Domain Adaptation: Learning Bounds and Algorithms

Yishay Mansour
Google Research and
Tel Aviv Univ.
mansour@tau.ac.il

Mehryar Mohri
Courant Institute and
Google Research
mohri@cims.nyu.edu

Afshin Rostamizadeh
Courant Institute
New York University
rostami@cs.nyu.edu

Abstract

This paper addresses the general problem of domain adaptation which arises in a variety of applications where the distribution of the labeled sample available somewhat differs from that of the test data. Building on previous work by Ben-David et al. (2007), we introduce a novel distance between distributions, *discrepancy distance*, that is tailored to adaptation problems with arbitrary loss functions. We give Rademacher complexity bounds for estimating the discrepancy distance from finite samples for different loss functions. Using this distance, we derive new generalization bounds for domain adaptation for a wide family of loss functions. We also present a series of novel adaptation bounds for large classes of regularization-based algorithms, including support vector machines and kernel ridge regression based on the empirical discrepancy. This motivates our analysis of the problem of minimizing the empirical discrepancy for various loss functions for which we also give several algorithms. We report the results of preliminary experiments that demonstrate the benefits of our discrepancy minimization algorithms for domain adaptation.

1 Introduction

In the standard PAC model (Valiant, 1984) and other theoretical models of learning, training and test instances are assumed to be drawn from the same distribution. This is a natural assumption since, when the training and test distributions substantially differ, there can be no hope for generalization. However, in practice, there are several crucial scenarios where the two distributions are more similar and learning can be more effective. One such scenario is that of *domain adaptation*, the main topic of our analysis.

The problem of domain adaptation arises in a variety of applications in natural language processing (Dredze et al., 2007; Blitzer et al., 2007; Jiang & Zhai, 2007; Chelba & Acero, 2006; Daumé III & Marcu, 2006), speech processing (Legetter & Woodland, 1995; Gauvain & Chin-Hui, 1994; Pietra et al., 1992; Rosenfeld, 1996; Jelinek, 1998; Roark & Bacchiani, 2003), computer vision (Martínez, 2002), and many other areas. Quite often, little or no labeled data is available from the *target domain*, but labeled data from a *source domain* somewhat similar to the target as well as large amounts of unlabeled data from the target domain are at one's disposal. The domain adaptation problem then consists of leveraging the source labeled and target unlabeled data to derive a hypothesis performing well on the target domain.

A number of different adaptation techniques have been introduced in the past by the publications just mentioned and other similar work in the context of specific applications. For example, a standard technique used in statistical language modeling and other generative models for part-of-speech tagging or parsing is based on the maximum a posteriori adaptation which uses the source data as prior knowledge to estimate the model parameters (Roark & Bacchiani, 2003). Similar techniques and other more refined ones have been used for training maximum entropy models for language modeling or conditional models (Pietra et al., 1992; Jelinek, 1998; Chelba & Acero, 2006; Daumé III & Marcu, 2006).

The first theoretical analysis of the domain adaptation problem was presented by Ben-David et al. (2007), who gave VC-dimension-based generalization bounds for adaptation in classification tasks. Perhaps, the most significant contribution of this work was the definition and application of a distance between distributions, the d_A distance, which is particularly relevant to the problem of domain adaptation and can be estimated from finite samples for a finite VC dimension, as previously shown by Kifer et al. (2004). This work was later extended by Blitzer et al. (2008) who also gave a bound on the error rate of a hypothesis derived from a weighted combination of the source data sets for the specific case of empirical risk minimization. A theoretical study of domain adaptation was also presented by Mansour et al. (2009), where the analysis deals with the related but distinct case of adaptation with multiple sources, and where the target is a mixture of the source distributions.

This paper presents a new theoretical and algorithmic analysis of the problem of domain adaptation. It builds on the work of Ben-David et al. (2007) and extends it in several ways. We introduce a novel distance, the *discrepancy distance*, that is tailored to comparing distributions in adaptation. This distance coincides with the d_A distance for 0-1 classification, but it can be used to compare distributions for more general tasks, including regression, and with other loss functions. As already pointed out, a crucial advantage of the

d_A distance is that it can be estimated from finite samples when the set of regions used has finite VC-dimension. We prove that the same holds for the discrepancy distance and in fact give data-dependent versions of that statement with sharper bounds based on the Rademacher complexity.

We give new generalization bounds for domain adaptation and point out some of their benefits by comparing them with previous bounds. We further combine these with the properties of the discrepancy distance to derive data-dependent Rademacher complexity learning bounds. We also present a series of novel results for large classes of regularization-based algorithms, including support vector machines (SVMs) (Cortes & Vapnik, 1995) and kernel ridge regression (KRR) (Saunders et al., 1998). We compare the pointwise loss of the hypothesis returned by these algorithms when trained on a sample drawn from the target domain distribution, versus that of a hypothesis selected by these algorithms when training on a sample drawn from the source distribution. We show that the difference of these pointwise losses can be bounded by a term that depends directly on the empirical discrepancy distance of the source and target distributions.

These learning bounds motivate the idea of replacing the empirical source distribution with another distribution with the same support but with the smallest discrepancy with respect to the target empirical distribution, which can be viewed as reweighting the loss on each labeled point. We analyze the problem of determining the distribution minimizing the discrepancy in both 0-1 classification and square loss regression. We show how the problem can be cast as a linear program (LP) for the 0-1 loss and derive a specific efficient combinatorial algorithm to solve it in dimension one. We also give a polynomial-time algorithm for solving this problem in the case of the square loss by proving that it can be cast as a semi-definite program (SDP). Finally, we report the results of preliminary experiments showing the benefits of our analysis and discrepancy minimization algorithms.

In section 2, we describe the learning set-up for domain adaptation and introduce the notation and Rademacher complexity concepts needed for the presentation of our results. Section 3 introduces the discrepancy distance and analyzes its properties. Section 4 presents our generalization bounds and our theoretical guarantees for regularization-based algorithms. Section 5 describes and analyzes our discrepancy minimization algorithms. Section 6 reports the results of our preliminary experiments.

2 Preliminaries

2.1 Learning Set-Up

We consider the familiar supervised learning setting where the learning algorithm receives a sample of m labeled points $\mathcal{S} = (z_1, \ldots, z_m) = ((x_1, y_1), \ldots, (x_m, y_m)) \in (X \times Y)^m$, where X is the input space and Y the label set, which is $\{0, 1\}$ in classification and some measurable subset of $\mathbb{R}$ in regression.

In the *domain adaptation problem*, the training sample $\mathcal{S}$ is drawn according to a *source distribution* Q, while test points are drawn according to a *target distribution* P that may somewhat differ from Q. We denote by $f\colon X \to Y$ the target labeling function. We shall also discuss cases where the source labeling function f_Q differs from the target domain labeling function f_P. Clearly, this dissimilarity will need to be small for adaptation to be possible.

We will assume that the learner is provided with an unlabeled sample $\mathcal{T}$ drawn i.i.d. according to the target distribution P. We denote by $L\colon Y \times Y \to \mathbb{R}$ a loss function defined over pairs of labels and by $\mathcal{L}_Q(f, g)$ the expected loss for any two functions $f, g\colon X \to Y$ and any distribution Q over X: $\mathcal{L}_Q(f, g) = \mathrm{E}_{x \sim Q}[L(f(x), g(x))]$.

The domain adaptation problem consists of selecting a hypothesis h out of a hypothesis set H with a small expected loss according to the target distribution P, $\mathcal{L}_P(h, f)$.

2.2 Rademacher Complexity

Our generalization bounds will be based on the following data-dependent measure of the complexity of a class of functions.

Definition 1 (Rademacher Complexity) *Let H be a set of real-valued functions defined over a set X. Given a sample $S \in X^m$, the empirical Rademacher complexity of H is defined as follows:*

$$\widehat{\mathfrak{R}}_S(H) = \frac{2}{m} \mathop{\mathrm{E}}_{\sigma} \Big[\sup_{h \in H} \Big| \sum_{i=1}^{m} \sigma_i h(x_i) \Big| \,\Big| S = (x_1, \ldots, x_m) \Big]. \tag{1}$$

The expectation is taken over $\sigma = (\sigma_1, \ldots, \sigma_n)$ where σ_is are independent uniform random variables taking values in $\{-1, +1\}$. The Rademacher complexity of a hypothesis set H is defined as the expectation of $\widehat{\mathfrak{R}}_S(H)$ over all samples of size m:

$$\mathfrak{R}_m(H) = \mathop{\mathrm{E}}_{S} \big[\widehat{\mathfrak{R}}_S(H) \big| |S| = m \big]. \tag{2}$$

The Rademacher complexity measures the ability of a class of functions to fit noise. The empirical Rademacher complexity has the added advantage that it is data-dependent and can be measured from finite samples. It can lead to tighter bounds than those based on other measures of complexity such as the VC-dimension (Koltchinskii & Panchenko, 2000).

