Journal of Machine Learning Research 9 (2008) 1757-1774 Submitted 7/07; Revised 4/08; Published 8/08

Learning from Multiple Sources*

Koby Crammer CRAMMER@CIS.UPENN.EDU
Michael Kearns MKEARNS@CIS.UPENN.EDU
Jennifer Wortman WORTMANJ@SEAS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA

Editor: Peter Bartlett

Abstract

We consider the problem of learning accurate models from multiple sources of “nearby” data. Given distinct samples from multiple data sources and estimates of the dissimilarities between these sources, we provide a general theory of which samples should be used to learn models for each source. This theory is applicable in a broad decision-theoretic learning framework, and yields general results for classification and regression. A key component of our approach is the development of approximate triangle inequalities for expected loss, which may be of independent interest. We discuss the related problem of learning parameters of a distribution from multiple data sources. Finally, we illustrate our theory through a series of synthetic simulations.

Keywords: error bounds, multi-task learning

1. Introduction

We introduce and analyze a theoretical model for the problem of learning from multiple sources of “nearby” data. As a hypothetical example of where such problems might arise, consider the following scenario: For each web user in a large population, we wish to learn a classifier for what sites that user is likely to find “interesting.” Assuming we have at least a small amount of labeled data for each user (as might be obtained either through direct feedback, or via indirect means such as click-throughs following a search), one approach would be to apply standard learning algorithms to each user’s data in isolation. However, if there are natural and accessible measures of similarity between the interests of pairs of users (as might be obtained through their mutual labelings of common web sites), an appealing alternative is to *aggregate* the data of “nearby” users when learning a classifier for each particular user. This alternative is intuitively subject to a trade-off between the increased sample size and how different the aggregated users are.

We treat this problem in some generality and provide a bound addressing the aforementioned trade-off. In our model there are K unknown data sources, with source i generating a distinct sample S_i of n_i observations. We assume we are given only the samples S_i, and a *disparity*[1] matrix D whose entry $D(i,j)$ bounds the difference between source i and source j. Given these inputs, we wish to

[*] A preliminary version of this work appeared in *Advances in Neural Information Processing Systems 19* (Crammer et al., 2007).

[1] We avoid using the term distance since our results include settings in which the underlying loss measures may not be formal distances.

decide which subset of the samples S_j will result in the best model for each source i. Our framework includes settings in which the sources produce data for classification, regression, and density estimation (and more generally any additive-loss learning problem obeying certain conditions).

Our main result is a general theorem establishing a bound on the expected loss incurred by using all data sources within a given disparity of the target source. Optimization of this bound then yields a recommended subset of the data to be used in learning a model of each source. Our bound clearly expresses a trade-off between three quantities: the sample size used (which increases as we include data from more distant models), a weighted average of the disparities of the sources whose data is used, and a model complexity term. It can be applied to any learning setting in which the underlying loss function obeys an *approximate* triangle inequality, and in which the class of hypothesis models under consideration obeys uniform convergence of empirical estimates of loss to expectations. For classification problems, the standard triangle inequality holds. For regression we prove a 2-approximation to the triangle inequality. Uniform convergence bounds for the settings we consider may be obtained via standard data-independent model complexity measures such as VC dimension and pseudo-dimension, or via more recent measures such as Rademacher complexity.

Recent work by Crammer et al. (2006) examines the considerably more limited problem of learning a model when all data sources are corrupted versions of a *single, fixed* source, for instance when each data source provides noisy samples of a fixed binary function, but with varying levels of noise. In the current work, the labels on each source may be entirely unrelated to those on other source except as constrained by the bounds on disparities, requiring us to develop new techniques. Blitzer et al. (2007) study the related problem of training classifiers using multiple sources of data drawn from different *underlying* domains but labeled using identical or similar labeling functions. Wu and Dietterich (2004) study similar problems experimentally in the context of SVMs. The framework examined here can also be viewed in the context of multi-task learning, or as a type of transfer learning (Baxter, 1995; Ben-David, 2003; Maurer, 2005).

In Section 2 we introduce a decision-theoretic framework for probabilistic learning that includes classification, regression, and many other settings as special cases, and then give our multiple source generalization of this model. In Section 3 we provide our main result, which is a general bound on the expected loss incurred by using all data within a given disparity of a target source. Section 4 discusses the most simple application of this bound to binary classification using VC theory. In Section 5, we give applications of our general theory to classification and regression using Rademacher complexity, and show more generally how the theory can be applied for any Lipschitz loss function. In Section 6 we discuss how to empirically estimate the disparity matrix from data. In Section 7, we discuss the related problem of learning parameters of a distribution from multiple data sources. Finally, in Section 8, we illustrate the theory through synthetic simulations.

2. Learning Models

Before detailing our multiple-source learning model, we first introduce a standard decision-theoretic learning framework in which our goal is to find a model minimizing a generalized notion of empirical loss (Haussler, 1992). Let the *hypothesis class* $\mathcal{H}$ be a set of models (which might be classifiers, real-valued functions, densities, etc.), and let f be the *target model*, which may or may not lie in the class $\mathcal{H}$. Let z be a (generalized) data point or observation. For instance, in (noise-free) classification and regression, z will consist of a pair $\langle x,y \rangle$ where $y = f(x)$. We assume that the target model f induces some underlying distribution P_f over observations z. In the case of classification or regression, P_f is induced by drawing the inputs x according to some underlying distribution P, and then setting $y = f(x)$ (possibly corrupted by noise).

Each setting we consider has an associated *loss function* $\mathcal{L}(h,z)$. For example, in classification we typically consider the 0/1 loss: $\mathcal{L}(h,\langle x,y\rangle) = 0$ if $h(x) = y$, and 1 otherwise. In regression we might consider the squared loss function $\mathcal{L}(h,\langle x,y\rangle) = (y-h(x))^2$. In each case, we are interested in the expected loss of a model g_2 on target g_1, $e(g_1,g_2) = \mathrm{E}_{z\sim P_{g_1}}[\mathcal{L}(g_2,z)]$. Expected loss is not necessarily symmetric.

In our multiple source model, we are presented with K distinct mutually independent samples or *sources* of data $S_1,...,S_K$, and a symmetric $K\times K$ matrix D. Each source S_i contains n_i observations that are generated from a fixed and unknown model f_i, and D satisfies $\max(e(f_i,f_j),e(f_j,f_i)) \leq D(i,j)$. When D is unknown, it often can be estimated from a small amount of data; see Section 6 for more details. Our goal is to decide which sources S_j to use in order to learn the best approximation (in terms of expected loss) to each f_i.

