Information-Theoretic Bounds on Transfer Generalization Gap Based on Jensen-Shannon Divergence

Sharu Theresa Jose and Osvaldo Simeone

***Abstract*—In transfer learning, training and testing data sets are drawn from different data distributions. The transfer generalization gap is the difference between the population loss on the target data distribution and the training loss. The training data set generally includes data drawn from both source and target distributions. This work presents novel information-theoretic upper bounds on the average transfer generalization gap that capture (i) the domain shift between the target data distribution P'_Z and the source distribution P_Z through Nielsen's family of α-Jensen-Shannon (JS) divergences $D^{\alpha}_{\rm JS}(P'_Z||P_Z)$; and (ii) the sensitivity of the transfer learner output W to each individual sample of the data set Z_i via the mutual information $I(W;Z_i)$. The α-JS divergence is bounded even when the support of P_Z is not included in that of P'_Z. This contrasts the Kullback-Leibler (KL) divergence $D_{\rm KL}(P_Z||P'_Z)$-based bounds of Wu et al. [1], which are vacuous under this assumption. Moreover, the obtained bounds hold for unbounded loss functions with bounded cumulant generating functions, unlike the ϕ-divergence based bound of Wu et al. We also obtain new upper bounds on the average transfer excess risk in terms of the α-JS divergence for empirical weighted risk minimization (EWRM), which minimizes the weighted average training losses over source and target data sets. Finally, we provide a numerical example to illustrate the merits of the introduced bounds.**

I. Introduction

In conventional learning, data sets for training and testing are drawn from the same underlying data distribution. *Transfer learning*, or domain adaptation, considers the scenario where a learning algorithm trained using a data set drawn from a source data distribution, or *source domain*, is tested on a data set drawn from a generally different target data distribution, or *target domain*. In practice, while abundant labelled data from source domain are available during training, few or no labelled data from the target domain are available. The goal of transfer learning is to infer a model parameter w from observation of the data from the source domain and possibly also from target domain, so that it generalizes well on test data from the target domain [2].

The objective of the transfer learner is to minimize the generalization, or population, loss $L_g(w)$, which is the average loss of model parameter w over the test data drawn from the target data distribution. However, this is not available at the learner since the target domain distribution is unknown. Instead, the learner can compute the empirical training loss $L_t(w|Z^M)$ of the parameter w on the data set Z^M, which is comprised of data from source and, possibly, target domains. The difference between the generalization loss and the training loss, known as the *transfer generalization gap*, is a key metric to evaluate the performance of a transfer learning algorithm. Specifically, if the transfer generalization gap is small, on average or with high probability, the performance of the model parameter w on the training loss can be taken as a reliable estimate of the generalization loss.

Theoretical work on transfer learning has largely concentrated on upper bounding the generalization loss $L_g(w)$ in the target domain in terms of (i) the generalization loss within the source domain, and (ii) the performance degradation due to the *domain shift* between source and target domains [3], [4]. The main goal of these studies has been to define appropriate distance measures to capture the impact of the domain shift that can be estimated from finite data with reasonable accuracy. Notably, Ben et al. [3] introduces the $\mathscr{H}$-divergence for classification tasks, and obtains high-probability gaurentee on empirical estimates of this divergence in terms of the VC (Vapnik-Chervonenkis) dimension. Various refinements of the $\mathscr{H}$-divergence have been studied including the discrepancy distance in [4], which can account for loss functions beyond the detection loss, the integral probability metric in [5], and the $\mathscr{H}\Delta\mathscr{H}$ divergence in [6].

