Generalization Bounds for Domain Adaptation

Chao Zhang^1, **Lei Zhang**^{2}, **Jieping Ye**^{1,3}
^{1}Center for Evolutionary Medicine and Informatics, The Biodesign Institute,
and ^{3}Computer Science and Engineering, Arizona State University, Tempe, USA
{czhan117,jieping.ye}@asu.edu
^{2}School of Computer Science and Technology,
Nanjing University of Science and Technology, Nanjing, P.R. China
zhanglei.njust@yahoo.com.cn

Abstract

In this paper, we provide a new framework to study the generalization bound of the learning process for domain adaptation. We consider two kinds of representative domain adaptation settings: one is domain adaptation with multiple sources and the other is domain adaptation combining source and target data. In particular, we use the integral probability metric to measure the difference between two domains. Then, we develop the specific Hoeffding-type deviation inequality and symmetrization inequality for either kind of domain adaptation to achieve the corresponding generalization bound based on the uniform entropy number. By using the resultant generalization bound, we analyze the asymptotic convergence and the rate of convergence of the learning process for domain adaptation. Meanwhile, we discuss the factors that affect the asymptotic behavior of the learning process. The numerical experiments support our results.

1 Introduction

In statistical learning theory, one of the major concerns is to obtain the generalization bound of a learning process, which measures the probability that a function, chosen from a function class by an algorithm, has a sufficiently small error (*cf.* [1,2]). Generalization bounds have been widely used to study the consistency of the learning process [3], the asymptotic convergence of empirical process [4] and the learnability of learning models [5]. Generally, there are three essential aspects to obtain generalization bounds of a specific learning process: complexity measures of function classes, deviation (or concentration) inequalities and symmetrization inequalities related to the learning process (*cf.* [3, 4, 6, 7]).

It is noteworthy that the aforementioned results of statistical learning theory are all built under the assumption that training and test data are drawn from the same distribution (or briefly called the assumption of same distribution). This assumption may not be valid in many practical applications such as speech recognition [8] and natural language processing [9] in which training and test data may have different distributions. Domain adaptation has recently been proposed to handle this situation and it is aimed to apply a learning model, trained by using the samples drawn from a certain domain (*source domain*), to the samples drawn from another domain (*target domain*) with a different distribution (*cf.* [10, 11, 12, 13]).

This paper is mainly concerned with two variants of domain adaptation. In the first variant, the learner receives training data from several source domains, known as *domain adaptation with multiple sources* (*cf.* [14, 15, 16, 17]). In the second variant, the learner minimizes a convex combination of empirical source and target risk, termed as *domain adaptation combining source and target data* (*cf.* [13, 18])[1].

1.1 Overview of Main Results

In this paper, we present a new framework to study generalization bounds of the learning processes for domain adaptation with multiple sources and domain adaptation combining source and target data, respectively. Based on the resultant bounds, we then study the asymptotic behavior of the learning process for the two kinds of domain adaptation, respectively. There are three major aspects in the framework: the quantity that measures the difference between two domains, the deviation inequalities and the symmetrization inequalities that are both designed for the situation of domain adaptation[2].

Generally, in order to obtain the generalization bounds of a learning process, it is necessary to obtain the corresponding deviation (or concentration) inequalities. For either kind of domain adaptation, we use a martingale method to develop the related Hoeffding-type deviation inequality. Moreover, in the situation of domain adaptation, since the source domain differs from the target domain, the desired symmetrization inequality for domain adaptation should incorporate some quantity to reflect the difference. We then obtain the related symmetrization inequality incorporating the integral probability metric that measures the difference between the distributions of the source and the target domains.

Next, we present the generalization bounds based on the uniform entropy number for both kinds of domain adaptation. Finally, based on the resultant bounds, we give a rigorous theoretic analysis to the asymptotic convergence and the rate of convergence of the learning processes for both types of domain adaptation. Meanwhile, we give a comparison with the related results under the assumption of same distribution. We also present numerical experiments to support our results.

1.2 Organization of the Paper

The rest of this paper is organized as follows. Section 2 introduces the problems studied in this paper. Section 3 introduces the integral probability metric that measures the difference between the distributions of two domains. We introduce the uniform entropy number for the situation of multiple sources in Section 4. In Section 5, we present the generalization bounds for domain adaptation with multiple sources, and then analyze the asymptotic behavior of the learning process for this type of domain adaptation. The last section concludes the paper. In the supplement (part A), we discuss the relationship between the integral probability metric $D_{\mathcal{F}}(S,T)$ and the other quantities proposed in the existing works including the $\mathcal{H}$-divergence and the discrepancy distance. Proofs of main results of this paper are provided in the supplement (part B). We study domain adaptation combining source and target data in the supplement (part C) and then give a comparison with the existing works on the theoretical analysis of domain adaptation in the supplement (part D).

