Adaptation Algorithm and Theory Based on Generalized Discrepancy

Corinna Cortes
Google Research
111 8th Avenue
New York, NY 10011

Mehryar Mohri
Courant Institute and Google
251 Mercer Street
New York, NY 10012

Andrés Muñoz Medina
Courant Institute
251 Mercer Street
New York, NY 10012

ABSTRACT

We present a new algorithm for domain adaptation improving upon the discrepancy minimization algorithm (DM), which was previously shown to outperform a number of popular algorithms designed for this task. Unlike most previous approaches adopted for domain adaptation, our algorithm does not consist of a fixed reweighting of the losses over the training sample. Instead, it uses a reweighting that depends on the hypothesis considered and is based on the minimization of a new measure of *generalized discrepancy*. We give a detailed description of our algorithm and show that it can be formulated as a convex optimization problem. We also present 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 the DM algorithm in several tasks.

1. INTRODUCTION

A standard assumption in much of learning theory and algorithms is that the training and test data are sampled from the same distribution. In practice, however, this assumption often does not hold. The learner then faces the more challenging problem of *domain adaptation* where the source and target distributions are distinct. This problem arises in a variety of applications such as natural language processing and computer vision [Dredze et al., 2007, Blitzer et al., 2007b, Jiang and Zhai, 2007, Leggetter and Woodland, 1995, Martínez, 2002, Hoffman et al., 2014] and many other others.

The theory of domain adaptation has been developed in recent years. Early generalization bounds were presented for this problem by Ben-David et al. [2006] and Blitzer et al. [2007a] using a d_A-*distance*. In previous work [Mansour, Mohri, and Rostamizadeh, 2009a, Cortes and Mohri, 2011], we introduced the notion of *discrepancy*, which generalizes the d_A-distance to arbitrary loss functions. We further showed that the discrepancy measure can be accurately estimated from data and proved data-dependent Rademacher complexity bounds for its estimation. We also gave new generalization bounds for domain adaptation based on the discrepancy measure, which we proved to be, under some plausible assumptions, superior to those previously derived by Ben-David et al. [2006] or Blitzer et al. [2007a] (which we showed in fact suffer a factor of 3 of the error that can make them vacuous). We also gave a series of pointwise loss guarantees for the broad class of kernel-based regularized empirical risk minimization algorithms in terms of the empirical discrepancy. In [Mohri and Muñoz, 2012] we further introduced and used the related notion of $\mathcal{Y}$-discrepancy (later rediscovered as *integral probability metric* [Zhang, Zhang, and Ye, 2012]) to derive guarantees for the problem of learning with drifting distributions. This notion was later used by Germain, Habrard, Laviolette, and Morvant [2013] to study the problem of domain adaptation in a PAC-Bayesian setting. Altogether, these theoretical results suggest that the discrepancy is a key quantity in the analysis of adaptation appearing both in upper bounds and lower bounds.

Clearly, domain adaptation cannot always succeed. This depends on the discrepancy between the source and target distribution and some related properties of the labeling functions. This is also corroborated by some negative examples given by Ben-David et al. [2010] and Ben-David and Urner [2012]. As pointed out by these authors, the problem becomes trivially intractable where the hypothesis set contains no candidate with good performance on the training set. However, the adaptation tasks found in applications seem to be often more favorable than such worst cases and several empirical results suggest that adaptation can indeed succeed. Recent work by Wen et al. [2014] also uses a game-theoretic approach to characterize some scenarios where domain adaptation is beneficial.

We can distinguish two broad families of adaptation algorithms. Some consist of finding 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 both to this family of algorithms. While some empirical evidence has been reported in the literature for the effectiveness of these algorithms, we are not aware of any theoretical guarantees in support of these techniques.

Many other adaptation algorithms can be viewed as reweighting techniques. Originated in the Statistics literature on sample bias correction, these techniques attempt to correct the difference between distributions by multiplying every training example by a positive weight. Most of the classical algorithms such as KMM [Huang et al., 2006], KLIEP [Sugiyama et al., 2007] and discrepancy minimization (DM) [Mansour et al., 2009b, Cortes and Mohri, 2011] fall in this category.

The underlying idea behind common reweighting techniques is that of minimizing the *distance* between the reweighted empirical source and target distribution. A crucial component of these learn-

ing algorithms is thus the choice of divergence distance between measures. The KLIEP algorithm is based on the minimization of the KL-divergence, while algorithms such as KMM or the algorithm of Zhang et al. [2013] use the maximum mean discrepancy distance as the divergence to be minimized. It is worth noting that, under some realizability assumptions the algorithm of Zhang et al. [2013] can also be used for the case when the labeling functions shift. The aforementioned algorithms do not provide any learning guarantees. Instead, if the source and target distributions admit densities $q(x)$ and $p(x)$ respectively, the authors show that the weight on the sample point x_i will converge to the importance ratio $p(x_i)/q(x_i)$. The use of this ratio is commonly known as *importance weighting* and it provides and unbiased estimate for the expected loss on the target distribution. While this unbiasedness makes it a natural approach, it has been shown both empirically and theoretically that importance weighting algorithms can fail for the common case where the importance ratio becomes unbounded unless the second-moment bounded, an assumption that cannot be tested in general [Cortes, Mansour, and Mohri, 2010].