We will denote by $\widehat{R}_S(h)$ the empirical average of a hypothesis $h\colon X \to \mathbb{R}$ and by $R(h)$ its expectation over a sample S drawn according to the distribution considered. The following is a version of the Rademacher complexity bounds by Koltchinskii and Panchenko (2000) and Bartlett and Mendelson (2002). For completeness, the full proof is given in the Appendix.

Theorem 2 (Rademacher Bound) *Let H be a class of functions mapping $Z = X \times Y$ to $[0, 1]$ and $\mathcal{S} = (z_1, \ldots, z_m)$ a finite sample drawn i.i.d. according to a distribution Q. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{S}$ of size m, the following inequality holds for all $h \in H$:*

$$R(h) \leq \widehat{R}(h) + \widehat{\mathfrak{R}}_{\mathcal{S}}(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{3}$$

3 Distances between Distributions

Clearly, for generalization to be possible, the distribution Q and P must not be too dissimilar, thus some measure of the similarity of these distributions will be critical in the derivation of our generalization bounds or the design of our algorithms. This section discusses this question and introduces a *discrepancy* distance relevant to the context of adaptation.

The l_1 distance yields a straightforward bound on the difference of the error of a hypothesis h with respect to Q versus its error with respect to P.

Proposition 1 *Assume that the loss L is bounded, $L \leq M$ for some $M > 0$. Then, for any hypothesis $h \in H$,*

$$|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| \leq M\, l_1(Q, P). \tag{4}$$

This provides us with a first adaptation bound suggesting that for small values of the l_1 distance between the source and target distributions, the average loss of hypothesis h tested on the target domain is close to its average loss on the source domain. However, in general, this bound is not informative since the l_1 distance can be large even in favorable adaptation situations. Instead, one can use a distance between distributions better suited to the learning task.

Consider for example the case of classification with the 0-1 loss. Fix $h \in H$, and let a denote the support of $|h - f|$. Observe that $|\mathcal{L}_Q(h, f) - \mathcal{L}_P(h, f)| = |Q(a) - P(a)|$. A natural distance between distributions in this context is thus one based on the supremum of the right-hand side over all regions a. Since the target hypothesis f is not known, the region a should be taken as the support of $|h - h'|$ for any two $h, h' \in H$.

This leads us to the following definition of a distance originally introduced by Devroye et al. (1996) [pp. 271-272] under the name of *generalized Kolmogorov-Smirnov distance*, later by Kifer et al. (2004) as *the d_A distance*, and introduced and applied to the analysis of adaptation in classification by Ben-David et al. (2007) and Blitzer et al. (2008).

Definition 3 (d_A-Distance) *Let $A \subseteq 2^{|X|}$ be a set of subsets of X. Then, the d_A-*distance *between two distributions Q_1 and Q_2 over X, is defined as*

$$d_A(Q_1, Q_2) = \sup_{a \in A} |Q_1(a) - Q_2(a)|. \tag{5}$$

As just discussed, in 0-1 classification, a natural choice for A is $A = H\Delta H = \{|h' - h| \colon h, h' \in H\}$. We introduce a distance between distributions, *discrepancy distance*, that can be used to compare distributions for more general tasks, e.g., regression. Our choice of the terminology is partly motivated by the relationship of this notion with the discrepancy problems arising in combinatorial contexts (Chazelle, 2000).

Definition 4 (Discrepancy Distance) *Let H be a set of functions mapping X to Y and let $L \colon Y \times Y \to \mathbb{R}_+$ define a loss function over Y. The discrepancy distance disc_L between two distributions Q_1 and Q_2 over X is defined by*

$$\mathrm{disc}_L(Q_1, Q_2) = \max_{h, h' \in H} \left| \mathcal{L}_{Q_1}(h', h) - \mathcal{L}_{Q_2}(h', h) \right|.$$

The discrepancy distance is clearly symmetric and it is not hard to verify that it verifies the triangle inequality, regardless of the loss function used. In general, however, it does not define a *distance*: we may have $\mathrm{disc}_L(Q_1, Q_2) = 0$ for $Q_1 \neq Q_2$, even for non-trivial hypothesis sets such as that of bounded linear functions and standard continuous loss functions.

Note that for the 0-1 classification loss, the discrepancy distance coincides with the d_A distance with $A = H\Delta H$. But the discrepancy distance helps us compare distributions for other losses such as $L_q(y, y') = |y - y'|^q$ for some q and is more general.

As shown by Kifer et al. (2004), an important advantage of the d_A distance is that it can be estimated from finite samples when A has finite VC-dimension. We prove that the same holds for the disc_L distance and in fact give data-dependent versions of that statement with sharper bounds based on the Rademacher complexity.

The following theorem shows that for a bounded loss function L, the discrepancy distance disc_L between a distribution and its empirical distribution can be bounded in terms of the empirical Rademacher complexity of the class of functions $L_H = \{x \mapsto L(h'(x), h(x)) \colon h, h' \in H\}$. In particular, when L_H has finite pseudo-dimension, this implies that the discrepancy distance converges to zero as $O(\sqrt{\log m / m})$.

Proposition 2 *Assume that the loss function L is bounded by $M > 0$. Let Q be a distribution over X and let $\widehat{Q}$ denote the corresponding empirical distribution for a sample $\mathcal{S} = (x_1, \ldots, x_m)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{S}$ of size m drawn according to Q:*

$$\mathrm{disc}_L(Q, \widehat{Q}) \leq \widehat{\mathfrak{R}}_{\mathcal{S}}(L_H) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{6}$$

Proof: We scale the loss L to $[0, 1]$ by dividing by M, and denote the new class by L_H/M. By Theorem 2 applied to L_H/M, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $h, h' \in H$:

$$\frac{\mathcal{L}_Q(h', h)}{M} \leq \frac{\mathcal{L}_{\widehat{Q}}(h', h)}{M} + \widehat{\mathfrak{R}}_{\mathcal{S}}(L_H/M) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

The empirical Rademacher complexity has the property that $\widehat{\mathfrak{R}}(\alpha H) = \alpha \widehat{\mathfrak{R}}(H)$ for any hypothesis class H and positive real number α (Bartlett & Mendelson, 2002). Thus, $\mathfrak{R}_{\mathcal{S}}(L_H/M) = \frac{1}{M}\mathfrak{R}_{\mathcal{S}}(L_H)$, which proves the proposition. ■

For the specific case of L_q regression losses, the bound can be made more explicit.

Corollary 5 *Let H be a hypothesis set bounded by some $M > 0$ for the loss function L_q: $L_q(h, h') \leq M$, for all $h, h' \in H$. Let Q be a distribution over X and let $\widehat{Q}$ denote the corresponding empirical distribution for a sample $\mathcal{S} = (x_1, \ldots, x_m)$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{S}$ of size m drawn according to Q:*

$$\mathrm{disc}_{L_q}(Q, \widehat{Q}) \leq 4q\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + 3M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{7}$$

Proof: The function $f\colon x \mapsto x^q$ is q-Lipschitz for $x \in [0,1]$:

$$|f(x') - f(x)| \leq q|x' - x|, \tag{8}$$

and $f(0) = 0$. For $L = L_q$, $L_H = \{x \mapsto |h'(x) - h(x)|^q \colon h, h' \in H\}$. Thus, by Talagrand's contraction lemma (Ledoux & Talagrand, 1991), $\widehat{\mathfrak{R}}(L_H)$ is bounded by $2q\widehat{\mathfrak{R}}(H')$ with $H' = \{x \mapsto (h'(x) - h(x)) \colon h, h' \in H\}$. Then, $\widehat{\mathfrak{R}}_{\mathcal{S}}(H')$ can be written and bounded as follows

$$\begin{aligned}
\widehat{\mathfrak{R}}_{\mathcal{S}}(H') &= \mathop{\mathrm{E}}_{\sigma}\big[\sup_{h,h'} \frac{1}{m}|\sum_{i=1}^{m} \sigma_i(h(x_i) - h'(x_i))|\big] \\
&\leq \mathop{\mathrm{E}}_{\sigma}[\sup_{h} \frac{1}{m}|\sum_{i=1}^{m} \sigma_i h(x_i)|] + \mathop{\mathrm{E}}_{\sigma}[\sup_{h'} \frac{1}{m}|\sum_{i=1}^{m} \sigma_i h'(x_i)|] \\
&= 2\widehat{\mathfrak{R}}_{\mathcal{S}}(H),
\end{aligned}$$

using the definition of the Rademacher variables and the subadditivity of the supremum function. This proves the inequality $\widehat{\mathfrak{R}}(L_H) \leq 4q\widehat{\mathfrak{R}}(H)$ and the corollary. ∎

A very similar proof gives the following result for classification.