While we are interested in accomplishing this goal for each f_i, it suffices and is convenient to examine the problem from the perspective of a fixed f_i. Thus without loss of generality let us suppose that we are given sources $S_1,...,S_K$ of size $n_1,\ldots,n_K$ from models $f_1,\ldots,f_K$ such that $\varepsilon_1 \equiv D(1,1) \leq \varepsilon_2 \equiv D(1,2) \leq \cdots \leq \varepsilon_K \equiv D(1,K)$, and our goal is to learn f_1. Here we have simply taken the problem in the preceding paragraph, focused on the problem for f_1, and reordered the other models according to our estimations or their proximity to f_1. To highlight the distinguished role of the target f_1 we shall denote it f. We denote the observations in S_j by $z_1^j,\ldots,z_{n_j}^j$. In all cases we will analyze, for any $k \leq K$, the hypothesis $\hat{h}_k$ minimizing the empirical loss $\hat{e}_k(h)$ on the first k sources $S_1,\ldots,S_k$, that is

$$\hat{h}_k = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\,\hat{e}_k(h) = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\,\frac{1}{n_{1:k}}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathcal{L}(h,z_i^j)\ ,$$

where $n_{1:k} = n_1+\cdots+n_k$. We also denote the expected error of function h with respect to the first k sources of data as

$$e_k(h) = \mathrm{E}\left[\hat{e}_k(h)\right] = \sum_{i=1}^{k}\left(\frac{n_i}{n_{1:k}}\right)e(f_i,h).$$

3. General Theory for the Multiple Source Problem

In this section we provide the first of our main results: a general bound on the expected loss of the model minimizing the empirical loss on the nearest k sources. Optimization of this bound leads to a recommended set of sources to incorporate when learning $f = f_1$. The key ingredients needed to apply this bound are an approximate triangle inequality and a uniform convergence bound, which we define below. In the subsequent sections we demonstrate that these ingredients can indeed be provided for a variety of natural learning problems.

Definition 1 *For $\alpha \geq 1$, we say that the **α-triangle inequality** holds for a class of models $\mathcal{F}$ and expected loss function e if for all $g_1,g_2,g_3 \in \mathcal{F}$ we have*

$$e(g_1,g_2) \leq \alpha(e(g_1,g_3)+e(g_3,g_2)).$$

The parameter $\alpha \geq 1$ is a constant that depends on $\mathcal{F}$ and e.

The choice $\alpha = 1$ yields the standard triangle inequality. We note that the restriction to models in the class $\mathcal{F}$ may in some cases be quite weak—for instance, when $\mathcal{F}$ is all possible classifiers or real-valued functions with bounded range—or stronger, as in densities from the exponential family. Our results will require only that the unknown *source* models $f_1, \ldots, f_K$ lie in $\mathcal{F}$, even when our *hypothesis* models are chosen from some possibly much more restricted class $\mathcal{H} \subseteq \mathcal{F}$. For now we simply leave $\mathcal{F}$ as a parameter of the definition.

Definition 2 *A* ***uniform convergence bound*** *for a hypothesis space $\mathcal{H}$ and loss function $\mathcal{L}$ is a bound that states that for any $0 < \delta < 1$, with probability at least $1 - \delta$ for any $h \in \mathcal{H}$*

$$|\hat{e}(h) - e(h)| \leq \beta(n, \delta) ,$$

where $\hat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(h, z_i)$ for n observations $z_1, \ldots, z_n$ generated independently according to distributions $P_1, \ldots P_n$, and $e(h) = \mathrm{E}[\hat{e}(h)]$ where the expectation is taken with respect to $z_1, \ldots, z_n$. Here β is a function of the number of observations n and the confidence δ, and depends on $\mathcal{H}$ and $\mathcal{L}$.

This definition simply asserts that for every model in $\mathcal{H}$, its empirical loss on a sample of size n and the expectation of this loss will be "close" when $\beta(n, \delta)$ is small. In general the function β will incorporate standard measures of the complexity of $\mathcal{H}$, and will be a decreasing function of the sample size n, as in the classical $O(\sqrt{d/n})$ bounds of VC theory. Our bounds will be derived from the rich literature on uniform convergence. The only twist to our setting is the fact that the observations are no longer necessarily identically distributed, since they are generated from multiple sources. However, generalizing the standard uniform convergence results to this setting is mostly straightforward as we will see in the upcoming sections on applications of the bound.

We are now ready to present our general bound.

Theorem 3 *Let e be the expected loss function for loss $\mathcal{L}$, and let $\mathcal{F}$ be a class of models for which the α-triangle inequality holds with respect to e. Let $\mathcal{H} \subseteq \mathcal{F}$ be a class of hypothesis models for which there is a uniform convergence bound β for $\mathcal{L}$. Let K, $f = f_1, f_2, \ldots, f_K \in \mathcal{F}$, $\{\epsilon_i\}_{i=1}^{K}$, $\{n_i\}_{i=1}^{K}$, and $\hat{h}_k$ be defined as above. For any δ such that $0 < \delta < 1$, with probability at least $1 - \delta$, for any $k \in \{1, \ldots, K\}$*

$$e(f, \hat{h}_k) \leq \alpha^2 \min_{h \in \mathcal{H}}\{e(f, h)\} + (\alpha + \alpha^2) \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}}\right) \epsilon_i + 2\alpha\beta(n_{1:k}, \delta/2K) .$$

Before providing the proof, let us examine the bound of Theorem 3, which expresses a natural and intuitive trade-off. The first term in the bound is simply the *approximation error*, the residual loss that we incur by limiting our hypothesis model to fall in the restricted class $\mathcal{H}$. The second term is a weighted sum of the disparities of the $k \leq K$ models whose data is used with respect to the target model $f = f_1$. We expect this term to *increase* as we increase k to include more distant sources. The final term is determined by the uniform convergence bound. We expect this term to *decrease* with added sources due to the increased sample size. All three terms are influenced by the strength of the approximate triangle inequality that we have, as quantified by α.

The bound given in Theorem 3 can be loose, but provides an upper bound necessary for optimization and suggests a natural choice for the number of sources k^* to use to estimate the target

f:

$$k^* = \operatorname*{argmin}_k \left((\alpha + \alpha^2) \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \varepsilon_i + 2\alpha\beta(n_{1:k}, \delta/2K) \right) .$$

Theorem 3 and this optimization make the implicit assumption that the best subset of sources to use will be a prefix of the sources—that is, that we should not “skip” a nearby source in favor of more distant ones. This assumption will be true for typical data-independent uniform convergence such as VC dimension bounds, and will be true on average for data-dependent bounds, where we expect uniform convergence bounds to improve with increased sample size.

We now give the proof of Theorem 3.

Proof: (Theorem 3) By Definition 1, for any $h \in \mathcal{H}$, any $k \in \{1, \dots K\}$, and any $i \in \{1, \dots, k\}$,

$$\left(\frac{n_i}{n_{1:k}} \right) e(f,h) \le \left(\frac{n_i}{n_{1:k}} \right) (\alpha e(f, f_i) + \alpha e(f_i, h)) \ .$$

Summing over all $i \in \{1, \dots, k\}$, we find

$$\begin{aligned} e(f,h) &\le \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) (\alpha e(f, f_i) + \alpha e(f_i, h)) \\ &= \alpha \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) e(f, f_i) + \ \alpha \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) e(f_i, h) \le \alpha \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \varepsilon_i + \alpha e_k(h) \ . \end{aligned}$$

In the first line above we have used the α-triangle inequality to deliberately introduce a weighted summation involving the f_i. In the second line, we have broken up the summation using the fact that $e(f, f_i) \le \varepsilon_i$ and the definition of $e_k(h)$. Notice that the first summation is a weighted average of the expected loss of each f_i, while the second summation is the expected loss of h on the data. Using the uniform convergence bound, we may assert that with high probability $e_k(h) \le \hat{e}_k(h) + \beta(n_{1:k}, \delta/2K)$, and with high probability

$$\hat{e}_k(\hat{h}_k) = \min_{h \in \mathcal{H}} \{\hat{e}_k(h)\} \le \min_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) e(f_i, h) + \beta(n_{1:k}, \delta/2K) \right\} \ .$$

