Mach Learn (2010) 79: 151–175
DOI 10.1007/s10994-009-5152-4

A theory of learning from different domains

Shai Ben-David · John Blitzer · Koby Crammer · Alex Kulesza · Fernando Pereira · Jennifer Wortman Vaughan

Received: 28 February 2009 / Revised: 12 September 2009 / Accepted: 18 September 2009 /
Published online: 23 October 2009

Abstract Discriminative learning methods for classification perform well when training and test data are drawn from the same distribution. Often, however, we have plentiful labeled training data from a *source* domain but wish to learn a classifier which performs well on a *target* domain with a different distribution and little or no labeled training data. In this work we investigate two questions. First, under what conditions can a classifier trained from source data be expected to perform well on target data? Second, given a small amount of labeled target data, how should we combine it during training with the large amount of labeled source data to achieve the lowest target error at test time?

Editors: Nicolo Cesa-Bianchi, David R. Hardoon, and Gayle Leen.

Preliminary versions of the work contained in this article appeared in *Advances in Neural Information Processing Systems* (Ben-David et al. 2006; Blitzer et al. 2007a).

S. Ben-David
David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada
e-mail: shai@cs.uwaterloo.ca

J. Blitzer (✉)
Department of Computer Science, UC Berkeley, Berkeley, CA, USA
e-mail: blitzer@cs.berkeley.edu

K. Crammer
Department of Electrical Engineering, The Technion, Haifa, Israel
e-mail: koby@ee.technion.ac.il

A. Kulesza
Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA
e-mail: kulesza@cis.upenn.edu

F. Pereira
Google Research, Mountain View, CA, USA
e-mail: pereira@google.com

J.W. Vaughan
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
e-mail: jenn@seas.harvard.edu

We address the first question by bounding a classifier's target error in terms of its source error and the divergence between the two domains. We give a classifier-induced divergence measure that can be estimated from finite, *unlabeled* samples from the domains. Under the assumption that there exists some hypothesis that performs well in both domains, we show that this quantity together with the empirical source error characterize the target error of a source-trained classifier.

We answer the second question by bounding the target error of a model which minimizes a convex combination of the empirical source and target errors. Previous theoretical work has considered minimizing just the source error, just the target error, or weighting instances from the two domains equally. We show how to choose the optimal combination of source and target error as a function of the divergence, the sample sizes of both domains, and the complexity of the hypothesis class. The resulting bound generalizes the previously studied cases and is always at least as tight as a bound which considers minimizing only the target error or an equal weighting of source and target errors.

Keywords Domain adaptation · Transfer learning · Learning theory · Sample-selection bias

1 Introduction

Most research in machine learning, both theoretical and empirical, assumes that models are trained and tested using data drawn from some fixed distribution. This single domain setting has been well studied, and uniform convergence theory guarantees that a model's empirical training error is close to its true error under such assumptions. In many practical cases, however, we wish to train a model in one or more *source* domains and then apply it to a different *target* domain. For example, we might have a spam filter trained from a large email collection received by a group of current users (the source domain) and wish to adapt it for a new user (the target domain). Intuitively this should improve filtering performance for the new user, under the assumption that users generally agree on what is spam and what is not. The challenge is that each user receives a unique distribution of email.

Many other examples arise in natural language processing. In general, labeled data for tasks like part-of-speech tagging (Ratnaparkhi 1996), parsing (Collins 1999), information extraction (Bikel et al. 1997), and sentiment analysis (Pang et al. 2002) are drawn from a limited set of document types and genres in a given language due to availability, cost, and specific goals of the project. However, useful applications for the trained systems may involve documents of different types or genres. We can hope to successfully adapt the systems in these cases since parts-of-speech, syntactic structure, entity mentions, and positive or negative sentiment are to a large extent stable across different domains, as they depend on general properties of language.

In this work we investigate the problem of *domain adaptation*. We analyze a setting in which we have plentiful labeled training data drawn from one or more *source* distributions but little or no labeled training data drawn from the *target* distribution of interest. This work answers two main questions. First, under what conditions on the source and target distributions can we expect to learn well? We give a bound on a classifier's target domain error in terms of its source domain error and a divergence measure between the two domains. In a distribution-free setting, we cannot obtain accurate estimates of common measures of divergence such as L_1 or Kullback-Leibler from finite samples. Instead, we show that when learning a hypothesis from a class of finite complexity, it is sufficient to use a classifier-induced divergence we call the $\mathcal{H}\Delta\mathcal{H}$-divergence (Kifer et al. 2004;

Ben-David et al. 2006). Finite sample estimates of the $\mathcal{H}\Delta\mathcal{H}$-divergence converge uniformly to the true $\mathcal{H}\Delta\mathcal{H}$-divergence, allowing us to estimate the domain divergence from *unlabeled* data in both domains. Our final bound on the target error is in terms of the empirical source error, the empirical $\mathcal{H}\Delta\mathcal{H}$-divergence between unlabeled samples from the domains, and the combined error of the best single hypothesis for both domains.

A second important question is how to learn when the large quantity of labeled source data is augmented with a small amount of labeled target data, for example, when our new email user has begun to manually mark a few received messages as spam. Given a source domain S and a target domain T, we consider hypotheses h which minimize a convex combination of empirical source and target error ($\hat{\epsilon}_T(h)$ and $\hat{\epsilon}_S(h)$, respectively), which we refer to as the empirical α-error:

$$\alpha\hat{\epsilon}_T(h) + (1-\alpha)\hat{\epsilon}_S(h).$$

Setting α involves trading off the ideal but small target dataset against the large (but less relevant) source dataset. Baseline choices for α include $\alpha = 0$ (using only source data) (Ben-David et al. 2006), $\alpha = 1$ (using only target data), and the equal weighting of source and target instances (Crammer et al. 2008), setting α to the fraction of the instances that are from the target domain. We give a bound on a classifier's target error in terms of its empirical α error. The α that minimizes the bound depends on the divergence between the domains as well as the size of the source and target training datasets. The optimal bound is always at least as tight as the bounds using only source, only target, or equally-weighted source and target instances. We show that for a real-world problem of sentiment classification, non-trivial settings of α perform better than the three baseline settings.

In the next section, we give a brief overview of related work. We then specify precisely our model of domain adaptation. Section 4 shows how to bound the target error of a hypothesis in terms of its source error and the source-target divergence. Section 5 gives our main result, a bound on the target error of a classifier which minimizes a convex combination of empirical errors on the two domains, and in Sect. 6 we investigate the properties of the best convex combination of that bound. In Sect. 7, we illustrate experimentally the above bounds on sentiment classification data. Section 8 describes how to extend the previous results to the case of multiple data sources. Finally, we conclude with a brief discussion of future directions for research in Sect. 9.

2 Related work

Crammer et al. (2008) introduced a PAC-style model of learning from multiple sources in which the distribution over input points is assumed to be the same across sources but each source may have its own deterministic labeling function. They derive bounds on the target error of the function that minimizes the empirical error on (uniformly weighted) data from any subset of the sources. As discussed in Sect. 8.2, the bounds that they derive are equivalent to ours in certain restricted settings, but their theory is significantly less general.

Daumé (2007) and Finkel (2009) suggest an empirically successful method for domain adaptation based on multi-task learning. The crucial difference between our domain adaptation setting and analyses of multi-task methods is that multi-task bounds require labeled data from each task, and make no attempt to exploit unlabeled data. Although these bounds have a more limited scope than ours, they can sometimes yield useful results even when the optimal predictors for each task (or domain in the case of Daumé 2007) are quite different (Baxter 2000; Ando and Zhang 2005).

Li and Bilmes (2007) give PAC-Bayesian learning bounds for adaptation using "divergence priors." In particular, they place a source-centered prior on the parameters of a model learned in the target domain. Like our model, the divergence prior emphasizes the tradeoff between source hypotheses trained on large (but biased) data sets and target hypotheses trained from small (but unbiased) data sets. In our model, however, we measure the divergence (and consequently the bias) of the source domain from unlabeled data. This allows us to choose a tradeoff parameter for source and target labeled data before training begins.

More recently, Mansour et al. (2009a, 2009b) introduced a theoretical model for the "multiple source adaptation problem." This model operates under assumptions very similar to our multiple source analysis (Sect. 8), and we address their work in more detail there.

Finally, domain adaptation is closely related to the setting of sample selection bias (Heckman 1979). A well-studied variant of this is covariate shift, which has seen significant work in recent years (Huang et al. 2007; Sugiyama et al. 2008; Cortes et al. 2008). This line of work leads to algorithms based on instance weighting, which have also been explored empirically in the machine learning and natural language processing communities (Jiang and Zhai 2007; Bickel et al. 2007). Our work differs from covariate shift primarily in two ways. First, we do not assume the labeling rule is identical for the source and target data (although there must exist some good labeling rule for both in order to achieve low error). Second, our $\mathcal{H}\Delta\mathcal{H}$-divergence can be computed from finite samples of unlabeled data, allowing us to directly estimate the error of a source-trained classifier on the target domain.

A point of general contrast is that we work in an agnostic setting in which we do not make strong assumptions about the data generation model, such as a specific relationship between the source and target data distributions, which would be needed to obtain absolute error bounds. Instead, we assume only that the samples from each of the two domains are generated i.i.d. according to the respective data distributions, and as a result our bounds must be relative to the error of some benchmark predictor rather than absolute, specifically, relative to the combined error on both domains of an optimal joint predictor.