Corollary 6 *Let H be a set of classifiers mapping X to $\{0,1\}$ and let L_{01} denote the 0-1 loss. Then, with the notation of Corollary 5, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{S}$ of size m drawn according to Q:*

$$\mathrm{disc}_{L_{01}}(Q, \widehat{Q}) \leq 4\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{9}$$

The factor of 4 can in fact be reduced to 2 in these corollaries when using a more favorable constant in the contraction lemma. The following corollary shows that the discrepancy distance can be estimated from finite samples.

Corollary 7 *Let H be a hypothesis set bounded by some $M > 0$ for the loss function L_q: $L_q(h, h') \leq M$, for all $h, h' \in H$. Let Q be a distribution over X and $\widehat{Q}$ the corresponding empirical distribution for a sample $\mathcal{S}$, and let P be a distribution over X and $\widehat{P}$ the corresponding empirical distribution for a sample $\mathcal{T}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over samples $\mathcal{S}$ of size m drawn according to Q and samples $\mathcal{T}$ of size n drawn according to P:*

$$\mathrm{disc}_{L_q}(P, Q) \leq \mathrm{disc}_{L_q}(\widehat{P}, \widehat{Q}) + 4q\Big(\widehat{\mathfrak{R}}_{\mathcal{S}}(H) + \widehat{\mathfrak{R}}_{\mathcal{T}}(H)\Big) + 3M\left(\sqrt{\frac{\log \frac{4}{\delta}}{2m}} + \sqrt{\frac{\log \frac{4}{\delta}}{2n}}\right).$$

Proof: By the triangle inequality, we can write

$$\mathrm{disc}_{L_q}(P, Q) \leq \mathrm{disc}_{L_q}(P, \widehat{P}) + \mathrm{disc}_{L_q}(\widehat{P}, \widehat{Q}) + \mathrm{disc}_{L_q}(Q, \widehat{Q}). \tag{10}$$

The result then follows by the application of Corollary 5 to $\mathrm{disc}_{L_q}(P, \widehat{P})$ and $\mathrm{disc}_{L_q}(Q, \widehat{Q})$. ∎

As with Corollary 6, a similar result holds for the 0-1 loss in classification.

4 Domain Adaptation: Generalization Bounds

This section presents generalization bounds for domain adaptation given in terms of the discrepancy distance just defined. In the context of adaptation, two types of questions arise:

(1) we may ask, as for standard generalization, how the average loss of a hypothesis on the target distribution, $\mathcal{L}_P(h, f)$, differs from $\mathcal{L}_{\widehat{Q}}(h, f)$, its empirical error based on the empirical distribution $\widehat{Q}$;

(2) another natural question is, given a specific learning algorithm, by how much does $\mathcal{L}_P(h_Q, f)$ deviate from $\mathcal{L}_P(h_P, f)$ where h_Q is the hypothesis returned by the algorithm when trained on a sample drawn from Q and h_P the one it would have returned by training on a sample drawn from the true target distribution P.

We will present theoretical guarantees addressing both questions.

4.1 Generalization bounds

Let $h^*_Q \in \mathrm{argmin}_{h \in H} \mathcal{L}_Q(h, f_Q)$ and similarly let h^*_P be a minimizer of $\mathcal{L}_P(h, f_P)$. Note that these minimizers may not be unique. For adaptation to succeed, it is natural to assume that the average loss $\mathcal{L}_Q(h^*_Q, h^*_P)$ between the best-in-class hypotheses is small. Under that assumption and for a small discrepancy distance, the following theorem provides a useful bound on the error of a hypothesis with respect to the target domain.

Theorem 8 *Assume that the loss function L is symmetric and obeys the triangle inequality. Then, for any hypothesis $h \in H$, the following holds*

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_P(h^*_P, f_P) + \mathcal{L}_Q(h, h^*_Q) + \mathrm{disc}(P, Q) + \mathcal{L}_Q(h^*_Q, h^*_P). \tag{11}$$

Proof: Fix $h \in H$. By the triangle inequality property of L and the definition of the discrepancy $\mathrm{disc}_L(P, Q)$, the following holds

$$\begin{aligned}
\mathcal{L}_P(h, f_P) &\leq \mathcal{L}_P(h, h^*_Q) + \mathcal{L}_P(h^*_Q, h^*_P) + \mathcal{L}_P(h^*_P, f_P) \\
&\leq \mathcal{L}_Q(h, h^*_Q) + \mathrm{disc}_L(P, Q) + \mathcal{L}_P(h^*_Q, h^*_P) \\
&\quad + \mathcal{L}_P(h^*_P, f_P).
\end{aligned}$$ ∎

We compare (11) with the main adaptation bound given by Ben-David et al. (2007) and Blitzer et al. (2008):

$$\mathcal{L}_P(h, f_P) \leq \mathcal{L}_Q(h, f_Q) + \mathrm{disc}_L(P, Q) + \min_{h \in H}\big(\mathcal{L}_Q(h, f_Q) + \mathcal{L}_P(h, f_P)\big). \tag{12}$$

It is very instructive to compare the two bounds. Intuitively, the bound of Theorem 8 has only one error term that involves the target function, while the bound of (12) has three terms involving the target function. One extreme case is when there is a single hypothesis h in H and a single target function f. In this case, Theorem 8 gives a bound of $\mathcal{L}_P(h, f) + \mathrm{disc}(P, Q)$, while the bound supplied by (12) is $2\mathcal{L}_Q(h, f) + \mathcal{L}_P(h, f) + \mathrm{disc}(P, Q)$, which is larger than $3\mathcal{L}_P(h, f) +$

$\mathrm{disc}(P,Q)$ when $\mathcal{L}_Q(h,f) \leq \mathcal{L}_P(h,f)$. One can even see that the bound of (12) might become vacuous for moderate values of $\mathcal{L}_Q(h,f)$ and $\mathcal{L}_P(h,f)$. While this is clearly an extreme case, an error with a factor of 3 can arise in more realistic situations, especially when the distance between the target function and the hypothesis class is significant.

While in general the two bounds are incomparable, it is worthwhile to compare them using some relatively plausible assumptions. Assume that the discrepancy distance between P and Q is small and so is the average loss between h^*_Q and h^*_P. These are natural assumptions for adaptation to be possible. Then, Theorem 8 indicates that the regret $\mathcal{L}_P(h,f_P) - \mathcal{L}_P(h^*_P,f_P)$ is essentially bounded by $\mathcal{L}_Q(h,h^*_Q)$, the average loss with respect to h^*_Q on Q. We now consider several special cases of interest.

(i) When $h^*_Q = h^*_P$ then $h^* = h^*_Q = h^*_P$ and the bound of Theorem 8 becomes

$$\mathcal{L}_P(h,f_P) \leq \mathcal{L}_P(h^*,f_P) + \mathcal{L}_Q(h,h^*) + \mathrm{disc}(P,Q). \tag{13}$$

The bound of (12) becomes

$$\mathcal{L}_P(h,f_P) \leq \mathcal{L}_P(h^*,f_P) + \mathcal{L}_Q(h,f_Q) + \mathcal{L}_Q(h^*,f_Q) + \mathrm{disc}(P,Q),$$

where the right-hand side essentially includes the sum of 3 errors and is always larger than the right-hand side of (13) since by the triangle inequality $\mathcal{L}_Q(h,h^*) \leq \mathcal{L}_Q(h,f_Q) + \mathcal{L}_Q(h^*,f_Q)$.

(ii) When $h^*_Q = h^*_P = h^* \wedge \mathrm{disc}(P,Q) = 0$, the bound of Theorem 8 becomes

$$\mathcal{L}_P(h,f_P) \leq \mathcal{L}_P(h^*,f_P) + \mathcal{L}_Q(h,h^*),$$

which coincides with the standard generalization bound. The bound of (12) does not coincide with the standard bound and leads to:

$$\mathcal{L}_P(h,f_P) \leq \mathcal{L}_P(h^*,f_P) + \mathcal{L}_Q(h,f_Q) + \mathcal{L}_Q(h^*,f_Q).$$

(iii) When $f_P \in H$ (consistent case), the bound of (12) simplifies to,

$$|\mathcal{L}_P(h,f_P) - \mathcal{L}_Q(h,f_P)| \leq \mathrm{disc}_L(Q,P),$$

and it can also be derived using the proof of Theorem 8.