Recently, inspired by the works in [7], [8], Wu et al. in [1] considered an information-theoretic framework to obtain upper bounds on the average transfer generalization gap when data from both source and target domains are available for training. The resulting bound captures the impact of the domain shift via the Kullback-Leibler (KL) divergence $D_{\rm KL}(P_Z||P'_Z)$ between the source-domain data distribution P_Z and target-domain data distribution P'_Z. The KL divergence based measure of domain shift suffers from a serious disadvantage: it is well-defined only when the source distribution P_Z is absolutely continuous with respect to P'_Z ($P_Z \ll P'_Z$), and takes value ∞ otherwise. This results in vacuous bounds under various practical conditions, such as for supervised learning problems where the data labels Y are deterministic functions of the feature X within data samples $Z=(X,Y)$; and when the support of the source data distribution includes that of the target data distribution.

In this work, we mitigate the above drawback of KL divergence based bounds on average transfer generalization gap,

The authors are with King's Communications, Learning, and Information Processing (KCLIP) lab at the Department of Engineering of Kings College London, UK (emails: sharu.jose@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). The authors have received funding from the European Research Council (ERC) under the European Unions Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

by introducing the α-family Jensen-Shannon (JS) divergence of Nielsen [9] to capture the domain shift. Precisely, for the setting when data from both source and target distributions are available for training, we obtain new information-theoretic upper bounds on the average transfer generalization gap that capture (i) the impact of the domain shift via the α-JS divergence $D^{\alpha}_{\mathrm{JS}}(P'_Z||P_Z)$ between source P_Z and target P'_Z distributions; and (ii) the generalization capability of the transfer learning algorithm through the mutual information between algorithm output and each individual sample of data set. The α-JS divergence is bounded for $\alpha \in (0,1)$, and gives non-vacuous bounds when $P_Z \not\ll P'_Z$. Moreover, the obtained α-JS divergence-based bound holds for unbounded loss functions with bounded cumulant generating function (CGF)[1]. In contrast, the ϕ-divergence based bound with $\phi(x)=|x-1|$ in [1, Corollary 3], which also holds when $P_Z \not\ll P'_Z$, requires loss functions to have bounded L_∞-norm.

Our work is motivated by the recent study [10] that employs JS divergence, a member of α-JS divergence corresponding to $\alpha = 0.5$, with the aim of upper bounding the target domain generalization loss $L_g(w)$ as a function of the source-domain generalization loss for a fixed model parameter w. The JS divergence was favoured since it can be estimated using binary classification given data drawn both from distributions.

Moving beyond [10], which focus on the performance of a given model parameter w, in this work, we consider the performance of a training algorithm that chooses model parameter w by minimizing the weighted average of training losses over source and target data [1]– an approach referred to as *empirical weighted risk minimization* (EWRM). We specialize the α-JS divergence-based bounds on average transfer generalization gap to EWRM, and obtain new upper bounds on the average transfer excess risk for EWRM. We show that by choosing the mixing ratio α, the α-JS divergence can better capture the relative impact of source and target data sets on the EWRM, yielding tighter bounds than the one with $\alpha = 0.5$.

A. Notation

We use upper case letters, e.g. X, to denote random variables and lower case letters, e.g. x to represent their realizations. For a discrete or continuous random variable X taking values in a set or vector space $\mathcal{X}$, P_X denotes its probability distribution, with $P_X(x)$ being the probability mass or density value at $X = x$. We denote as P_X^N the $N-$fold product distribution induced by P_X. The conditional distribution of a random variable X given random variable Y is similarly defined as $P_{X|Y}$. We define as $\mathrm{supp}(P_X)$ the support of distribution P_X.