3 A rigorous model of domain adaptation

We formalize the problem of domain adaptation for binary classification as follows. We define a *domain*[1] as a pair consisting of a distribution $\mathcal{D}$ on inputs $\mathcal{X}$ and a labeling function $f : \mathcal{X} \to [0, 1]$, which can have a fractional (expected) value when labeling occurs nondeterministically. Initially, we consider two domains, a *source* domain and a *target* domain. We denote by $\langle \mathcal{D}_S, f_S \rangle$ the source domain and $\langle \mathcal{D}_T, f_T \rangle$ the target domain.

A *hypothesis* is a function $h : \mathcal{X} \to \{0, 1\}$. The probability according to the distribution $\mathcal{D}_S$ that a hypothesis h disagrees with a labeling function f (which can also be a hypothesis) is defined as

$$\epsilon_S(h, f) = \mathrm{E}_{\mathbf{x} \sim \mathcal{D}_S}\left[|h(\mathbf{x}) - f(\mathbf{x})|\right].$$

When we want to refer to the source error (sometimes called *risk*) of a hypothesis, we use the shorthand $\epsilon_S(h) = \epsilon_S(h, f_S)$. We write the empirical source error as $\hat{\epsilon}_S(h)$. We use the parallel notation $\epsilon_T(h, f)$, $\epsilon_T(h)$, and $\hat{\epsilon}_T(h)$ for the target domain.

[1] Note that this notion of domain is *not* the domain of a function. We always mean a specific distribution and function pair when we say "domain."

4 A bound relating the source and target error

We now proceed to develop bounds on the target domain generalization performance of a classifier trained in the source domain. We first show how to bound the target error in terms of the source error, the difference between labeling functions f_S and f_T, and the divergence between the distributions $\mathcal{D}_S$ and $\mathcal{D}_T$. Since we expect the labeling function difference to be small in practice, we focus here on measuring distribution divergence, and especially on how to estimate it with finite samples of unlabeled data from $\mathcal{D}_S$ and $\mathcal{D}_T$. That is the role of the $\mathcal{H}$-divergence introduced in Sect. 4.1.

A natural measure of divergence for distributions is the L^1 or variation divergence

$$d_1(\mathcal{D}, \mathcal{D}') = 2 \sup_{B \in \mathcal{B}} |\Pr_{\mathcal{D}}[B] - \Pr_{\mathcal{D}'}[B]|,$$

where $\mathcal{B}$ is the set of measurable subsets under $\mathcal{D}$ and $\mathcal{D}'$. We make use of this measure to state an initial bound on the target error of a classifier.

Theorem 1 *For a hypothesis h,*

$$\epsilon_T(h) \leq \epsilon_S(h) + d_1(\mathcal{D}_S, \mathcal{D}_T) + \min\left\{\mathrm{E}_{\mathcal{D}_S}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right], \mathrm{E}_{\mathcal{D}_T}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right]\right\}.$$

Proof See Appendix. □

In this bound, the first term is the source error, which a training algorithm might seek to minimize, and the third is the difference in labeling functions across the two domains, which we expect to be small. The problem is the remaining term. Bounding the error in terms of the L^1 divergence between distributions has two disadvantages. First, it cannot be accurately estimated from finite samples of arbitrary distributions (Batu et al. 2000; Kifer et al. 2004) and therefore has limited usefulness in practice. Second, for our purposes the L^1 divergence is an overly strict measure that unnecessarily inflates the bound, since it involves a supremum over all measurable subsets. We are only interested in the error of a hypothesis from some class of finite complexity, thus we can restrict our attentions to the subsets on which this type of hypothesis can commit errors. The divergence measure introduced in the next section addresses both of these concerns.

4.1 The $\mathcal{H}$-divergence

Definition 1 (Based on Kifer et al. 2004) Given a domain $\mathcal{X}$ with $\mathcal{D}$ and $\mathcal{D}'$ probability distributions over $\mathcal{X}$, let $\mathcal{H}$ be a hypothesis class on $\mathcal{X}$ and denote by $I(h)$ the set for which $h \in \mathcal{H}$ is the characteristic function; that is, $\mathbf{x} \in I(h) \Leftrightarrow h(\mathbf{x}) = 1$. The *$\mathcal{H}$-divergence* between $\mathcal{D}$ and $\mathcal{D}'$ is

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{h \in \mathcal{H}} |\Pr_{\mathcal{D}}[I(h)] - \Pr_{\mathcal{D}'}[I(h)]|.$$

The $\mathcal{H}$-divergence resolves both problems associated with the L^1 divergence. First, for hypothesis classes $\mathcal{H}$ of finite VC dimension, the $\mathcal{H}$-divergence can be estimated from finite samples (see Lemma 1 below). Second, the $\mathcal{H}$-divergence for any class $\mathcal{H}$ is never larger than the L^1 divergence, and is in general smaller when $\mathcal{H}$ has finite VC dimension.

Since it plays an important role in the rest of this work, we now state a slight modification of Theorem 3.4 of Kifer et al. (2004) as a lemma.

Lemma 1 *Let $\mathcal{H}$ be a hypothesis space on $\mathcal{X}$ with VC dimension d. If $\mathcal{U}$ and $\mathcal{U}'$ are samples of size m from $\mathcal{D}$ and $\mathcal{D}'$ respectively and $\hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}')$ is the empirical $\mathcal{H}$-divergence between samples, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$d_{\mathcal{H}}(\mathcal{D},\mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') + 4\sqrt{\frac{d\log(2m)+\log(\frac{2}{\delta})}{m}}.$$

Lemma 1 shows that the empirical $\mathcal{H}$-divergence between two samples from distributions $\mathcal{D}$ and $\mathcal{D}'$ converges uniformly to the true $\mathcal{H}$-divergence for hypothesis classes $\mathcal{H}$ of finite VC dimension.

The next lemma shows that we can compute the $\mathcal{H}$-divergence by finding a classifier which attempts to separate one domain from the other. Our basic plan of attack will be as follows: Label each unlabeled source instance with 0 and unlabeled target instance as 1. Then train a classifier to discriminate between source and target instances. The $\mathcal{H}$-divergence is immediately computable from the error.

Lemma 2 *For a symmetric hypothesis class $\mathcal{H}$ (one where for every $h \in \mathcal{H}$, the inverse hypothesis $1-h$ is also in $\mathcal{H}$) and samples $\mathcal{U}, \mathcal{U}'$ of size m*

$$\hat{d}_{\mathcal{H}}(\mathcal{U},\mathcal{U}') = 2\left(1 - \min_{h\in\mathcal{H}}\left[\frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x}\in\mathcal{U}] + \frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=1} I[\mathbf{x}\in\mathcal{U}']\right]\right),$$

where $I[\mathbf{x}\in\mathcal{U}]$ is the binary indicator variable which is 1 when $\mathbf{x}\in\mathcal{U}$.

Proof See Appendix. □

This lemma leads directly to a procedure for computing the $\mathcal{H}$-divergence. We first find a hypothesis in $\mathcal{H}$ which has minimum error for the binary classification problem of distinguishing source from target instances. The error of this hypothesis is related to the $\mathcal{H}$-divergence by Lemma 2. Of course, minimizing error for most reasonable hypothesis classes is a computationally intractable problem. Nonetheless, as we shall see in Sect. 7, the error of hypotheses trained to minimize convex upper bounds on error are useful in approximating the $\mathcal{H}$-divergence.

4.2 Bounding the difference in error using the $\mathcal{H}$-divergence

The $\mathcal{H}$-divergence allows us to estimate divergence from unlabeled data, but in order to use it in a bound we must have tools to represent error relative to other hypotheses in our class. We introduce two new definitions.

Definition 2 The *ideal joint hypothesis* is the hypothesis which minimizes the combined error

$$h^* = \operatorname*{argmin}_{h\in\mathcal{H}} \epsilon_S(h) + \epsilon_T(h).$$

We denote the combined error of the ideal hypothesis by

$$\lambda = \epsilon_S(h^*) + \epsilon_T(h^*).$$

The ideal joint hypothesis explicitly embodies our notion of adaptability. When this hypothesis performs poorly, we cannot expect to learn a good target classifier by minimizing source error. On the other hand, we will show that if the ideal joint hypothesis performs well, we can measure adaptability of a source-trained classifier by using the $\mathcal{H}$-divergence between the marginal distributions $\mathcal{D}_S$ and $\mathcal{D}_T$.

Next we define the *symmetric difference hypothesis space* $\mathcal{H}\Delta\mathcal{H}$ for a hypothesis space $\mathcal{H}$, which will be very useful in reasoning about error.

Definition 3 For a hypothesis space $\mathcal{H}$, the *symmetric difference hypothesis space* $\mathcal{H}\Delta\mathcal{H}$ is the set of hypotheses

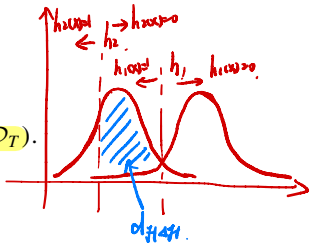
$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(\mathbf{x}) = h(\mathbf{x}) \oplus h'(\mathbf{x}) \quad \text{for some } h, h' \in \mathcal{H},$$

where $\oplus$ is the XOR function. In words, every hypothesis $g \in \mathcal{H}\Delta\mathcal{H}$ is the set of disagreements between two hypotheses in $\mathcal{H}$.

