \mathbf{z} \\ \mathbf{z}^\top & 1 \end{pmatrix} \succeq 0 \\
&\qquad \begin{pmatrix} \lambda\mathbf{K}_t + \mathbf{K}_t^2 & \frac{1}{2}\mathbf{K}_t\mathbf{K}_{st}\mathbf{z} \\ \frac{1}{2}\mathbf{z}^\top\mathbf{K}_{st}^\top\mathbf{K}_t & \beta \end{pmatrix} \succeq 0 \ \wedge\ \operatorname{Tr}(\mathbf{K}_s^2\mathbf{Z}) - 2\mathbf{y}^\top\mathbf{K}_s\mathbf{z} + \|\mathbf{y}\|^2 \leq r^2 \ \wedge\ \nu \geq 0,
\end{aligned}$$

where $\widetilde{\mathbf{K}} = \mathbf{K}_{st}^\top\mathbf{K}_t(\lambda\mathbf{K}_t + \mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}$.

Proof

By Lemma 16, we may rewrite (23) as

$$\begin{aligned}
&\min_{\mathbf{a},\gamma,\eta,\mathbf{b}} \mathbf{b}^\top(\lambda\mathbf{K}_t + \mathbf{K}_t^2)\mathbf{b} + \frac{1}{2}\mathbf{a}^\top\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{a} - \mathbf{a}^\top\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} + \gamma \qquad (27) \\
&\text{s. t.} \quad \begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} + \eta\mathbf{K}_s^2 & \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} - \eta\mathbf{K}_s\mathbf{y} \\ \frac{1}{2}\mathbf{b}^\top\mathbf{K}_t\mathbf{K}_{st} - \eta\mathbf{y}^\top\mathbf{K}_s & \eta(\|\mathbf{y}\|^2 - r^2) + \gamma \end{pmatrix} \quad \wedge \quad \eta \geq 0 \\
&\qquad \|\mathbf{K}_s\mathbf{a} - \mathbf{y}\|^2 \leq r^2.
\end{aligned}$$

Let us apply the change of variables $\mathbf{b} = \frac{1}{2}(\lambda\mathbf{K}_t + \mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}\mathbf{a} + \mathbf{v}$. The following equalities can be easily verified.

$$\begin{aligned}
\mathbf{b}^\top(\lambda\mathbf{K}_t + \mathbf{K}_t^2)\mathbf{b} &= \frac{1}{4}\mathbf{a}^\top\mathbf{K}_{st}^\top\mathbf{K}_t(\lambda\mathbf{K}_t + \mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}\mathbf{a} + \mathbf{v}^\top\mathbf{K}_t\mathbf{K}_{st}\mathbf{a} + \mathbf{v}^\top(\lambda\mathbf{K}_t + \mathbf{K}_t^2)\mathbf{v}. \\
\mathbf{a}^\top\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{b} &= \frac{1}{2}\mathbf{a}^\top\mathbf{K}_{st}^\top\mathbf{K}_t(\lambda\mathbf{K}_t + \mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}\mathbf{a} + \mathbf{v}^\top\mathbf{K}_t\mathbf{K}_{st}\mathbf{a}.
\end{aligned}$$

Thus, replacing $\mathbf{b}$ on (27) yields

$$\begin{aligned}
&\min_{\mathbf{a},\mathbf{v},\gamma,\eta} \ \mathbf{v}^\top(\lambda\mathbf{K}_t+\mathbf{K}_t^2)\mathbf{v}+\ \mathbf{a}^\top\Big(\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}}\Big)\mathbf{a}+\gamma\\
&\text{s. t.}\quad \begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}+\eta\mathbf{K}_s^2 & \frac{1}{4}\widetilde{\mathbf{K}}\mathbf{a}+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{v}-\eta\mathbf{K}_s\mathbf{y}\\ \frac{1}{4}\mathbf{a}^\top\widetilde{\mathbf{K}}+\frac{1}{2}\mathbf{v}^\top\mathbf{K}_t\mathbf{K}_{st}-\eta\mathbf{y}^\top\mathbf{K}_s & \eta(\|\mathbf{y}\|^2-r^2)+\gamma\end{pmatrix}\succeq 0\quad\wedge\quad\eta\geq 0\\
&\qquad\|\mathbf{K}_s\mathbf{a}-\mathbf{y}\|^2\leq r^2.
\end{aligned}$$