II. Problem Formulation

In transfer learning, we are given a data set that consists of: (i) data points from a *source domain* with an underlying *unknown* data distribution, $P_Z \in \mathcal{P}(\mathcal{Z})$, defined in a subset or vector space $\mathcal{Z}$; as well as (ii) data from a *target domain* with a generally different data distribution $P'_Z \in \mathcal{P}(\mathcal{Z})$. Specifically, the learner has access to a training data set $Z^M = (Z_1, Z_2, \ldots, Z_M)$, which consists of βM, for some fixed $\beta \in (0,1]$, independent and identically distributed (i.i.d.) samples $Z^{\beta M} = (Z_1, \ldots, Z_{\beta M}) \sim P_Z^{\beta M}$ drawn from the source domain P_Z, and $(1-\beta)M$ i.i.d. samples $Z^{(1-\beta)M} = (Z_{\beta M+1}, \ldots Z_M) \sim P_Z'^{(1-\beta)M}$ from the target domain P'_Z. The learner does not know the distributions P_Z and P'_Z. The learner uses the training data set Z^M to choose a model, or hypothesis, W from the model class $\mathcal{W}$ by using a *randomized* learning algorithm defined by a conditional distribution $P_{W|Z^M} \in \mathcal{P}(\mathcal{W})$ as $W \sim P_{W|Z^M}$. The conditional distribution $P_{W|Z^M}$ defines a stochastic mapping from the training data set Z^M to the model class $\mathcal{W}$.

The performance of a model parameter vector $w \in \mathcal{W}$ on a data sample $z \in \mathcal{Z}$ is measured by a loss function $l(w,z)$ where $l : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}_+$. The *generalization loss*, also known as population loss, for a model parameter vector $w \in \mathcal{W}$ is evaluated on the target domain, and is defined as

$$L_g(w) = \mathbb{E}_{P'_Z}[l(w,Z)], \tag{1}$$

where the average is taken over a test example Z drawn independently of Z^M from the target task data distribution P'_Z.

The generalization loss cannot be computed by the learner, given that the data distribution P'_Z is unknown. A typical solution is for the learner to minimize instead the *weighted average training loss* on the data set Z^M, which is defined as the empirical average $L_t(w|Z^M) =$

$$\frac{\gamma}{\beta M}\sum_{i=1}^{\beta M} l(w,Z_i) + \frac{1-\gamma}{(1-\beta)M}\sum_{i=\beta M+1}^{M} l(w,Z_i), \tag{2}$$

where $\gamma \in [0,1]$ is a hyperparameter [6], [1]. We call the algorithm that minimizes (2) as the empirical weighted risk minimization (EWRM) algorithm. In formulation, EWRM algorithm outputs

$$W^{\mathrm{EWRM}}(Z^M) = \arg\min_{w \in \mathcal{W}} L_t(w|Z^M) \tag{3}$$

for input training set Z^M.

The difference between generalization loss (1) and training loss (2), known as *transfer generalization gap*, is defined as

$$\Delta L(w|Z^M) = L_g(w) - L_t(w|Z^M), \tag{4}$$

and is a key metric that relates to the performance of the learner. As mentioned, this is because a small transfer generalization gap ensures that the training loss (2) is a reliable estimate of the generalization loss (1).

III. α-JS Divergence-Based Bounds on Average Transfer Generalization Gap

In this section, we obtain bounds on the average transfer generalization gap $\Delta L^{\mathrm{avg}} := \mathbb{E}_{P_{Z^M}P_{W|Z^M}}[\Delta L(W|Z^M)]$, where the training set distribution is given as $P_{Z^M} = P_Z^{\beta M} \times P_Z'^{(1-\beta)M}$. Towards this goal, we assume the following.

[1] The cumulant generating function (CGF) of a random variable $X \sim P_X$ is defined as $\log \mathbb{E}_{P_X}[\exp(\lambda(X - \mathbb{E}_{P_X}[X]))]$ for $\lambda \in \mathbb{R}$.

Assumption 3.1: The loss function $l(W,Z)$ is σ^2-sub-Gaussian[2] under $(W,Z) \sim P_W R_Z^\alpha$, where P_W is the marginal of the joint distribution $P_{W|Z^M} P_{Z^M}$ and

$$R_Z^\alpha(z) = \alpha P_Z(z) + (1-\alpha) P'_Z(z), \tag{5}$$

for some $\alpha \in [0,1]$, is a mixture of the source and target data distributions.

Note that if the loss function is bounded, i.e., $0 \leq a \leq l(\cdot,\cdot) \leq b < \infty$, Assumption 3.1 is satisfied with $\sigma^2 = (b-a)^2/4$ under any data distribution R_Z^α for $\alpha \in [0,1]$.