The following simple lemma shows how we can make use of the $\mathcal{H}\Delta\mathcal{H}$-divergence in bounding the error of our hypothesis.

Lemma 3 *For any hypotheses* $h, h' \in \mathcal{H}$,

$$|\epsilon_S(h, h') - \epsilon_T(h, h')| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T).$$



Proof By the definition of $\mathcal{H}\Delta\mathcal{H}$-distance,

$$\begin{aligned} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) &= 2 \sup_{h,h' \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S}\left[h(x) \neq h'(x)\right] - \Pr_{x \sim \mathcal{D}_T}\left[h(x) \neq h'(x)\right] \right| \\ &= 2 \sup_{h,h' \in \mathcal{H}} |\epsilon_S(h, h') - \epsilon_T(h, h')| \geq 2|\epsilon_S(h, h') - \epsilon_T(h, h')|. \end{aligned}$$

□

We are now ready to give a bound on target error in terms of the new divergence measure we have defined.

Theorem 2 *Let* $\mathcal{H}$ *be a hypothesis space of VC dimension* d. *If* $\mathcal{U}_S, \mathcal{U}_T$ *are unlabeled samples of size* m' *each, drawn from* $\mathcal{D}_S$ *and* $\mathcal{D}_T$ *respectively, then for any* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$ (*over the choice of the samples*), *for every* $h \in \mathcal{H}$:

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda.$$

Proof This proof relies on Lemma 3 and the triangle inequality for classification error (Ben-David et al. 2006; Crammer et al. 2008), which implies that for any labeling functions f_1, f_2, and f_3, we have $\epsilon(f_1, f_2) \leq \epsilon(f_1, f_3) + \epsilon(f_2, f_3)$. Then

$$\begin{aligned} \epsilon_T(h) &\leq \epsilon_T(h^*) + \epsilon_T(h, h^*) \\ &\leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + \left|\epsilon_T(h, h^*) - \epsilon_S(h, h^*)\right| \\ &\leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \end{aligned}$$

$$\leq \epsilon_T(h^*) + \epsilon_S(h) + \epsilon_S(h^*) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$$

$$= \epsilon_S(h) + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda$$

$$\leq \epsilon_S(h) + \frac{1}{2} \hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda.$$

The last step is an application of Lemma 1, together with the observation that since we can represent every $g \in \mathcal{H}\Delta\mathcal{H}$ as a linear threshold network of depth 2 with 2 hidden units, the VC dimension of $\mathcal{H}\Delta\mathcal{H}$ is at most twice the VC dimension of $\mathcal{H}$ (Anthony and Bartlett 1999). □

The bound in Theorem 2 is relative to λ, and we briefly comment that the form $\lambda = \epsilon_S(h^*) + \epsilon_T(h^*)$ comes from the use of the triangle inequality for classification error. Other losses result in other forms for this bound (Crammer et al. 2008). When the combined error of the ideal joint hypothesis is large, then there is no classifier that performs well on both the source and target domains, so we cannot hope to find a good target hypothesis by training only on the source domain. On the other hand, for small λ (the most relevant case for domain adaptation), the bound shows that source error and unlabeled $\mathcal{H}\Delta\mathcal{H}$-divergence are important quantities in computing the target error.

5 A learning bound combining source and target training data

Theorem 2 shows how to relate source and target error. We now proceed to give a learning bound for empirical risk minimization using combined source and target training data.

At train time a learner receives a sample $S = (S_T, S_S)$ of m instances, where S_T consists of βm instances drawn independently from $\mathcal{D}_T$ and S_S consists of $(1-\beta)m$ instances drawn independently from $\mathcal{D}_S$. The goal of a learner is to find a hypothesis that minimizes target error $\epsilon_T(h)$. When β is small, as in domain adaptation, minimizing empirical target error may not be the best choice. We analyze learners that instead minimize a convex combination of empirical source and target error,

$$\hat{\epsilon}_\alpha(h) = \alpha\hat{\epsilon}_T(h) + (1-\alpha)\hat{\epsilon}_S(h),$$

for some $\alpha \in [0, 1]$. We denote as $\epsilon_\alpha(h)$ the corresponding weighted combination of true source and target errors, measured with respect to $\mathcal{D}_S$ and $\mathcal{D}_T$.

We bound the target error of a domain adaptation algorithm that minimizes $\hat{\epsilon}_\alpha(h)$. The proof of the bound has two main components, which we state as lemmas below. First we bound the difference between the target error $\epsilon_T(h)$ and weighted error $\epsilon_\alpha(h)$. Then we bound the difference between the true and empirical weighted errors $\epsilon_\alpha(h)$ and $\hat{\epsilon}_\alpha(h)$.

Lemma 4 *Let h be a hypothesis in class $\mathcal{H}$. Then*

$$|\epsilon_\alpha(h) - \epsilon_T(h)| \leq (1-\alpha)\left(\frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda\right).$$

Proof See Appendix. □

The lemma shows that as α approaches 1, we rely increasingly on the target data, and the distance between domains matters less and less. The uniform convergence bound on the α-error is nearly identical to the standard uniform convergence bound for hypothesis classes of finite VC dimension (Vapnik 1998; Anthony and Bartlett 1999), only with target and source errors weighted differently. The key part of the proof relies on a slight modification of Hoeffding's inequality for our setup, which we state here:

Lemma 5 *For a fixed hypothesis h, if a random labeled sample of size m is generated by drawing βm points from $\mathcal{D}_T$ and $(1-\beta)m$ points from $\mathcal{D}_S$, and labeling them according to f_S and f_T respectively, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of the samples),*

$$\Pr\left[|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| \geq \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\right).$$

Before giving the proof, we first restate Hoeffding's inequality for completeness.

Proposition 1 (Hoeffding's inequality) *If $X_1, \ldots, X_n$ are independent random variables with $a_i \leq X_i \leq b_i$ for all i, then for any $\epsilon > 0$,*

$$\Pr\left[|\bar{X} - E[\bar{X}]| \geq \epsilon\right] \leq 2e^{-2n^2\epsilon^2 / \sum_{i=1}^n (b_i - a_i)^2},$$

where $\bar{X} = (X_1 + \cdots + X_n)/n$.

We are now ready to prove the lemma.

Proof (Lemma 5) Let $X_1, \ldots, X_{\beta m}$ be random variables that take on the values

$$\frac{\alpha}{\beta}|h(x) - f_T(x)|$$

for the βm instances $x \in S_T$. Similarly, let $X_{\beta m+1}, \ldots, X_m$ be random variables that take on the values

$$\frac{1-\alpha}{1-\beta}|h(x) - f_S(x)|$$

for the $(1-\beta)m$ instances $x \in S_S$. Note that $X_1, \ldots, X_{\beta m} \in [0, \alpha/\beta]$ and $X_{\beta m+1}, \ldots, X_m \in [0, (1-\alpha)/(1-\beta)]$. Then

$$\hat{\epsilon}_\alpha(h) = \alpha\hat{\epsilon}_T(h) + (1-\alpha)\hat{\epsilon}_S(h)$$

$$= \alpha\frac{1}{\beta m}\sum_{x \in S_T}|h(x) - f_T(x)| + (1-\alpha)\frac{1}{(1-\beta)m}\sum_{x \in S_S}|h(x) - f_S(x)| = \frac{1}{m}\sum_{i=1}^m X_i.$$

Furthermore, by linearity of expectations

$$E[\hat{\epsilon}_\alpha(h)] = \frac{1}{m}\left(\beta m\frac{\alpha}{\beta}\epsilon_T(h) + (1-\beta)m\frac{1-\alpha}{1-\beta}\epsilon_S(h))\right)$$

$$= \alpha\epsilon_T(h) + (1-\alpha)\epsilon_S(h) = \epsilon_\alpha(h).$$

So by Hoeffding's inequality the following holds for every h.

$$\begin{aligned}\Pr\left[|\hat{\epsilon}_\alpha(h) - \epsilon_\alpha(h)| \geq \epsilon\right] &\leq 2\exp\left(\frac{-2m^2\epsilon^2}{\sum_{i=1}^m \text{range}^2(X_i)}\right)\\ &= 2\exp\left(\frac{-2m^2\epsilon^2}{\beta m(\frac{\alpha}{\beta})^2 + (1-\beta)m(\frac{1-\alpha}{1-\beta})^2}\right)\\ &= 2\exp\left(\frac{-2m\epsilon^2}{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\right).\end{aligned}$$

□

This lemma shows that as α moves away from β (where each instance is weighted equally), our finite sample approximation to $\epsilon_\alpha(h)$ becomes less reliable. We can now move on to the main theorem of this section.