Finally, clearly Theorem 8 leads to bounds based on the empirical error of h on a sample drawn according to Q. We give the bound related to the 0-1 loss, others can be derived in a similar way from Corollaries 5-7 and other similar corollaries. The result follows Theorem 8 combined with Corollary 7, and a standard Rademacher classification bound (Bartlett & Mendelson, 2002).

Theorem 9 *Let H be a family of functions mapping X to $\{0,1\}$ and let the rest of the assumptions be as in Corollary 7. Then, for any hypothesis $h \in H$, with probability at least $1-\delta$, the following adaptation generalization bound holds for the 0-1 loss:*

$$\mathcal{L}_P(h,f_P) - \mathcal{L}_P(h^*_P,f_P) \leq \mathcal{L}_{\widehat{Q}}(h,h^*_Q) + \mathrm{disc}_{L_{01}}(\widehat{P},\widehat{Q}) + (4q+\frac{1}{2})\widehat{\mathfrak{R}}_S(H) + 4q\widehat{\mathfrak{R}}_T(H) + 4\sqrt{\frac{\log\frac{8}{\delta}}{2m}} + 3\sqrt{\frac{\log\frac{8}{\delta}}{2n}} + \mathcal{L}_Q(h^*_Q,h^*_P). \tag{14}$$

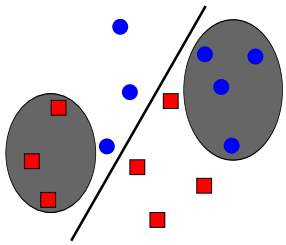


Figure 1: In this example, the gray regions are assumed to have zero support in the target distribution P. Thus, there exist consistent hypotheses such as the linear separator displayed. However, for the source distribution Q no linear separation is possible.

4.2 Guarantees for regularization-based algorithms

In this section, we first assume that the hypothesis set H includes the target function f_P. Note that this does not imply that f_Q is in H. Even when f_P and f_Q are restrictions to $\mathrm{supp}(P)$ and $\mathrm{supp}(Q)$ of the same labeling function f, we may have $f_P \in H$ and $f_Q \notin H$ and the source problem could be non-realizable. Figure 1 illustrates this situation.

For a fixed loss function L, we denote by $\widehat{R}_{\widehat{Q}}(h)$ the empirical error of a hypothesis h with respect to an empirical distribution $\widehat{Q}$: $R_{\widehat{Q}}(h) = \mathcal{L}_{\widehat{Q}}(h,f)$. Let $N\colon H \to \mathbb{R}_+$ be a function defined over the hypothesis set H. We will assume that H is a convex subset of a vector space and that the loss function L is convex with respect to each of its arguments. Regularization-based algorithms minimize an objective of the form

$$F_{\widehat{Q}}(h) = \widehat{R}_{\widehat{Q}}(h) + \lambda N(h), \tag{15}$$

where $\lambda \geq 0$ is a trade-off parameter. This family of algorithms includes support vector machines (SVM) (Cortes & Vapnik, 1995), support vector regression (SVR) (Vapnik, 1998), kernel ridge regression (Saunders et al., 1998), and other algorithms such as those based on the relative entropy regularization (Bousquet & Elisseeff, 2002).

We denote by B_F the Bregman divergence associated to a convex function F,

$$B_F(f\|g) = F(f) - F(g) - \langle f-g, \nabla F(g)\rangle \tag{16}$$

and define Δh as $\Delta h = h' - h$.

Lemma 10 *Let the hypothesis set H be a vector space. Assume that N is a proper closed convex function and that N and L are differentiable. Assume that $F_{\widehat{Q}}$ admits a minimizer $h \in H$ and $F_{\widehat{P}}$ a minimizer $h' \in H$ and that f_P and f_Q coincide on the support of $\widehat{Q}$. Then, the following bound holds,*

$$B_N(h'\|h) + B_N(h\|h') \leq \frac{2\mathrm{disc}_L(\widehat{P},\widehat{Q})}{\lambda}. \tag{17}$$

Proof: Since $B_{F_{\widehat{Q}}} = B_{\widehat{R}_{\widehat{Q}}} + \lambda B_N$ and $B_{F_{\widehat{P}}} = B_{\widehat{R}_{\widehat{P}}} + \lambda B_N$, and a Bregman divergence is non-negative, the following inequality holds:

$$\lambda\big(B_N(h'\|h) + B_N(h\|h')\big) \leq B_{F_{\widehat{Q}}}(h'\|h) + B_{F_{\widehat{P}}}(h\|h').$$

By the definition of h and h' as the minimizers of $F_{\widehat{Q}}$ and $F_{\widehat{P}}$, $\nabla_{\widehat{Q}} F(h) = \nabla_{\widehat{P}} F(h') = 0$ and

$$\begin{aligned} &B_{F_{\widehat{Q}}}(h' \| h) + B_{F_{\widehat{P}}}(h \| h') \\ &\quad = \widehat{R}_{\widehat{Q}}(h') - \widehat{R}_{\widehat{Q}}(h) + \widehat{R}_{\widehat{P}}(h) - \widehat{R}_{\widehat{P}}(h') \\ &\quad = \big(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{Q}}(h, f_P)\big) \\ &\qquad - \big(\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{Q}}(h', f_P)\big) \leq 2\mathrm{disc}_L(\widehat{P}, \widehat{Q}). \end{aligned}$$

This last inequality holds since by assumption f_P is in H. ■

We shall consider loss functions L for which there exists $\sigma \in \mathbb{R}_+$ such that $L(\cdot, y)$ is σ-Lipschitz for all $y \in Y$. This assumption holds for the hinge loss with $\sigma = 1$ and for the L_q loss with $\sigma = q(2M)^{q-1}$ when the hypothesis set and the set of output labels are bounded by some $M \in \mathbb{R}_+$: $\forall h \in H, \forall x \in X, |h(x)| \leq M$ and $\forall y \in Y, |y| \leq M$.

Theorem 11 *Let $K: X \times X \to \mathbb{R}$ be a positive-definite symmetric kernel such that $K(x, x) \leq \kappa^2 < \infty$ for all $x \in X$, and let H be the reproducing kernel Hilbert space associated to K. Assume that $L(\cdot, y)$ is σ-Lipschitz for all $y \in Y$. Let h' be the hypothesis returned by the regularization algorithm based on $N(\cdot) = \|\cdot\|_K^2$ for the empirical distribution $\widehat{P}$, and h the one returned for the empirical distribution $\widehat{Q}$, and assume that f_P and f_Q coincide on $\mathrm{supp}(\widehat{Q})$. Then, for all $x \in X$, $y \in Y$,*

$$\big|L(h'(x), y) - L(h(x), y)\big| \leq \kappa\sigma \sqrt{\frac{\mathrm{disc}_L(\widehat{P}, \widehat{Q})}{\lambda}}. \quad (18)$$

Proof: For $N(\cdot) = \|\cdot\|_K^2$, N is a proper closed convex function and is differentiable. We have $B_N(h' \| h) = \|h' - h\|_K^2$, thus $B_N(h' \| h) + B_N(h \| h') = 2\|\Delta h\|_K^2$. When L is differentiable, by Lemma 10,

$$2\|\Delta h\|_K^2 \leq \frac{2\mathrm{disc}_L(\widehat{P}, \widehat{Q})}{\lambda}. \quad (19)$$

This result can also be shown directly without assuming that L is differentiable by using the convexity of N and the minimizing properties of h and h' with a proof that is longer than that of Lemma 10.

Now, by the reproducing property of H, for all $x \in H$, $\Delta h(x) = \langle \Delta h, K(x, \cdot) \rangle$ and by the Cauchy-Schwarz inequality, $|\Delta h(x)| \leq \|\Delta h\|_K (K(x, x))^{1/2} \leq \kappa \|\Delta h\|_K$. Since for all $y \in Y$ $L(\cdot, y)$ is σ-Lipschitz, for all $x \in X$, $y \in Y$,

$$|L(h'(x), y) - L(h(x), y)| \leq \sigma |\Delta h(x)| \leq \kappa\sigma \|\Delta h\|_K,$$

which, combined with (19), proves the statement of the theorem. ■

Theorem 11 provides a strong guarantee on the pointwise difference of the loss for h' and h with probability one. The result, as well as the proof, also suggests that the discrepancy distance is the "right" measure of difference of distributions for this context. The theorem applies to a variety of algorithms, in particular SVMs combined with arbitrary PDS kernels and kernel ridge regression.

A similar result can be derived for the difference between expected losses by bounding the expectation of $\Delta h(x)$ in the proof, instead of its maximum. But, the resulting upper bound only differs from that of theorem by $\mathrm{E}_P[K(x, x)^{1/2}]$ versus $\max_x K(x, x)^{1/2}$, which, for a fixed kernel, are both constant terms and cannot be minimized.