2 Problem Setup

We denote $\mathcal{Z}^{(S_k)} := \mathcal{X}^{(S_k)} \times \mathcal{Y}^{(S_k)} \subset \mathbb{R}^I \times \mathbb{R}^J$ $(1 \leq k \leq K)$ and $\mathcal{Z}^{(T)} := \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^I \times \mathbb{R}^J$ as the k-th source domain and the target domain, respectively. Set $L = I + J$. Let $\mathcal{D}^{(S_k)}$ and $\mathcal{D}^{(T)}$ stand for the distributions of the input spaces $\mathcal{X}^{(S_k)}$ $(1 \leq k \leq K)$ and $\mathcal{X}^{(T)}$, respectively. Denote $g_*^{(S_k)} : \mathcal{X}^{(S_k)} \rightarrow \mathcal{Y}^{(S_k)}$ and $g_*^{(T)} : \mathcal{X}^{(T)} \rightarrow \mathcal{Y}^{(T)}$ as the labeling functions of $\mathcal{Z}^{(S_k)}$ $(1 \leq k \leq K)$ and $\mathcal{Z}^{(T)}$, respectively. In the situation of domain adaptation with multiple sources, the distributions $\mathcal{D}^{(S_k)}$ $(1 \leq k \leq K)$ and $\mathcal{D}^{(T)}$ differ from each other, or $g_*^{(S_k)}$ $(1 \leq k \leq K)$ and $g_*^{(T)}$ differ from each other, or both of the cases occur. There are sufficient amounts of i.i.d. samples

[1] Due to the page limitation, the discussion on domain adaptation combining source and target data is provided in the supplement (part C).

[2] Due to the page limitation, we only present the generalization bounds for domain adaptation with multiple sources and the discussions of the corresponding deviation inequalities and symmetrization inequalities are provided in the supplement (part B) along with the proofs of main results.

$\mathbf{Z}_1^{N_k} = \{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}$ drawn from each source domain $\mathcal{Z}^{(S_k)}$ $(1 \leq k \leq K)$ but little or no labeled samples drawn from the target domain $\mathcal{Z}^{(T)}$.

Given $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, let $g_{\mathbf{w}} \in \mathcal{G}$ be the function that minimizes the empirical risk

$$\mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) = \sum_{k=1}^K w_k \mathrm{E}_{N_k}^{(S_k)}(\ell \circ g) = \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} \ell(g(\mathbf{x}_n^{(k)}), \mathbf{y}_n^{(k)}) \tag{1}$$

over $\mathcal{G}$ with respect to sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$, and it is expected that $g_{\mathbf{w}}$ will perform well on the target expected risk:

$$\mathrm{E}^{(T)}(\ell \circ g) := \int \ell(g(\mathbf{x}^{(T)}), \mathbf{y}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}), \;\; g \in \mathcal{G}, \tag{2}$$

i.e., $g_{\mathbf{w}}$ approximates the labeling $g_*^{(T)}$ as precisely as possible.

In the learning process of domain adaptation with multiple sources, we are mainly interested in the following two types of quantities:

- $\mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}})$, which corresponds to the estimation of the expected risk in the target domain $\mathcal{Z}^{(T)}$ from a weighted combination of the empirical risks in the multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$;
- $\mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*)$, which corresponds to the performance of the algorithm for domain adaptation with multiple sources,

where $\widetilde{g}_* \in \mathcal{G}$ is the function that minimizes the expected risk $\mathrm{E}^{(T)}(\ell \circ g)$ over $\mathcal{G}$.