In contrast, in [Mansour, Mohri, and Rostamizadeh, 2009b] and [Cortes and Mohri, 2011], we derived generalization bounds for domain adaptation and showed that these bounds directly depend on the discrepancy. We further derived a discrepancy minimization (DM) algorithm that seeks to minimize this generalization bound [Cortes and Mohri, 2011]. 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 a two stage algorithm by Bickel et. al [Bickel et al., 2007]. The main advantage of the DM algorithm is that it takes into account the hypothesis set and the loss function which were previously ignored by other reweighting techniques even though these are crucial elements of any learning algorithm.

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. 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 first describe the learning scenario of domain adaptation in Section 2. Then, we give a detailed description of our algorithm and prove that it can be formulated as a convex optimization problem (Section 3). Next, we analyze the theoretical properties of our algorithm, which guide us to choose 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 Blitzer et al. [2007a], Mansour et al. [2009a], or Cortes and Mohri [2013] and introduces 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 $\mathcal{X}$ while $f_Q, f_P \colon \mathcal{X} \to \mathcal{Y}$, are 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 $x_1, \ldots, x_m$ 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, a small amount of labeled data from the target domain $\mathcal{T}' = ((x''_1, y''_1), \ldots, (x''_s, y''_s)) \in (\mathcal{X} \times \mathcal{Y})^s$ is received by the learner.

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: $\mathcal{L}_{\mathsf{q}}(h, h') = \sum_{x \in \mathcal{X}} q(x) L(h(x), h'(x))$.

3. ALGORITHM

In this section, we introduce our adaptation algorithm. We first review related previous work, next we present the key idea behind the 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. [2009a] and Cortes and Mohri [2011] (see also the *d_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:

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

The discrepancy has several advantages over a measure such as the L_1 or total variation distance [Cortes and Mohri, 2013]: it is a finer measure than the L_1 distance, it takes into account the loss function and the hypothesis set, it can be accurately estimated from finite samples for common hypothesis sets such as kernel-based ones, it is symmetric and verifies the triangle inequality. It further defines a distance in the case of an L_p loss used with a universal kernel such as a Gaussian kernel.

Several generalization bounds for adaptation in terms of the discrepancy have been given in the past [Mansour et al., 2009a, Cortes and Mohri, 2011, 2013], including pointwise guarantees in the case of kernel-based regularized empirical risk minimization, which includes algorithms such as support vector machines (SVM), kernel ridge regression, or support vector regression (SVR). The bounds

given in [Mansour et al., 2009a] 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 on 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 $\mathsf{q}_{\min} = \operatorname{argmin}_{\mathsf{q} \in \mathcal{Q}} \operatorname{disc}(\mathsf{q}, \widehat{P})$, where $\mathcal{Q}$ is the set of all distributions defined over the support of $\widehat{Q}$. Using $\mathsf{q}_{\min}$ instead of $\widehat{Q}$ amounts to reweighting the loss on the training samples to minimize the discrepancy between the empirical distribution and $\widehat{P}$. 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 considered by the learning algorithm or available to it. Thus, a learning algorithm based on discrepancy minimization tends to be too conservative.

3.2 Main idea

Assume as in several previous studies [Mansour et al., 2009a, Cortes and Mohri, 2013] that the standard algorithm selected by the learner is regularized risk minimization over the Hilbert space $\mathbb{H}$ induced by a PSD kernel K. This covers a broad family of algorithms frequently used in applications. Ideally, that is in the absence of a domain adaptation problem, the learner would have access to the labels of the points in $\mathcal{T}$. Therefore, he would return the hypothesis h^* solution of the optimization problem $\min_{h \in \mathbb{H}} F(h)$, where F is the convex function defined for all $h \in \mathbb{H}$ by

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

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 $\widehat{\mathcal{L}}_P(h, f_P)$ with respect to the target domain is as close as possible to $\widehat{\mathcal{L}}_P(h^*, f_P)$. 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. Allowing the weights to be in a richer space than the space of probabilities over $\mathcal{S}_{\mathcal{X}}$ could raise over-fitting concerns but, we will later see that this in fact does not affect our learning guarantees and leads to good empirical results.

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

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

where $\mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})$ 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 and 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, in practice, $\mathcal{L}_{\widehat{P}}(h, f_P)$ is not available 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):

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

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

3.3 Formulation of optimization problem

The following result provides a more explicit expression for $\mathcal{L}_{\mathsf{Q}_h}(h, f_Q)$ leading to a simpler formulation of the optimization problem defining our algorithm.

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 $\mathsf{q} \in \mathcal{F}(\mathcal{S}_{\mathbf{X}}, \mathbb{R})$. Thus, for any $h \in \mathbb{H}$, we can write

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

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:

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

The function $h \mapsto \max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')$ is convex as a pointwise maximum of the convex functions $h \mapsto \mathcal{L}_{\widehat{P}}(h, h'')$. Since the loss function L is jointly convex, so is $\mathcal{L}_{\widehat{P}}$, therefore, the function derived by partial minimization over a non-empty convex set H'' for one of the arguments, $h \mapsto \min_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')$, also defines a convex function [Boyd and Vandenberghe, 2004]. Thus, G is a convex function as a sum of convex functions.