Theorem 3 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. Let $\mathcal{U}_S$ and $\mathcal{U}_T$ be unlabeled samples of size m' each, drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively. Let S be a labeled sample of size m generated by drawing βm points from $\mathcal{D}_T$ and $(1-\beta)m$ points from $\mathcal{D}_S$ and labeling them according to f_S and f_T, respectively. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_\alpha(h)$ on S and $h_T^* = \min_{h\in\mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0,1)$, with probability at least $1-\delta$ (over the choice of the samples),*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1)) + 2\log(\frac{8}{\delta})}{m}}$$
$$+ 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{8}{\delta})}{m'}} + \lambda\right).$$

The proof follows the standard set of steps for proving learning bounds (Anthony and Bartlett 1999), using Lemma 4 to bound the difference between target and weighted errors and Lemma 5 for the uniform convergence of empirical and true weighted errors. The full proof is in Appendix.

When $\alpha = 0$ (that is, we ignore target data), the bound is identical to that of Theorem 2, but with an empirical estimate for the source error. Similarly when $\alpha = 1$ (that is, we use only target data), the bound is the standard learning bound using only target data. At the optimal α (which minimizes the right hand side), the bound is always at least as tight as either of these two settings. Finally note that by choosing different values of α, the bound allows us to effectively trade off the small amount of target data against the large amount of less relevant source data.

We remark that when it is known that $\lambda = 0$, the dependence on m in Theorem 3 can be improved; this corresponds to the restricted or realizable setting.

6 Optimal mixing value

We examine now the bound of Theorem 3 in more detail to illustrate some interesting properties. Writing the bound as a function of α and omitting additive constants, we obtain

$$f(\alpha) = 2B\sqrt{\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + 2(1-\alpha)A, \tag{1}$$

where

$$A = \left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda\right),$$

is the total divergence between source and target, and

$$B = 4\sqrt{\frac{2d\log(2(m+1)) + 2\log(\frac{8}{\delta})}{m}}$$

is the complexity term, which is approximately $\sqrt{d/m}$. The optimal value α^* is a function of the number of target examples $m_T = \beta m$, the number of source examples $m_S = (1-\beta)m$, and the ratio $D = \sqrt{d}/A$:

$$\alpha^*(m_T, m_S; D) = \begin{cases} 1 & m_T \geq D^2 \\ \min\{1, \nu\} & m_T \leq D^2, \end{cases} \tag{2}$$

where

$$\nu = \frac{m_T}{m_T + m_S}\left(1 + \frac{m_S}{\sqrt{D^2(m_S + m_T) - m_S m_T}}\right).$$

Several observations follow from this analysis. First, if $m_T = 0$ ($\beta = 0$) then $\alpha^* = 0$ and if $m_S = 0$ ($\beta = 1$) then $\alpha^* = 1$. That is, if we have only source or only target data, the best combination is to use exactly what we have. Second, if we are certain that the source and target are the same, that is if $A = 0$ (or $D \to \infty$), then $\alpha^* = \beta$, that is, the optimal combination is to use the training data with uniform weighting of the examples across all examples, as in Crammer et al. (2008), who always enforce such a uniform weighing. Finally, two phase transitions occur in the value of α^*. First, if there are enough target data (specifically, if $m_T \geq D^2 = d/A^2$) then no source data are needed, and in fact using any source data will yield suboptimal performance. This is because the possible reduction in error due to additional source data is always less than the increase in error caused by the source data being too far from the target data. Second, even if there are few target examples, it might be the case that we do not have enough source data to justify using it, and this small amount of source data should be ignored. Once we have enough source data then we get a non-trivial value for α^*.

These two phase transitions are illustrated in Fig. 1. The intensity of a point reflects the value α^* and ranges from 0 (white) to 1 (black). In this plot α^* is a function of m_S (x axis) and m_T (y axis), and we fix the complexity to $d = 1{,}601$ and the divergence between source and target to $A = 0.715$. We chose these values to correspond more closely to real data (see Sect. 7). Observe first that $D^2 = 1{,}601/(0.715)^2 \approx 3{,}130$. When $m_T \geq D^2$, the first case of (2) predicts that $\alpha^* = 1$ for all values of m_S, which is illustrated by the black region above the line $m_T = 3{,}130$. Furthermore, fixing the value of $m_T \leq 3{,}130$, the second case of (2) predicts that α^* will be either one (1) if m_S is small enough, or go smoothly to zero as m_S increases. This is illustrated by any horizontal line with $m_T \leq 3{,}130$. Each such line is black for small values of m_S and then gradually becomes white as m_S increases (left to right).

7 Results on sentiment classification

In this section we illustrate our theory on the natural language processing task of sentiment classification (Pang et al. 2002). The point of these experiments is not to instantiate the

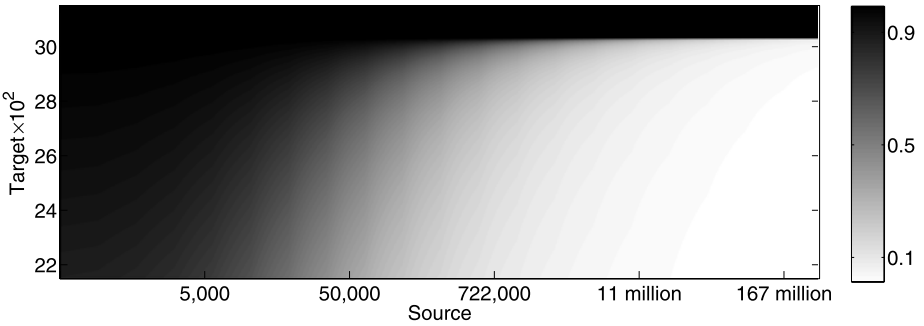


Fig. 1 An illustration of the phase transition in the balance between source and target training data. The value of α minimizing the bound is indicated by the intensity, where *black* means $\alpha = 1$. We fix $d = 1{,}601$ and $A = 0.715$, approximating the empirical setup in Fig. 3. The x-axis shows the number of source instances (log-scale). The y-axis shows the number of target instances. A phase transition occurs at 3,130 target instances. With more target instances than this, it is more effective to ignore even an infinite amount of source data

bound from Theorem 3 directly, since the amount of data we use here is much too small for the bound to yield meaningful numerical results. Instead, we show how the two main principles of our theory from Sects. 4 and 5 can be applied on a real-world problem. First, we show that an approximation to the $\mathcal{H}$-distance, obtained by training a linear model to discriminate between instances from different domains, correlates well with the loss incurred by training in one domain and testing in another. Second, we investigate minimizing the α-error as suggested by Theorem 3. We show that the optimal value of α for a given amount of source and target data is closely related to our approximate $\mathcal{H}$-distance.

The next subsection describes the problem of sentiment classification, along with our dataset, features, and learning algorithms. Then we show experimentally how our approximate $\mathcal{H}$-distance is related to the adaptation performance and the optimal value of α.

7.1 Sentiment classification

Given a piece of text (usually a review or essay), automatic sentiment classification is the task of determining whether the sentiment expressed by the text is positive or negative (Pang et al. 2002; Turney 2002). While movie reviews are the most commonly studied domain, sentiment analysis has been extended to a number of new domains, ranging from stock message boards to congressional floor debates (Das and Chen 2001; Thomas et al. 2006). Research results have been deployed industrially in systems that gauge market reaction and summarize opinion from Web pages, discussion boards, and blogs.

We used the publicly available data set from (Blitzer et al. 2007b) to examine our theory.[2] The data set consists of reviews from the Amazon website for several different types of products. We chose reviews from the domains *apparel*, *books*, *DVDs*, *kitchen & housewares*, and *electronics*. Each review consists of a rating (1–5 stars), a title, and review text. We created a binary classification problem by binning reviews with 1–2 stars as "negative" and 4–5 stars as "positive". Reviews with 3 stars were discarded.

Classifying product reviews as having either positive or negative sentiment fits well into our theory of domain adaptation. We note that reviews for different products have widely

[2]Available at http://www.cs.jhu.edu/~mdredze/.

Positive *books* review

Title: A great find during an annual summer shopping trip
Review: I found this novel at a bookstore on the boardwalk I visit every summer....The narrative was brilliantly told, the dialogue completely believable and the plot totally heartwrenching. If I had made it to the end without some tears, I would believe myself made of stone!

Negative *books* review

Title: The Hard Way
Review: I am not sure whatever possessed me to buy this book. Honestly, it was a complete waste of my time. To quote a friend, it was not the best use of my entertainment dollar. If you are a fan of pedestrian writing, lack-luster plots and hackneyed character development, this is your book.

Positive *kitchen & housewares* review

Title: no more doggy feuds with neighbor
Review: i absolutely love this product. my neighbor has four little yippers and my shepard/chow mix was antagonized by the yipping on our side of the fence. I hung the device on my side of the fence and the noise keeps the neighbors dog from picking "arguments" with my dog. all barking and fighting has ceased.

Negative *kitchen & housewares* review

Title: cooks great, lid does not work well...
Review: I Love the way the Tefal deep fryer cooks, however, I am returning my second one due to a defective lid closure. The lid may close initially, but after a few uses it no longer stays closed. Since I have small children in my home, I will not be purchasing this one again.

Fig. 2 Some sample product reviews for sentiment classification. The *top row* shows reviews from the *books* domain. The *bottom row* shows reviews from *kitchen & housewares*

different vocabularies, so classifiers trained on one domain are likely to miss out on important lexical cues in a different domain. On the other hand, a single good universal sentiment classifier is likely to exist—namely the classifier that assigns high positive weight to all positive words and high negative weight to all negative words, regardless of product type. We illustrate the type of text in this dataset in Fig. 2, which shows one positive and one negative review each from the domains *books* and *kitchen & housewares*.