Putting these pieces together, we find that with high probability

$$\begin{aligned} e(f, \hat{h}_k) &\le \alpha \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \varepsilon_i + 2\alpha\beta(n_{1:k}, \delta/2K) + \alpha \min_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) e(f_i, h) \right\} \\ &\le \alpha \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \varepsilon_i + 2\alpha\beta(n_{1:k}, \delta/2K) \\ &\qquad + \alpha \min_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \alpha e(f_i, f) + \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \alpha e(f, h) \right\} \\ &= (\alpha + \alpha^2) \sum_{i=1}^{k} \left(\frac{n_i}{n_{1:k}} \right) \varepsilon_i + 2\alpha\beta(n_{1:k}, \delta/2K) + \alpha^2 \min_{h \in \mathcal{H}} \{e(f,h)\} \ . \end{aligned}$$

■

4. Simple Application to Binary Classification

We demonstrate the applicability of the general theory given by Theorem 3 to several standard learning settings. As a warm-up, we begin with the most straightforward application, classification using VC bounds.

In (noise-free) binary classification, we assume that our target model is a fixed, unknown and arbitrary function f from some input set $\mathcal{X}$ to $\{0,1\}$, and that there is a fixed and unknown distribution P on the $\mathcal{X}$. Note that the distribution P over input does not depend on the target function f. The observations are of the form $z = \langle x,y\rangle$ where $y \in \{0,1\}$. The loss function $\mathcal{L}(h,\langle x,y\rangle)$ is defined as 0 if $y = h(x)$ and 1 otherwise, and the corresponding expected loss is $e(g_1,g_2) = \mathrm{E}_{\langle x,y\rangle\sim P_{g_1}}[\mathcal{L}(g_2,\langle x,y\rangle)] = \mathrm{Pr}_{x\sim P}[g_1(x) \neq g_2(x)]$.

For 0/1 loss it is well-known and easy to see that the (standard) 1-triangle inequality holds. Classical VC theory (Vapnik, 1998) provides us with uniform convergence as follows.

Lemma 4 *Let $\mathcal{H} : \mathcal{X} \to \{0,1\}$ be a class of functions with VC dimension d, and let $\mathcal{L}(h,\langle x,y\rangle) = |y-h(x)|$ be the 0/1-loss. The following function β is a uniform convergence bound for $\mathcal{H}$ and $\mathcal{L}$ when $n \geq d/2$:*

$$\beta(n,\delta) = \sqrt{\frac{8(d\ln(2en/d)+\ln(4/\delta))}{n}} \,.$$

The proof is analogous to the standard proof of uniform convergence using the VC Dimension (see, for example, Chapters 2–4 of Anthony and Bartlett (1999)), requiring only minor modifications to the symmetrization argument to handle the fact that the samples need not be uniformly distributed. It relies heavily on Hoeffding's inequality (Hoeffding, 1963), stated here for completeness.

Lemma 5 (Hoeffding's Inequality) *Let X be a set, $D_1,\cdots,D_m$ be probability distributions on X, and $f_1,\cdots,f_m$ be real-valued functions on X such that $f_i : X \to [a_i,b_i]$ for $i = 1,\cdots,m$. Then*

$$\Pr\left(\left|\left(\frac{1}{m}\sum_{i=1}^{m} f_i(x_i)\right) - \left(\frac{1}{m}\sum_{i=1}^{m} E_{x\sim D_i}[f_i(x)]\right)\right| \geq \varepsilon\right) \leq 2\exp\left(\frac{-2\varepsilon^2 m^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\right) \,,$$

where the probability is over the sequence of values x_i distributed according to D_i for all $i = 1,\cdots,m$.

With Lemma 4 in place, the conditions of Theorem 3 are easily satisfied, yielding the following result.

Theorem 6 *Let $\mathcal{F}$ be the set of all functions from an input set $\mathcal{X}$ into {0,1} and let d be the VC dimension of $\mathcal{H} \subseteq \mathcal{F}$. Let e be the expected 0/1 loss. Let K, $f = f_1, f_2, \ldots, f_K \in \mathcal{F}$, $\{\varepsilon_i\}_{i=1}^{K}$, $\{n_i\}_{i=1}^{K}$, and $\hat{h}_k$ be defined as above in the multi-source learning model, and assume that $n_1 \geq d/2$. For any δ such that $0 < \delta < 1$, with probability at least $1-\delta$, for any $k \in \{1,\ldots,K\}$*

$$e(f,\hat{h}_k) \leq \min_{h\in\mathcal{H}}\{e(f,h)\} + 2\sum_{i=1}^{k}\left(\frac{n_i}{n_{1:k}}\right)\varepsilon_i + \sqrt{\frac{32\,(d\ln(2en_{1:k}/d)+\ln(8K/\delta))}{n_{1:k}}} \,.$$

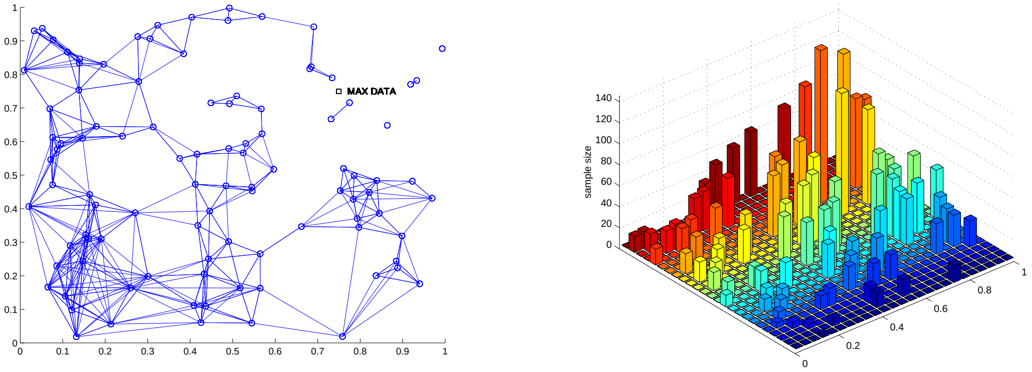

Figure 1: Visual illustration of Theorem 6.

In Figure 1 we provide a visual illustration of the behavior of Theorem 3 applied to a simple classification problem. In this problem there are $K = 100$ classifiers, each classifier f_i for $i = 1 \dots 100$ is defined by 2 parameters represented by a point in the unit square, such that the expected disagreement rate between two such classifiers is proportional the L_1 distance between their parameters.[2] We chose the 100 parameter vectors f_i uniformly at random from the unit square (the circles in the left panel). To generate varying source sizes, we let n_i decrease with the distance of f_i from a chosen "central" point at $(0.75, 0.75)$ (marked "MAX DATA" in the left panel); the resulting source sizes for each model are shown in the bar plot in the right panel, where the origin $(0,0)$ is in the near corner, $(1,1)$ in the far corner, and the source sizes clearly peak near $(0.75, 0.75)$. For every function f_i we used Theorem 6 to find the best sources j to be used to estimate its parameters. The undirected graph on the left includes an edge between f_i and f_j if and only if the data from f_j is used to learn f_i and/or the converse.