We now present our α-JS-divergence-based bounds on the average transfer generalization gap. The α-JS divergence between distributions P_Z and P'_Z is defined as [9]

$$D_{\mathrm{JS}}^\alpha(P'_Z||P_Z) = \frac{1}{2}(D_{\mathrm{KL}}(P'_Z||R_Z^\alpha) + D_{\mathrm{KL}}(P_Z||R_Z^\alpha)). \tag{6}$$

In particular, if $\alpha = 0.5$, $D_{\mathrm{JS}}^{0.5}(P'_Z||P_Z)$ yields the conventional JS divergence. Towards obtaining α-JS-divergence-based bounds, we decompose the transfer generalization gap (4) as

$$\begin{aligned} \mathit{\Delta} L(w|Z^M) = \gamma(L_g(w) - L_t(w|Z^{\beta M})) \\ + (1-\gamma)(L_g(w) - L_t(w|Z^{(1-\beta)M})), \end{aligned} \tag{7}$$

where $L_t(w|Z^{\beta M}) = \sum_{i=1}^{\beta M} l(w,Z_i)/(\beta M)$ is the training loss over the source-domain data and $L_t(w|Z^{(1-\beta)M}) = \sum_{i=\beta M+1}^{M} l(w,Z_i)/((1-\beta)M)$ is the training loss of the target-domain data. By separately bounding the average of the two differences in the above decomposition, we obtain the following bound.

Theorem 3.1: Under Assumption 3.1 and $\beta \in (0,1)$, the following upper bound on the average transfer generalization gap holds for any algorithm $P_{W|Z^M}$,

$$\begin{aligned} \mathit{\Delta} L^{\mathrm{avg}} \leq \frac{2\gamma}{\beta M} \sum_{i=1}^{\beta M} \sqrt{\sigma^2 \Big(2 D_{\mathrm{JS}}^\alpha(P'_Z||P_Z) + I(W;Z_i)\Big)} + \\ \frac{2(1-\gamma)}{(1-\beta)M} \sum_{i=\beta M+1}^{M} \sqrt{\sigma^2 \Big(2 D_{\mathrm{KL}}(P'_Z||R_Z^\alpha) + I(W;Z_i)\Big)}. \end{aligned} \tag{8}$$

Proof: See Appendix A. ■

The first term in (8) accounts for the contribution to the transfer generalization gap caused by the limited availability of the source-domain data. It comprises of (i) the sensitivity measure of the algorithm to the individual sample of the source-domain training set captured by the mutual information $I(W;Z_i)$; and (ii) the domain shift between source and target data distributions captured by the α-JS-divergence $D_{\mathrm{JS}}^\alpha(P'_Z||P_Z)$. The second term of (8) similarly accounts for the contribution of the limited data from the target-domain. It comprises of the mutual information $I(W;Z_i)$ which accounts for the sensitivity of the learning algorithm to individual sample of the target-domain training set; and of the KL divergence term $D_{\mathrm{KL}}(P'_Z||R_Z^\alpha)$, which quantify the distance between the target distribution P'_Z and the mixture distribution R_Z^α.

We note that the KL divergence term $D_{\mathrm{KL}}(P'_Z||R_Z^\alpha)$ arises here since the sub-Gaussianity of the loss function $l(W,Z)$ is assumed under $(W,Z) \sim P_W R_Z^\alpha$ (Assumption 3.1). We also note that, for $\alpha < 1$, we have $\mathrm{supp}(P'_Z) \subseteq \mathrm{supp}(R_Z^\alpha)$, and hence the KL divergence $D_{\mathrm{KL}}(P'_Z||R_Z^\alpha)$ is well-defined. Moreover, using mixture distribution R_Z^α with arbitrary mixing ratio α can yield tighter bounds than the one with $\alpha = 0.5$ which results in JS-divergence. For instance, for the extreme case when $\gamma = 0$, the bound in (8) is minimized by choosing $\alpha = \gamma = 0$.