For each domain, the data set contains 1,600 labeled documents and between 5,000 and 6,000 unlabeled documents. We follow Pang et al. (2002) and represent each instance (review) by a sparse vector containing the counts of its unigrams and bigrams, and we normalize the vectors in L_1. Finally, we discard all but the most frequent 1,600 unigrams and bigrams from each data set. In all of the learning problems of the next section, including those that require us to estimate an approximate $\mathcal{H}$-distance, we use signed linear classifiers. To estimate the parameters of these classifiers, we minimize a Huber loss with stochastic gradient descent (Zhang 2004).

7.2 Experiments

We explore Theorem 3 further by comparing its predictions to the predictions of an approximation that can be computed from finite labeled source and unlabeled source and target samples. As we shall see, our approximation is a finite-sample analog of (1). We first address λ, the error of the ideal hypothesis. Unfortunately, in general we cannot assume any relationship between the labeling functions f_S and f_T. Thus in order to estimate λ, we must

estimate $\epsilon_T(h^*)$ independently of the source data. If we had enough target data to do this accurately, we would not need to adapt a source classifier in the first place. For our sentiment task, however, λ is small enough to be a negligible term in the bound. Thus we ignore it here.

We approximate the divergence between two domains by training a linear classifier to discriminate between unlabeled instances from the source and target domains. Then we apply Lemma 2 to get an estimate of $\hat{d}_{\mathcal{H}}$ that we denote by $\zeta(\mathcal{U}_S, \mathcal{U}_T)$. $\zeta(\mathcal{U}_S, \mathcal{U}_T)$ is a lower bound on $\hat{d}_{\mathcal{H}}$, which is in turn a lower bound on $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$. For Theorem 3 to be valid, we need an upper bound on $\hat{d}_{\mathcal{H}\Delta\mathcal{H}}$. Unfortunately, this is computationally intractable for linear threshold classifiers, since finding a minimum error classifier is hard in general (Ben-David et al. 2003). We chose our $\zeta(\mathcal{U}_S, \mathcal{U}_T)$ estimate because it requires no new machinery beyond an algorithm for empirical risk minimization on $\mathcal{H}$. Finally, we note that the unlabeled sample size m' is large, so we leave out the finite sample error term for the $\mathcal{H}\Delta\mathcal{H}$-divergence. We set C to be 1,601, the VC dimension of a 1,600-dimensional linear classifier and ignore the $\log m$ term in the numerator of the bound. The complete approximation to the bound is

$$f(\alpha) = \sqrt{\frac{C}{m}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)} + (1-\alpha)\zeta(\mathcal{U}_S, \mathcal{U}_T). \tag{3}$$

Note that $\sqrt{\frac{C}{m}}$ in (3) corresponds to B from (1), and $\zeta(\mathcal{U}_S, \mathcal{U}_T)$ is a finite sample approximation to A when λ is negligible and we have large unlabeled samples from both the source and target domains.

We compare (3) to experimental results for the sentiment classification task. All of our experiments use the *apparel* domain as the target. We obtain empirical curves for the error

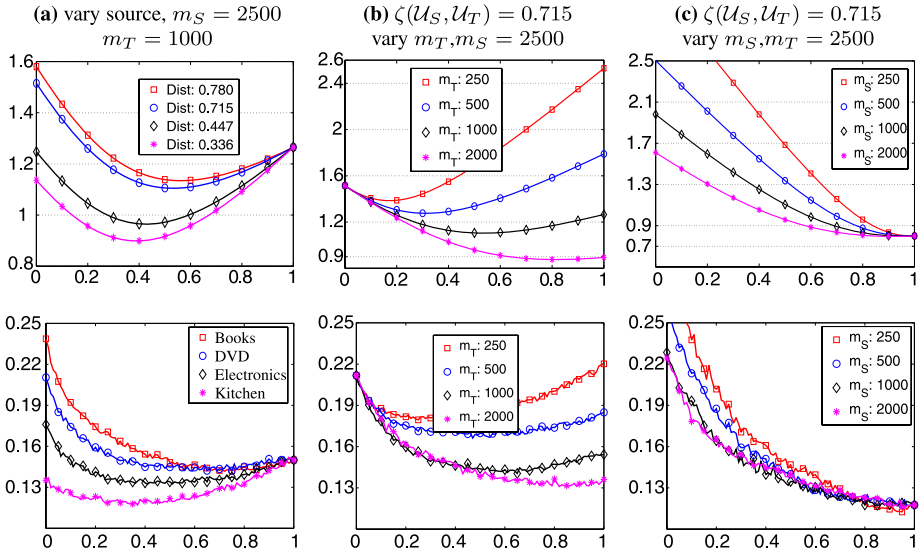


Fig. 3 Comparing the bound from Theorem 3 with test error for sentiment classification. Each column varies one component of the bound. For all plots, the y-axis shows the error and the x-axis shows α. Plots on the *top row* show the value given by our approximation to the bound, and plots on the *bottom row* show the empirical test set error. Column (**a**) depicts different distances among domains. Column (**b**) depicts different numbers of target instances, and column (**c**) represents different numbers of source instances

as a function of α by training a classifier using a weighted hinge loss. Suppose the target domain has weight α and there are βm target training instances. Then we scale the loss of target training instance by α/β and the loss of a source training instance by $(1-\alpha)/(1-\beta)$.

Figure 3 shows a series of plots of (3) (top row) coupled with corresponding plots of test error (bottom row) as a function of α for different amounts of source and target data and different distances between domains. In each column, a single parameter (distance, number of target instances m_T, or number of source instances m_S) is varied while the other two are held constant. Note that $\beta = m_T/(m_T+m_S)$. The plots on the top row of Fig. 3 are not meant to be numerical proxies for the true error. (For the source domains "books" and "dvd", the distance alone is well above 1/2.) However, they illustrate that the bound is similar in shape to the true error curve and that important relationships are preserved.

Note that in every pair of plots, the empirical error curves, like the bounds, have an essentially convex shape. Furthermore, the value of α that minimizes the bound also yields low empirical error in each case. This suggests that choosing α to minimize the bound of Theorem 3 and subsequently training a classifier to minimize the empirical error $\hat{\epsilon}_\alpha(h)$ can work well in practice, provided we have a reasonable measure of complexity and λ is small. Column (a) shows that more distant source domains result in higher target error. Column (b) illustrates that for more target data, we have not only lower error in general, but also a higher minimizing α. Finally, column (c) demonstrates the limitations of distant source data.

When enough labeled target data exists, we always prefer to use only the target data, no matter how much source data is available. Intuitively this is because any biased source domain cannot help to reduce error beyond some positive constant. When the target data alone is sufficient to surpass this level of performance, the source data ceases to be useful. Thus column (c) illustrates empirically one phase transition we discuss in Sect. 6.

8 Combining data from multiple sources

We now explore an extension of our theory to the case of multiple source domains. In this setting, the learner is presented with data from N distinct sources. Each source S_j is associated with an unknown distribution $\mathcal{D}_j$ over input points and an unknown labeling function f_j. The learner receives a total of m labeled samples, with $m_j = \beta_j m$ from each source S_j, and the objective is to use these samples to train a model to perform well on a target domain $\langle \mathcal{D}_T, f_T \rangle$, which may or may not be one of the sources. This setting is motivated by several domain adaptation algorithms (Huang et al. 2007; Bickel et al. 2007; Jiang and Zhai 2007; Dai et al. 2007) that weigh the loss from training instances depending on how "far" they are from the target domain. That is, each training instance is its own source domain.

As before, we examine algorithms that minimize convex combinations of training error over the labeled examples from each source domain. Given a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ of domain weights with $\sum_{j=1}^N \alpha_j = 1$, we define the empirical $\boldsymbol{\alpha}$-weighted error of function h as

$$\hat{\epsilon}_{\boldsymbol{\alpha}}(h) = \sum_{j=1}^N \alpha_j \hat{\epsilon}_j(h) = \sum_{j=1}^N \frac{\alpha_j}{m_j} \sum_{x \in S_j} |h(x) - f_j(x)|.$$

The true $\boldsymbol{\alpha}$-weighted error $\epsilon_{\boldsymbol{\alpha}}(h)$ is defined analogously. We use $\mathcal{D}_{\boldsymbol{\alpha}}$ to denote the mixture of the N source distributions with mixing weights equal to the components of $\boldsymbol{\alpha}$.

We present in turn two alternative generalizations of the bounds in Sect. 5. The first bound considers the quality and quantity of data available from each source individually,

ignoring the relationships between sources. In contrast, the second bound depends directly on the $\mathcal{H}\Delta\mathcal{H}$-distance between the target domain and the weighted *combination* of source domains. This dependence allows us to achieve significantly tighter bounds when there exists a mixture of sources that approximates the target better than any single source. Both results require the derivation of uniform convergence bounds for the empirical $\boldsymbol{\alpha}$-error. We begin with those.

8.1 Uniform convergence

The following lemma provides a uniform convergence bound for the empirical $\boldsymbol{\alpha}$-error.