The graph simultaneously displays the geometry implicit in Theorem 6 as well as its adaptivity to local circumstances. Near the central point, the graph is sparse and the edges quite short, corresponding to the fact that for such models we have enough direct data (represented with high bars in the right panel) that it is not advantageous to include data from distant models. Far from the central point the graph becomes dense and the edges long, as we are required to aggregate a larger neighborhood to learn the optimal model. In addition, decisions are affected locally by how many models are "nearby" a given model, when there are many close functions f_j to a given f_i there is no need to use "far" models, but when the neighborhood of a function is not populated with many examples, there is a need for data from models far-away.

5. Bounds Using Rademacher Complexity

Given the interest in tighter, potentially data-dependent convergence bounds (such as maximum margin bounds, PAC-Bayes, and others) in recent years, it is natural to ask how our multi-source theory can exploit these modern bounds. We examine one specific case here using Rademacher

[2] It is easy to create simple input distributions and classifiers that generate exactly this geometry. Let the input x be a pair $x = (p,b)$ where $p \in [0,1], b \in \{0,1\}$ and let the hypothesis class consist of functions defined as pairs of thresholds $f = (t_1,t_2)$ where $f(x) = 1$ if and only if $(p > t_1$ and $b = 0)$ or $(p > t_2$ and $b = 1)$. The distribution of $x = (p,b)$ is a product of a uniform distribution for p and a fair coin for b.

complexity (Bartlett and Mendelson, 2002; Bartlett et al., 2002; Koltchinskii, 2001; Koltchinskii and Panchenko, 2000); analogs can be derived in a similar manner for other complexity measures. We start by deriving bounds for settings in which generic Lipschitz loss functions are used, and then discuss specific applications to classification and to regression with squared loss.

5.1 Rademacher Complexity and General Lipschitz-loss Bounds

If $\mathcal{H}$ is a class of functions mapping from a set $\mathcal{X}$ to $\mathbb{R}$, the *empirical Rademacher complexity* of $\mathcal{H}$ on a fixed set of observations $x_1, \dots, x_n$ is defined as

$$\hat{R}_n(\mathcal{H}) = \mathrm{E}\left[\sup_{h \in \mathcal{H}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right| \right] ,$$

where the expectation is taken with respect to independent uniform $\{\pm 1\}$-valued random variables $\sigma_1, \dots, \sigma_n$. The *Rademacher complexity* for n observations can then be defined as $R_n(\mathcal{H}) = \mathrm{E}\left[\hat{R}_n(\mathcal{H})\right]$ where the expectation is with respect to observations $x_1, \dots, x_n$. In the standard setting, $x_1, \dots, x_n$ are drawn i.i.d. from a fixed distribution. In our setting, these observations will still be independent, but not necessarily identically distributed. We will show that the standard uniform convergence results still hold for this modified definition of Rademacher complexity.

Consider any setting in which each generalized data point $z = \langle x, y \rangle$ for some $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ with $y = f(x)$. A *cost function* for the loss $\mathcal{L}$ is a function $\phi(y, a) : \mathbb{R} \to \mathbb{R}$ such that $\mathcal{L}(h, \langle x, y \rangle) = \phi(y, h(x))$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $h \in \mathcal{H}$. We will consider cost functions ϕ that are Lipschitz in the second parameter. Define $\phi'(y, a) = \phi(y, a) - \phi(y, 0)$. Note that if ϕ is Lipschitz in the second parameter with constant L then ϕ' is also Lipschitz in the second parameter with the same constant L.

Lemma 8 below gives a uniform convergence bound for any loss function with a corresponding Lipschitz cost function. The proof of this lemma is in Appendix A. It is analogous to the proof of Theorem 8 in Bartlett and Mendelson (2002), which makes a similar claim in the i.i.d. setting, and uses the following lemma from Bartlett and Mendelson (2002).

Lemma 7 *If $\phi : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L and $\phi(0) = 0$, then $R_n(\phi \circ \mathcal{H}) \leq 2LR_n(\mathcal{H})$.*

Lemma 8 *Let $\mathcal{L}$ be a loss function bounded in $[0, 1]$, and $\phi : \mathbb{R} \to \mathbb{R}$ a cost function such that $\mathcal{L}(f, \langle x, y \rangle) = \phi(y, f(x))$ where ϕ is Lipschitz in the second parameter with constant L. Let $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$ be a class of functions and let $\{\langle x_i, y_i \rangle\}_{i=1}^{n}$ be sampled independently according to some probability distributed P. For any n, for any $0 < \delta < 1$, with probability $1 - \delta$ over samples of length n, every $h \in \mathcal{H}$ satisfies*

$$\beta(n, \delta) = 2LR_n(\mathcal{H}) + \sqrt{\frac{2 \ln(2/\delta)}{n}} .$$

5.2 Application to Classification Using Rademacher Complexity

Theorem 9 below follows from the application of Theorem 3 using the 1-triangle inequality and an application of Lemma 8 with

$$\phi(y, a) = \begin{cases} 1 & \text{if } ya \leq 0, \\ 1 - ya & \text{if } 0 < ya \leq 1, \\ 0 & \text{if } ya > 1. \end{cases}$$

Notice first that if $\mathcal{L}$ is the 0/1 loss, then for all $x \in \mathcal{X}$, $y \in \{-1,1\}$, and $h \in \mathcal{X} \to \{-1,1\}$, $\mathcal{L}(h,\langle x,y\rangle) = \phi(y,h(x))$, and furthermore that ϕ is Lipschitz with constant 1, so Lemma 8 can be applied immediately.

Theorem 9 *Let $\mathcal{F}$ be a set of functions from an input set $\mathcal{X}$ into {-1,1} and let $R_{n_{1:k}}(\mathcal{H})$ be the Rademacher complexity of $\mathcal{H} \subseteq \mathcal{F}$ on the first k sources of data. Let e be the expected 0/1 loss. Let K, $f = f_1, f_2, \ldots, f_K \in \mathcal{F}$, $\{\varepsilon_i\}_{i=1}^{K}$, $\{n_i\}_{i=1}^{K}$, and $\hat{h}_k$ be defined as in the multi-source learning model. For any δ such that $0 < \delta < 1$, with probability at least $1-\delta$, for any $k \in \{1,\ldots,K\}$*

$$e(f,\hat{h}_k) \le \min_{h \in \mathcal{H}}\{e(f,h)\} + 2\sum_{i=1}^{k}\left(\frac{n_i}{n_{1:k}}\right)\varepsilon_i + 2\sqrt{\frac{2\ln(4K/\delta)}{n_{1:k}}} + 4R_{n_{1:k}}(\mathcal{H}) \; .$$

Before moving on, let us briefly examine the behavior of this bound. Similarly to the VC-based bound given in Theorem 6, as k increases and more sources of data are combined, the second term will grow while the third will shrink. The behavior of the final term $R_{n_{1:k}}(\mathcal{H})$, however, is less predictable and may grow or shrink as more sources of data are combined.

