Author/s:
Wu, X;Manton, JH;Aickelin, U;Zhu, J

Title:
Information-theoretic analysis for transfer learning

Date:
2020

Citation:
Wu, X., Manton, J. H., Aickelin, U. & Zhu, J. (2020). Information-theoretic analysis for transfer learning. IEEE International Symposium on Information Theory - Proceedings, 2020-June, pp.2819-2824. IEEE. https://doi.org/10.1109/ISIT44484.2020.9173989.

Information-theoretic analysis for transfer learning

Xuetong Wu[1], Jonathan H. Manton[1], Uwe Aickelin[2], Jingge Zhu[1]
[1]Department of Electrical and Electronic Engineering
[2]Department of Computing and Information Systems
University of Melbourne
Parkville, Victoria, Australia
Email: xuetongw1@student.unimelb.edu, {jmanton, uwe.aickelin, jingge.zhu}.unimelb.edu.au

***Abstract*—Transfer learning, or domain adaptation, is concerned with machine learning problems in which training and testing data come from possibly different distributions (denoted as μ and μ', respectively). In this work, we give an information-theoretic analysis on the generalization error and excess risk of transfer learning algorithms, following a line of work initiated by Russo and Zhou. Our results suggest, perhaps as expected, that the Kullback-Leibler (KL) divergence $D(\mu||\mu')$ plays an important role in characterizing the generalization error in the settings of domain adaptation. Specifically, we provide generalization error upper bounds for general transfer learning algorithms, and extend the results to a specific empirical risk minimization (ERM) algorithm where data from both distributions are available in the training phase. We further apply the method to iterative, noisy gradient descent algorithms, and obtain upper bounds which can be easily calculated, only using parameters from the learning algorithms. A few illustrative examples are provided to demonstrate the usefulness of the results. In particular, our bound is tighter in specific classification problems than the bound derived using Rademacher complexity.**

I. Introduction

Most machine learning methods focus on the setup where the training and testing data are drawn from the same distribution. Transfer learning, or domain adaptation, is concerned with machine learning problems where training and testing data come from possibly different distributions. This setup is of particular interest in real-world applications, as in many cases we often have easy access to a substantial amount of data from one distribution, on which our learning algorithm trains, but wish to use the learnt hypothesis for data coming from a different distribution, from which we have limited data for training.

Generalization error is defined as the difference between the empirical loss and the population loss (defined as (1) and (2) in Section II) for a given hypothesis, and indicates if the hypothesis has been overfitted (or underfitted). Recently, [1] proposed an information-theoretic framework for analyzing generalization error of learning algorithms, and showed that the mutual information between the training data and the output hypothesis can be used to upper bound the generalization error. One nice property of this framework is that the mutual information bound explicitly explores the dependence between training data and the output hypothesis, in contrast to the bounds obtained by traditional methods with VC dimension and Rademacher complexity [2]. As pointed out by [3], the information-theoretic upper bound could be substantially tighter than the traditional bounds, if we could leverage specific properties of the learning algorithms, e.g. the loss function is assumed to be *subgaussian* and learning process forms a subgaussian process. While upper bounds on generalization error are classical results in statistical learning theory, only a relatively small number of papers are devoted to this problem for transfer learning algorithms. To mention a few, Ben-David *et al.* [4] gave VC dimension-style bounds for classification problems. Blitzer *et al.* [5] and Zhang [6] studied similar problems and obtained upper bounds in terms of Rademacher complexity. Specific error bounds for particular transfer learning algorithms and loss metrics are investigated in [7] and [8]. Long *et al.* [9] developed a more general framework for transfer learning where the error is bounded with the distribution difference and output hypothesis adaptability.

Compared with traditional learning problems, the generalization error of transfer learning additionally takes the distribution divergence between the source and target into account and how to evaluate this "domain shift" is non-trivial. We exploit the information-theoretic framework in the transfer learning settings to resolve this issue following the information-theoretic framework studied by [1], [10] and [11]. The main contributions are summarized as follows.

1. We give an information-theoretic upper bound on the generalization error of transfer learning algorithms where training and testing data come from different distributions and KL-divergence between the source and target distribution captures the effect of domain shift.
2. We give upper bounds to the excess risk of a specific ERM algorithm where data from both distributions are available to the learning algorithm. Our example shows that our bound is tighter than the existing bounds in specific classification problems which depend on the Rademacher complexity of the hypothesis space, as our bounds are data-algorithm dependent.
3. We further develop generalization error and excess risk upper bounds for noisy, iterative gradient descent algorithms. The results are useful in the sense that the bounds on the mutual information can be easily calculated only using parameters from the optimization algorithms.