Note that the bound in (8) does not account for the case $\beta = 0$, i.e., when only target-domain data set is available for training. In this case, the problem reduces to the conventional learning with $P_Z = P'_Z$. We now specialize the bound in (8) to the case when only data from source distribution is available for training, i.e., when $\beta = 1$.

Corollary 3.2: Under Assumption 3.1, the following bound holds when $\beta = 1$

$$\mathit{\Delta} L^{\mathrm{avg}} \leq \frac{1}{M} \sum_{i=1}^{M} 2 \sqrt{\sigma^2 \Big(2 D_{\mathrm{JS}}^\alpha(P'_Z||P_Z) + I(W;Z_i)\Big)}. \tag{9}$$

The bound in (8) can be proven to hold also under the following assumption, similar to the one considered in [7].

Assumption 3.2: The loss function $l(w,Z)$ is σ^2-sub-Gaussian under $Z \sim R_Z^\alpha$ for all $w \in \mathcal{W}$.

To see this, one can follow the steps in the derivation of the exponential inequalities in Lemma A.1 of Appendix A, starting from the additional step of averaging both sides of the inequality $\mathbb{E}_{R_Z^\alpha}[\exp(\lambda(l(w,Z) - \mathbb{E}_{R_Z^\alpha}[l(w,Z)]) - \lambda^2\sigma^2/2)] \leq 1$ over $W \sim P_W$. As discussed in [8], in general, Assumption 3.1 does not imply this assumption, and vice versa. However, both assumptions hold when $l(\cdot,\cdot)$ is bounded.

We finally note that the α-JS-divergence-based bounds on average transfer generalization gap can be similarly obtained (proof omitted) for a general class of loss functions with bounded CGF as stated in the following corollary.

Corollary 3.3: Assume that the CGF of the loss function $l(W,Z) \sim P_W R_Z^\alpha$ is upper bounded by a function $\mathit{\Psi}(\lambda)$ for $\lambda \in [b_-, b_+]$. Then, for $\beta \in (0,1)$, the following bound holds

$$\begin{aligned} \mathit{\Delta} L^{\mathrm{avg}} \leq \frac{\gamma}{\beta M} \sum_{i=1}^{\beta M} \inf_{\lambda_1 \in (0,b)} \bigg(\widehat{\mathit{\Psi}}(\lambda_1) + 2 D_{\mathrm{JS}}^\alpha(P'_Z||P_Z) \\ + I(W;Z_i)\bigg) + \frac{1-\gamma}{(1-\beta)M} \sum_{i=\beta M+1}^{M} \inf_{\lambda_2 \in (0,b)} \bigg(\widehat{\mathit{\Psi}}(\lambda_2) \\ + 2 D_{\mathrm{KL}}(P'_Z||R_Z^\alpha) + I(W;Z_i)\bigg), \end{aligned} \tag{10}$$

where $\widehat{\mathit{\Psi}}(\lambda) = \mathit{\Psi}(\lambda) + \mathit{\Psi}(-\lambda)$ and $b = \min\{b_+, -b_-\}$.

For sub-Gaussian loss function $l(W,Z)$, the inequality (8) then follows from (10) by setting $\mathit{\Psi}(\lambda) = \mathit{\Psi}(-\lambda) = \lambda^2\sigma^2/2$ and $b_+ = b_- = \infty$. Another class of loss functions that satisfy assumption in Corollary 3.3 is the sub-gamma loss $l(W,Z)$

[2] A random variable $X \sim P_X$ is said to be σ^2-sub-Gaussian if its CGF is upper bounded by $\lambda^2\sigma^2/2$ for all $\lambda \in \mathbb{R}$.

with variance parameter σ and scale parameter c, whose CGF is upper bounded by $\Psi(\lambda) = \lambda^2\sigma^2/2(1 - c|\lambda|)$ for $|\lambda| < 1/c$.