Recalling (1) and (2), since

$$\mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ \widetilde{g}_*) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}}) \geq 0,$$

we have

$$\begin{aligned} \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) =& \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) + \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) \\ \leq& \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ \widetilde{g}_*) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g_{\mathbf{w}}) + \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{T}(\ell \circ \widetilde{g}_*) + \mathrm{E}^{T}(\ell \circ \widetilde{g}_*) \\ \leq& 2 \sup_{g \in \mathcal{G}} \big| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \big| + \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*), \end{aligned} \tag{3}$$

and thus

$$0 \leq \mathrm{E}^{(T)}(\ell \circ g_{\mathbf{w}}) - \mathrm{E}^{(T)}(\ell \circ \widetilde{g}_*) \leq 2 \sup_{g \in \mathcal{G}} \big| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \big|.$$

This shows that the asymptotic behaviors of the aforementioned two quantities, when the sample numbers $N_1, \cdots, N_K$ go to *infinity*, can both be described by the supremum

$$\sup_{g \in \mathcal{G}} \big| \mathrm{E}^{(T)}(\ell \circ g) - \mathrm{E}_{\mathbf{w}}^{(S)}(\ell \circ g) \big|, \tag{4}$$

which is the so-called generalization bound of the learning process for domain adaptation with multiple sources.

For convenience, we define the loss function class

$$\mathcal{F} := \{\mathbf{z} \mapsto \ell(g(\mathbf{x}), \mathbf{y}) : g \in \mathcal{G}\}, \tag{5}$$

and call $\mathcal{F}$ as the function class in the rest of this paper. By (1) and (2), given sample sets $\{\mathbf{Z}_1^{N_k}\}_{k=1}^K$ drawn from the multiple sources $\{\mathcal{Z}^{(S_k)}\}_{k=1}^K$ respectively, we briefly denote for any $f \in \mathcal{F}$,

$$\mathrm{E}^{(T)} f := \int f(\mathbf{z}^{(T)}) d\mathrm{P}(\mathbf{z}^{(T)}) \; ; \quad \mathrm{E}_{\mathbf{w}}^{(S)} f := \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} f(\mathbf{z}_n^{(k)}). \tag{6}$$

Thus, we can equivalently rewrite the generalization bound (4) for domain adaptation with multiple sources as

$$\sup_{f \in \mathcal{F}} \big| \mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f \big|. \tag{7}$$

3 Integral Probability Metric

As shown in some prior works (*e.g.* [13, 16, 17, 18, 19, 20]), one of major challenges in the theoretical analysis of domain adaptation is how to measure the distance between the source domain $\mathcal{Z}^{(S)}$ and the target domain $\mathcal{Z}^{(T)}$. Recall that, if $\mathcal{Z}^{(S)}$ differs from $\mathcal{Z}^{(T)}$, there are three possibilities: $\mathcal{D}^{(S)}$ differs from $\mathcal{D}^{(T)}$, or $g_*^{(S)}$ differs from $g_*^{(T)}$, or both of them occur. Therefore, it is necessary to consider two kinds of distances: the distance between $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the distance between $g_*^{(S)}$ and $g_*^{(T)}$.

In [13, 18], the $\mathcal{H}$*-divergence* was introduced to derive the generalization bounds based on the VC dimension under the condition of "λ-close". Mansour *et al.* [20] obtained the generalization bounds based on the Rademacher complexity by using the *discrepancy distance*. Both quantities are aimed to measure the difference between two distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$. Moreover, Mansour *et al.* [17] used the Rényi divergence to measure the distance between two distributions. In this paper, we use the following quantity to measure the difference of the distributions between the source and the target domains:

Definition 3.1 *Given two domains $\mathcal{Z}^{(S)}, \mathcal{Z}^{(T)} \subset \mathbb{R}^L$, let $\mathbf{z}^{(S)}$ and $\mathbf{z}^{(T)}$ be the random variables taking value from $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively. Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be a function class. We define*

$$D_{\mathcal{F}}(S,T) := \sup_{f \in \mathcal{F}} |\mathrm{E}^{(S)} f - \mathrm{E}^{(T)} f|, \tag{8}$$

where the expectations $\mathrm{E}^{(S)}$ and $\mathrm{E}^{(T)}$ are taken with respect to the distributions $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$, respectively.

The quantity $D_{\mathcal{F}}(S,T)$ is termed as the integral probability metric that plays an important role in probability theory for measuring the difference between the two probability distributions (*cf.* [23, 24, 25, 26]). Recently, Sriperumbudur *et al.* [27] gave the further investigation and proposed the empirical methods to compute the integral probability metric in practice. As mentioned by Müller [page 432, 25], the quantity $D_{\mathcal{F}}(S,T)$ is a semimetric and it is a metric if and only if the function class $\mathcal{F}$ separates the set of all signed measures with $\mu(\mathcal{Z}) = 0$. Namely, according to Definition 3.1, given a non-trivial function class $\mathcal{F}$, the quantity $D_{\mathcal{F}}(S,T)$ is equal to *zero* if the domains $\mathcal{Z}^{(S)}$ and $\mathcal{Z}^{(T)}$ have the same distribution.