II. Problem formulation and main results

We consider an instance space $\mathcal{Z}$, a hypothesis space $\mathcal{W}$ and a non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}^+$. Let μ

and μ' be two probability distributions defined on $\mathcal{Z}$, and assume that μ is *absolute continuous* with respect to μ' ($\mu \ll \mu'$). In the sequel, the distribution μ is referred to as the *source distribution*, and μ' as the *target distribution*. We are given a set of training data $\{Z_1, \ldots, Z_n\}$. More precisely, for a fixed number $\beta \in [0, 1)$, we assume that the samples $S' = \{Z_1, \ldots, Z_{\beta n}\}$ are drawn IID from the target distribution, and the samples $S = \{Z_{\beta n+1}, \ldots, Z_n\}$ are drawn IID from the source distribution.

In the setup of transfer learning, a learning algorithm is a (randomized) mapping from the training data S, S' to a hypothesis $w \in \mathcal{W}$, characterized by a conditional distribution $P_{W|SS'}$, with the goal to find a hypothesis w that minimizes the population risk with respect to the *target distribution*

$$L_{\mu'}(w) := \mathbb{E}_{Z\sim\mu'}\{\ell(w, Z)\} \tag{1}$$

where Z is distributed according to μ'. Notice that $\beta = 0$ corresponds to the important case when we do not have any samples from the target distribution. Obviously, $\beta = 1$ takes us back to the classical setup where training data comes from the same distribution as test data, which is not our focus.

A. Empirical risk minimization

In this section, we focus on one particular *empirical risk minimization* (ERM) algorithm. For a hypothesis $w \in \mathcal{W}$, the empirical risk of w on a training sequence $\tilde{S} := \{Z_1, \ldots, Z_m\}$ is defined as

$$\hat{L}(w, \tilde{S}) := \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i). \tag{2}$$

Given samples S and S' from both distributions, it is natural to form an empirical risk function as a convex combination of the empirical risk induced by S and S' [4] defined as

$$\hat{L}_\alpha(w, S, S') := \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n} \ell(w, Z_i) + \frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n} \ell(w, Z_i)$$

for some weight parameter $\alpha \in [0, 1]$ to be determined. We define $W_{\mathsf{ERM}} := \operatorname{argmin}_w \hat{L}_\alpha(w)$ as the ERM solution, and also define the optimal hypothesis (with respect to the distribution μ') as $w^* = \operatorname{argmin}_{w\in\mathcal{W}} L_{\mu'}(w)$.

We are interested in two quantities for this ERM algorithm. The first one is the *generalization error* defined as

$$\operatorname{gen}(W_{\mathsf{ERM}}, S, S') := L_{\mu'}(W_{\mathsf{ERM}}) - \hat{L}_\alpha(W_{\mathsf{ERM}}, S, S') \tag{3}$$

namely the difference between the minimized empirical risk and the population risk of the ERM solution under the target distribution. We are also interested in the *excess risk* as

$$R_{\text{excess}}(W_{\mathsf{ERM}}) := L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*)$$

which is the difference between the population risk of W_{ERM} compared to that of the optimal hypothesis. Notice that the excess risk is related to the generalization error via the following upper bound

$$\begin{aligned} L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*) \le \operatorname{gen}(W_{\mathsf{ERM}}, S, S') + \hat{L}_\alpha(w^*, S, S') \\ - L_\alpha(w^*) + (1-\alpha)(L_\mu(w^*) - L_{\mu'}(w^*)) \end{aligned} \tag{4}$$

where we have used the fact $\hat{L}_\alpha(W_{\mathsf{ERM}}, S, S') - \hat{L}_\alpha(w^*, S, S') \le 0$ by the definition of W_{ERM}. For any $w \in \mathcal{W}$, the quantity $L_\alpha(w)$ in the above expression is defined as

$$L_\alpha(w) := (1-\alpha)\mathbb{E}_{Z\sim\mu}\{\ell(w, Z)\} + \alpha\mathbb{E}_{Z\sim\mu'}\{\ell(w, Z)\}.$$