Lemma 6 *For each $j \in \{1, \ldots, N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. For any fixed weight vector $\boldsymbol{\alpha}$, let $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ be the empirical $\boldsymbol{\alpha}$-weighted error of some fixed hypothesis h on this sample, and let $\epsilon_{\boldsymbol{\alpha}}(h)$ be the true $\boldsymbol{\alpha}$-weighted error. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$:*

$$\Pr\left[|\hat{\epsilon}_{\boldsymbol{\alpha}}(h) - \epsilon_{\alpha}(h)| \geq \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}}\right).$$

Proof See Appendix. □

Note that this bound is minimized when $\alpha_j = \beta_j$ for all j. In other words, convergence is fastest when all data instances are weighted equally.

8.2 A bound using pairwise divergence

The first bound we present considers the pairwise $\mathcal{H}\Delta\mathcal{H}$-distance between each source and the target, and illustrates the trade-off that exists between minimizing the average divergence of the training data from the target and weighting all points equally to encourage faster convergence. The term $\sum_{j=1}^{N} \alpha_j \lambda_j$ that appears in this bound plays a role corresponding to λ in the previous section. Somewhat surprisingly, this term can be small even when there is not a single hypothesis that works well for all heavily weighted sources.

Theorem 4 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For each $j \in \{1, \ldots, N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ for a fixed weight vector $\boldsymbol{\alpha}$ on these samples and $h_T^* = \min_{h \in \mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 2\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{d\log(2m) - \log(\delta)}{2m}\right)} + \sum_{j=1}^{N} \alpha_j \left(2\lambda_j + d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right),$$

where $\lambda_j = \min_{h \in \mathcal{H}}\{\epsilon_T(h) + \epsilon_j(h)\}$.

Proof See Appendix. □

In the special case where the $\mathcal{H}\Delta\mathcal{H}$-divergence between each source and the target is 0 and all data instances are weighted equally, the bound in Theorem 4 becomes

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 2\sqrt{\frac{2d\log(2(m+1)) + 2\log(\frac{4}{\delta})}{m}} + 2\sum_{j=1}^{N}\alpha_j\lambda_j.$$

This bound is nearly identical to the multiple source classification bound given in Theorem 6 of Crammer et al. (2008). Aside from the constants in the complexity term, the only difference is that the quantity λ_i that appears here is replaced by an alternate measure of the label error between source S_j and the target. Furthermore, these measures are equivalent when the true target function is a member of $\mathcal{H}$. However, the bound of Crammer et al. (2008) is less general. In particular, it does not handle positive $\mathcal{H}\Delta\mathcal{H}$-divergence or non-uniform weighting of the data.

8.3 A bound using combined divergence

In the previous bound, divergence between domains is measured only on pairs, so it is not necessary to have a single hypothesis that is good for every source domain. However, this bound does not give us the flexibility to take advantage of domain structure when calculating unlabeled divergence. The alternate bound given in Theorem 5 allows us to effectively alter the source distribution by changing $\boldsymbol{\alpha}$. This has two consequences. First, we must now demand that there exists a hypothesis h^* which has low error on both the $\boldsymbol{\alpha}$-weighted convex combination of sources and the target domain. Second, we measure $\mathcal{H}\Delta\mathcal{H}$-divergence between the target and a mixture of sources, rather than between the target and each single source.

Theorem 5 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For each $j \in \{1,\ldots,N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ for a fixed weight vector $\boldsymbol{\alpha}$ on these samples and $h_T^* = \min_{h\in\mathcal{H}}\epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 4\sqrt{\left(\sum_{j=1}^{N}\frac{\alpha_j^2}{\beta_j}\right)\left(\frac{d\log(2m) - \log(\delta)}{2m}\right)} + 2\gamma_{\boldsymbol{\alpha}} + d_{\mathcal{H}\Delta\mathcal{H}}(D_{\boldsymbol{\alpha}}, D_T),$$

where $\gamma_{\boldsymbol{\alpha}} = \min_h\{\epsilon_T(h) + \epsilon_{\boldsymbol{\alpha}}(h)\} = \min_h\{\epsilon_T(h) + \sum_{j=1}^{N}\alpha_j\epsilon_j(h)\}$.

Proof See Appendix. □

Theorem 5 reduces to Theorem 3 when $N = 2$ and one of the two source domains is the target domain (that is, we have some small number of target instances).

8.4 Discussion

One might ask whether there exist settings where a non-uniform weighting can lead to a significantly lower value of the bound than a uniform weighting. Indeed, this can happen if some non-uniform weighting of sources accurately approximates the target distribution. This is true, for example, in the setting studied by Mansour et al. (2009a, 2009b), who derive results for combining pre-computed hypotheses. In particular, they show that for arbitrary convex losses, if the Rényi divergence between the target and a mixture of sources is small, it is possible to combine low-error source hypotheses to create a low-error target hypothesis. They then show that if for each domain j there exists a hypothesis h_j with error less than ϵ, it is possible to achieve an error less than ϵ on the target by weighting the predictions of $h_1, \ldots, h_N$ appropriately.

The Rényi divergence is not directly comparable to the $\mathcal{H}\Delta\mathcal{H}$-divergence in general; however it is possible to exhibit source and target distributions which have low $\mathcal{H}\Delta\mathcal{H}$-divergence and high (even infinite) Rényi divergence. For example, the Rényi divergence is infinite when the source and target distributions do not share support, but the $\mathcal{H}\Delta\mathcal{H}$-divergence is only large when these regions of differing support also coincide with classifier disagreement regions. On the other hand, we require that a single hypothesis be trained on the mixture of sources. Mansour et al. (2009a, 2009b) give algorithms which do not require the original training data at all, but only a single hypothesis from each source.

9 Conclusion

We presented a theoretical investigation of the task of domain adaptation, a task in which we have a large amount of training data from a source domain, but we wish to apply a model in a target domain with a much smaller amount of training data. Our main result is a uniform convergence learning bound for algorithms which minimize convex combinations of empirical source and target error. Our bound reflects the trade-off between the size of the source data and the accuracy of the target data, and we give a simple approximation to it that is computable from finite labeled and unlabeled samples. This approximation makes correct predictions about model test error for a sentiment classification task. Our theory also extends in a straightforward manner to a multi-source setting, which we believe helps to explain the success of recent empirical work in domain adaptation.

There are two interesting open problems that deserve future exploration. First, our bounds on the divergence between source and target distribution are in terms of VC dimension. We do not yet know whether our divergence measure admits tighter data-dependent bounds (McAllester 2003; Bartlett and Mendelson 2002), or if there are other, more appropriate divergence measures which do. Second, it would be interesting to investigate algorithms that choose a convex combination of multiple sources to minimize the bound in Theorem 5 as possible approaches to adaptation from multiple sources.

Acknowledgements This material is based upon work partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. NBCHD030010 (CALO), by the National Science Foundation under grants ITR 0428193 and RI 0803256, and by a gift from Google, Inc. to the University of Pennsylvania. Koby Crammer is a Horev fellow, supported by the Taub Foundations. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the DARPA, Department of Interior-National Business Center (DOI-NBC), NSF, the Taub Foundations, or Google, Inc.

Appendix: Proofs

Theorem 1 *For a hypothesis h,*

$$\epsilon_T(h) \leq \epsilon_S(h) + d_1(\mathcal{D}_S, \mathcal{D}_T) + \min\left\{\mathrm{E}_{\mathcal{D}_S}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right], \mathrm{E}_{\mathcal{D}_T}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right]\right\}.$$

Proof Recall that $\epsilon_T(h) = \epsilon_T(h, f_T)$ and $\epsilon_S(h) = \epsilon_S(h, f_S)$. Let ϕ_S and ϕ_T be the density functions of $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively.

$$\begin{aligned}
\epsilon_T(h) &= \epsilon_T(h) + \epsilon_S(h) - \epsilon_S(h) + \epsilon_S(h, f_T) - \epsilon_S(h, f_T) \\
&\leq \epsilon_S(h) + |\epsilon_S(h, f_T) - \epsilon_S(h, f_S)| + |\epsilon_T(h, f_T) - \epsilon_S(h, f_T)| \\
&\leq \epsilon_S(h) + \mathrm{E}_{\mathcal{D}_S}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right] + |\epsilon_T(h, f_T) - \epsilon_S(h, f_T)| \\
&\leq \epsilon_S(h) + \mathrm{E}_{\mathcal{D}_S}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right] + \int |\phi_S(\mathbf{x}) - \phi_T(\mathbf{x})||h(\mathbf{x}) - f_T(\mathbf{x})| d\mathbf{x} \\
&\leq \epsilon_S(h) + \mathrm{E}_{\mathcal{D}_S}\left[|f_S(\mathbf{x}) - f_T(\mathbf{x})|\right] + d_1(\mathcal{D}_S, \mathcal{D}_T).
\end{aligned}$$

In the first line, we could instead choose to add and subtract $\epsilon_T(h, f_S)$ rather than $\epsilon_S(h, f_T)$, which would result in the same bound only with the expectation taken with respect to $\mathcal{D}_T$ instead of $\mathcal{D}_S$. Choosing the smaller of the two gives us the bound. □

Lemma 2 *For a symmetric hypothesis class $\mathcal{H}$ (one where for every $h \in \mathcal{H}$, the inverse hypothesis $1 - h$ is also in $\mathcal{H}$) and samples $\mathcal{U}, \mathcal{U}'$ of size m, the empirical $\mathcal{H}$-distance is*

$$d_{\mathcal{H}}(\mathcal{U}, \mathcal{U}') = 2\left(1 - \min_{h \in \mathcal{H}}\left[\frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x} \in \mathcal{U}] + \frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=1} I[\mathbf{x} \in \mathcal{U}']\right]\right)$$

where $I[\mathbf{x} \in \mathcal{U}]$ is the binary indicator variable which is 1 when $\mathbf{x} \in \mathcal{U}$.