4. LEARNING GUARANTEES

Our description of the algorithm leaves the choice of the hypothesis set H'' unspecified. Our choice will be guided by the theoretical analysis of this section. This will be carried out in two stages. First, we prove a pointwise loss guarantee and a generalization bound for an arbitrary choice of H''. Next, we seek to minimize that bound by choosing H'' out of a family of hypothesis sets $\mathcal{H}$ parametrized by a single parameter r. Our choice of $\mathcal{H}$ is motivated by the proof of existence of parameter values r for which the bound we present is more favorable than that of the DM algorithm.

As in previous work, we assume that the loss function L is μ-*admissible*: there exists $\mu > 0$ such that

$$|L(h(x), y) - L(h'(x), y)| \leq \mu |h(x) - h'(x)| \quad (8)$$

holds for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $h', h \in H$, a condition that is somewhat weaker than μ-Lipschitzness with respect to the first argument. The L_p losses commonly used in regression, $p \geq 1$, verify this condition [Cortes and Mohri, 2013].

4.1 Generalization bounds

The existing pointwise guarantees for the DM algorithm are directly derived from a bound on the norm of the difference of the ideal function h^* and the hypothesis obtained after reweighting the sample losses using a distribution q. The bound is expressed in terms of the discrepancy and a term $\eta_H(f_P, f_Q)$ measuring the difference of the source and target labeling functions defined by

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

and is given by the following proposition.

Theorem 1 ([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:*

$$\lambda\|h^* - h_\mathsf{q}\|_K^2 \leq \mu\, \eta_H(f_P, f_Q) + \mathrm{disc}(\widehat{P}, \mathsf{q}). \quad (9)$$

The DM algorithm is defined by selecting the distribution q minimizing the right-hand side of the bound (9), that is $\mathrm{disc}(\widehat{P}, \mathsf{q})$.

We will show a result of the same nature for our hypothesis-dependent reweighting Q_h by showing that its choice also coincides with that of minimizing an upper bound on $\lambda\|h^* - h'\|_K^2$. Let $\mathcal{A}(H)$ be 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. $\mathcal{A}(H)$ contains all constant functions U such that $\mathsf{U}_h = \mathsf{q}$ for all $h \in H$, where q is a distribution over $\mathcal{S}_\mathcal{X}$. By Proposition 1, $\mathcal{A}(H)$ also includes the function $\mathsf{Q}\colon h \to \mathsf{Q}_h$ used by our algorithm.

Definition 1 (generalized discrepancy)**.** *For any $\mathsf{U} \in \mathcal{A}(H)$, we define the* generalized discrepancy *between $\widehat{P}$ and U as the quantity* $\mathrm{DISC}(\widehat{P}, \mathsf{U})$ *given by*

$$\mathrm{DISC}(\widehat{P}, \mathsf{U}) = \max_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{U}_h}(h, f_Q)|. \quad (10)$$

We also denote by $d_\infty^{\widehat{P}}(f_P, H'')$ the following distance of f_P to H'' over the support of $\widehat{P}$:

$$d_\infty^{\widehat{P}}(f_P, H'') = \min_{h_0 \in H''} \max_{x \in \mathrm{supp}(\widehat{P})} |h_0(x) - f_P(x)|. \quad (11)$$

The following theorem gives an upper bound on the norm of the difference of the minimizing hypotheses in terms of the generalized discrepancy and $d_\infty^{\widehat{P}}(f_P, H'')$.

Theorem 2. *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 $H'' \subseteq H$:*

$$\lambda\|h^* - h_\mathsf{U}\|_K^2 \leq \mu\, d_\infty^{\widehat{P}}(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{U}). \quad (12)$$

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

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

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:

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

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

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

which concludes the proof. □

The following pointwise guarantee for the solution h_{Q} returned by our algorithm is a direct corollary.

Corollary 1. *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 $(x, y) \in \mathcal{X} \times \mathcal{Y}$:*

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

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 $y \in \mathcal{Y}$:

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

Upper bounding $\|h' - h^*\|_K$ using Theorem 2 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.,

$$\begin{aligned}
&\mathcal{L}_P(h_{\mathsf{Q}}, f_P) \\
&\leq \mathcal{L}_P(h^*, f_P)| + \mu R \sqrt{\frac{\mu\, d_\infty^{\widehat{P}}(f_P, H'') + \mathrm{DISC}(\widehat{P}, \mathsf{Q})}{\lambda}}. \quad (13)
\end{aligned}$$

4.2 Choice of H''

In this section, we assume that L is the L_p loss for some $p \geq 1$. The results of the previous section suggest choosing H'' to minimize the generalization bound (13). We will seek to do precisely that by selecting H'' out of the family $\mathcal{H}$ defined by

$$\mathcal{H} = \{B(r)\colon r \geq 0\},$$

where $B(r) = \{h'' \in H | \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^p\}$. Thus, $\mathcal{H}$ is the set of all balls in H centered in f_Q defined in terms of $\mathcal{L}_{\mathsf{q}}$, which is parametrized only by the radius $r \geq 0$. We provide a strong justification for this choice of $\mathcal{H}$ by proving that it contains balls H'' that lead to a generalization bound more favorable than that of the DM algorithm. Our algorithm is defined by selecting the radius r minimizing the generalization bound (13). This can be done by using as validation set a small amount of labeled data from the target domain, which is typically available in practice.