Introducing the scalar multipliers $\mu,\nu\geq 0$ and the matrix

$$\begin{pmatrix}\mathbf{Z} & \mathbf{z}\\ \mathbf{z}^\top & \widetilde{z},\end{pmatrix}\succeq 0$$

as a multiplier for the matrix constraint, we can form the Lagrangian:

$$\begin{aligned}
\mathfrak{L}:=\ &\mathbf{v}^\top(\lambda\mathbf{K}_t+\mathbf{K}_t^2)\mathbf{v}+\mathbf{a}^\top\Big(\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}}\Big)\mathbf{a}+\gamma-\mu\eta+\nu(\|\mathbf{K}_s\mathbf{a}-\mathbf{y}\|^2-r^2)\\
&-\operatorname{Tr}\left(\begin{pmatrix}\mathbf{Z} & \mathbf{z}\\ \mathbf{z} & \widetilde{z}\end{pmatrix}\begin{pmatrix} -\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}+\eta\mathbf{K}_s^2 & \frac{1}{4}\widetilde{\mathbf{K}}\mathbf{a}+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{v}-\eta\mathbf{K}_s\mathbf{y}\\ \frac{1}{4}\mathbf{a}^\top\widetilde{\mathbf{K}}+\frac{1}{2}\mathbf{v}^\top\mathbf{K}_t\mathbf{K}_{st}-\eta\mathbf{y}^\top\mathbf{K}_s & \eta(\|\mathbf{y}\|^2-r^2)+\gamma\end{pmatrix}\right).
\end{aligned}$$

The KKT conditions $\frac{\partial\mathfrak{L}}{\partial\eta}=\frac{\partial\mathfrak{L}}{\partial\gamma}=0$ trivially imply $\widetilde{z}=1$ and $\operatorname{Tr}(\mathbf{K}_s^2\mathbf{Z})-2\mathbf{y}^\top\mathbf{K}_s\mathbf{z}+\|\mathbf{y}\|^2-r^2+\mu=0$. These constraints on the dual variables guarantee that the primal variables η and γ will vanish from the Lagrangian, thus yielding

$$\begin{aligned}
\mathfrak{L}=\ &\frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{Z})+\nu(\|\mathbf{y}\|^2-r^2)+\mathbf{v}^\top(\lambda\mathbf{K}_t+\mathbf{K}_t^2)\mathbf{v}^\top-\mathbf{z}^\top\mathbf{K}_{st}^\top\mathbf{K}_t\mathbf{v}\\
&+\mathbf{a}^\top\Big(\nu\mathbf{K}_s^2+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}}\Big)\mathbf{a}-\Big(2\nu\mathbf{K}_s\mathbf{y}+\frac{1}{2}\widetilde{\mathbf{K}}\mathbf{z}\Big)^\top\mathbf{a}.
\end{aligned}$$

This is a quadratic function on the primal variables $\mathbf{a}$ and $\mathbf{v}$ with minimizing solutions

$$\mathbf{a}=\frac{1}{2}\Big(\nu\mathbf{K}_s^2+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}}\Big)^\dagger\Big(2\nu\mathbf{K}_s\mathbf{y}+\frac{1}{2}\widetilde{\mathbf{K}}\mathbf{z}\Big)\qquad\text{and}\qquad\mathbf{v}=\frac{1}{2}(\lambda\mathbf{K}_t+\mathbf{K}_t^2)^\dagger\mathbf{K}_t\mathbf{K}_{st}\mathbf{z},$$

and optimal value equal to the objective of the Lagrangian dual:

$$\begin{aligned}
&\frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{Z})+\nu(\|\mathbf{y}\|^2-r^2)-\frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}}\mathbf{z}\\
&\qquad-\frac{1}{4}\Big(2\nu\mathbf{K}_s\mathbf{y}+\frac{1}{2}\widetilde{\mathbf{K}}\mathbf{z}\Big)^\top\Big(\nu\mathbf{K}_s^2+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}}\Big)^\dagger\Big(2\nu\mathbf{K}_s\mathbf{y}+\frac{1}{2}\widetilde{\mathbf{K}}\mathbf{z}\Big).
\end{aligned}$$