Proof We will show that for any hypothesis h and corresponding set $I(h)$ of positively labeled instances,

$$1 - \left[\frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x} \in \mathcal{U}] + \frac{1}{m}\sum_{\mathbf{x}:h(\mathbf{x})=1} I[\mathbf{x} \in \mathcal{U}']\right] = \Pr_{\mathcal{U}}[I(h)] - \Pr_{\mathcal{U}'}[I(h)].$$

We have

$$\begin{aligned}
&1 - \frac{1}{m}\left(\sum_{\mathbf{x}:h(\mathbf{x})=0} I[\mathbf{x} \in \mathcal{U}] + \sum_{\mathbf{x}:h(\mathbf{x})=1} I\left[\mathbf{x} \in \mathcal{U}'\right]\right) \\
&= \frac{1}{2m}\sum_{\mathbf{x}:h(\mathbf{x})=0}\left(I[\mathbf{x} \in \mathcal{U}] + I\left[\mathbf{x} \in \mathcal{U}'\right]\right) + \frac{1}{2m}\sum_{\mathbf{x}:h(\mathbf{x})=1}\left(I[\mathbf{x} \in \mathcal{U}] + I\left[\mathbf{x} \in \mathcal{U}'\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{m}\Bigg(\sum_{\mathbf{x}:h(\mathbf{x})=0} I\left[\mathbf{x}\in\mathcal{U}\right] + \sum_{\mathbf{x}:h(\mathbf{x})=1} I\left[\mathbf{x}\in\mathcal{U}'\right]\Bigg)\\
&= \frac{1}{2m}\sum_{\mathbf{x}:h(\mathbf{x})=0}\left(I\left[\mathbf{x}\in\mathcal{U}'\right] - I\left[\mathbf{x}\in\mathcal{U}\right]\right) + \frac{1}{2m}\sum_{\mathbf{x}:h(\mathbf{x})=1}\left(I\left[\mathbf{x}\in\mathcal{U}\right] - I\left[\mathbf{x}\in\mathcal{U}'\right]\right)\\
&= \frac{1}{2}\left(1-\Pr_{\mathcal{U}'}\left[I(h)\right] - \left(1-\Pr_{\mathcal{U}}\left[I(h)\right]\right)\right) + \frac{1}{2}\left(\Pr_{\mathcal{U}}\left[I(h)\right] - \Pr_{\mathcal{U}'}\left[I(h)\right]\right)\\
&= \Pr_{\mathcal{U}}\left[I(h)\right] - \Pr_{\mathcal{U}'}\left[I(h)\right]. \qquad (4)
\end{aligned}
$$

The absolute value in the statement of the lemma follows from the symmetry of $\mathcal{H}$. □

Lemma 4 *Let h be a hypothesis in class $\mathcal{H}$. Then*

$$
|\epsilon_\alpha(h) - \epsilon_T(h)| \le (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T) + \lambda\right).
$$

Proof Similarly to the proof of Theorem 2, this proof relies heavily on the triangle inequality for classification error.

$$
\begin{aligned}
&|\epsilon_\alpha(h) - \epsilon_T(h)|\\
&\quad= (1-\alpha)|\epsilon_S(h) - \epsilon_T(h)|\\
&\quad\le (1-\alpha)\left[|\epsilon_S(h) - \epsilon_S(h,h^*)| + |\epsilon_S(h,h^*) - \epsilon_T(h,h^*)| + |\epsilon_T(h,h^*) - \epsilon_T(h)|\right]\\
&\quad\le (1-\alpha)\left[\epsilon_S(h^*) + |\epsilon_S(h,h^*) - \epsilon_T(h,h^*)| + \epsilon_T(h^*)\right]\\
&\quad\le (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T) + \lambda\right).
\end{aligned}
$$

□

Theorem 3 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. Let $\mathcal{U}_S$ and $\mathcal{U}_T$ be unlabeled samples of size m' each, drawn from $\mathcal{D}_S$ and $\mathcal{D}_T$ respectively. Let S be a labeled sample of size m generated by drawing βm points from $\mathcal{D}_T$ and $(1-\beta)m$ points from $\mathcal{D}_S$, labeling them according to f_S and f_T, respectively. If $\hat{h}\in\mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_\alpha(h)$ on S and $h_T^* = \min_{h\in\mathcal{H}}\epsilon_T(h)$ is the target error minimizer, then for any $\delta\in(0,1)$, with probability at least $1-\delta$ (over the choice of the samples),*

$$
\begin{aligned}
\epsilon_T(\hat{h}) \le \epsilon_T(h_T^*) &+ 4\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1)) + 2\log\left(\frac{8}{\delta}\right)}{m}}\\
&+ 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S,\mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{4}{\delta})}{m'}} + \lambda\right).
\end{aligned}
$$

Proof The complete proof is mostly identical to the standard proof of uniform convergence for empirical risk minimizers. We show here the steps that are different. Below we use L4, and Thm2 to indicate that a line of the proof follows by application of Lemma 4 or Theorem 2 respectively. L5 indicates that the proof follows by Lemma 5, but also relies on sample symmetrization and bounding the growth function by the VC dimension (Anthony

and Bartlett 1999).

$$\begin{aligned}
\epsilon_T(\hat{h}) &\le \epsilon_\alpha(\hat{h}) + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \quad \text{(L4)}\\
&\le \hat{\epsilon}_\alpha(\hat{h}) + 2\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(\frac{8}{\delta})}{m}}\\
&\quad + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \quad \text{(L5)}\\
&\le \hat{\epsilon}_\alpha(h_T^*) + 2\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(\frac{8}{\delta})}{m}}\\
&\quad + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \quad \left(\hat{h}=\arg\min_{h\in\mathcal{H}}\hat{\epsilon}_\alpha(h)\right)\\
&\le \epsilon_\alpha(h_T^*) + 4\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(\frac{8}{\delta})}{m}}\\
&\quad + (1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \quad \text{(L5)}\\
&\le \epsilon_T(h_T^*) + 4\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(\frac{8}{\delta})}{m}}\\
&\quad + 2(1-\alpha)\left(\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\lambda\right) \quad \text{(L4)}\\
&\le \epsilon_T(h_T^*) + 4\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{1-\beta}}\sqrt{\frac{2d\log(2(m+1))+2\log(\frac{8}{\delta})}{m}}\\
&\quad + 2(1-\alpha)\left(\frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S,\mathcal{U}_T)+4\sqrt{\frac{2d\log(2m')+\log(\frac{4}{\delta})}{m'}}+\lambda\right) \quad \text{(Thm 2)}
\end{aligned}$$

□

Lemma 6 *For each $j \in \{1,\dots,N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. For any fixed weight vector $\boldsymbol{\alpha}$, let $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ be the empirical $\boldsymbol{\alpha}$-weighted error of some fixed hypothesis h on this sample, and let $\epsilon_{\boldsymbol{\alpha}}(h)$ be the true $\boldsymbol{\alpha}$-weighted error. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\Pr\left[|\hat{\epsilon}_{\boldsymbol{\alpha}}(h)-\epsilon_\alpha(h)|\ge\epsilon\right] \le 2\exp\left(\frac{-2m\epsilon^2}{\sum_{j=1}^N \frac{\alpha_j^2}{\beta_j}}\right).$$

Proof Due to its similarity to the proof of Lemma 5, we omit some details of this proof, and concentrate only on the parts that differ.

For each source j, let $X_{j,1},\dots,X_{j,\beta_j m}$ be random variables that take on the values $(\alpha_j/\beta_j)|h(x)-f_j(x)|$ for the $\beta_j m$ instances $x \in S_j$. Note that $X_{j,1},\dots,X_{j,\beta_j m} \in [0,\alpha_j/\beta_j]$.

Then

$$\hat{\epsilon}_{\boldsymbol{\alpha}}(h) = \sum_{j=1}^{N} \alpha_j \hat{\epsilon}_j(h) = \sum_{j=1}^{N} \alpha_j \frac{1}{\beta_j m} \sum_{x \in S_j} |h(x) - f_j(x)| = \frac{1}{m} \sum_{j=1}^{N} \sum_{i=1}^{\beta_j m} X_{j,i}.$$

By linearity of expectations, we have that $E[\hat{\epsilon}_{\boldsymbol{\alpha}}(h)] = \epsilon_{\boldsymbol{\alpha}}(h)$, and so by Hoeffding's inequality, for every $h \in \mathcal{H}$,

$$\Pr\left[|\hat{\epsilon}_{\boldsymbol{\alpha}}(h) - \epsilon_{\boldsymbol{\alpha}}(h)| \geq \epsilon\right] \leq 2\exp\left(\frac{-2m^2\epsilon^2}{\sum_{j=1}^{N}\sum_{i=1}^{\beta_j m} \text{range}^2(X_{j,i})}\right) = 2\exp\left(\frac{-2m\epsilon^2}{\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}}\right). \quad \square$$

Theorem 4 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For each $j \in \{1, \ldots, N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ for a fixed weight vector $\boldsymbol{\alpha}$ on these samples and $h_T^* = \min_{h \in \mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + 4\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} + \sum_{j=1}^{N} \alpha_j \left(2\lambda_j + d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right),$$

where $\lambda_j = \min_{h \in \mathcal{H}}\{\epsilon_T(h) + \epsilon_j(h)\}$.