B. Upper bound on generalization errors

We view the ERM solution W_{ERM} as a random variable induced by the random samples S, S' and the (possibly random) ERM algorithm, characterized by a conditional distribution $P_{W|SS'}$. We will first study the expectation of the generalization error

$$\mathbb{E}_{WSS'}\{L_{\mu'}(W_{\mathsf{ERM}}) - \hat{L}_\alpha(W_{\mathsf{ERM}}, S, S')\} \tag{5}$$

where the expectation is taken with respect to the distribution $P_{WSS'}$ defined as

$$P_{WSS'}(w, z^n) := P_{W|SS'}(w|z^n)\prod_{i=1}^{\beta n}\mu'(z_i)\prod_{i=\beta n+1}^{n}\mu(z_i).$$

Furthermore we use P_W to denote the marginal distribution of W induced by the joint distribution $P_{WSS'}$.

Following the characterization used in [11], the following theorem provides an upper bound on the expectation of the generalization error in terms of the mutual information between individual samples Z_i and the any solution W, as well as the KL-divergence between the source and target distributions. As pointed out in [11], using mutual information between the hypothesis and individual samples $I(W; Z_i)$ in general gives a tighter upper bounds than using $I(W; S)$.

Theorem 1 (Generalization error of generic hypothesis)**.** *Assume we have a hypothesis W distributed over P_W (In particular, W is not necessarily the same as $W_{\mathsf{ERM}} := \operatorname{argmin}_w \hat{L}(w, S)$). Assume that the cumulant generating function of the random variable $\ell(W, Z) - \mathbb{E}\{\ell(W, Z)\}$ is upper bounded by $\psi(\lambda)$ in the interval (b_-, b_+) under the product distribution $P_W \otimes \mu'$ for some $b_- < 0$ and $b_+ > 0$. Then for any $\beta > 0$, the expectation of the generalization error in (5) is upper bounded as*

$$\begin{aligned} \mathbb{E}_{WSS'}\{\operatorname{gen}(W, S, S')\} &\le \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\psi_-^{*-1}(I(W; Z_i)) \\ &+ \frac{(1-\alpha)}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\psi_-^{*-1}(I(W; Z_i) + D(\mu||\mu')) \\ -\mathbb{E}_{WSS'}\{\operatorname{gen}(W, S, S')\} &\le \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\psi_+^{*-1}(I(W; Z_i)) \\ &+ \frac{(1-\alpha)}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\psi_+^{*-1}(I(W; Z_i) + D(\mu||\mu')) \end{aligned}$$

where we define

$$\begin{aligned} \psi_-^{*-1}(x) &:= \inf_{\lambda\in[0,-b_-)} \frac{x + \psi(-\lambda)}{\lambda} \\ \psi_+^{*-1}(x) &:= \inf_{\lambda\in[0,b_+)} \frac{x + \psi(\lambda)}{\lambda} \end{aligned}$$

All the proofs in this paper can be found in [12]. In fact, the bound above is not specific to the ERM algorithm, but applicable to any hypothesis generated by a learning algorithm characterized by the conditional distribution $P_{W|S,S'}$. From a stability point of view [13], good algorithms (ERM, for example) should ensure that $I(W;Z_i)$ vanishes as $n \to \infty$. On the other hand, the domain shift is reflected in the KL-divergence $D(\mu||\mu')$, as this term does not vanish when n goes to infinity.

Optimizing α in the above expression is non-trivial as W_{ERM} inexplicitly involves α. However, if we care about the generalization error with respect to the population risk under the target distribution for $n \to \infty$ (the number of samples S' from the target distribution also goes to infinity), the intuition says that we should choose $\alpha = 1$, i.e. only using S' from the target domain in the training process. On the other hand, if we only have limited data samples, α can be set to be β as suggested in [4], [6] that this choice is shown to achieve the faster convergence rate and tighter bound. Overall, we suggest that α should approach 1 with n increasing, say, $\alpha = 1 - O(\frac{1}{n})$.

The result in Theorem 1 does not cover the case $\beta = 0$ (no samples from the target distribution). However, it is easy to see that in this case we should choose $\alpha = 0$ in our ERM algorithm, and a corresponding upper bound is given as in the following corollary under generic hypothesis.