As in Lemma 16, we apply the properties of the Schur complement to show that the dual is given by

$$\begin{aligned}
&\max_{\alpha,\beta,\nu,\mathbf{Z},\mathbf{z}}\ \frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top\mathbf{K}_{st}\mathbf{Z})-\beta-\alpha\\
&\quad\text{s. t}\ \begin{pmatrix}\nu\mathbf{K}_s^2+\frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st}-\frac{1}{4}\widetilde{\mathbf{K}} & \nu\mathbf{K}_s\mathbf{y}+\frac{1}{4}\widetilde{\mathbf{K}}\mathbf{z}\\ \nu\mathbf{y}^\top\mathbf{K}_s+\frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}} & \alpha+\nu(\|\mathbf{y}\|^2-r^2)\end{pmatrix}\succeq 0\quad\wedge\quad\begin{pmatrix}\mathbf{Z} & \mathbf{z}\\ \mathbf{z}^\top & 1\end{pmatrix}\succeq 0\\
&\qquad\operatorname{Tr}(\mathbf{K}_s^2\mathbf{Z})-2\mathbf{y}^\top\mathbf{K}_s\mathbf{z}+\|\mathbf{y}\|^2\leq r^2\quad\wedge\quad\beta\geq\frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}}\mathbf{z}\quad\wedge\quad\nu\geq 0.
\end{aligned}$$

Finally, recalling the definition of $\widetilde{\mathbf{K}}$ and using the Schur complement one more time we arrive to the final SDP formulation

$$\max_{\alpha,\beta,\nu,\mathbf{Z},\mathbf{z}} \frac{1}{2}\operatorname{Tr}(\mathbf{K}_{st}^\top \mathbf{K}_{st}\mathbf{Z}) - \beta - \alpha$$
$$\text{s. t } \begin{pmatrix} \nu\mathbf{K}_s^2 + \frac{1}{2}\mathbf{K}_{st}^\top\mathbf{K}_{st} - \frac{1}{4}\widetilde{\mathbf{K}} & \nu\mathbf{K}_s\mathbf{y} + \frac{1}{4}\widetilde{\mathbf{K}}\mathbf{z} \\ \nu\mathbf{y}^\top\mathbf{K}_s + \frac{1}{4}\mathbf{z}^\top\widetilde{\mathbf{K}} & \alpha + \nu(\|\mathbf{y}\|^2 - r^2) \end{pmatrix} \succeq 0 \quad \wedge \quad \begin{pmatrix} \mathbf{Z} & \mathbf{z} \\ \mathbf{z}^\top & 1 \end{pmatrix} \succeq 0$$
$$\begin{pmatrix} \lambda\mathbf{K}_t + \mathbf{K}_t^2 & \frac{1}{2}\mathbf{K}_t\mathbf{K}_{st}\mathbf{z} \\ \frac{1}{2}\mathbf{z}^\top\mathbf{K}_{st}^\top\mathbf{K}_t & \beta \end{pmatrix} \succeq 0 \ \wedge \ \operatorname{Tr}(\mathbf{K}_s^2\mathbf{Z}) - 2\mathbf{y}^\top\mathbf{K}_s\mathbf{z} + \|\mathbf{y}\|^2 \leq r^2 \ \wedge \ \nu \geq 0.$$