Proof Let $h_j^* = \text{argmin}_h\{\epsilon_T(h) + \epsilon_j(h)\}$. Then

$$\begin{aligned}
&|\epsilon_{\boldsymbol{\alpha}}(h) - \epsilon_T(h)| \\
&= \left|\sum_{j=1}^{N} \alpha_j \epsilon_j(h) - \epsilon_T(h)\right| \leq \sum_{j=1}^{N} \alpha_j \left|\epsilon_j(h) - \epsilon_T(h)\right| \\
&\leq \sum_{j=1}^{N} \alpha_j \left(\left|\epsilon_j(h) - \epsilon_j(h, h_j^*)\right| + \left|\epsilon_j(h, h_j^*) - \epsilon_T(h, h_j^*)\right| + \left|\epsilon_T(h, h_j^*) - \epsilon_T(h)\right|\right) \\
&\leq \sum_{j=1}^{N} \alpha_j \left(\epsilon_j(h_j^*) + \left|\epsilon_j(h, h_j^*) - \epsilon_T(h, h_j^*)\right| + \epsilon_T(h_j^*)\right) \\
&\leq \sum_{j=1}^{N} \alpha_j \left(\lambda_j + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right).
\end{aligned}$$

The third line follows from the triangle inequality. The last line follows from the definition of λ_j and Lemma 3. Putting this together with Lemma 6, we find that for any $\delta \in (0, 1)$,

with probability $1-\delta$,

$$\begin{aligned}
\epsilon_T(\hat{h}) &\leq \epsilon_{\boldsymbol{\alpha}}(\hat{h}) + \sum_{j=1}^{N} \alpha_j \left(\lambda_j + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right) \\
&\leq \hat{\epsilon}_{\boldsymbol{\alpha}}(\hat{h}) + 2\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} \\
&\quad + \sum_{j=1}^{N} \alpha_j \left(\lambda_j + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right) \\
&\leq \hat{\epsilon}_{\boldsymbol{\alpha}}(h_T^*) + 2\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} \\
&\quad + \sum_{j=1}^{N} \alpha_j \left(\lambda_j + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right) \\
&\leq \epsilon_{\boldsymbol{\alpha}}(h_T^*) + 4\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} \\
&\quad + \sum_{j=1}^{N} \alpha_j \left(\lambda_j + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right) \\
&\leq \epsilon_T(h_T^*) + 4\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} \\
&\quad + \sum_{j=1}^{N} \alpha_j \left(2\lambda_j + d_{\mathcal{H}\Delta\mathcal{H}}(D_j, D_T)\right).
\end{aligned}$$

□

Theorem 5 *Let $\mathcal{H}$ be a hypothesis space of VC dimension d. For each $j \in \{1, \ldots, N\}$, let S_j be a labeled sample of size $\beta_j m$ generated by drawing $\beta_j m$ points from $\mathcal{D}_j$ and labeling them according to f_j. If $\hat{h} \in \mathcal{H}$ is the empirical minimizer of $\hat{\epsilon}_{\boldsymbol{\alpha}}(h)$ for a fixed weight vector $\boldsymbol{\alpha}$ on these samples and $h_T^* = \min_{h\in\mathcal{H}} \epsilon_T(h)$ is the target error minimizer, then for any $\delta \in (0, 1)$, with probability at least $1-\delta$,*

$$\begin{aligned}
\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) &+ 2\sqrt{\left(\sum_{j=1}^{N} \frac{\alpha_j^2}{\beta_j}\right)\left(\frac{2d\log(2(m+1)) + \log(\frac{4}{\delta})}{m}\right)} \\
&+ 2\left(\gamma_{\boldsymbol{\alpha}} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{\boldsymbol{\alpha}}, D_T)\right),
\end{aligned}$$

where $\gamma_{\boldsymbol{\alpha}} = \min_h\{\epsilon_T(h) + \epsilon_{\boldsymbol{\alpha}}(h)\} = \min_h\{\epsilon_T(h) + \sum_{j=1}^{N} \alpha_j \epsilon_j(h)\}$.

Proof The proof is almost identical to that of Theorem 4 with minor modifications to the derivation of the bound on $|\epsilon_{\boldsymbol{\alpha}}(h) - \epsilon_T(h)|$. Let $h^* = \operatorname{argmin}_h\{\epsilon_T(h) + \epsilon_{\boldsymbol{\alpha}}(h)\}$. By the triangle

inequality and Lemma 3,

$$\begin{aligned}|\epsilon_{\alpha}(h) - \epsilon_T(h)| &\leq \left|\epsilon_{\alpha}(h) - \epsilon_{\alpha}(h, h^*)\right| + \left|\epsilon_{\alpha}(h, h^*) - \epsilon_T(h, h^*)\right| + \left|\epsilon_T(h, h^*) - \epsilon_T(h)\right| \\ &\leq \epsilon_{\alpha}(h^*) + \left|\epsilon_{\alpha}(h, h^*) - \epsilon_T(h, h^*)\right| + \epsilon_T(h^*) \\ &\leq \gamma + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(D_{\alpha}, D_T).\end{aligned}$$

The remainder of the proof is unchanged. □

References

Ando, R., & Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, *6*, 1817–1853.

Anthony, M., & Bartlett, P. (1999). *Neural network learning: theoretical foundations*. Cambridge: Cambridge University Press.

Bartlett, P., & Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, *3*, 463–482.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W., & White, P. (2000). Testing that distributions are close. In: *IEEE symposium on foundations of computer science* (Vol. 41, pp. 259–269).

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, *12*, 149–198.

Ben-David, S., Eiron, N., & Long, P. (2003). On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, *66*, 496–514.

Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2006). Analysis of representations for domain adaptation. In: *Advances in neural information processing systems*.

Bickel, S., Brückner, M., & Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In: *Proceedings of the international conference on machine learning*.

Bikel, D., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: a high-performance learning name-finder. In: *Conference on applied natural language processing*.

Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Wortman, J. (2007a). Learning bounds for domain adaptation. In: *Advances in neural information processing systems*.

Blitzer, J., Dredze, M., & Pereira, F. (2007b) Biographies, Bollywood, boomboxes and blenders: domain adaptation for sentiment classification. In: *ACL*.

Collins, M. (1999). *Head-driven statistical models for natural language parsing*. PhD thesis, University of Pennsylvania.

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008). Sample selection bias correction theory. In: *Proceedings of the 19th annual conference on algorithmic learning theory*.

Crammer, K., Kearns, M., & Wortman, J. (2008). Learning from multiple sources. *Journal of Machine Learning Research*, *9*, 1757–1774.

Dai, W., Yang, Q., Xue, G., & Yu, Y. (2007). Boosting for transfer learning. In: *Proceedings of the international conference on machine learning*.

Das, S., & Chen, M. (2001). Yahoo! for Amazon: extracting market sentiment from stock message boards. In: *Proceedings of the Asia pacific finance association annual conference*.

Daumé, H. (2007). Frustratingly easy domain adaptation. In: *Association for computational linguistics (ACL)*.

Finkel, J. R. Manning, C. D. (2009). Hierarchical Bayesian domain adaptation. In: *Proceedings of the north American association for computational linguistics*.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153–161.

Huang, J., Smola, A., Gretton, A., Borgwardt, K., & Schoelkopf, B. (2007). Correcting sample selection bias by unlabeled data. In: *Advances in neural information processing systems*.

Jiang, J., & Zhai, C. (2007). Instance weighting for domain adaptation. In: *Proceedings of the association for computational linguistics*.

Kifer, D., Ben-David, S., & Gehrke, J. (2004). Detecting change in data streams. In: *Ver large databases*.

Li, X., & Bilmes, J. (2007). A Bayesian divergence prior for classification adaptation. In: *Proceedings of the international conference on artificial intelligence and statistics*.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009a). Domain adaptation with multiple sources. In: *Advances in neural information processing systems*.

Mansour, Y., Mohri, M., & Rostamizadeh, A. (2009b). Multiple source adaptation and the rényi divergence. In: *Proceedings of the conference on uncertainty in artificial intelligence*.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. In: *Proceedings of the sixteenth annual conference on learning theory*.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In: *Proceedings of empirical methods in natural language processing*.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In: *Proceedings of empirical methods in natural language processing*.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., & Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, *60*, 699–746.

Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: determining support or opposition from congressional floor-debate transcripts. In: *Proceedings of empirical methods in natural language processing*.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the association for computational linguistics*.

Vapnik, V. (1998). *Statistical learning theory*. New York: Wiley.

Zhang, T. (2004). Solving large-scale linear prediction problems with stochastic gradient descent. In: *Proceedings of the international conference on machine learning*.