Corollary 1 (Generalization error with source only)**.** *Let $\beta = 0$ so that we only have samples S from the source distribution μ. Let $P_{W|S}$ be the conditional distribution characterizing the learning algorithm which maps samples S to a hypothesis W. Under the same assumption as in Theorem 1, the expected generalization error of W is upper bounded as*

$$\mathbb{E}_{WS}\{L_{\mu'}(W) - \hat{L}(W,S)\} \leq \frac{1}{n}\sum_{i=1}^{n} \psi_{-}^{*-1}(I(W;Z_i) + D(\mu||\mu'))$$

$$-\mathbb{E}_{WS}\{L_{\mu'}(W) - \hat{L}(W,S)\} \leq \frac{1}{n}\sum_{i=1}^{n} \psi_{+}^{*-1}(I(W;Z_i) + D(\mu||\mu'))$$

If the loss function $\ell(W,Z)$ is r^2-subgaussian, namely

$$\log \mathbb{E}\{e^{\lambda(\ell(W,Z) - \mathbb{E}\{\ell(W,Z)\})}\} \leq \frac{r^2\lambda^2}{2}$$

for any $\lambda \in \mathbb{R}$ under the distribution $P_W \otimes \mu'$, the bound in Theorem 1 can be further simplified with $\psi^{*-1}(y) = \sqrt{2r^2 y}$. In particular, if the loss function takes value in $[a,b]$, then $\ell(W,Z)$ is $\frac{(b-a)^2}{4}$-subgaussian. We give the following corollary for the subgaussian loss function.

Corollary 2 (Generalization error for subgaussian loss functions)**.** *If $\ell(w,Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu'$, then the expectation of the generalization error of the ERM solution in (5) is upper bounded as*

$$|\mathbb{E}_{WSS'}\{\mathrm{gen}(W_{\mathsf{ERM}}, S, S')\}| \leq \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i)}$$
$$+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i) + D(\mu||\mu'))}$$

If $\beta = 0$, for any hypothesis $\hat{W}$ (not necessarily the ERM solution) induced by S and a learning algorithm $P_{\hat{W}|S}$, we have the upper bound

$$|\mathbb{E}_{\hat{W}S}\{L_{\mu'}(\hat{W}) - \hat{L}(\hat{W},S)\}| \leq \frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{I(\hat{W};Z_i) + D(\mu||\mu')} \tag{6}$$

The above result follows directly from Theorem 1 and by noticing that we can set $\psi(\lambda) = \frac{r^2\lambda^2}{2}, , b_- = -\infty, b_+ = \infty$ with the assumption that $\ell(W,Z)$ is r^2-subgaussian.

Remark 1. *Using the chain rule of mutual information and the fact that Z_i's are IID, we can relax the upper bound in (6) as*

$$\mathbb{E}_{\hat{W}S}\{L_{\mu'}(\hat{W}) - \hat{L}(\hat{W},S)\} \leq \sqrt{2r^2\left(\frac{I(\hat{W};S)}{n} + D(\mu||\mu')\right)}$$

which recovers the result in the [10] if $\mu = \mu'$. Moreover, we see that the effect of the "change of domain" is simply captured by the KL divergence between the source and the target distribution.

C. Upper bound on the excess risk of ERM

In this section we focus on the case $\beta > 0$ and give a data-dependent upper bound on the excess risk defined in (4). To do this, we first define a L^1 distance quantity between the two divergent distributions as

$$d_{\mathcal{W}}(\mu,\mu') = \sup_{w\in\mathcal{W}} |L_{\mu}(w) - L_{\mu'}(w)|. \tag{7}$$

The following theorem gives a bound for the excess risk.

Theorem 2 (Excess risk of ERM)**.** *Assume that for any $w \in \mathcal{W}$, the loss function $\ell(w,Z)$ is r^2-subgaussian under the distribution $P_w \otimes \mu'$. Then for any $\epsilon > 0$ and $\delta > 0$, there exists an n_0 (depending on δ and ϵ) such that for all $n \geq n_0$, the following inequality holds with probability at least $1-\delta$ (over the randomness of samples and the learning algorithm),*

$$L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*) \leq \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i)}$$
$$+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i) + D(\mu||\mu'))}$$
$$+\sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}} + (1-\alpha)d_{\mathcal{W}}(\mu,\mu') + \epsilon \tag{8}$$