In general, the functions f_P and f_Q may not coincide on $\mathrm{supp}(\widehat{Q})$. For adaptation to be possible, it is reasonable to assume however that

$$L_{\widehat{Q}}(f_Q(x), f_P(x)) \ll 1 \quad \text{and} \quad L_{\widehat{P}}(f_Q(x), f_P(x)) \ll 1.$$

This can be viewed as a condition on the proximity of the labeling functions (the Ys), while the discrepancy distance relates to the distributions on the input space (the Xs). The following result generalizes Theorem 11 to this setting in the case of the square loss.

Theorem 12 *Under the assumptions of Theorem 11, but with f_Q and f_P potentially different on $\mathrm{supp}(\widehat{Q})$, when L is the square loss L_2 and $\delta^2 = L_{\widehat{Q}}(f_Q(x), f_P(x)) \ll 1$, then, for all $x \in X$, $y \in Y$,*

$$\big|L(h'(x), y) - L(h(x), y)\big| \leq \frac{2\kappa M}{\lambda}\Big(\kappa\delta + \sqrt{\kappa^2\delta^2 + 4\lambda \mathrm{disc}_L(\widehat{P}, \widehat{Q})}\Big). \quad (20)$$

Proof: Proceeding as in the proof of Lemma 10 and using the definition of the square loss and the Cauchy-Schwarz inequality give

$$\begin{aligned} &B_{F_{\widehat{Q}}}(h' \| h) + B_{F_{\widehat{P}}}(h \| h') \\ &= \widehat{R}_{\widehat{Q}}(h') - \widehat{R}_{\widehat{Q}}(h) + \widehat{R}_{\widehat{P}}(h) - \widehat{R}_{\widehat{P}}(h') \\ &= \big(\mathcal{L}_{\widehat{P}}(h, f_P) - \mathcal{L}_{\widehat{Q}}(h, f_P)\big) \\ &\quad - \big(\mathcal{L}_{\widehat{P}}(h', f_P) - \mathcal{L}_{\widehat{Q}}(h', f_P)\big) \\ &\quad + 2 \operatorname*{E}_{\widehat{Q}}[(h'(x) - h(x))(f_P(x) - f_Q(x)] \\ &\leq 2\mathrm{disc}_L(\widehat{P}, \widehat{Q}) + 2\sqrt{\operatorname*{E}_{\widehat{Q}}[\Delta h(x)^2] \operatorname*{E}_{\widehat{Q}}[L(f_P(x), f_Q(x))]} \\ &\leq 2\mathrm{disc}_L(\widehat{P}, \widehat{Q}) + 2\kappa \|\Delta h\|_K \delta. \end{aligned}$$

Since $N(\cdot) = \|\cdot\|_K^2$, the inequality can be rewritten as

$$\lambda \|\Delta h\|_K^2 \leq \mathrm{disc}_L(\widehat{P}, \widehat{Q}) + \kappa\delta \|\Delta h\|_K. \quad (21)$$

Solving the second-degree polynomial in $\|\Delta h\|_K$ leads to the equivalent constraint

$$\|\Delta h\|_K \leq \frac{1}{2\lambda}\Big(\kappa\delta + \sqrt{\kappa^2\delta^2 + 4\lambda \mathrm{disc}_L(\widehat{P}, \widehat{Q})}\Big). \quad (22)$$

The result then follows by the σ-Lipschitzness of $L(\cdot, y)$ as in the proof of Theorem 11, with $\sigma = 4M$. ■

Using the same proof schema, similar bounds can be derived for other loss functions.

When the assumption $f_P \in H$ is relaxed, the following theorem holds.

Theorem 13 *Under the assumptions of Theorem 11, but with f_P not necessarily in H and f_Q and f_P potentially different on $\mathrm{supp}(\widehat{Q})$, when L is the square loss L_2 and $\delta' =$*

$L_{\widehat{Q}}(h^*_P(x), f_Q(x))^{1/2} + L_{\widehat{P}}(h^*_P(x), f_P(x))^{1/2} \ll 1$, *then, for all* $x \in X$, $y \in Y$,

$$\big|L(h'(x), y) - L(h(x), y)\big| \leq \frac{2\kappa M}{\lambda}\Big(\kappa\delta' + \sqrt{\kappa^2\delta'^2 + 4\lambda \mathrm{disc}_L(\widehat{P}, \widehat{Q})}\Big). \quad (23)$$

Proof: Proceeding as in the proof of Theorem 12 and using the definition of the square loss and the Cauchy-Schwarz inequality give

$$\begin{aligned}
&B_{F_{\widehat{Q}}}(h' \| h) + B_{F_{\widehat{P}}}(h \| h') \\
&= \big(\mathcal{L}_{\widehat{P}}(h, h^*_P) - \mathcal{L}_{\widehat{Q}}(h, h^*_P)\big) \\
&\quad - \big(\mathcal{L}_{\widehat{P}}(h', h^*_P) - \mathcal{L}_{\widehat{Q}}(h', h^*_P)\big) \\
&\quad - 2\operatorname*{E}_{\widehat{P}}[(h'(x) - h(x))(h^*_P(x) - f_P(x)] \\
&\quad + 2\operatorname*{E}_{\widehat{Q}}[(h'(x) - h(x))(h^*_P(x) - f_Q(x)] \\
&\leq 2\mathrm{disc}_L(\widehat{P}, \widehat{Q}) + 2\sqrt{\operatorname*{E}_{\widehat{P}}[\Delta h(x)^2] \operatorname*{E}_{\widehat{P}}[L(h^*_P(x), f_P(x))]} \\
&\quad + 2\sqrt{\operatorname*{E}_{\widehat{Q}}[\Delta h(x)^2] \operatorname*{E}_{\widehat{Q}}[L(h^*_P(x), f_Q(x))]} \\
&\leq 2\mathrm{disc}_L(\widehat{P}, \widehat{Q}) + 2\kappa \|\Delta h\|_K \delta'.
\end{aligned}$$

The rest of the proof is identical to that of Theorem 12. ■

5 Discrepancy Minimization Algorithms

The discrepancy distance $\mathrm{disc}_L(\widehat{P}, \widehat{Q})$ appeared as a critical term in several of the bounds in the last section. In particular, Theorems 11 and 12 suggest that if we could select, instead of $\widehat{Q}$, some other empirical distribution $\widehat{Q}'$ with a smaller empirical discrepancy $\mathrm{disc}_L(\widehat{P}, \widehat{Q}')$ and use that for training a regularization-based algorithm, a better guarantee would be obtained on the difference of pointwise loss between h' and h. Since h' is fixed, a sufficiently smaller discrepancy would actually lead to a hypothesis h with pointwise loss closer to that of h'.

The training sample is given and we do not have any control over the support of $\widehat{Q}$. But, we can search for the distribution $\widehat{Q}'$ with the minimal empirical discrepancy distance:

$$\widehat{Q}' = \operatorname*{argmin}_{\widehat{Q}' \in \mathcal{Q}} \mathrm{disc}_L(\widehat{P}, \widehat{Q}'), \quad (24)$$

where $\mathcal{Q}$ denotes the set of distributions with support $\mathrm{supp}(\widehat{Q})$. This leads to an optimization problem that we shall study in detail in the case of several loss functions.

Note that using $\widehat{Q}'$ instead of $\widehat{Q}$ for training can be viewed as *reweighting* the cost of an error on each training point. The distribution $\widehat{Q}'$ can be used to emphasize some points or de-emphasize others to reduce the empirical discrepancy distance. This bears some similarity with the reweighting or *importance weighting* ideas used in statistics and machine learning for sample bias correction techniques (Elkan, 2001; Cortes et al., 2008) and other purposes. Of course, the objective optimized here based on the discrepancy distance is distinct from that of previous reweighting techniques.

We will denote by S_Q the support of $\widehat{Q}$, by S_P the support of $\widehat{P}$, and by S their union $\mathrm{supp}(\widehat{Q}) \cup \mathrm{supp}(\widehat{P})$, with $|S_Q| = m_0 \leq m$ and $|S_P| = n_0 \leq n$.