In the supplement (part A), we discuss the relationship between the quantity $D_{\mathcal{F}}(S,T)$ and other quantities proposed in the previous works, and then show that the quantity $D_{\mathcal{F}}(S,T)$ can be bounded by the summation of the *discrepancy distance* and another quantity, which measure the difference between the input-space distributions $\mathcal{D}^{(S)}$ and $\mathcal{D}^{(T)}$ and the difference between the labeling functions $g_*^{(S)}$ and $g_*^{(T)}$, respectively.

4 The Uniform Entropy Number

Generally, the generalization bound of a certain learning process is achieved by incorporating the complexity measure of the function class, *e.g.*, the covering number, the VC dimension and the Rademacher complexity. The results of this paper are based on the uniform entropy number that is derived from the concept of the covering number and we refer to [22] for more details.

The covering number of a function class $\mathcal{F}$ is defined as follows:

Definition 4.1 *Let $\mathcal{F}$ be a function class and d is a metric on $\mathcal{F}$. For any $\xi > 0$, the covering number of $\mathcal{F}$ at radius ξ with respect to the metric d, denoted by $\mathcal{N}(\mathcal{F}, \xi, d)$ is the minimum size of a cover of radius ξ.*

In some classical results of statistical learning theory, the covering number is applied by letting d be the distribution-dependent metric. For example, as shown in Theorem 2.3 of [22], one can set d as the norm $\ell_1(\mathbf{Z}_1^N)$ and then derive the generalization bound of the i.i.d. learning process by incorporating the expectation of the covering number, *i.e.,* $\mathrm{E}\mathcal{N}(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N))$. However, in the situation of domain adaptation, we only know the information of source domain, while the expectation $\mathrm{E}\mathcal{N}(\mathcal{F}, \xi, \ell_1(\mathbf{Z}_1^N))$ is dependent on both distributions of source and target domains because $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. Therefore, the covering number is no longer applicable to our framework to obtain the generalization bounds for domain adaptation. By contrast, uniform entropy number is distribution-free and thus we choose it as the complexity measure of function class to derive the generalization bounds for domain adaptation.

For clarity of presentation, we give a useful notation for the following discussion. For any $1 \leq k \leq K$, given a sample set $\mathbf{Z}_1^{N_k} := \{\mathbf{z}_n^{(k)}\}_{n=1}^{N_k}$ drawn from $\mathcal{Z}^{(S_k)}$, we denote $\mathbf{Z'}_1^{N_k} := \{\mathbf{z'}_n^{(k)}\}_{n=1}^{N_k}$ as the ghost-sample set drawn from $\mathcal{Z}^{(S_k)}$ such that the ghost sample $\mathbf{z'}_n^{(k)}$ has the same distribution as $\mathbf{z}_n^{(k)}$ for any $1 \leq n \leq N_k$ and any $1 \leq k \leq K$. Denoting $\mathbf{Z}_1^{2N_k} := \{\mathbf{Z}_1^{N_k}, \mathbf{Z'}_1^{N_k}\}$. Moreover, given any $f \in \mathcal{F}$ and any $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$, we introduce a variant of the ℓ_1 norm:

$$\|f\|_{\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)} := \sum_{k=1}^K \frac{w_k}{N_k} \sum_{n=1}^{N_k} \left(|f(\mathbf{z}_n^{(k)})| + |f(\mathbf{z'}_n^{(k)})| \right). \tag{9}$$

It is noteworthy that the variant $\ell_1^{\mathbf{w}}$ of the ℓ_1 norm is still a norm on the functional space, which can be easily verified by using the definition of norm, so we omit it here. In the situation of domain adaptation with multiple sources, setting the metric d as $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$, we then define the uniform entropy number of $\mathcal{F}$ with respect to the metric $\ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)$ as

$$\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi, 2\sum_{k=1}^K N_k\big) := \sup_{\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K} \ln \mathcal{N}\left(\mathcal{F}, \xi, \ell_1^{\mathbf{w}}(\{\mathbf{Z}_1^{2N_k}\}_{k=1}^K)\right). \tag{10}$$

5 Domain Adaptation with Multiple Sources

In this section, we present the generalization bound for domain adaptation with multiple sources. Based on the resultant bound, we then analyze the asymptotic convergence and the rate of convergence of the learning process for such kind of domain adaptation.