Note that for the special case of classification with 0/1 loss, it is possible to get tighter bounds with better dependence on $R_{n_{1:k}}$ by using a more careful analysis than the one in the proof of Lemma 8. Such bounds are given in an earlier version of this paper (Crammer et al., 2007); we choose not to present these alternate bounds here to simplify presentation.

5.3 Regression

We now turn to (noise-free) regression with squared loss. Here our target model f is any function from an input class $\mathcal{X}$ into some bounded subset of $\mathbb{R}$. (Frequently we will have $\mathcal{X} \subseteq \mathbb{R}^d$, but this is not required.) Our loss function is $\mathcal{L}(h,\langle x,y\rangle) = (y-h(x))^2$, and the expected loss is thus $e(g_1,g_2) = \mathrm{E}_{\langle x,y\rangle \sim P_{g_1}}\left[\mathcal{L}(g_2,\langle x,y\rangle)\right] = \mathrm{E}_{x\sim P}\left[(g_1(x)-g_2(x))^2\right]$.

For regression it is known that the standard 1-triangle inequality does not hold. However, a 2-triangle inequality does hold and is stated in the following lemma.

Lemma 10 *Given any three functions $g_1, g_2, g_3 : \mathcal{X} \to \mathbb{R}$, a fixed and unknown distribution P on the inputs $\mathcal{X}$, and the expected loss $e(g_1,g_2) = \mathrm{E}_{x\sim P}\left[(g_1(x)-g_2(x))^2\right]$,*

$$e(g_1,g_2) \le 2\,(e(g_1,g_3) + e(g_3,g_1))\,.$$

Proof: By Jensen's inequality and the convexity of $x \mapsto x^2$, for any g_1, g_2, and g_3,

$$\begin{aligned}
e(g_1,g_2) &= \mathrm{E}_{x\sim P}\left[(g_1(x)-g_2(x))^2\right] \\
&= \mathrm{E}_{x\sim P}\left[4\left(\frac{1}{2}(g_1(x)-g_3(x)) + \frac{1}{2}(g_3(x)-g_2(x))\right)^2\right] \\
&\le \mathrm{E}_{x\sim P}\left[2(g_1(x)-g_3(x))^2 + 2(g_3(x)-g_2(x))^2\right] = 2\,(e(g_1,g_3)+e(g_3,g_1)) \; .
\end{aligned}$$

■

We can derive a uniform convergence bound for squared loss using Rademacher complexity as long as the region $\mathcal{Y}$ is bounded.

Lemma 11 *Let $\mathcal{H} : \mathcal{X} \to [-B,B]$ be a class of functions, and let $\mathcal{L}(h,\langle x,y\rangle) = (y-h(x))^2$ be the squared loss. The following function β is a uniform convergence bound for $\mathcal{H}$ and $\mathcal{L}$:*

$$\beta(n,\delta) = 8BR_n(\mathcal{H}) + 4B^2\sqrt{\frac{2\ln(2/\delta)}{n}} \ .$$

Proof: We cannot apply Lemma 8 directly using the squared loss function, since it may output values outside of the range $[0,1]$. Instead, we apply the Lemma 8 using the alternate loss function $\mathcal{L}'(h,\langle x,y\rangle) = \phi(y,h(x))$ where

$$\phi(y,a) = \begin{cases} \frac{1}{4B^2}(y+B)^2 & \text{if } a < -B, \\ \frac{1}{4B^2}(y-a)^2 & \text{if } -B \le a \le B, \\ \frac{1}{4B^2}(y+B)^2 & \text{if } a > B. \end{cases}$$

It is easy to see that ϕ always outputs values in the range $[0,1]$. Furthermore, for any $y \in [-B,B]$, ϕ is Lipschitz in the second parameter with parameter $1/B$. For any $[a,b] \in [-B,B]$,

$$\begin{aligned} |\phi(y,a)-\phi(y,b)| &= \frac{1}{4B^2}\left|(y-a)^2-(y-b)^2\right| = \frac{1}{4B^2}\left|a^2-b^2+2y(b-a)\right| \\ &\le \frac{1}{4B^2}\left|a^2-b^2\right| + \frac{1}{2B^2}\left|y(a-b)\right| \\ &\le \frac{1}{4B^2}|a+b|\,|a-b| + \frac{1}{2B^2}|y(a-b)| \le \frac{1}{B}|a-b| \ . \end{aligned}$$

Applying Lemma 8 gives a uniform convergence bound of $(2/B)R_n(\mathcal{H}) + \sqrt{2\ln(2/\delta)/n}$ for $\mathcal{L}'$. Scaling by $4B^2$ yields the bound for $\mathcal{L}$. ■

Combining this with Lemma 10 and applying Theorem 3 yields the following.

Theorem 12 *Let $\mathcal{F}$ be the set of functions from $\mathcal{X}$ into $[-B,B]$, and $\mathcal{H} \subseteq \mathcal{F}$. Let e be the expected squared loss. Let K, $f = f_1, f_2, \ldots, f_K \in \mathcal{F}$, $\{\varepsilon_i\}_{i=1}^K$, $\{n_i\}_{i=1}^K$, and $\hat{h}_k$ be defined as in the multi-source learning model. For any δ such that $0 < \delta < 1$, with probability at least $1-\delta$, for any $k \in \{1,\ldots,K\}$*

$$e(f,\hat{h}_k) \le 4\min_{h\in\mathcal{H}}\{e(f,h)\} + 6\sum_{i=1}^{k}\left(\frac{n_i}{n_{1:k}}\right)\varepsilon_i + 32BR_{n_{1:k}}(\mathcal{H}) + 16B^2\sqrt{\frac{2\ln(4K/\delta)}{n_{1:k}}} \ .$$

5.4 Remarks on the Use of Data-Dependent Complexity Measures

The following lemma, which relates the true Rademacher complexity of a function class to its empirical Rademacher complexity, follows directly from Theorem 11 of Bartlett and Mendelson (2002), the proof of which does not require samples to be identically distributed.

Lemma 13 *Let $\mathcal{H}$ be a class of functions mapping to $[-1,1]$. For any integer n, for any $0 < \delta < 1$, with probability $1-\delta$,*

$$\left|R_n(\mathcal{H}) - \hat{R}_n(\mathcal{H})\right| \le \sqrt{\frac{8\ln(2/\delta)}{n}}.$$

This lemma immediately allows us to replace $R_n(\mathcal{H})$ with that data-dependent quantity $\hat{R}_n$ in any of the bounds above for only a small penalty.

While the use of data-dependent complexity measures can be expected to yield more accurate bounds and thus better decisions about the number k^* of sources to use, it is not without its costs in comparison to the more standard data-independent approaches. In particular, in principle the optimization of a data-dependent version of the bound given in Theorem 9 to choose k^* may actually involve running the learning algorithm on all possible prefixes of the sources, since we cannot know the data-dependent complexity term for each prefix without doing so. In contrast, the data-independent bounds can be computed and optimized for k^* without examining the data at all, and the learning performed only once on the first k^* sources. This is especially useful in the case that labels are not free but must be purchased at a price.

6. Estimating the Disparity Matrix

A potential drawback of the theory presented here is the need to estimate the disparity matrix D when it is unknown. However, it is often the case that this matrix can be estimated with many fewer labeled samples than are required for learning. In this section, we discuss how D can be estimated in the classification setting.