The following theorem proves the existence of a ball $H'' \in \mathcal{H}$ for which (12) is a uniformly tighter upper bound than (9). The result is expressed in terms of the *local discrepancy* defined by:

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

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

Theorem 3. *There exists $H'' \in \mathcal{H}$ such that the following holds:*

$$\mu\, d_\infty^{\widehat{P}}(f_P, H'') + \max_{h \in H, h'' \in H''} |\mathcal{L}_{\widehat{P}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \leq \mu\, \eta_H(f_P, f_Q) + \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}).$$

Proof. Let h_0^* be the minimizer in the definition of $\eta_H(f_P, f_Q)$:

$$h_0^* = \operatorname*{argmin}_{h_0 \in H} \big(\max_{x \in \mathrm{supp}(\widehat{P})} |f_P(x) - h_0(x)| + \max_{x \in \mathrm{supp}(\widehat{Q})} |f_Q(x) - h_0(x)|\big),$$

and let $r = \max_{x \in \mathrm{supp}(\widehat{Q})} |f_Q(x) - h_0^*(x)|$. Let q be a distribution over $\mathcal{S}_{\mathcal{X}}$ and choose $H'' \in \mathcal{H}$ as $H'' = \{h'' \in H | \mathcal{L}_{\mathsf{q}}(h'', f_Q) \leq r^p\}$. Then, h_0^* is in H'' since

$$\begin{aligned}
\mathcal{L}_{\mathsf{q}}(h_0^*, f_Q) &= \mathop{\mathrm{E}}_{x \sim \mathsf{q}} \left[|h_0^*(x) - f_Q(x)|^p\right] \\
&\leq \max_{x \in \mathrm{supp}(\widehat{Q})} |h_0^*(x) - f_Q(x)|^p = r^p.
\end{aligned}$$

For the L_p loss, it is not hard to show [Cortes et al., 2014][Lemma 14] that for all $h, h'' \in H$, $|\mathcal{L}_{\mathsf{q}}(h, h'') - \mathcal{L}_{\mathsf{q}}(h, f_Q)| \leq \mu[\mathcal{L}_{\mathsf{q}}(h'', f_Q)]^{\frac{1}{p}}$. In view of this inequality, we can write:

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

Using this inequality and the fact that $h_0^* \in H''$, we can write

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

which concludes the proof. □

The theorem shows that for that particular choice of H'', for any constant function $\mathsf{U}_h \in \mathcal{A}(H)$ 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 1 is lower bounded by the right-hand side of the bound of Theorem 2, 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, our algorithm benefits from a more favorable guarantee than the DM algorithm for that particular choice 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 result readily follows from Theorem 3.

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

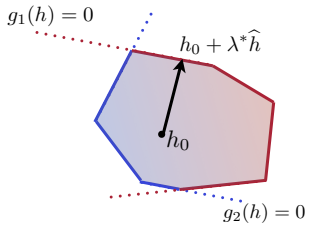

Figure 1: Illustration of the sampling process on the set H''.

Then, there exists a choice of $H'' \in \mathcal{H}$ for which the following inequality holds uniformly over over $(x, y) \in \mathcal{X} \times \mathcal{Y}$:

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

We conclude this section by briefly discussing the effect of the sample sizes on our guarantees. Clearly, a larger source sample, that is a larger $\mathrm{supp}(\widehat{Q})$, results in a smaller minimal discrepancy $\mathrm{disc}_{H''}(\widehat{P}, \mathsf{q}) = \min_{\mathsf{q} \in \mathrm{supp}(\widehat{Q})} \mathrm{disc}_{H''}(\widehat{P}, \mathsf{q})$, thereby leading to a more beneficial pointwise guarantee, in view of Corollary 2. A larger target sample, improves the guarantee on the expected loss $\mathrm{E}[L(h^*(x), y)]$ via standard supervised learning bounds, which, by Corollary 2 further improves the guarantee on the expected loss $\mathrm{E}[L(h_{\mathsf{Q}}(x), y)]$.

5. OPTIMIZATION SOLUTION

As shown in Section 3.3, 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 the term $\max_{h'' \in H''} \mathcal{L}_{\widehat{P}}(h, h'')$ that it contains requires solving a non-convex optimization problem. Here, we present an approximation to this problem based on a QP that can be efficiently solved. We have also derived an exact but less efficient solution by giving a semi-definite programming (SDP) formulation for the problem. Due to space limitations, we do not include that solution here, but it can be found in the full version of this paper [Cortes et al., 2014].