Furthermore in the case when $\beta = 0$ (no samples from the distribution μ'), the inequality becomes

$$L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*) \leq \sqrt{\frac{2r^2\log\frac{2}{\delta}}{n}} + |L_{\mu}(w^*) - L_{\mu'}(w^*)|$$
$$+\frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i) + D(\mu||\mu'))} + \epsilon$$

Note that $d_{\mathcal{W}}(\mu,\mu')$ is normally known as the integral probability metric, which is challenging to evaluate. Sriperumbudur *et al.* [14] investigated the data-dependent estimation

to compute the quantity using Kantorovich metric, Dudley metric and kernel distance, respectively. Ben-David *et al.* [4] proposed another evaluation method to resolve the issue for classification problem. We point out that the result in Theorem 2 is not effective for a class of supervised machine learning problems if μ is not absolutely continuous with respect to μ'. Specifically when the label Y is determined by the features X, the KL divergence is $D(\mu||\mu') = \infty$, leading to a useless bound. To develop an appropriate upper bound to handle such scenarios, we follow the methods in [15] to extend the results by using other types of ϕ-divergence. In particular, we choose $\phi(x) = |x-1|$, which do not impose the absolute continuity restriction.

Corollary 3. *(Generalization error bound of ERM using ϕ-divergence) Assume that for any $w \in \mathcal{W}$, the loss function $\ell(w, Z)$ is L_∞-norm bounded by σ under the distribution $P_W \otimes \mu'$. Then for any $\epsilon > 0$ and $\delta > 0$, there exists an n_0 (depending on δ and ϵ) such that for all $n \geq n_0$, the following inequality holds with probability at least $1-\delta$ (over the randomness of samples and the learning algorithm) that*

$$L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*) \leq \frac{\alpha\|\sigma\|_\infty}{\beta n}\sum_{i=1}^{\beta n} I_\phi(W_{\mathsf{ERM}}; z_i) + \frac{(1-\alpha)\|\sigma\|_\infty}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I_\phi(W_{\mathsf{ERM}}; z_i) + 2TV(\mu||\mu')\right) + \epsilon$$

where $I_\phi(W_{\mathsf{ERM}}; z_i) = D_\phi(P_{W_{\mathsf{ERM}}, z_i}||P_{W_{\mathsf{ERM}}} \otimes P_{z_i})$ is the ϕ-divergence between the distribution $P_{W_{\mathsf{ERM}}, z_i}$ and $P_{W_{\mathsf{ERM}}} \otimes P_{z_i}$ with $D_\phi(P||Q) = \int |dP - dQ|$ and $TV(\mu||\mu') = \frac{1}{2}D_\phi(\mu||\mu')$ denotes the total variation distance between the distribution μ and μ'.

D. Generalization error bound for noisy gradient descent algorithm

The upper bound obtained in previous section cannot be evaluated directly as it depends on the distribution of the data, which is in general assumed unknown in learning problems. Furthermore, in most cases, W_{ERM} does not have a closed-form solution, but obtained by using an optimization algorithm. In this section, we study the class of optimization algorithms that iteratively update its optimization variable based on both source S and target dataset S'. The upper bound derived in this section are useful in the sense that the bound can be easily calculated if the relative learning parameters are given. Specifically, the hypothesis W is represented by the optimization variable of the optimization algorithm, and we use $W(t)$ to denote the variable at iteration t. In particular, we consider the following noisy gradient descent algorithm

$$W(t) = W(t-1) - \eta_t \nabla \hat{L}_\alpha(W(t-1), S, S') + n(t) \quad (9)$$

where $W(t)$ is initialized to be $W(0) \in \mathcal{W}$ arbitrarily, $\nabla \hat{L}_\alpha$ denotes the gradient of $\hat{L}_\alpha$ with respect to W, and $n(t)$ can be any noises with the mean value of 0 and variance of $\sigma_t^2 I_d \in \mathbb{R}^d$. A typical example is $n(t) \sim \mathcal{N}(0, \sigma_t^2 I_d)$.