As before, consider the scenario in which each target function is a fixed, unknown and arbitrary function from some input set $\mathcal{X}$ to $\{-1,1\}$, and assume that there is a fixed and unknown distribution P over $\mathcal{X}$. Suppose we are given m data points labeled by a pair of functions f_i and f_j, and let $\hat{e}(f_i,f_j)$ be the fraction of points on which the labels disagree. By Hoeffding's inequality, with probability $1-\delta'$,

$$|\hat{e}(f_i,f_j)-e(f_i,f_j)| \leq \sqrt{\frac{\ln(2/\delta')}{2m}}\ .$$

Thus in order to approximate $e(f_i,f_j)$ with an error no more than ε, only $\ln(2/\delta')/(2\varepsilon^2)$ commonly labeled points are needed. Applying the union bound gives us the following lemma.

Lemma 14 *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ into $\{-1,1\}$, and suppose $f_1,\ldots,f_K \in \mathcal{F}$. Let e be the expected 0/1 loss. Suppose that for each pair $i,j \in \{1,\cdots,K\}$, there exist $m_{i,j} \geq m_0$ examples distributed according to P commonly labeled by f_i and f_j, where*

$$m_0 = \frac{2\ln(K)+\ln(2/\delta)}{2\varepsilon^2}$$

for any δ such that $0 \leq \delta \leq 1$, and let $\hat{e}(f_i,f_j)$ be the fraction of commonly labeled examples on which f_i and f_j disagree. Then with probability $1-\delta$, for all $i,j \in \{1,\cdots,K\}$, $|\hat{e}(f_i,f_j)-e(f_i,f_j)| \leq \varepsilon$.

Using the lemma we set the upper bound on the mutual error $e(f_i,f_j)$ between the pair of function f_i and f_j to be $D_{i,j} = \hat{e}(f_i,f_j)+\varepsilon$. With probability at least $1-\delta$ these bound holds simultaneously for all i,j.

Note that in general, $\log(K)$ will be significantly smaller than the dimension d of $\mathcal{H}$. Thus many fewer labeled examples are required to estimate the disparity matrix than to actually learn the best function in the class.

The assumption that there exist commonly labeled points for each pair of functions is natural in many settings. Consider, for example, the problem of predicting whether or not users will enjoy certain movies using ratings from other users. It is often the case that pairs of users will have seen many of the same movies. These commonly rated movies can be used to determine how similar each pair of users are, while ratings of additional movies can be reserved to learn the prediction functions.

7. Estimating the Parameters of a Distribution

We now proceed with the study of the related problem of estimating the unknown parameters of a distribution from multiple sources of data. As in the previous sections, we provide a bound on the diversity of an estimator based on the first k sources from the target. Up until this point, we have measured the diversity between two functions by using the expected value of a loss function. The loss is a function of two *specific* observations. Thus, although two functions may not agree on many points, the diversity between them could be zero (if the measure of their disagreement points is zero). In this section we use a more direct way to measure the diversity between two functions by computing the distance between the parameters used to specify these distributions.

Before stating the problem formally we provide with some illustrative examples for intuition.

Example 1 *We wish to estimate the bias* θ *of a coin given* K *sources of training observations* $N_1, ..., N_K$*. Each source* N_k *contains* n_k *outcomes of flips of a coin with bias* θ_k*. The only information we are given is that* $\theta_k \in [\theta - \varepsilon_k, \theta + \varepsilon_k]$.

In the next example we consider the simple generalization to the multinomial distribution, which involves more than a single parameter.

Example 2 *We wish to estimate the probability* $\Theta^{(p)}$ *of a die to fall on its pth side (out of D possible outcomes) given K sources of training observations* $N_1, ..., N_K$*. Each source* N_k *contains* n_k *outcomes using a die with parameters* $\Theta_k^{(p)}$*. The only information provided is a bound on the* ℓ_∞ *distance between the parameter sets,* $\max_p |\Theta_k^{(p)} - \Theta^{(p)}| = \|\Theta_k - \Theta\|_\infty \leq \varepsilon_k$.

Formally, let $\Pr[X|\Theta]$ be a parametric family of distributions such that $X \in \mathbb{R}^d$ and $\Theta \in \mathbb{R}^D$. We assume that there exists a vector function Ψ such that

$$\mathrm{E}\left[\Psi^{(p)}(X)\right] = \Theta^{(p)} \qquad \text{for } p = 1, \ldots, D .$$

This assumption is met, for example, by any member of the exponential family. In the two examples we have discussed, the function Ψ is simply an identity or indicator. This function is useful because it allows us to estimate the parameters of the distribution from data. Let $X_1, \cdots, X_n$ be a sequence of n i.i.d. samples from such a distribution, where the function Ψ is known. Then the estimator obtained by the method of moments is given by the empirical mean

$$\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} \Psi(X_i) .$$

In our setting, we wish to estimate the parameters Θ of a parametric distribution $\Pr[X|\Theta]$ given K sources of training observations $N_1, ..., N_K$. Each source N_k contains n_k outcomes from a distribution with parameters Θ_k, that is, $\Pr[X|\Theta_k]$. The only information we are given is a bound on the ℓ_∞ distance between the parameter sets, $\|\Theta - \Theta_k\|_\infty \leq \varepsilon_k$.

We first bound the deviation of this estimation from the true parameters using Hoeffding's inequality. Fix the value of the index $p = 1,\ldots,D$. We assume that there exist A and $B > 0$ such that,

$$\Psi^{(p)}(X_i) \in [A, A+B] \quad \text{for } i = 1,\ldots,n \; .$$

Then,

$$\Pr\left[\left|\mathrm{E}\left[\hat{\Theta}^{(p)}\right] - \hat{\Theta}^{(p)}\right| \geq \varepsilon\right] \leq 2\exp\left(-2\frac{n\varepsilon^2}{B^2}\right) \; .$$

Setting the right hand-side of the inequality equal to δ and solving for ε, we get

$$\Pr\left[\left|\mathrm{E}\left[\hat{\Theta}^{(p)}\right] - \hat{\Theta}^{(p)}\right| \geq \sqrt{\frac{B^2 \ln(\frac{2}{\delta})}{2n}}\right] \leq \delta \; .$$

We can use the union bound to bound on this difference for all D parameters at once and get

$$\Pr\left[\exists p \; : \; \left|\mathrm{E}\left[\hat{\Theta}^{(p)}\right] - \hat{\Theta}^{(p)}\right| \geq \sqrt{\frac{B^2 \ln(\frac{2D}{\delta})}{2n}}\right] \leq \sum_{p=1}^{D} \frac{\delta}{D} = \delta \; .$$

This proves the following lemma.