■

Appendix B. QP Formulation

Proposition 21 *Let $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{n\times k}$ be the matrix defined by $Y_{ij} = n^{-1/2}h_j(x'_i)$ and $\mathbf{y}' = (y'_1, \ldots, y'_k)^\top \in \mathbb{R}^k$ the vector defined by $y'_i = n^{-1}\sum_{j=1}^n h_i(x'_j)^2$. Then, the dual problem of (24) is given by*

$$\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\beta} \ -\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big)^\top \mathbf{K}_t\Big(\lambda\mathbf{I} + \frac{1}{2}\mathbf{K}_t\Big)^{-1}\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big) - \frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{K}_t\mathbf{K}_t^\dagger\boldsymbol{\gamma} + \boldsymbol{\alpha}^\top\mathbf{y}' - \beta \qquad (28)$$
$$s.t.\ \mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2}, \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top\boldsymbol{\gamma}, \qquad \boldsymbol{\alpha} \geq 0,$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^k$ with all components equal to 1. Furthermore, the solution h of (24) can be recovered from a solution $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta)$ of (28) by $\forall x, h(x) = \sum_{i=1}^n a_i K(x_i, x)$, where $\boldsymbol{a} = \big(\lambda\mathbf{I} + \frac{1}{2}\mathbf{K}_t\big)^{-1}(\mathbf{Y}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\gamma})$.

We will first prove a simplified version of the proposition for the case of linear hypotheses, i.e. we can represent hypotheses in $\mathbb{H}$ and elements of $\mathcal{X}$ as vectors $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$ respectively. Define $\mathbf{X}' = n^{-1/2}(\mathbf{x}'_1, \ldots, \mathbf{x}'_n)$ to be the matrix whose columns are the normalized sample points from the target distribution. Let also $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$ be a sample taken from $\partial H''$ and define $\boldsymbol{W} := (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d\times k}$. Under this notation, problem (24) may be rewritten as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|^2 + \frac{1}{2}\max_{i=1,\ldots,k}\|\mathbf{X}'^\top(\mathbf{w} - \mathbf{w}_i)\|^2 + \frac{1}{2}\min_{\mathbf{w}'\in\mathcal{C}}\|\mathbf{X}'^\top(\mathbf{w} - \mathbf{w}')\|^2. \qquad (29)$$

Lemma 22 *The Lagrange dual of problem (29) is given by*

$$\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\beta} -\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big)^\top \mathbf{X}'^\top\Big(\lambda\mathbf{I} + \frac{\mathbf{X}'\mathbf{X}'^\top}{2}\Big)^{-1}\mathbf{X}'\Big(\boldsymbol{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big) - \frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{X}'^\top(\mathbf{X}'\mathbf{X}'^\top)^\dagger\mathbf{X}'\boldsymbol{\gamma} + \boldsymbol{\alpha}^\top\mathbf{y}' - \beta$$
$$s.\ t.\ \mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2} \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top\boldsymbol{\gamma} \qquad \boldsymbol{\alpha} \geq 0,$$

where $\mathbf{Y} = \mathbf{X}'^\top\boldsymbol{W}$ and $\mathbf{y}'_i = \|\mathbf{X}'^\top\mathbf{w}_i\|^2$.

Proof By applying the change of variable $\mathbf{u} = \mathbf{w}' - \mathbf{w}$, problem (29) is can be made equivalent to

$$\min_{\mathbf{w}\in\mathbb{R}^d \mathbf{u}\in\mathcal{C}-\mathbf{w}} \lambda\|\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{u}\|^2 + \frac{1}{2}\max_{i=1,\ldots,k} \|\mathbf{X}'^\top\mathbf{w}_i\|^2 - 2\mathbf{w}_i^\top\mathbf{X}'\mathbf{X}'^\top\mathbf{w}.$$

By making the constraints on $\mathbf{u}$ explicit and replacing the maximization term with the variable r the above problem becomes

$$\begin{aligned}\min_{\mathbf{w},\mathbf{u},r,\boldsymbol{\mu}} \quad & \lambda\|\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{u}\|^2 + \frac{1}{2}r \\ \text{s. t.} \quad & \mathbf{1}r \geq \mathbf{y}' - 2\mathbf{Y}^\top\mathbf{X}'^\top\mathbf{w} \quad \wedge \quad \mathbf{1}^\top\boldsymbol{\mu} = 1 \quad \wedge \quad \boldsymbol{\mu} \geq 0 \quad \wedge \quad \boldsymbol{W}\boldsymbol{\mu} - \mathbf{w} = \mathbf{u}.\end{aligned}$$