5.1 Generalization Bounds for Domain Adaptation with Multiple Sources

Based on the aforementioned uniform entropy number, the generalization bound for domain adaptation with multiple sources is presented in the following theorem:

Theorem 5.1 *Assume that $\mathcal{F}$ is a function class consisting of the bounded functions with the range $[a,b]$. Let $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$. Then, given an arbitrary $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, we have for any $\big(\prod_{k=1}^K N_k\big) \geq \frac{8(b-a)^2}{(\xi')^2}$ and any $\epsilon > 0$, with probability at least $1-\epsilon$,*

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_{\mathbf{w}}^{(S)} f - \mathrm{E}^{(T)} f\right| \leq D_{\mathcal{F}}^{(\mathbf{w})}(S,T) + \left(\frac{\Big(\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k\big) - \ln(\epsilon/8)\Big)}{\frac{\left(\prod_{k=1}^K N_k\right)}{32(b-a)^2\left(\sum_{k=1}^K w_k^2 (\prod_{i\neq k} N_i)\right)}} \right)^{\frac{1}{2}}, \tag{11}$$

where $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ and

$$D_{\mathcal{F}}^{(\mathbf{w})}(S,T) := \sum_{k=1}^K w_k D_{\mathcal{F}}(S_k, T). \tag{12}$$

In the above theorem, we show that the generalization bound $\sup_{f\in\mathcal{F}} |\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f|$ can be bounded by the right-hand side of (11). Compared to the classical result under the assumption of same distribution (*cf.* Theorem 2.3 and Definition 2.5 of [22]): with probability at least $1-\epsilon$,

$$\sup_{f\in\mathcal{F}} \left|\mathrm{E}_N f - \mathrm{E} f\right| \leq O\left(\left(\frac{\ln \mathcal{N}_1\big(\mathcal{F}, \xi, N\big) - \ln(\epsilon/8)}{N} \right)^{\frac{1}{2}} \right) \tag{13}$$

with $\mathrm{E}_N f$ being the empirical risk with respect to the sample set $\mathbf{Z}_1^N$, there is a discrepancy quantity $D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$ that is determined by the two factors: the choice of $\mathbf{w}$ and the quantities $D_{\mathcal{F}}(S_k,T)$ $(1 \leq k \leq K)$. The two results will coincide if any source domain and the target domain match, *i.e.*, $D_{\mathcal{F}}(S_k,T) = 0$ holds for any $1 \leq k \leq K$.

In order to prove this result, we develop the related Hoeffding-type deviation inequality and the symmetrization inequality for domain adaptation with multiple sources, respectively. The detailed proof is provided in the supplement (part B). By using the resultant bound (11), we can analyze the asymptotic behavior of the learning process for domain adaptation with multiple sources.

5.2 Asymptotic Convergence

In statistical learning theory, it is well-known that the complexity of the function class is the main factor to the asymptotic convergence of the learning process under the assumption of same distribution (*cf.* [3, 4, 22]).

Theorem 5.1 can directly lead to the concerning theorem showing that the asymptotic convergence of the learning process for domain adaptation with multiple sources:

Theorem 5.2 *Assume that $\mathcal{F}$ is a function class consisting of bounded functions with the range $[a,b]$. Let $\mathbf{w} = (w_1, \cdots, w_K) \in [0,1]^K$ with $\sum_{k=1}^K w_k = 1$. If the following condition holds:*

$$\lim_{N_1,\cdots,N_K \to +\infty} \frac{\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k\big)}{\frac{\big(\prod_{k=1}^K N_k\big)}{32(b-a)^2\big(\sum_{k=1}^K w_k^2(\prod_{i\neq k} N_i)\big)}} < +\infty \tag{14}$$

with $\xi' = \xi - D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, then we have for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$,

$$\lim_{N_1,\cdots,N_K \to +\infty} \Pr\Big\{ \sup_{f\in\mathcal{F}} \big|\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f\big| > \xi \Big\} = 0. \tag{15}$$