Lemma 15 *Let $X_1,\ldots,X_n$ be a sequence of i.i.d. random variables. Let $\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} \Psi(X_i)$ and $\Theta = \mathrm{E}\left[\hat{\Theta}\right]$, where both $\hat{\Theta}$ and Θ are D-dimensional vectors. Assume that $\Psi^{(p)}(X_i) \in [A, A+B]$ for $i = 1,\ldots,n$, $p = 1,\ldots,D$, for some A and $B > 0$. Then, for any $\delta \in (0,1)$ the following bound holds.*

$$\Pr\left[\|\Theta - \hat{\Theta}\|_\infty \geq \sqrt{\frac{B^2 \ln(\frac{2D}{\delta})}{2n}}\right] \leq \delta \; .$$

We now turn our attention to the problem of choosing the best sources. We define the estimator using the first k sources to be,

$$\hat{\Theta}_k = \frac{1}{n_{1:k}} \sum_{i=1}^{k} \sum_{X \in N_i} \Psi(X) \; ,$$

where as before $n_{1:k} = \sum_{i=1}^{k} n_i$. We denote the expectation of this estimate by

$$\bar{\Theta}_k = \mathrm{E}\left[\hat{\Theta}_k\right] = \frac{1}{n_{1:k}} \sum_{i=1}^{k} n_i \Theta_i \; .$$

We now bound the deviation of the estimate $\hat{\Theta}_k$ from the true set of parameters Θ using the expectation $\bar{\Theta}_k$,

$$\begin{aligned}
\|\Theta - \hat{\Theta}_k\|_\infty &= \|\Theta - \bar{\Theta}_k + \bar{\Theta}_k - \hat{\Theta}_k\|_\infty \\
&\leq \|\Theta - \bar{\Theta}_k\|_\infty + \|\bar{\Theta}_k - \hat{\Theta}_k\|_\infty \\
&\leq \sum_{i=1}^{k} \frac{n_i \|\Theta - \Theta_i\|_\infty}{n_{1:k}} + \|\bar{\Theta}_k - \hat{\Theta}_k\|_\infty \\
&\leq \sum_{i=1}^{k} \frac{n_i}{n_{1:k}} \varepsilon_k + \|\bar{\Theta}_k - \hat{\Theta}_k\|_\infty \; .
\end{aligned}$$

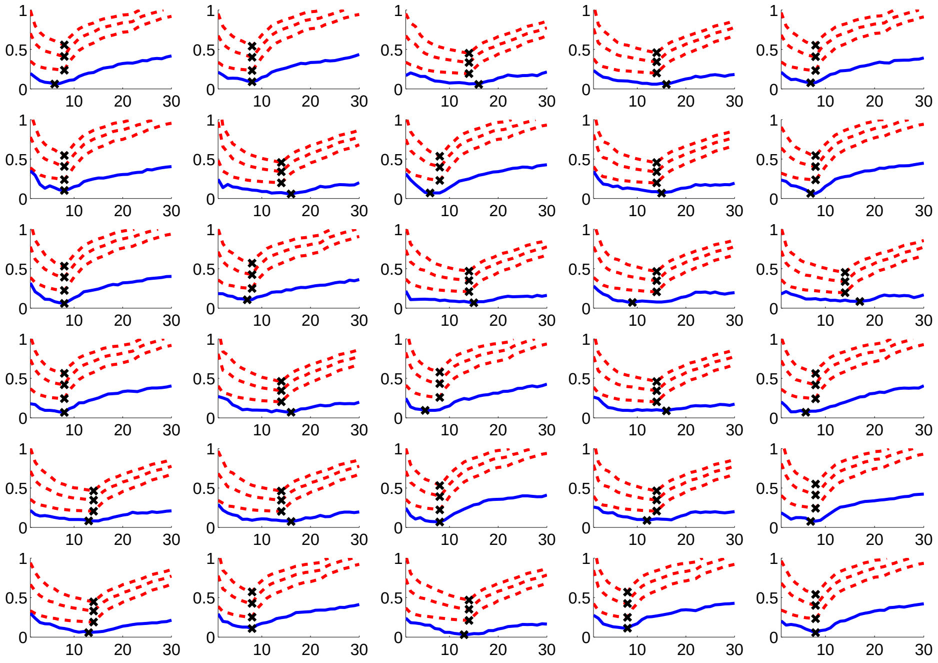

Figure 2: Simulation of the multiple source error bounds.

Let $B = \max_{k=1\ldots K} \sup X \|\Psi(X)\|_\infty$. We can then use Lemma 15 to bound the second term above, yielding the following theorem.

Theorem 16 *Let $\hat{\Theta}_k$ be the estimate of Θ obtained by using only the data from the first k sources, where both $\hat{\Theta}$ and Θ are D-dimensional vectors. Assume that $-B \leq \Psi^{(p)}(X_i) \leq B$. Then for any $\delta > 0$, with probability $\geq 1 - \delta$ we have*

$$\|\Theta - \hat{\Theta}_k\|_\infty \leq \sum_{i=1}^{k} \frac{n_i}{n_{1:k}} \varepsilon_i + \sqrt{\frac{4B^2 \ln(\frac{2DK}{\delta})}{2n_{1:k}}}$$

simultaneously for all $k = 1, \ldots, K$.

As we did with Theorem 3, we can convert Theorem 16 into an algorithm for selecting data sources. Given the K sources of data we simply compute the bounds provided by these theorems for each prefix of the sources of length k and select the subset of sources that yields the smallest bound. A bound for the special case of Example 1 was developed and presented in previous work (Crammer et al., 2006). That bound has the same form as the bound given here in Theorem 16 but with better constants.

8. Synthetic Simulations

In this section, we illustrate the bounds of our main theorem through a simple synthetic simulation. Our hypothesis class $\mathcal{H}$ consists of all linear separators through the origin in 15 dimensions. The goal is to learn thirty classifiers from this class using only limited amounts of data. These data points are drawn uniformly at random from inside the 15-dimensional unit sphere. In this restricted setting, it is easy to calculate the disparity between two functions. Representing each function f by a unit weight vector w such that $f(x) = \text{sign}(w \cdot x)$, the distance between functions w and w' is simply θ/π where $\theta = \arccos(w \cdot w')$ is the angle between w and w'.

In each simulation we ran, the linear classifiers were generated as follows. First, three base classifiers were generated by choosing weight vectors uniformly at random from the surface of the 15-dimensional sphere. Each of the thirty classifiers was then generated by randomly choosing one of the base classifiers, perturbing each coordinate of its weight vector, and renormalizing the perturbed weights.

The number of training samples available for each function was generated from a Poisson distribution with a mean of 8. Each data instance was then sampled from inside the 15-dimensional unit sphere via rejection sampling and labeled by the corresponding classifier, and 500 test samples for each function were generated in the same manner.

To predict the optimal set of training data sources to use for each model, we calculated an approximation of the multiple-source VC bound for classification. It is well known that the constants in the VC-based uniform convergence bounds are not tight. Thus for the purpose of illustrating how these bounds might be used in practice, we have chosen to show approximations of our bounds with a variety of constants. In particular, we have chosen to approximate the bound with

$$2\sum_{i=1}^{k}\left(\frac{n_k}{n_{1:K}}\right)\varepsilon_k + C\sqrt{\frac{(d\ln(2en_{1:K}/d)+\ln(8K/\delta))}{n_{1:K}}}$$

with $\delta = 0.001$ for different values of C. These approximations yield curves that are closer in shape and magnitude to the actual error than a curve generated using the precise, overly conservative constants of Theorem 6.