For $\boldsymbol{\alpha}, \boldsymbol{\delta} \geq 0$, the Lagrangian of this problem is defined as

$$\begin{aligned}\mathfrak{L}(\mathbf{w}, \mathbf{u}, \boldsymbol{\mu}, r, \boldsymbol{\alpha}, \beta, \boldsymbol{\delta}, \boldsymbol{\gamma}') = \lambda\|\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{w}\|^2 + \frac{1}{2}\|\mathbf{X}'^\top\mathbf{u}\|^2 + \frac{1}{2}r + \beta(\mathbf{1}^\top\boldsymbol{\mu} - 1) \\ + \boldsymbol{\alpha}^\top(\mathbf{y}' - 2(\mathbf{X}'\mathbf{Y})^\top\mathbf{w} - \mathbf{1}r) - \boldsymbol{\delta}^\top\boldsymbol{\mu} + \boldsymbol{\gamma}'^\top(\boldsymbol{W}\boldsymbol{\mu} - \mathbf{w} - \mathbf{u}).\end{aligned}$$

Minimizing with respect to the primal variables yields the following KKT conditions:

$$\mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2} \qquad \mathbf{1}\beta = \boldsymbol{\delta} - \boldsymbol{W}^\top\boldsymbol{\gamma}'. \tag{30}$$

$$\mathbf{X}'\mathbf{X}'^\top\mathbf{u} = \boldsymbol{\gamma}' \qquad 2\left(\lambda\mathbf{I} + \frac{\mathbf{X}'\mathbf{X}'^\top}{2}\right)\mathbf{w} = 2(\mathbf{X}'\boldsymbol{Y})\boldsymbol{\alpha} + \boldsymbol{\gamma}'. \tag{31}$$

Condition (30) implies that the terms involving r and $\boldsymbol{\mu}$ will vanish from the Lagrangian. Furthermore, the first equation in (31) implies that any feasible $\boldsymbol{\gamma}'$ must satisfy $\boldsymbol{\gamma}' = \mathbf{X}'\boldsymbol{\gamma}$ for some $\gamma \in \mathbb{R}^n$. Finally, it is immediate that $\boldsymbol{\gamma}'^\top\mathbf{u} = \mathbf{u}^\top\mathbf{X}'\mathbf{X}'^\top\mathbf{u}$ and $2\mathbf{w}^\top\left(\lambda\mathbf{I} + \frac{\mathbf{X}'\mathbf{X}'^\top}{2}\right)\mathbf{w} = 2\boldsymbol{\alpha}^\top(\mathbf{X}'\mathbf{Y})^\top\mathbf{w} + \boldsymbol{\gamma}'^\top\mathbf{w}$. Thus, at the optimal point, the Lagrangian becomes

$$\begin{aligned}& -\mathbf{w}^\top\Big(\lambda\mathbf{I} + \frac{1}{2}\mathbf{X}'\mathbf{X}'^\top\Big)\mathbf{w} - \frac{1}{2}\mathbf{u}^\top\mathbf{X}'\mathbf{X}'^\top\mathbf{u} + \boldsymbol{\alpha}^\top\mathbf{y}' - \beta \\ \text{s. t.} \quad & \mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2} \qquad \mathbf{1}\beta = \boldsymbol{\delta} - \boldsymbol{W}^\top\boldsymbol{\gamma}' \qquad \boldsymbol{\alpha} \geq 0 \wedge \boldsymbol{\delta} \geq 0.\end{aligned}$$