As shown in Theorem 5.2, if the choice of $\mathbf{w} \in [0,1]^K$ and the uniform entropy number $\ln \mathcal{N}_1^{\mathbf{w}}\big(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k\big)$ satisfy the condition (14) with $\sum_{k=1}^K w_k = 1$, the probability of the event that "$\sup_{f\in\mathcal{F}} \big|\mathrm{E}^{(T)} f - \mathrm{E}_{\mathbf{w}}^{(S)} f\big| > \xi$" will converge to *zero* for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$, when the sample numbers $N_1, \cdots, N_K$ of multiple sources go to *infinity*, respectively. This is partially in accordance with the classical result of the asymptotic convergence of the learning process under the assumption of same distribution (*cf.* Theorem 2.3 and Definition 2.5 of [22]): the probability of the event that "$\sup_{f\in\mathcal{F}} \big|\mathrm{E} f - \mathrm{E}_N f\big| > \xi$" will converge to *zero* for any $\xi > 0$, if the uniform entropy number $\ln \mathcal{N}_1(\mathcal{F}, \xi, N)$ satisfies the following:

$$\lim_{N\to+\infty} \frac{\ln \mathcal{N}_1(\mathcal{F}, \xi, N)}{N} < +\infty. \tag{16}$$

Note that in the learning process of domain adaptation with multiple sources, the uniform convergence of the empirical risk on the source domains to the expected risk on the target domain may not hold, because the limit (15) does not hold for any $\xi > 0$ but for any $\xi > D_{\mathcal{F}}^{(\mathbf{w})}(S,T)$. By contrast, the limit (15) holds for all $\xi > 0$ in the learning process under the assumption of same distribution, if the condition (16) is satisfied.

By Cauchy-Schwarz inequality, setting $w_k = \frac{N_k}{\sum_{k=1}^K N_k}$ $(1 \leq k \leq K)$ minimizes the second term of the right-hand side of (11) and then we arrive at

$$\begin{aligned} &\sup_{f\in\mathcal{F}} \big|\mathrm{E}_{\mathbf{w}}^{(S)} f - \mathrm{E}^{(T)} f\big| \\ \leq & \frac{\sum_{k=1}^K N_k D_{\mathcal{F}}(S_k,T)}{\sum_{k=1}^K N_k} + \left(\frac{(\ln \mathcal{N}_1^{\mathbf{w}}(\mathcal{F}, \xi'/8, 2\sum_{k=1}^K N_k) - \ln(\epsilon/8)}{\frac{\big(\sum_{k=1}^K N_k\big)}{32(b-a)^2}} \right)^{\frac{1}{2}}, \end{aligned} \tag{17}$$

which implies that setting $w_k = \frac{N_k}{\sum_{k=1}^K N_k}$ $(1 \leq k \leq K)$ can result in the fastest rate of convergence and our numerical experiments presented in the next section also support this point (*cf.* Fig. 1).

6 Numerical Experiments

We have performed the numerical experiments to verify the theoretic analysis of the asymptotic convergence of the learning process for domain adaptation with multiple sources. Without loss of generality, we only consider the case of $K = 2$, *i.e.,* there are two source domains and one target domain. The experiment data are generated in the following way:

For the target domain $\mathcal{Z}^{(T)} = \mathcal{X}^{(T)} \times \mathcal{Y}^{(T)} \subset \mathbb{R}^{100} \times \mathbb{R}$, we consider $\mathcal{X}^{(T)}$ as a Gaussian distribution $N(0, 1)$ and draw $\{\mathbf{x}_n^{(T)}\}_{n=1}^{N_T}$ ($N_T = 4000$) from $\mathcal{X}^{(T)}$ randomly and independently. Let $\beta \in \mathbb{R}^{100}$ be a random vector of a Gaussian distribution $N(1, 5)$, and let the random vector $R \in \mathbb{R}^{100}$ be a noise term with $R \sim N(0, 0.5)$. For any $1 \leq n \leq N_T$, we randomly draw β and R from $N(1, 5)$ and $N(0, 0.01)$ respectively, and then generate $y_n^{(T)} \in \mathcal{Y}$ as follows:

$$y_n^{(T)} = \langle \mathbf{x}_n^{(T)}, \beta \rangle + R. \tag{18}$$

The derived $\{(\mathbf{x}_n^{(T)}, y_n^{(T)})\}_{n=1}^{N_T}$ ($N_T = 4000$) are the samples of the target domain $\mathcal{Z}^{(T)}$ and will be used as the test data.

In the similar way, we derive the sample set $\{(\mathbf{x}_n^{(1)}, y_n^{(1)})\}_{n=1}^{N_1}$ ($N_1 = 2000$) of the source domain $\mathcal{Z}^{(S_1)} = \mathcal{X}^{(1)} \times \mathcal{Y}^{(1)} \subset \mathbb{R}^{100} \times \mathbb{R}$: for any $1 \leq n \leq N_1$,

$$y_n^{(1)} = \langle \mathbf{x}_n^{(1)}, \beta \rangle + R, \tag{19}$$

where $\mathbf{x}_n^{(1)} \sim N(0.5, 1)$, $\beta \sim N(1, 5)$ and $R \sim N(0, 0.5)$.