5.1 QP formulation

The analysis presented in this section holds for an arbitrary 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 [Welzl, 1991, Kumar et al., 2003, Schőnherr, 2002, Fischer et al., 2003, 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 in $\partial H''$ 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:

$$\min_{h \in \mathbb{H}} \lambda \|h\|_K^2 + \frac{1}{2} \max_{i=1,\ldots,k} \mathcal{L}_{\widehat{P}}(h, h_i) + \frac{1}{2} \min_{h' \in \mathcal{C}} \mathcal{L}_{\widehat{P}}(h, h'). \quad (14)$$

Proposition 2. *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 (14) is given by*

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta} -\Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big)^\top \mathbf{K}_t \Big(\lambda \mathbf{I} + \frac{1}{2}\mathbf{K}_t\Big)^{-1} \Big(\mathbf{Y}\boldsymbol{\alpha} + \frac{\boldsymbol{\gamma}}{2}\Big) - \frac{1}{2}\boldsymbol{\gamma}^\top \mathbf{K}_t \mathbf{K}_t^\dagger \boldsymbol{\gamma} + \boldsymbol{\alpha}^\top \mathbf{y}' - \beta \quad (15)$$

$$\text{s.t. } \mathbf{1}^\top \boldsymbol{\alpha} = \frac{1}{2}, \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top \boldsymbol{\gamma}, \qquad \boldsymbol{\alpha} \geq 0,$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^k$ with all components equal to 1. Furthermore, the solution h of (14) can be recovered from a solution $(\boldsymbol{\alpha}, \boldsymbol{\gamma}, \beta)$ of (15) by $\forall x, h(x) = \sum_{i=1}^n a_i K(x_i, x)$, where $\boldsymbol{a} = \big(\lambda \mathbf{I} + \frac{1}{2}\mathbf{K}_t\big)^{-1}\big(\mathbf{Y}\boldsymbol{\alpha} + \frac{1}{2}\boldsymbol{\gamma}\big)$.

The proof of the proposition is given in Appendix A. The result shows that, given a finite sample $h_1, \ldots, h_k$ on the boundary of H'', (14) is in fact equivalent to a standard QP and therefore can be efficiently with one of the many off-the-shelf QP algorithms.

We now describe the process of sampling from the boundary of H'', a necessary step for defining problem (14). Let $H'' := \{h'' \in \mathbb{H} \mid g_i(h'') \leq 0\}$ be a compact set, where the functions g_i are continuous and convex. For instance, we can consider a family of sets $H''_p = \{h'' \in H| \mid \sum_{i=1}^m \mathsf{q}_{\min}(x_i)|h(x_i) - y_i|^p \leq r^p\}$.

Assume h_0 is given, where $g_i(h_0) < 0$. 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 $\lambda^* = \min_i \lambda_i$. The compactness of H'' guarantees $\lambda^* < \infty$. Moreover, 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 $g_i(h_0 + \lambda^* \widehat{h}) > 0$ 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 $g_i(h_0 + \lambda'_i \widehat{h}) = 0$ contradicting the choice of λ_i. Thus, $g_i(h_0 + \lambda^* \widehat{h}) \leq 0$ 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''.

In the next section, we report the results of experiments with our algorithm in several tasks in which it outperforms the DM algorithm.

5.2 Implementation for the L_2 loss

We now describe how to fully implement our 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 begin by describing the necessary steps 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 now immediate that the point $h_0 = \frac{1}{2}(h_r + h_\Lambda)$ is in the interior of H''. Of course, finding the value of λ_r with the desired properties may not be easy. However, since r is chosen through cross-validation, we do not need to find λ_r as a function of r. Instead, we can simply select λ_r through cross-validation too.

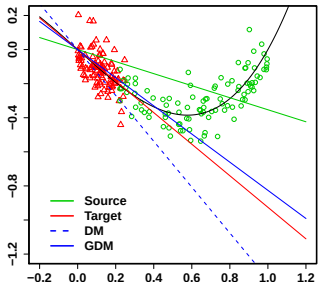

Figure 2: Linear hypotheses obtained by training on the source (green circles), target (red triangles) and by using the DM (solid blue) and GDM algorithms (dashed blue).

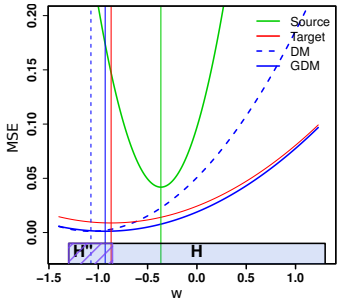

Figure 3: Objective functions associated with training on the source distribution, training on the target distribution, as well as the GDM and DM algorithms. The hypothesis set H and surrogate hypothesis set H'' are shown at the bottom of the plot.