The set of plots shown in Figure 2 illustrates the results of a single multiple source simulation. (Results from repeated versions of this experiment and experiments with different source sizes were similar.) Each individual plot represents a particular target function. On the x axis is the number of data sources used in training. On the y axis is error. The solid blue curves show test error of a model trained using logistic regression. Dashed red curves show our multiple source error bound with C set to $1/4$ in the lowest curve, $1/2$ in the middle curve, and $1/\sqrt{2}$ in the highest curve. The × on each curve marks the minimum value.

These plots clearly show the trade-off that exists. When too few sources are used, there is not enough data available to learn a 15-dimensional function. When too many sources are used, the labels on the training data often will not correspond to the labels that would have been assigned by the target function. The optimal amount of data lies somewhere in between.

Although the VC bounds remain loose even after constants have been dropped, the bounds tend to maintain the appropriate shape and thus predict the optimal set of sources quite well. In general, when C is set to small values, the predicted error values for small amounts of data (low k) tend to be quite accurate, while predicted values for larger amounts of data overestimate the true error. As C is set to larger values, the predictions become much larger in magnitude than the true error curves, but the shape of the prediction curves become more similar to the true error. In both cases, although the bounds are loose, they can still prove useful in determining the optimal set of sources to consider.

Acknowledgments

We thank the anonymous reviewers for many valuable suggestions, especially on the simplified presentation of the application to regression.

Appendix A. Proof of Lemma 8

The proof relies on McDiarmid's inequality (McDiarmid, 1989), which is stated here for completeness.

Lemma 17 (McDiarmid's inequality) *Let $x_1,\ldots,x_n$ be independent random variables taking on values in a set A and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x'_i \in A} |f(x_1,\ldots,x_n) - f(x_1,\ldots,x_{i-1},x_{i'},x_{i+1},\ldots,x_n)| \le c_i$$

for every $1 \le i \le n$. Then for every $t > 0$,

$$\Pr\left[f(x_1,\ldots,x_n) - \mathrm{E}\left[f(x_1,\ldots,x_n)\right] \ge t\right] \le \exp^{-2t^2/\sum_{i=1}^n c_i^2} .$$

Here we show one direction of the bound, namely that with probability $1-\delta/2$, for all $h \in \mathcal{H}$,

$$e(h) \le \hat{e}(h) + 2LR_n(\mathcal{H}) + \sqrt{\frac{2\ln(2/\delta)}{n}} .$$

The proof of the other direction is nearly identical. For $i \in \{1,\ldots,n\}$, let $\langle x_i, y_i\rangle$ be the ith training instance, distributed according to P_i, and let $\langle x'_i, y'_i\rangle$ be independent random variables drawn according to P_i. Note that for all $h \in \mathcal{H}$,

$$\begin{aligned}
e(h) &= e(h) + \hat{e}(h) - \hat{e}(h) \le \hat{e}(h) + \sup_{h' \in \mathcal{H}} \left(e(h') - \hat{e}(h')\right) \\
&= \hat{e}(h) + \sup_{h' \in \mathcal{H}} \left(\mathrm{E}_{\{\langle x'_i,y'_i\rangle\}_{i=1}^n} \left[\frac{1}{n} \sum_{i=1}^n \phi(y'_i, h'(x'_i)) \right] - \frac{1}{n} \sum_{i=1}^n \phi(y_i, h'(x_i)) \right) \\
&= \hat{e}(h) + \sup_{h' \in \mathcal{H}} \left(\mathrm{E}_{\{\langle x'_i,y'_i\rangle\}_{i=1}^n} \left[\frac{1}{n} \sum_{i=1}^n \phi'(y'_i, h'(x'_i)) + \phi(y'_i, 0) \right] \right. \\
&\qquad\qquad\qquad \left. - \frac{1}{n} \sum_{i=1}^n \phi'(y_i, h'(x_i)) + \phi(y_i, 0) \right) .
\end{aligned}$$

When only one instance $\langle x_i, y_i\rangle$ changes, the sup term can change by at most $2/n$. Thus we can apply McDiarmid's inequality to see that with probability at least $1-\delta/2$,

$$e(h) \le \hat{e}(h) + \mathrm{E}\left[\sup_{h' \in \mathcal{H}} \left(\mathrm{E}\left[\frac{1}{n} \sum_{i=1}^n \phi'(y'_i, h'(x'_i)) \right] - \frac{1}{n} \sum_{i=1}^n \phi'(y_i, h'(x_i)) \right) \right] + \sqrt{\frac{2\ln(2/\delta)}{n}} ,$$

where the outer expectation is with respect to set of training instances $\{\langle x_i, y_i\rangle\}_{i=1}^n$ and the inner expectation is with respect to the set of random variables $\{\langle x'_i, y'_i\rangle\}_{i=1}^n$. Now it suffices to show that this middle term is bounded by $2LR_n(\mathcal{H})$. Using the fact that the supremum of an expectation is less than or equal to the expectation of a supremum, we find that

$$\begin{aligned}
&\mathrm{E}_{\{\langle x_i,y_i\rangle\}_{i=1}^n}\left[\sup_{h'\in\mathcal{H}}\left(\mathrm{E}_{\{\langle x'_i,y'_i\rangle\}_{i=1}^n}\left[\frac{1}{n}\sum_{i=1}^n \phi'(y'_i,h'(x'_i))\right]-\frac{1}{n}\sum_{i=1}^n \phi'(y_i,h'(x_i))\right)\right]\\
\leq\quad &\mathrm{E}_{\{\langle x_i,y_i\rangle\}_{i=1}^n,\{\langle x'_i,y'_i\rangle\}_{i=1}^n}\left[\sup_{h'\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\left(\phi'(y'_i,h'(x'_i))-\phi'(y_i,h'(x_i))\right)\right]\\
=\quad &\mathrm{E}_{\{\langle x_i,y_i\rangle\}_{i=1}^n,\{\langle x'_i,y'_i\rangle\}_{i=1}^n,\{\sigma_i\}_{i=1}^n}\left[\sup_{h'\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i\left(\phi'(y'_i,h'(x'_i))-\phi'(y_i,h'(x_i))\right)\right]\\
\leq\quad &\mathrm{E}_{\{\langle x_i,y_i\rangle\}_{i=1}^n,\{\sigma_i\}_{i=1}^n}\left[\sup_{h'\in\mathcal{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\phi'(y_i,h'(x_i))\right]=R_n(\phi'\circ\mathcal{H})\ .
\end{aligned}$$

Lemma 7 implies that $R_n(\phi'\circ\mathcal{H})\leq 2LR_n(\mathcal{H})$ since ϕ is Lipschitz with parameter L. The result follows.

References

M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

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

P. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

J. Baxter. Learning internal representations. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, 1995.

S. Ben-David. Exploiting task relatedness for multiple task learning. In *Proceedings of the Sixteenth Annual Conference on Computational Learning Theory*, 2003.

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems 20*, 2007.

K. Crammer, M. Kearns, and J. Wortman. Learning from data of variable quality. In *Advances in Neural Information Processing Systems 18*, 2006.

K. Crammer, M. Kearns, and J. Wortman. Learning from multiple sources. In *Advances in Neural Information Processing Systems 19*, 2007.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459, 2000.

A. Maurer. Algorithmic stability and meta-learning. *Journal of Machine Learning Research*, 6: 967–994, 2005.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

P. Wu and T. Dietterich. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004.