For the source domain $\mathcal{Z}^{(S_2)} = \mathcal{X}^{(2)} \times \mathcal{Y}^{(2)} \subset \mathbb{R}^{100} \times \mathbb{R}$, the samples $\{(\mathbf{x}_n^{(2)}, y_n^{(2)})\}_{n=1}^{N_2}$ ($N_2 = 2000$) are generated in the following way: for any $1 \leq n \leq N_2$,

$$y_n^{(2)} = \langle \mathbf{x}_n^{(2)}, \beta \rangle + R, \tag{20}$$

where $\mathbf{x}_n^{(2)} \sim N(2, 5)$, $\beta \sim N(1, 5)$ and $R \sim N(0, 0.5)$.

In this experiment, we use the method of Least Square Regression to minimize the empirical risk

$$\mathrm{E}_w^{(S)}(\ell \circ g) = \frac{w}{N_1} \sum_{n=1}^{N_1} \ell(g(\mathbf{x}_n^{(1)}), y_n^{(1)}) + \frac{(1-w)}{N_2} \sum_{n=1}^{N_2} \ell(g(\mathbf{x}_n^{(2)}), y_n^{(2)}) \tag{21}$$

for different combination coefficients $w \in \{0.1, 0.3, 0.5, 0.9\}$ and then compute the discrepancy $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ for each $N_1 + N_2$. The initial N_1 and N_2 both equal to 200. Each test is repeated 30 times and the final result is the average of the 30 results. After each test, we increment both N_1 and N_2 by 200 until $N_1 = N_2 = 2000$. The experiment results are shown in Fig. 1.

From Fig. 1, we can observe that for any choice of w, the curve of $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ is decreasing when $N_1 + N_2$ increases, which is in accordance with the results presented in Theorems 5.1 & 5.2. Moreover, when $w = 0.5$, the discrepancy $|E_w^{(S)} f - E_{N_T}^{(T)} f|$ has the fastest rate of convergence, and the rate becomes slower as w is further away from 0.5. In this experiment, we set $N_1 = N_2$ that implies that $N_2/(N_1 + N_2) = 0.5$. Recalling (17), we have shown that $w = N_2/(N_1 + N_2)$ will provide the fastest rate of convergence and this proposition is supported by the experiment results shown in Fig. 1.

7 Conclusion

In this paper, we present a new framework to study the generalization bounds of the learning process for domain adaptation. We use the integral probability metric to measure the difference between the distributions of two domains. Then, we use a martingale method to develop the specific deviation inequality and the symmetrization inequality incorporating the integral probability metric. Next, we utilize the resultant deviation inequality and symmetrization inequality to derive the generalization bound based on the uniform entropy number. By using the resultant generalization bound, we analyze the asymptotic convergence and the rate of convergence of the learning process for domain adaptation.

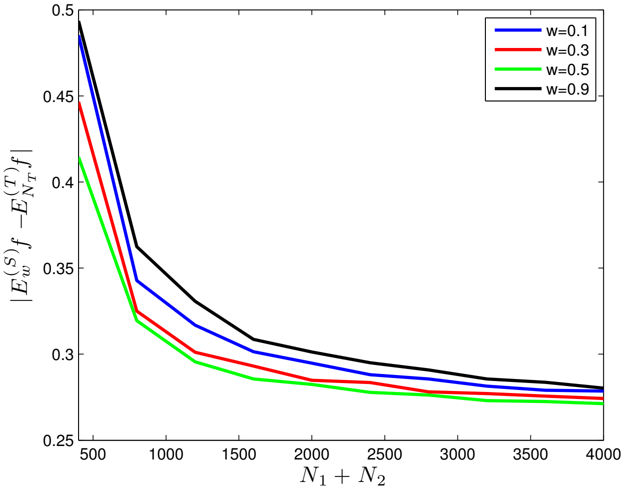

Figure 1: Domain Adaptation with Multiple Sources

We point out that the asymptotic convergence of the learning process is determined by the complexity of the function class $\mathcal{F}$ measured by the uniform entropy number. This is partially in accordance with the classical result of the asymptotic convergence of the learning process under the assumption of same distribution (*cf.* Theorem 2.3 and Definition 2.5 of [22]). Moreover, the rate of convergence of this learning process is equal to that of the learning process under the assumption of same distribution. The numerical experiments support our results. Finally, we give a comparison with the previous works [13, 14, 15, 16, 17, 18, 20] (*cf.* supplement, part D).