In view of the definition of the discrepancy distance, problem (24) can be written as a min-max problem:

$$\widehat{Q}' = \operatorname*{argmin}_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)|. \quad (25)$$

As with all min-max problems, the problem has a natural game theoretical interpretation. However, here, in general, we cannot permute the $\min$ and $\max$ operators since the convexity-type assumptions of the minimax theorems do not hold. Nevertheless, since the max-min value is always a lower bound for the min-max, it provides us with a lower bound on the value of the game, that is the minimal discrepancy:

$$\begin{aligned}
&\max_{h, h' \in H} \min_{\widehat{Q}' \in \mathcal{Q}} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)| \leq \\
&\qquad \min_{\widehat{Q}' \in \mathcal{Q}} \max_{h, h' \in H} |\mathcal{L}_{\widehat{P}}(h', h) - \mathcal{L}_{\widehat{Q}'}(h', h)|. \quad (26)
\end{aligned}$$

We will later make use of this inequality. Let us now examine the minimization problem (24) and its algorithmic solutions in the case of classification with the 0-1 loss and regression with the L_2 loss.

5.1 Classification, 0-1 Loss

For the 0-1 loss, the problem of finding the best distribution $\widehat{Q}'$ can be reformulated as the following min-max program:

$$\min_{Q'} \max_{a \in H\Delta H} \big|\widehat{Q}'(a) - \widehat{P}(a)\big| \quad (27)$$

$$\text{subject to} \quad \forall x \in S_Q, \widehat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \widehat{Q}'(x) = 1, \quad (28)$$

where we have identified $H\Delta H = \{|h' - h| \colon h, h' \in H\}$ with the set of regions $a \subseteq X$ that are the support of an element of $H\Delta H$. This problem is similar to the min-max resource allocation problem that arises in task optimization (Karabati et al., 2001). It can be rewritten as the following linear program (LP):

$$\min_{Q'} \quad \delta \quad (29)$$

$$\begin{aligned}
\text{subject to} \quad & \forall a \in H\Delta H, \widehat{Q}'(a) - \widehat{P}(a) \leq \delta & (30)\\
& \forall a \in H\Delta H, \widehat{P}(a) - \widehat{Q}'(a) \leq \delta & (31)\\
& \forall x \in S_Q, \widehat{Q}'(x) \geq 0 \wedge \sum_{x \in S_Q} \widehat{Q}'(x) = 1. & (32)
\end{aligned}$$

The number of constraints is proportional to $|H\Delta H|$ but it can be reduced to a finite number by observing that two subsets $a, a' \in H\Delta H$ containing the same elements of S lead to redundant constraints, since

$$\big|\widehat{Q}'(a) - \widehat{P}(a)\big| = \big|\widehat{Q}'(a') - \widehat{P}(a')\big|. \quad (33)$$

Thus, it suffices to keep one canonical member a for each such equivalence class. The necessary number of constraints to be considered is proportional to $\Pi_{H\Delta H}(m_0 + n_0)$, the shattering coefficient of order $(m_0 + n_0)$ of the hypothesis

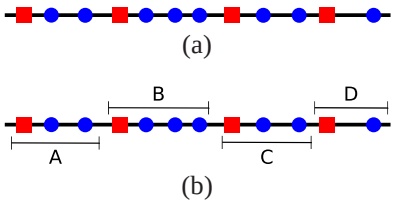


Figure 2: Illustration of the discrepancy minimization algorithm in dimension one. (a) Sequence of labeled (red) and unlabeled (blue) points. (b) The weight assigned to each labeled point is the sum of the weights of the consecutive blue points on its right.

class $H\Delta H$. By the Sauer's lemma, this is bounded in terms of the VC-dimension of the class $H\Delta H$, $\Pi_{H\Delta H}(m_0+n_0) \leq O((m_0+n_0)^{VC(H\Delta H)})$, which can be bounded by $O((m_0+n_0)^{2VC(H)})$ since it is not hard to see that $VC(H\Delta H) \leq 2VC(H)$.

In cases where we can test efficiently whether there exists a consistent hypothesis in H, e.g., for half-spaces in $\mathbb{R}^d$, we can generate in time $O((m_0+n_0)^{2d})$ all consistent labeling of the sample points by H. (We remark that computing the discrepancy with the 0-1 loss is closely related to agnostic learning. The implications of this fact will be described in a longer version of this paper.)

5.2 Computing the Discrepancy in 1D

We consider the case where $X = [0,1]$ and derive a simple algorithm for minimizing the discrepancy for 0-1 loss. Let H be the class of all prefixes (i.e., $[0,z]$) and suffixes (i.e., $[z,1]$). Our class of $H\Delta H$ includes all the intervals (i.e., $(z_1,z_2]$) and their complements (i.e., $[0,z_1] \cup (z_2,1]$). We start with a general lower bound on the discrepancy.

Let U denote the set of *unlabeled regions*, that is the set of regions a such that $a \cap S_Q = \emptyset$ and $a \cap S_P \neq \emptyset$. If a is an unlabeled region, then $|\widehat{Q}'(a) - \widehat{P}(a)| = \widehat{P}(a)$ for any $\widehat{Q}'$. Thus, by the max-min inequality (26), the following lower bound holds for the minimum discrepancy:

$$\max_{a \in U} \widehat{P}(a) \leq \min_{\widehat{Q}' \in \mathcal{Q}} \max_{h,h' \in H} |\mathcal{L}_{\widehat{P}}(h',h) - \mathcal{L}_{\widehat{Q}'}(h',h)|. \quad (34)$$

In particular, if there is a large unlabeled region a, we cannot hope to achieve a small empirical discrepancy.

In the one-dimensional case, we give a simple linear-time algorithm that does not require an LP and show that the lower bound (34) is reached. Thus, in that case, the $\min$ and $\max$ operators commute and the minimal discrepancy distance is precisely $\min_{a \in U} \widehat{P}(a)$.

Given our definition of H, the unlabeled regions are open intervals, or complements of these sets, containing only points from S_P with endpoints defined by elements of S_Q.

Let us denote by $s_1, \ldots, s_{m_0}$ the elements of S_Q, by n_i, $i \in [1, m_0]$, the number of consecutive unlabeled points to the right of s_i and $n = \sum n_i$. We will make an additional technical assumption that there are no unlabeled points to the left of s_1. Our algorithm consists of defining the weight $\widehat{Q}'(s_i)$ as follows:

$$\widehat{Q}'(s_i) = n_i/n. \quad (35)$$

This requires first sorting $S_Q \cup S_P$ and then computing n_i for each s_i. Figure 2 illustrates the algorithm.

Proposition 3 *Assume that X consists of the set of points on the real line and H the set of half-spaces on X. Then, for any $\widehat{Q}$ and $\widehat{P}$, $\widehat{Q}'(s_i) = n_i/n$ minimizes the empirical discrepancy and can be computed in time $O((m+n)\log(m+n))$.*

The proof is given in the Appendix.

5.3 Regression, L_2 loss

For the square loss, the problem of finding the best distribution can be written as

$$\min_{\widehat{Q}' \in \mathcal{Q}} \max_{h,h' \in H} \Big| \mathop{\mathrm{E}}_{\widehat{P}}[(h'(x)-h(x))^2] - \mathop{\mathrm{E}}_{\widehat{Q}'}[(h'(x)-h(x))^2] \Big|.$$

If X is a subset of $\mathbb{R}^N$, $N > 1$, and the hypothesis set H is a set of bounded linear functions $H = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \colon \|\mathbf{w}\| \leq 1\}$, then, the problem can be rewritten as

$$\begin{aligned}
&\min_{\widehat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \Big| \mathop{\mathrm{E}}_{\widehat{P}}[((\mathbf{w}'-\mathbf{w})^\top \mathbf{x})^2] - \mathop{\mathrm{E}}_{\widehat{Q}'}[((\mathbf{w}'-\mathbf{w})^\top \mathbf{x})^2] \Big| \\
&= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\substack{\|\mathbf{w}\| \leq 1 \\ \|\mathbf{w}'\| \leq 1}} \Big| \sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x}))[(\mathbf{w}'-\mathbf{w})^\top \mathbf{x}]^2 \Big| \\
&= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \Big| \sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x}))[\mathbf{u}^\top \mathbf{x}]^2 \Big| \\
&= \min_{\widehat{Q}' \in \mathcal{Q}} \max_{\|\mathbf{u}\| \leq 2} \Big| \mathbf{u}^\top \Big(\sum_{\mathbf{x} \in S} (\widehat{P}(\mathbf{x}) - \widehat{Q}'(\mathbf{x})) \mathbf{x}\mathbf{x}^\top \Big) \mathbf{u} \Big|. \quad (36)
\end{aligned}$$