Theorem 3 (Generalization error of noisy gradient descent)**.** *Assume that $W(T)$ is obtained from (9) at T iteration, and assume that $\ell(w, Z)$ is r^2-subgaussian over $P_W \otimes \mu'$, and the gradient is bounded, e.g., $\|\nabla(\ell(w(t), Z))\|_2 \leq K_{ST}$ for any $w(t)$. then*

$$\mathbb{E}_{wSS'}\left\{\mathrm{gen}\left(W(T), S, S'\right)\right\} \leq \alpha\sqrt{\frac{2r^2}{\beta n}\hat{I}(S)} + (1-\alpha)\sqrt{2r^2\left(\frac{\hat{I}(S)}{(1-\beta)n} + D\left(\mu\|\mu'\right)\right)} \quad (10)$$

where we define

$$\hat{I}(S) := \frac{d}{2}\sum_{t=1}^{T}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right) - \sum_{t=1}^{T} h(n_t) \quad (11)$$

In this bound, we observe that if the optimization parameters (such as $\alpha, \beta, n(t), w(0), T, d$) and loss function are fixed, the generalization error bound is easy to calculate by using the parameters given above. Also note that our assumptions do not require that the noise is Gaussian distributed or the loss function $\ell(w; z)$ is convex, this generality provides a possibility to tackle a wider range of optimization problems. However, in many cases the generalization error does not fully reflect the effectiveness of the hypothesis if $W(T) \neq W_{\mathsf{ERM}}$. One can further provide an excess risk upper bound by utilizing the proposition 3 in [16] with the assumption of strongly convex loss function, which guarantees the convergence of hypothesis.

III. Examples

In this section, we provide two simple examples to illustrate the upper bounds we obtained in previous sections.

A. Estimating the mean of Gaussian

We consider an example studied in [11]. Assume that S comes from the source distribution $\mu = \mathcal{N}(m, \sigma^2)$ and S' comes form the target distribution $\mu' = \mathcal{N}(m', \sigma^2)$ where $m \neq m'$. We define the loss function as

$$\ell(w, z) = (w - z)^2.$$

For simplicity we assume here that $\beta = 0$. The empirical risk minimization (ERM) solution is obtained by minimizing $\hat{L}(w, S) := \frac{1}{n}\sum_{i=1}^{n}(w - Z_i)^2$, where the solution is given by

$$W_{\mathsf{ERM}} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

To obtain the upper bound, we first notice that in this case

$$I(W_{\mathsf{ERM}}; Z_i) = \frac{1}{2}\log\frac{n}{n-1}$$

for all i. It is easy to see that the loss function $\ell(W; Z_i)$ is non-central chi-square distribution $\chi^2(1)$ of 1 degree of freedom with the variance of $\sigma_\ell^2 = \frac{n+1}{n}\sigma^2$. Furthermore, the cumulant generating function can be bounded as

$$\log\mathbb{E}e^{\lambda(\ell(W;Z_i) - \mathbb{E}\ell(W;Z_i))} \leq \sigma_\ell^4\lambda^2 + \frac{2\lambda^2\sigma_\ell^2(m-m')^2}{1 + 2\lambda\sigma_\ell^2}, \text{ for } \lambda > 0$$

By Corollary 1, the generalization error bound is given as

$$\mathbb{E}\{\mathrm{gen}(W_{\mathsf{ERM}})\} \leq \frac{1}{n}\sum_{i=1}^{n}\psi^{*-1}(I(W_{\mathsf{ERM}}; Z_i) + D(\mu||\mu'))$$

By the definition of $\psi^{*-1}(x)$,

$$\psi^{*-1}(x) \geq (m-m')^2 + \sigma_\ell^4 \lambda + \frac{I(W;Z)}{\lambda}$$

We set $\lambda = \sqrt{\frac{I(W;Z_i)}{\sigma_\ell^4}}$ and substitute $I(W_{\mathsf{ERM}}; Z_i)$ in the generalization error above, we reach

$$\mathbb{E}\{\text{gen}(W_{\mathsf{ERM}})\} \leq 2\left(\frac{n+1}{n}\right)\sigma^2\sqrt{\frac{1}{2}\log\frac{n}{n-1}} + 2\sigma^2 D(\mu\|\mu')$$

where $D(\mu||\mu') = \frac{(m-m')^2}{2\sigma^2}$. In this case, the generalization error of W_{ERM} can be calculated exactly to be

$$\mathbb{E}\{\hat{L}(W_{\mathsf{ERM}}, S) - L_{\mu'}(W_{\mathsf{ERM}})\} = \frac{2\sigma^2}{n} + 2\sigma^2 D(\mu||\mu')$$

The derived bound approaches $2\sigma^2 D(\mu||\mu')$ as $n \to \infty$ with a decay rate $O(1/\sqrt{n})$. The derived bound captures the bound asymptotically well with a lower rate, which is often the results using Rademacher complexity bound [6].