It is noteworthy that by Theorem 2.18 of [22], the generalization bound (11) can lead to the result based on the fat-shattering dimension. According to Theorem 2.6.4 of [4], the bound based on the VC dimension can also be obtained from the result (11). In our future work, we will attempt to find a new distance between two distributions and develop the generalization bounds based on other complexity measures, *e.g.*, Rademacher complexities, and analyze other theoretical properties of domain adaptation.

Acknowledgments

This research is sponsored in part by NSF (IIS-0953662, CCF-1025177), NIH (LM010730), and ONR (N00014-1-1-0108).

References

[1] V.N. Vapnik (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks* **10**(5):988-999.

[2] O. Bousquet, S. Boucheron, and G. Lugosi (2004). Introduction to Statistical Learning Theory. In O. Bousquet *et al.* (ed.), *Advanced Lectures on Machine Learning*, 169-207.

[3] V.N. Vapnik (1998). *Statistical Learning Theory.* New York: John Wiley and Sons.

[4] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM* **36**(4):929-965.

[5] A. van der Vaart, and J. Wellner (2000). *Weak Convergence and Empirical Processes With Applications to Statistics (Hardcover).* Springer.

[6] P.L. Bartlett, O. Bousquet, and S. Mendelson (2005). Local Rademacher Complexities. *Annals of Statistics* **33**:1497-1537.

[7] Z. Hussain, and J. Shawe-Taylor (2011). Improved Loss Bounds for Multiple Kernel Learning. *Journal of Machine Learning Research - Proceedings Track* **15**:370-377.

[8] J. Jiang, and C. Zhai (2007). Instance Weighting for Domain Adaptation in NLP. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 264-271.

[9] J. Blitzer, M. Dredze, and F. Pereira (2007). Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 440-447.

[10] S. Bickel, M. Brückner, and T. Scheffer (2007). Discriminative learning for differing training and test distributions. *Proceedings of the 24th international conference on Machine learning (ICML)*, 81-88.

[11] P. Wu, and T.G. Dietterich (2004). Improving SVM accuracy by training on auxiliary data sources. *Proceedings of the twenty-first international conference on Machine learning (ICML),* 871-878.

[12] J. Blitzer, R. McDonald, and F. Pereira (2006). Domain adaptation with structural correspondence learning. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 120-128.

[13] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman (2010). A Theory of Learning from Different Domains. *Machine Learning* **79**:151-175.

[14] K. Crammer, M. Kearns, and J. Wortman (2006). Learning from Multiple Sources. *Advances in Neural Information Processing Systems (NIPS).*

[15] K. Crammer, M. Kearns, and J. Wortman (2008). Learning from Multiple Sources. *Journal of Machine Learning Research* **9**:1757-1774.

[16] Y. Mansour, M. Mohri, and A. Rostamizadeh (2008). Domain adaptation with multiple sources. *Advances in Neural Information Processing Systems (NIPS)*, 1041-1048.

[17] Y. Mansour, M. Mohri, and A. Rostamizadeh (2009). Multiple Source Adaptation and The Rényi Divergence. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI).*

[18] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman (2007). Learning Bounds for Domain Adaptation. *Advances in Neural Information Processing Systems (NIPS).*

[19] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, F (2006). Analysis of Representations for Domain Adaptation. *Advances in Neural Information Processing Systems (NIPS)*, 137-144.

[20] Y. Mansour, M. Mohri, and A. Rostamizadeh (2009). Domain Adaptation: Learning Bounds and Algorithms. *Conference on Learning Theory (COLT).*

[21] W. Hoeffding (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* **58**(301):13-30.

[22] S. Mendelson (2003). A Few Notes on Statistical Learning Theory. *Lecture Notes in Computer Science* **2600**:1-40.

[23] V.M. Zolotarev (1984). Probability Metrics. *Theory of Probability and its Application* **28**(1):278-302.

[24] S.T. Rachev (1991). *Probability Metrics and the Stability of Stochastic Models.* John Wiley and Sons.

[25] A. Müller (1997). Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability* **29**(2):429-443.

[26] M.D. Reid and R.C. Williamson (2011). Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research* **12**:731-817.

[27] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G.R.G. Lanckriet and B. Schölkopf (2009). A Note on Integral Probability Metrics and ϕ-Divergences. *CoRR* abs/0901.2698.