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 may only consider hypotheses $h = \sum_{i=1}^m \alpha_i K(x_i, \cdot)$. Therefore, we can sample a direction $\widehat{h}$ by letting $\widehat{h} = \sum_{i=1}^m \alpha'_i K(x_i, \cdot)$ where the vector $\boldsymbol{\alpha'} = (\alpha'_1, \ldots, \alpha'_m)$ is sampled uniformly from the unit sphere in $\mathbb{R}^m$. A full implementation of our algorithm then consists of the following steps:

- compute 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 5.1

6. EXPERIMENTS

In this section, we report the results of extensive comparisons between GDM and several other adaptation algorithms, which show favorable results for our algorithm.

6.1 Synthetic data set

To give an empirical comparison of the GDM and DM algorithms, we adopted the following setup inspired by Huang et al. [2006]: we sampled source distribution examples from the uniform distribution over the interval $[.2, 1]$ and target samples from the uniform distribution 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 and our hypothesis set was chosen to be that of linear functions.

Figure 2(b) shows the regression hypotheses obtained by training the DM and GDM algorithms as well as those obtained by training on the source and the target distributions. Notice how the GDM solution approaches the ideal solution better than DM. These results can be better explained by Figure 3 which plots the objective function minimized by each algorithm as a function of the slope w of the linear function, the only variable of the hypothesis. 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 $H = \{w \colon |w| \leq \Lambda\}$ is shown at the bottom of this plot. The shaded region represents the set $H'' = H \cap \{h'' | \mathcal{L}_{\mathsf{q}_{\min}}(h'') \leq r\}$. 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 designed precisely for this purpose.

6.2 Adaptation Data Sets

We now present the results of evaluating the performance of our algorithm and comparing with several others. GDM is compared to DM and to training on the source distribution. The following algorithms were also used:

1. The KMM algorithm [Huang et al., 2006] reweights data samples to match empirical target and source means on the feature space induced by Gaussian kernels. 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] minimizes the KL-divergence between the source and target empirical distributions. Distributions are modeled as a mixture of Gaussians. The bandwidth of the kernel for both KLIEP and KMM was selected from the set $\{\sigma d \colon \sigma = 2^{-5}, \ldots, 1\}$ 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 to a common high-dimensional feature space where the difference of the distributions is hoped to be smaller

We refrained from comparing against the two-stage algorithm of Bickel et al. [2007], as it was already shown to perform similarly to KMM and KLIEP [Cortes and Mohri, 2013].

The hypothesis set H was given by linear functions. 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]. For all adaptation algorithms, we selected the parameter λ via 10-fold cross validation over the training data for $\lambda \in \boldsymbol{\Lambda} = \{2^{-10}, \ldots, 2^{10}\}$. 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 in Section 5.1. The parameter $\lambda_r \in \boldsymbol{\Lambda}$ was chosen via cross-validation on a small amount of data from the target distribution. To be complete, we also report the results of training only on the small amount of target data.

To make a fair comparison, we allowed the use of the small amount of labeled data to all other baselines. To do so, we simply added this data to the training set and ran the algorithms on the extended source data.

Our first task is given by the 4 `kin-8xy` Delve data sets [Rasmussen et al., 1996]. These data sets are variations of the same

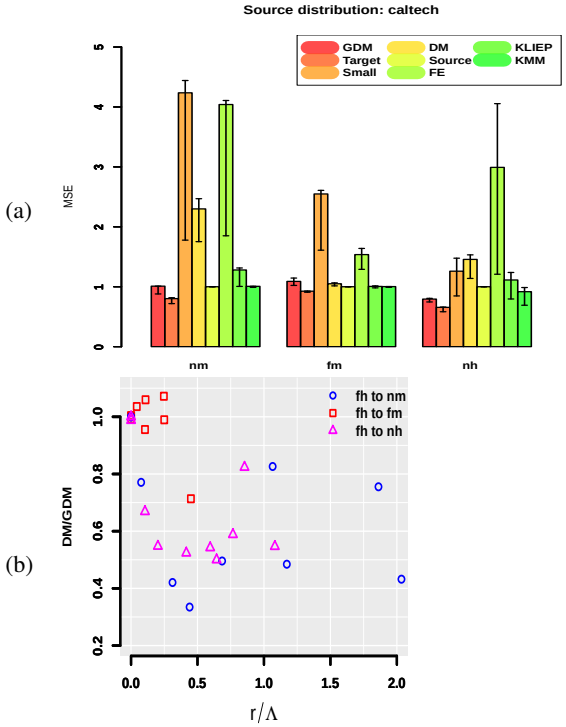

Figure 4: (a) Normalized MSE performance for different adaptation algorithms when adapting from `kin-8fh` to the three other `kin-8xy` domains. `Small` denotes training on small labeled target sample. (b) Relative error of DM over GDM as a function of the ratio $\frac{r}{\Lambda}$.