The positivity of $\boldsymbol{\delta}$ implies that $\mathbf{1}\beta \geq -\boldsymbol{W}^\top\boldsymbol{\gamma}'$. Solving for $\mathbf{w}$ and $\mathbf{u}$ on (31) and applying the change of variable $\mathbf{X}'\boldsymbol{\gamma} = \boldsymbol{\gamma}'$ we obtain the final expression for the dual problem:

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\beta} & -\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big)^\top\mathbf{X}'^\top\Big(\lambda\mathbf{I} + \frac{\mathbf{X}'\mathbf{X}'^\top}{2}\Big)^{-1}\mathbf{X}'\Big(\boldsymbol{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big) - \frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{X}'^\top(\mathbf{X}'\mathbf{X}'^\top)^\dagger\mathbf{X}'\boldsymbol{\gamma} + \boldsymbol{\alpha}^\top\mathbf{y}' - \beta \\ \text{s. t.} & \ \mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2} \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top\boldsymbol{\gamma} \qquad \boldsymbol{\alpha} \geq 0,\end{aligned}$$

where we have used the fact that $\mathbf{Y}^\top\boldsymbol{\gamma} = \mathbf{W}\mathbf{X}'^\top\boldsymbol{\gamma}$ to simplify the constraints. Notice also that we can recover the solution $\mathbf{w}$ of problem (29) as $\mathbf{w} = (\lambda\mathbf{I} + \frac{1}{2}\mathbf{X}'^\top\mathbf{X}')^{-1}\mathbf{X}'(\mathbf{Y}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\gamma})$ ■

The proof of Proposition 18 follows from a straightforward application of the well known matrix identities $\mathbf{X}'(\lambda\mathbf{I} + \mathbf{X}'^\top\mathbf{X}')^{-1} = (\lambda\mathbf{I} + \mathbf{X}'\mathbf{X}'^\top)^{-1}\mathbf{X}'$ and $\mathbf{X}'^\top\mathbf{X}'(\mathbf{X}'^\top\mathbf{X}')^\dagger = \mathbf{X}'^\top(\mathbf{X}'\mathbf{X}'^\top)^\dagger\mathbf{X}'$, and by the fact that the kernel matrix $\mathbf{K}_t$ is equal to $\mathbf{X}'^\top\mathbf{X}'$.

Appendix C. μ-admissibility

Lemma 23 *Assume that $L_p(h(x), y) \leq M$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, then L_p is μ-admissible with $\mu = pM^{p-1}$.*

Proof Since $x \mapsto x^p$ is p-Lipschitz over $[0, 1]$ we can write

$$\begin{aligned}|L(h(x), y) - L(h'(x), y)| &= M^p \left| \Big(\frac{|h(x) - y|}{M}\Big)^p - \Big(\frac{|h'(x) - y|}{M}\Big)^p \right| \\ &\leq pM^{p-1}|h(x) - y + y - h'(x)| = pM^{p-1}|h(x) - h'(x)|.\end{aligned}$$

■

Lemma 24 *Let L be the L_p loss for some $p \geq 1$ and let h, h', h'' be functions satisfying $L_p(h(x), h'(x)) \leq M$ and $L_p(h''(x), h'(x)) \leq M$ for all $x \in \mathcal{X}$, for some $M \geq 0$. Then, for any distribution $\mathcal{D}$ over $\mathcal{X}$, the following inequality holds:*

$$|\mathcal{L}_{\mathcal{D}}(h, h') - \mathcal{L}_{\mathcal{D}}(h'', h')| \leq pM^{p-1}[\mathcal{L}_{\mathcal{D}}(h, h'')]^{\frac{1}{p}}. \tag{32}$$

Proof Proceeding as in the proof of Lemma 23, we obtain

$$\begin{aligned}|\mathcal{L}_{\mathcal{D}}(h, h') - \mathcal{L}_{\mathcal{D}}(h'', h')| &= |\operatorname*{E}_{x \in \mathcal{D}}\big[L_p(h(x), h'(x)) - L_p(h''(x), h'(x)\big]| \\ &\leq pM^{p-1} \operatorname*{E}_{x \in \mathcal{D}}\big[|h(x) - h''(x)|\big].\end{aligned}$$

Since $p \geq 1$, by Jensen's inequality, we can write $\operatorname{E}_{x \in \mathcal{D}}\big[|h(x) - h''(x)|\big] \leq \operatorname{E}_{x \in \mathcal{D}}\big[|h(x) - h''(x)|^p\big]^{1/p} = [\mathcal{L}_{\mathcal{D}}(h, h'')]^{\frac{1}{p}}$. ■

References

Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *Proceedings of ALT*, pages 139–153, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Proceedings of NIPS*, pages 137–144, 2006.

Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. *JMLR - Proceedings Track*, 9:129–136, 2010.

Christopher Berlind and Ruth Urner. Active nearest neighbors in changing environments. In *Proceedings of ICML*, pages 1870–1879, 2015.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of ICML*, pages 81–88, 2007.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, 2007.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

Corinna Cortes and Mehryar Mohri. Domain adaptation in regression. In *Proceedings of ALT*, 2011.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 9474, 2013.

Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. Sample selection bias correction theory. In *Proceedings of ALT*, pages 38–53, 2008.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Proceedings of NIPS*, pages 442–450, 2010.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of ACL*, 2007.

Luc Devroye, Lázló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Mark Dredze, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graça, and Fernando Pereira. Frustratingly hard domain adaptation for dependency parsing. In *EMNLP-CoNLL*, 2007.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of ICML*, 2013.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

Judy Hoffman, Trevor Darrell, and Kate Saenko. Continuous manifold based adaptation for evolving visual domains. In *Proceedings of IEEE CVPR*, pages 867–874, 2014.

Jiayuan Huang, Alexander J. Smola, Arthur Gretton, Karsten M. Borgwardt, and Bernhard Schölkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of NIPS*, volume 19, pages 601–608, 2006.

Jing Jiang and ChengXiang Zhai. Instance Weighting for Domain Adaptation in NLP. In *Proceedings of ACL*, pages 264–271, 2007.

Piyush Kumar, Joseph S. B. Mitchell, and E. Alper Yildirim. Computing core-sets and approximate smallest enclosing hyperspheres in high dimensions. In *ALENEX, Lecture Notes Comput. Sci*, pages 45–55, 2003.

C. J. Leggetter and Philip C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*. Omnipress, 2009.

Aleix M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal.*, 24(6), 2002.

Andrés Muñoz Medina. *Learning Theory and Algorithms for Auctioning and Adaptation Problems*. PhD thesis, New York University, 2015.

Mehryar Mohri and Andres Muñoz. New analysis and algorithm for learning with drifting distributions. In *Proceedings of ALT*. Springer, 2012.

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Trans. on Neural Networks*, 22(2):199–210, 2011.

Carl Edward Rasmussen, Radford M. Neal, Geoffrey Hinton, Drew van Camp, Michael Revow Zoubin Ghahramani, Rafal Kustra, and Rob Tibshirani. The delve project. `http://www.cs.toronto.edu/~delve/data/datasets.html`, 1996. version 1.0.

Sven Schőnherr. *Quadratic Programming in Geometric Optimization: Theory, Implementation, and Applications*. PhD thesis, Swiss Federal Institute of Technology, 2002.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of NIPS*, pages 1433–1440, 2007.

Tatiana Tommasi, Tinne Tuytelaars, and Barbara Caputo. A testbed for cross-dataset analysis. *CoRR*, abs/1402.5923, 2014. URL `http://arxiv.org/abs/1402.5923`.

Leslie G. Valiant. *A Theory of the Learnable*. ACM Press New York, NY, USA, 1984.

Vladimir N. Vapnik. *Statistical Learning Theory*. J. Wiley & Sons, 1998.

E. Alper Yildirim. Two algorithms for the minimum enclosing ball problem. *SIAM Journal on Optimization*, 19(3):1368–1391, 2008.

Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. In *Proceedings of NIPS*, pages 1790–1798. MIT Press, 2012.

Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *Proceedings of ECML PKDD 2010 Part III*, pages 547–562, 2010.