B. Logistic regression transfer

In this section, we apply our bound in a typical classification problem. Consider the following logistic regression problem in a 2-dimensional space shown in Figure 1. For each $w \in \mathbb{R}^2$ and $z_i = (x_i, y_i) \in \mathbb{R}^2 \times \{0, 1\}$, the loss function is given by

$$\ell(w, z_i) := -(y_i \log(\sigma(w^T x_i)) + (1-y_i)\log(1-\sigma(w^T x_i)))$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$.

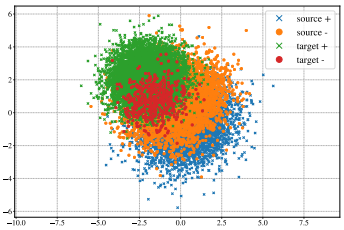

Fig. 1. The source data x_i are sampled from the **truncated** Gaussian distribution $\mathcal{N}_{tc} \sim (\mathbf{0}, 2\mathbf{I})$ while the target data are sampled from the **truncated** Gaussian distribution $\mathcal{N}_{tc} \sim ((-2, 2), \mathbf{I})$. The according label $y \in \{0, 1\}$, is generated from the Bernoulli distribution with probability $p(1) = \frac{1}{1+e^{-w^T x}}$, where $w_s = (0.5, -1)$ for the source and $w_t = (-0.5, 1.5)$ for the target.

Here we truncate the Gaussian random variables $x_i = \{(x_1, x_2)\big|\|x_1\|_2 < 6, \|x_2\|_2 < 6\}$, for $i = 1, \cdots, n$. We also restrict hypothesis space as $\mathcal{W} = \{w : \|w\|_2 < 3\}$ where W_{ERM} falls in this area with high probability. It can be easily checked that $\mu \ll \mu'$ and the loss function is bounded, hence we can upper bound generalization error using Corollary 2. To this end, we firstly fix the source samples $n_s = 10000$, while the target samples n_t varies from 100 to 100000 and $\alpha = \beta = \frac{n_t}{n_s+n_t}$ following the guideline from [4], [6]. We give the empirical estimation for r^2 within the according hypothesis space such that

$$r^2 = \frac{(\max_{Z\in S', w\in\mathcal{W}} \ell(w, Z) - \min_{Z\in S', w\in\mathcal{W}} \ell(w, Z))^2}{4}$$

To evaluate the mutual information $I(W_{\mathsf{ERM}}, Z_i)$ efficiently, we follow the work [17] by repeatedly generating W_{ERM} and Z_i. As $\mu \ll \mu'$, we decompose $D(\mu(X,Y)\|\mu'(X',Y')) = D(P_X\|P_{X'}) + \mathbb{E}_{X\sim P_X}\{D(P_{Y|X=x}\|P_{Y'|X=x})\}$ in terms of the feature distributions and conditional distributions of the labels. The first term $D(P_X\|P_{X'})$ can be calculated using the parameters of Gaussian distributions. The latter term denotes the expected KL-divergence over P_X between two Bernoulli distributions, which can be evaluated by generating abundant samples from the source domain. Further we apply Theorem 2 to upper bound the excess risk, where we give a data-dependent estimation for the term $d_{\mathcal{W}}(\mu, \mu')$ as

$$\hat{d}_{\mathcal{W}}(\mu, \mu') = \sup_{w\in\mathcal{W}} |\hat{L}(w, S) - \hat{L}(w, S')|.$$

To demonstrate the usefulness of our algorithm, we compare the bound in the following theorem using the Rademacher complexity under the same domain adaptation framework. Detailed experiment settings can be found in [12].