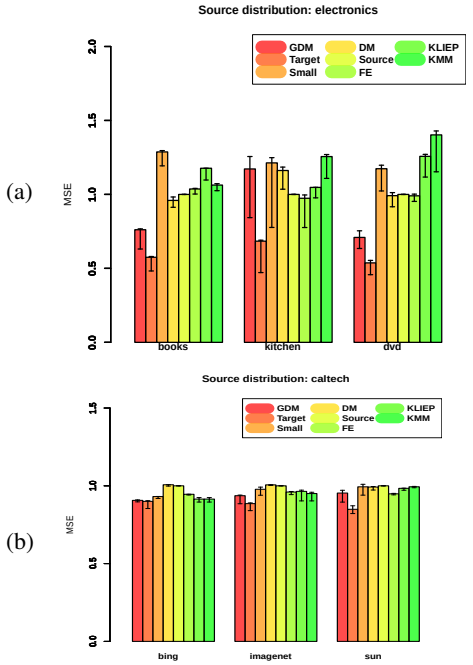

Figure 5: (a) Normalized MSE for the sentiment adaptation task from the `electronics` domain to all others. (b) Normalized MSE of different algorithms adapting from the `caltech256` dataset to all other datasets.

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. 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`). 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. The results of testing on a sample of 400 points from the target domain are reported in Figure 4(a). The bars represent the median performance of each algorithm and error bars show the inter-quartile range. All results were normalized in such a way that training on the source had error constantly 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 significantly approach the performance on training on target for all three tasks. 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 4(b). Notice that when the ratio r/Λ is small, then both algorithms behave similarly which typically 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 can be interpreted as follows: a small ratio means that the size of H'' is small and the hypothesis returned by GDM will be close to that of DM, while for H'' large, GDM can find a better hypothesis.

For our next experiment, we considered the cross-domain sentiment analysis data set of Blitzer et al. [2007b]. This data set consists of consumer reviews from 4 different domains: `books`, `kitchen`, `electronics` and `dvds`. We used the top 1,000 unigrams and bigrams as features. 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 r. The final evaluation was done on a test set of 1,000 points. Figure 5(a) shows normalized MSE of all algorithms when adapting from `electronics` to all other domains.

Finally, we considered a key domain adaptation task in the computer vision community [Tommasi et al., 2014] where 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 used the data from the first 5 shared classes and sampled 800 labeled points from the source distribution and 800 unlabeled points from the target distribution, as well as 50 labeled target points used as validation to determine r. The results of testing on 1,000 points from the target domain are shown in Figure 5(b) where we trained on `caltech256`. Due to space limitations, we were not able to present the results of all possible adaptation tasks. They can be found in Cortes et al. [2014]. The results of this section show that GDM was the only algorithm that could consistently perform better than or on par with the best algorithm.

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 solution for the particular case of the L_2 loss which can be solved in polynomial time. Our empirical results show that our new algorithm is the only adaptation algorithm consistently achieving a performance close to that of training on the target distribution.

References

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

S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira. Analysis of representations for domain adaptation. In *Proceedings of NIPS*, pages 137–144, 2006.

S. Ben-David, T. Lu, T. Luu, and D. Pál. Impossibility theorems for domain adaptation. *JMLR - Proceedings Track*, 9:129–136, 2010.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of ICML*, pages 81–88, 2007.

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Proceedings of NIPS*, 2007a.

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL*, 2007b.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

C. Cortes and M. Mohri. Domain adaptation in regression. In *Proceedings of ALT*, 2011.

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

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Proceedings of NIPS*, pages 442–450, 2010.

C. Cortes, M. Mohri, and A. Muñoz. Adaptation algorithm and theory based on generalized discrepancy. *ArXiv:1405.1503*, May 2014.

H. Daumé III. Frustratingly easy domain adaptation. In *Proceedings of ACL*, Prague, Czech Republic, 2007.

M. Dredze, J. Blitzer, P. P. Talukdar, K. Ganchev, J. Graça, and F. Pereira. Frustratingly hard domain adaptation for dependency parsing. In *EMNLP-CoNLL*, 2007.

K. Fischer, B. Gärtner, and M. Kutz. Fast smallest-enclosing-ball computation in high dimensions. In *Algorithms-ESA 2003*, pages 630–641. Springer, 2003.

P. Germain, A. Habrard, F. Laviolette, and E. Morvant. A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *Proceedings of ICML*, 2013.

J. Hoffman, T. Darrell, and K. Saenko. Continuous manifold based adaptation for evolving visual domains. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Proceedings of NIPS*, volume 19, pages 601–608, 2006.

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

P. Kumar, J. S. B. Mitchell, and E. A. Yildirim. Computing coresets and approximate smallest enclosing hyperspheres in high dimensions. In *ALENEX, Lecture Notes Comput. Sci*, pages 45–55, 2003.

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

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Proceedings of COLT*. Omnipress, 2009a.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *Proceedings of NIPS*. MIT Press, 2009b.

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

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

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

C. E. Rasmussen, R. M. Neal, G. Hinton, D. van Camp, M. R. Z. Ghahramani, R. Kustra, and R. Tibshirani. The delve project. `http://www.cs.toronto.edu/~delve/data/datasets.html`, 1996. version 1.0.

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

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Proceedings of NIPS*, pages 1433–1440, 2007.

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

E. Welzl. Smallest enclosing disks (balls and ellipsoids). In *New results and new trends in computer science (Graz, 1991)*, volume 555 of *Lecture Notes in Comput. Sci.*, pages 359–370. Springer, Berlin, 1991.