We now simplify the notation and denote by $\mathbf{s}_1, \ldots, \mathbf{s}_{m_0}$ the elements of S_Q, by z_i the distribution weight at point $\mathbf{s}_i$: $z_i = \widehat{Q}'(\mathbf{s}_i)$, and by $\mathbf{M}(\mathbf{z}) \in \mathbb{S}^N$ a symmetric matrix that is an affine function of $\mathbf{z}$:

$$\mathbf{M}(\mathbf{z}) = \mathbf{M}_0 - \sum_{i=1}^{m_0} z_i \mathbf{M}_i, \quad (37)$$

where $\mathbf{M}_0 = \sum_{\mathbf{x} \in S} P(\mathbf{x}) \mathbf{x}\mathbf{x}^\top$ and $\mathbf{M}_i = \mathbf{s}_i \mathbf{s}_i^\top$. Since problem (36) is invariant to the non-zero bound on $\|\mathbf{u}\|$, we can equivalently write it with a bound of one and in view of the notation just introduced give its equivalent form

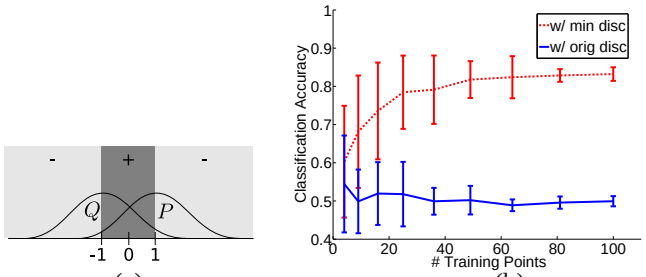
$$\min_{\substack{\|\mathbf{z}\|_1 = 1 \\ \mathbf{z} \geq 0}} \max_{\|\mathbf{u}\| = 1} |\mathbf{u}^\top \mathbf{M}(\mathbf{z}) \mathbf{u}|. \quad (38)$$

Since $\mathbf{M}(\mathbf{z})$ is symmetric, $\max_{\|\mathbf{u}\|=1} \mathbf{u}^\top \mathbf{M}(\mathbf{z}) \mathbf{u}$ is the maximum eigenvalue $\lambda_{\max}$ of $\mathbf{M}(\mathbf{z})$ and the problem is equivalent to the following maximum eigenvalue minimization for a symmetric matrix:

$$\min_{\substack{\|\mathbf{z}\|_1 = 1 \\ \mathbf{z} \geq 0}} \max\{\lambda_{\max}(\mathbf{M}(\mathbf{z})), \lambda_{\max}(-\mathbf{M}(\mathbf{z}))\}. \quad (39)$$

This is a convex optimization problem since the maximum eigenvalue of a matrix is a convex function of that matrix and $\mathbf{M}$ is an affine function of $\mathbf{z}$, and since $\mathbf{z}$ belongs to a simplex. The problem is equivalent to the following semi-definite programming (SDP) problem:

$$\begin{aligned}
\min_{\mathbf{z}, \lambda} \quad & \lambda & (40) \\
\text{subject to} \quad & \lambda \mathbf{I} - \mathbf{M}(\mathbf{z}) \succeq 0 & (41) \\
& \lambda \mathbf{I} + \mathbf{M}(\mathbf{z}) \succeq 0 & (42) \\
& \mathbf{1}^\top \mathbf{z} = 1 \wedge \mathbf{z} \geq 0. & (43)
\end{aligned}$$



(a) (b)

Figure 3: Example of application of the discrepancy minimization algorithm in dimensions one. (a) Source and target distributions Q and P. (b) Classification accuracy empirical results plotted as a function of the number of training points for both the *unweighted case* (using original empirical distribution $\widehat{Q}$) and the *weighted case* (using distribution $\widehat{Q}'$ returned by our discrepancy minimizing algorithm). The number of unlabeled points used was ten times the number of labeled. Error bars show ± 1 standard deviation.

SDP problems can be solved in polynomial time using general interior point methods (Nesterov & Nemirovsky, 1994). Thus, using the general expression of the complexity of interior point methods for SDPs, the following result holds.

Proposition 4 *Assume that X is a subset of $\mathbb{R}^N$ and that the hypothesis set H is a set of bounded linear functions $H = \{\mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x} \colon \|\mathbf{w}\| \leq 1\}$. Then, for any $\widehat{Q}$ and $\widehat{P}$, the discrepancy minimizing distribution $\widehat{Q}'$ for the square loss can be found in time $O(m_0^2 N^{2.5} + n_0 N^2)$.*

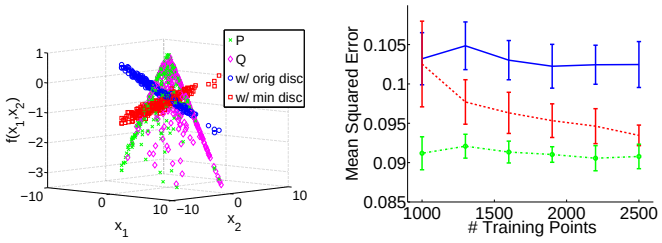
It is worth noting that the unconstrained version of this problem (no constraint on $\mathbf{z}$) and other close problems seem to have been studied by a number of optimization publications (Fletcher, 1985; Overton, 1988; Jarre, 1993; Helmberg & Oustry, 2000; Alizadeh, 1995). This suggests possibly more efficient specific algorithms than general interior point methods for solving this problem in the constrained case as well. Observe also that the matrices $\mathbf{M}_i$ have a specific structure in our case, they are rank-one matrices and in many applications quite sparse, which could be further exploited to improve efficiency.

As shown in a longer version of this paper, the results of this section can be extended to the case where H is a reproducing kernel Hilbert space associated to a positive definite symmetric kernel function K.

6 Experiments

This section reports the results of preliminary experiments showing the benefits of our discrepancy minimization algorithms. Our results confirm that our algorithm is effective in practice and produces a distribution that reduces the empirical discrepancy distance, which allows us to train on a sample closer to the target distribution with respect to this metric. They also demonstrate the accuracy benefits of this algorithm with respect to the target domain.

Figures 3(a)-(b) show the empirical advantages of using the distribution $\widehat{Q}'$ returned by the discrepancy minimizing algorithm described in Proposition 3 in a case where source and target distributions are shifted Gaussians: the source distribution is a Gaussian centered at -1 and the target distribution a Gaussian centered at $+1$, both with standard deviation 2. The hypothesis set used was the set of half-spaces and the target function selected to be the interval $[-1, 1]$. Thus, training on a sample drawn form Q generates a separator at -1 and errs on about half of the test points produced by P. In contrast, training with the distribution $\widehat{Q}'$ minimizing the empirical discrepancy yields a hypothesis separating the points at $+1$, thereby dramatically reducing the error rate.



(a) (b)

Figure 4: (a) An (x_1, x_2, y) plot of $\widehat{Q}$ (magenta), $\widehat{P}$ (green), weighted (red) and unweighted (blue) hypothesis. (b) Comparison of mean-squared error for the hypothesis trained on $\widehat{Q}$ (top), trained on $\widehat{Q}'$ (middle) and on $\widehat{P}$ (bottom) over a varying number of training points.

Figures 4(a)-(b) show the application of the SDP derived in (40) to determining the distribution minimizing the empirical discrepancy for ridge regression. In Figure 4(a), the distributions Q and P are Gaussians centered at $(\sqrt{2}, \sqrt{2})$ and $(-\sqrt{2}, -\sqrt{2})$, both with covariance matrix $2\mathbf{I}$. The target function is $f(x_1, x_2) = (1 - |x_1|) + (1 - |x_2|)$, thus the optimal linear prediction derived from Q has a negative slope, while the optimal prediction with respect to the target distribution P in fact has a positive slope. Figure 4(b) shows the performance of ridge regression when the example is extended to 16-dimensions, before and after minimizing the discrepancy. In this higher-dimension setting and even with several thousand points, using (http://sedumi.ie.lehigh.edu/), our SDP problem could be solved in about 15s using a single 3GHz processor with 2GB RAM. The SDP algorithm yields distribution weights that decrease the discrepancy and assist ridge regression in selecting a more appropriate hypothesis for the target distribution.

7 Conclusion

We presented an extensive theoretical and an algorithmic analysis of domain adaptation. Our analysis and algorithms are widely applicable and can benefit a variety of adaptation tasks. More efficient versions of these algorithms, in some instances efficient approximations, should further extend the applicability of our techniques to large-scale adaptation problems.

References

Alizadeh, F. (1995). Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5, 13–51.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and

Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 2002.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2007). Analysis of representations for domain adaptation. *Proceedings of NIPS 2006*.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2008). Learning bounds for domain adaptation. *Proceedings of NIPS 2007*.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *ACL 2007*.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *JMLR*, *2*, 499–526.

Chazelle, B. (2000). *The discrepancy method: randomness and complexity*. New York: Cambridge University Press.