Theorem 4. *(Generalization error of ERM with Rademacher complexity) [6, Theorem 6.2] Assume that for any $w \in \mathcal{W}$, the loss function $\ell(w, Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu$ or $P_W \otimes \mu'$. Then for any $\delta > 0$, the following inequality holds with probability at least $1-\delta$ (over the randomness of samples and the learning algorithm)*

$$\mathbb{E}\{\text{gen}(W_{\mathsf{ERM}})\} \leq (1-\alpha)d_{\mathcal{W}}(\mu, \mu') + 2\alpha E_{\sigma\otimes\mu}\{\sup_{w\in\mathcal{W}} \sigma\ell(Z, w)\}$$
$$+ \frac{2(1-\alpha)}{\beta n}\mathrm{E}_\sigma\{\sup_{w\in\mathcal{W}}\sum_{i=1}^{\beta n}\sigma_i\ell(z_i, w)\} + 3\alpha\sqrt{\frac{r\ln(4/\delta)}{\beta n}}$$
$$+ (1-\alpha)\sqrt{2r^2\ln(\frac{2}{\delta})\left(\frac{\alpha^2}{\beta n} + \frac{(1-\alpha)^2}{(1-\beta)n}\right)}$$

where σ is randomly selected from {-1,+1}.

The comparisons of generalization error bound and excess risk bound are shown in figure 2. It is obvious that the true losses are bounded by our developed upper bounds. The result also suggests that our bound is tighter than Rademacher complexity bound in terms of both generalization error and excess risk. This is possibly due to that the generalization error bound with Rademacher complexity is characterized by the domain difference in the whole hypothesis space, while our bound is data-algorithm dependent, which is only concerned with W_{ERM}. As expected, the data-algorithm dependent bound captures the true behaviour of generalization error while Rademacher complexity bound fails to do so. It is noteworthy that both bounds converge as n increases. The result confirms that the bounds captures the dependence of the input data and output hypothesis, as well as the stochasticity of the algorithm.

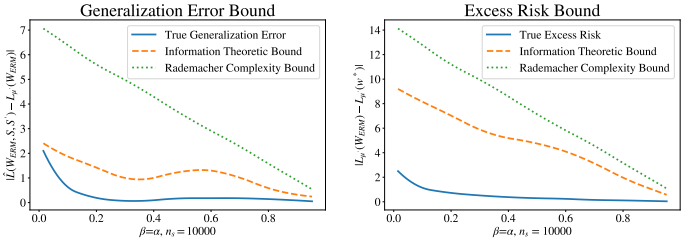

Fig. 2. Comparisons for generalization error and excess risk

REFERENCES

[1] D. Russo and J. Zou, "Controlling Bias in Adaptive Data Analysis Using Information Theory," in *Artificial Intelligence and Statistics*, May 2016, pp. 1232–1240.

[2] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning by Shai Shalev-Shwartz*. Cambridge Core, May 2014.

[3] A. Asadi, E. Abbe, and S. Verdu, "Chaining Mutual Information and Tightening Generalization Bounds," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 7234–7243.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine Learning*, vol. 79, no. 1, pp. 151–175, May 2010.

[5] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning Bounds for Domain Adaptation," in *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 129–136.

[6] C. Zhang, L. Zhang, and J. Ye, "Generalization bounds for domain adaptation," in *Advances in neural information processing systems*, 2012, pp. 3320–3328.

[7] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 193–200.

[8] Y. Zhang, T. Liu, M. Long, and M. I. Jordan, "Bridging theory and algorithm for domain adaptation," *arXiv preprint arXiv:1904.05801*, 2019.

[9] M. Long, J. Wang, G. Ding, S. J. Pan, and S. Y. Philip, "Adaptation regularization: A general framework for transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1076–1089, 2013.

[10] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 2524–2533.

[11] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening Mutual Information Based Bounds on Generalization Error," *arXiv:1901.04609 [cs, stat]*, Jan. 2019, arXiv: 1901.04609. [Online]. Available: http://arxiv.org/abs/1901.04609

[12] "Supplementary materials - proofs." [Online]. Available: https://github.com/wfyitf/Information-Theoretic-for-Domain-Adaptation/blob/master/Proof_ISIT2020.pdf

[13] O. Bousquet and A. Elisseeff, "Stability and Generalization," *Journal of Machine Learning Research*, vol. 2, no. Mar, pp. 499–526, 2002.

[14] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet *et al.*, "On the empirical estimation of integral probability metrics," *Electronic Journal of Statistics*, vol. 6, pp. 1550–1599, 2012.

[15] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *2017 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 1475–1479.

[16] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, 2011, pp. 1458–1466.

[17] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal processing*, vol. 16, no. 3, pp. 233–248, 1989.