J. Wen, C. Yu, and R. Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *Proceedings of ICML*, pages 631–639, 2014.

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

C. Zhang, L. Zhang, and J. Ye. Generalization bounds for domain adaptation. In *Proceedings of NIPS*, pages 1790–1798. MIT Press, 2012.

K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proceedings of ICML 2013*, pages 819–827, 2013.

APPENDIX

A. QP FORMULATION

Proposition 2. *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* (14) *is given by*

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\gamma},\beta} \; & -\Big(\mathbf{Y}\boldsymbol{\alpha}+\frac{\boldsymbol{\gamma}}{2}\Big)^\top \mathbf{K}_t\Big(\lambda\mathbf{I}+\frac{1}{2}\mathbf{K}_t\Big)^{-1}\Big(\mathbf{Y}\boldsymbol{\alpha}+\frac{\boldsymbol{\gamma}}{2}\Big) \\ & -\frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{K}_t\mathbf{K}_t^\dagger\boldsymbol{\gamma}+\boldsymbol{\alpha}^\top\mathbf{y}'-\beta \qquad (16)\\ \textit{s.t. } & \mathbf{1}^\top\boldsymbol{\alpha}=\frac{1}{2}, \qquad \mathbf{1}\beta \geq -\mathbf{Y}^\top\boldsymbol{\gamma}, \qquad \boldsymbol{\alpha}\geq 0,\end{aligned}$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^k$ with all components equal to 1. Furthermore, the solution h of (14) *can be recovered from a solution $(\boldsymbol{\alpha},\boldsymbol{\gamma},\beta)$ of* (16) *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 $\boldsymbol{W} := (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d\times k}$. With this notation, problem (14) can be rewritten as follows:

$$\begin{aligned}\min_{\mathbf{w}\in\mathbb{R}^d} \lambda\|\mathbf{w}\|^2 + \frac{1}{2}\max_{i=1,\ldots,k}\|\mathbf{X}'^\top(\mathbf{w}-\mathbf{w}_i)\|^2 \\ + \frac{1}{2}\min_{\mathbf{w}'\in\mathcal{C}}\|\mathbf{X}'^\top(\mathbf{w}-\mathbf{w}')\|^2. \qquad (17)\end{aligned}$$

Lemma 1. *The Lagrange dual of problem* (17) *is given by*

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

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

Proof. Using the change of variable $\mathbf{u} = \mathbf{w}' - \mathbf{w}$, we obtain the following problem equivalent to (17):

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

Making the constraints on $\mathbf{u}$ explicit and replacing the maximization term with the variable r yield:

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

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:

$$\mathbf{1}^\top\boldsymbol{\alpha} = \frac{1}{2} \qquad \mathbf{1}\beta = \boldsymbol{\delta} - \boldsymbol{W}^\top\boldsymbol{\gamma}'. \qquad (18)$$

$$\mathbf{X}'\mathbf{X}'^\top\mathbf{u} = \boldsymbol{\gamma}' \qquad 2\left(\lambda\mathbf{I}+\frac{\mathbf{X}'\mathbf{X}'^\top}{2}\right)\mathbf{w} = 2(\mathbf{X}'\boldsymbol{Y})\boldsymbol{\alpha}+\boldsymbol{\gamma}' \qquad (19)$$

Condition (18) implies that the terms involving r and $\boldsymbol{\mu}$ will vanish from the Lagrangian. Furthermore, the first equation in (19) 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 (19) 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 (17) as $\mathbf{w} = (\lambda\mathbf{I}+\frac{1}{2}\mathbf{X}'^\top\mathbf{X}')^{-1}\mathbf{X}'(\mathbf{Y}\boldsymbol{\alpha}+\frac{1}{2}\boldsymbol{\gamma})$ □

Using the matrix identities $\mathbf{X}'(\lambda\mathbf{I}+\mathbf{X}'^\top\mathbf{X}')^{-1} = (\lambda\mathbf{I}+\mathbf{X}'\mathbf{X}'^\top)\mathbf{X}'$ and $\mathbf{X}'^\top\mathbf{X}'(\mathbf{X}'^\top\mathbf{X}')^\dagger = \mathbf{X}'^\top(\mathbf{X}'\mathbf{X}'^\top)^\dagger\mathbf{X}'$, the proof of Proposition 2 is now immediate.

Proposition 2. We can rewrite the dual objective of the previous lemma in terms of the Gram matrix $\mathbf{X}'^\top\mathbf{X}'$ alone as follows:

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

By replacing $\mathbf{X}'^\top\mathbf{X}'$ by the more general kernel matrix $\mathbf{K_t}$ (which corresponds to the Gram matrix in the feature space) we obtain the desired expression for the dual. Additionally, the same matrix identities applied to condition (19) imply that the optimal hypothesis h is given by $h(x) = \sum_{i=1}^n a_i K(x'_i, x)$ where $\boldsymbol{a} = (\lambda\mathbf{I}+\frac{1}{2}\mathbf{K}_t)^{-1}(\mathbf{Y}\boldsymbol{\alpha}+\frac{\boldsymbol{\gamma}}{2})$. □