Chelba, C., & Acero, A. (2006). Adaptation of maximum entropy capitalizer: Little data can help a lot. *Computer Speech & Language*, *20*, 382–399.

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. *Proceedings of ALT 2008*. Springer, Heidelberg, Germany.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*.

Daumé III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, *26*, 101–126.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer.

Dredze, M., Blitzer, J., Talukdar, P. P., Ganchev, K., Graca, J., & Pereira, F. (2007). Frustratingly Hard Domain Adaptation for Parsing. *CoNLL 2007*.

Elkan, C. (2001). The foundations of cost-sensitive learning. *IJCAI* (pp. 973–978).

Fletcher, R. (1985). On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Control and Optimization*, *23*, 493–513.

Gauvain, J.-L., & Chin-Hui (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*, 291–298.

Helmberg, C., & Oustry, F. (2000). Bundle methods to minimize the maximum eigenvalue function. In *Handbook of semidefinite programming: Theory, algorithms, and applications*. Kluwer Academic Publishers, Boston, MA.

Jarre, F. (1993). An interior-point method for minimizing the maximum eigenvalue of a linear combination of matrices. *SIAM J. Control Optim.*, *31*, 1360–1377.

Jelinek, F. (1998). *Statistical Methods for Speech Recognition*. The MIT Press.

Jiang, J., & Zhai, C. (2007). Instance Weighting for Domain Adaptation in NLP. *Proceedings of ACL 2007* (pp. 264–271). Association for Computational Linguistics.

Karabati, S., Kouvelis, P., & Yu, G. (2001). A min-max-sum resource allocation problem and its application. *Operations Research*, *49*, 913–922.

Kifer, D., Ben-David, S., & Gehrke, J. (2004). Detecting change in data streams. *Proceedings of the 30th International Conference on Very Large Data Bases*.

Koltchinskii, V., & Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In *High dimensional probability ii*, 443–459. preprint.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces: isoperimetry and processes*. Springer.

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

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009). Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems (2008)*.

Martínez, A. M. (2002). Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *IEEE Trans. Pattern Anal. Mach. Intell.*, *24*, 748–763.

Nesterov, Y., & Nemirovsky, A. (1994). *Interior point polynomial methods in convex programming: Theory and applications*. SIAM.

Overton, M. L. (1988). On minimizing the maximum eigenvalue of a symmetric matrix. *SIAM J. Matrix Anal. Appl.*, *9*, 256–268.

Pietra, S. D., Pietra, V. D., Mercer, R. L., & Roukos, S. (1992). Adaptive language modeling using minimum discriminant estimation. *HLT '91: Proceedings of the workshop on Speech and Natural Language* (pp. 103–106).

Roark, B., & Bacchiani, M. (2003). Supervised and unsupervised PCFG adaptation to novel domains. *Proceedings of HLT-NAACL*.

Rosenfeld, R. (1996). A Maximum Entropy Approach to Adaptive Statistical Language Modeling. *Computer Speech and Language*, *10*, 187–228.

Saunders, C., Gammerman, A., & Vovk, V. (1998). Ridge Regression Learning Algorithm in Dual Variables. *ICML* (pp. 515–521).

Valiant, L. G. (1984). *A theory of the learnable*. ACM Press New York, NY, USA.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons.

A Proof of Theorem 2

Theorem 14 (Rademacher Bound) *Let H be a class of functions mapping $Z = X \times Y$ to $[0,1]$ and $\mathcal{S} = (z_1, \ldots, z_m)$ a finite sample drawn i.i.d. according to a distribution Q. Then, for any $\delta > 0$, with probability at least $1-\delta$ over samples $\mathcal{S}$ of size m, the following inequality holds for all $h \in H$:*

$$R(h) \leq \widehat{R}(h) + \widehat{\mathfrak{R}}_{\mathcal{S}}(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{44}$$

Proof: Let $\Phi(\mathcal{S})$ be defined by $\Phi(\mathcal{S}) = \sup_{h \in H} R(h) - \widehat{R}(h)$. Changing a point of $\mathcal{S}$ affects $\Phi(\mathcal{S})$ by at most $1/m$. Thus, by McDiarmid's inequality applied to $\Phi(\mathcal{S})$, for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$, the following holds for all $h \in H$:

$$\Phi(\mathcal{S}) \leq \mathop{\mathrm{E}}_{\mathcal{S} \sim D}[\Phi(\mathcal{S})] + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{45}$$

$\mathrm{E}_{\mathcal{S} \sim D}[\Phi(\mathcal{S})]$ can be bounded in terms of the empirical Rademacher complexity as follows:

$$\begin{aligned}
&\mathop{\mathrm{E}}_{\mathcal{S}}[\Phi(\mathcal{S})] \\
&= \mathop{\mathrm{E}}_{\mathcal{S}}\big[\sup_{h \in H} \mathop{\mathrm{E}}_{\mathcal{S}'}[R_{\mathcal{S}'}(h)] - R_{\mathcal{S}}(h)\big] \\
&= \mathop{\mathrm{E}}_{\mathcal{S}}\big[\sup_{h \in H} \mathop{\mathrm{E}}_{\mathcal{S}'}[R_{\mathcal{S}'}(h) - R_{\mathcal{S}}(h)]\big] \\
&\leq \mathop{\mathrm{E}}_{\mathcal{S},\mathcal{S}'}\big[\sup_{h \in H} R_{\mathcal{S}'}(h) - R_{\mathcal{S}}(h)\big] \\
&= \mathop{\mathrm{E}}_{\mathcal{S},\mathcal{S}'}\big[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m}(h(x'_i) - h(x_i))\big] \\
&= \mathop{\mathrm{E}}_{\sigma,\mathcal{S},\mathcal{S}'}\big[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x'_i) - h(x_i))\big] \\
&\leq \mathop{\mathrm{E}}_{\sigma,\mathcal{S}'}\big[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x'_i)\big] + \mathop{\mathrm{E}}_{\sigma,\mathcal{S}}\big[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)\big] \\
&= 2 \mathop{\mathrm{E}}_{\sigma,\mathcal{S}}\big[\sup_{h \in H} \frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\big] \\
&\leq 2 \mathop{\mathrm{E}}_{\sigma,\mathcal{S}}\big[\sup_{h \in H} \big|\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\big|\big] \\
&= \mathfrak{R}_m(H).
\end{aligned}$$

Changing a point of $\mathcal{S}$ affects $\mathfrak{R}_m(H)$ by at most $2/m$. Thus, by McDiarmid's inequality applied to $\mathfrak{R}_m(H)$, with probability at least $1 - \delta/2$, the following holds:

$$\mathfrak{R}_m(H) \leq \widehat{\mathfrak{R}}_{\mathcal{S}}(H) + \sqrt{\frac{2\log \frac{2}{\delta}}{m}}. \tag{46}$$

Combining this inequality with Inequality (45) and the bound on $\mathrm{E}_{\mathcal{S}}[\Phi(\mathcal{S})]$ above yields directly the statement of the theorem. ■

B Proof of Proposition 3

Proposition 5 *Assume that X consists of the set of points on the real line and H the set of half-spaces on X. Then, for any $\widehat{Q}$ and $\widehat{P}$, $\widehat{Q}'(s_i) = n_i/n$ minimizes the empirical discrepancy and can be computed in time $O((m+n)\log(m+n))$.*

Proof: Consider an interval $[z_1, z_2]$ that maximizes the discrepancy of $\widehat{Q}'$. The case of a complement of an interval is the same, since the discrepancy of a hypothesis and its negation are identical. Let $s_i, \ldots, s_j \in [z_1, z_2]$ be the subset of $\widehat{Q}$ in that interval, and $p_{i'}, \ldots, p_{j'} \in [z_1, z_2]$ the subset of $\widehat{P}$ in that interval. The discrepancy is $d = |\sum_{k=i}^{j} \widehat{Q}'(s_k) - \frac{j'-i'}{n}|$. By our definition of $\widehat{Q}'$, we can write $\sum_{k=i}^{j} \widehat{Q}'(s_k) = \frac{1}{n}\sum_{k=i}^{j} n_k$. Let $p_{i''}$ be the maximal point in $\widehat{P}$ which is less than s_i and j'' the minimal point in $\widehat{P}$ larger than s_j. We have that $j' - i' = (i'' - i') + \sum_{k=i}^{j-1} n_k + (j'' - j')$. Therefore $d = |(i''-i') + (j''-j') - n_j| = |(i''-i') - (n_j - (j''-j'))|$. Since d is maximal and both terms are non-negative, one of them is zero. Since $j' - j'' \leq n_j$ and $i'' - i' \leq n_i$, the discrepancy of $\widehat{Q}'$ meets the lower bound of (34) and is thus optimal. ■