A. Bound on Average Transfer Excess Risk for EWRM

In this section, we obtain an upper bound on the average transfer excess risk of EWRM. Let

$$w^* = \arg\min_{w\in\mathcal{W}} L_g(w) \tag{11}$$

be the optimizing model parameter of the transfer generalization loss $L_g(w)$. Then, the average transfer excess risk for the EWRM algorithm is defined as

$$\Delta L_g^* = \mathbb{E}_{P_{Z^M}}[L_g(W^{\mathrm{EWRM}})] - L_g(w^*), \tag{12}$$

where we have used W^{EWRM} to denote $W^{\mathrm{EWRM}}(Z^M)$ for notational convenience.

To obtain an upper bound on the average excess risk ΔL_g^*, we use the decomposition

$$\Delta L_g^* = \underbrace{\mathbb{E}_{P_{Z^M}}[L_g(W^{\mathrm{EWRM}})] - \mathbb{E}_{P_{Z^M}}[L_t(W^{\mathrm{EWRM}}|Z^M)]}_{A}$$
$$+ \underbrace{\mathbb{E}_{P_{Z^M}}[L_t(W^{\mathrm{EWRM}}|Z^M)] - L_g(w^*)}_{B}. \tag{13}$$

Term A in (13) corresponds to the average transfer generalization gap for the EWRM, and hence it can be upper bounded using (8). Using the definition (3) of EWRM, term B can be upper bounded as

$$\begin{aligned} B &\leq \mathbb{E}_{P_{Z^M}}[L_t(w^*|Z^M)] - L_g(w^*) \\ &= \gamma\Big[\mathbb{E}_{P_Z}[l(w^*, Z)] - \mathbb{E}_{P'_Z}[l(w^*, Z)]\Big], \end{aligned} \tag{14}$$

where the last equality follows from (2) and using the identity $\mathbb{E}_{P_{Z^{(1-\beta)M}}}[L_t(w^*|Z^{(1-\beta)M})] = L_g(w^*)$. Denoting the upper bound on term A which follows from (8) as $\mathrm{UB}(W^{\mathrm{EWRM}}) =$

$$\frac{2\gamma\sigma}{\beta M}\sum_{i=1}^{\beta M}\sqrt{2D^{\alpha}_{\mathrm{JS}}(P'_Z||P_Z) + I(W^{\mathrm{EWRM}}; Z_i)} +$$
$$\frac{2(1-\gamma)\sigma}{(1-\beta)M}\sum_{i=\beta M+1}^{M}\sqrt{2D_{\mathrm{KL}}(P'_Z||R^{\alpha}_Z) + I(W^{\mathrm{EWRM}}; Z_i)},$$

and combining this with an upper bound on term B yields the following bound on the average transfer excess risk for EWRM.

Theorem 3.4: Under Assumption 3.2, the following bound holds for $\beta \in (0, 1]$

$$\Delta L_g^* \leq \mathrm{UB}(W^{\mathrm{EWRM}}) + \gamma d_{\mathcal{W}}(P_Z, P'_Z), \quad \text{where} \tag{15}$$

$$d_{\mathcal{W}}(P_Z, P'_Z) = \sup_{w\in\mathcal{W}} |\mathbb{E}_{P_Z}[l(w, Z)] - \mathbb{E}_{P'_Z}[l(w, Z)]|. \tag{16}$$

In addition to the bound $\mathrm{UB}(W^{\mathrm{EWRM}})$ on the generalization gap, the excess-risk bound in (15) includes the term $d_{\mathcal{W}}(P_Z, P'_Z)$, which is the integral probability metric used in [5] to account for the effect of domain shift on the generalization loss. Since this term is challenging to evaluate, in the following, we show that an easier-to-estimate upper bound can be obtained by using the α-JS divergence.

Theorem 3.5: Under the same setting as in Theorem 3.4, we have the following bound

$$\Delta L_g^* \leq \mathrm{UB}(W^{\mathrm{EWRM}}) + 2\gamma\sqrt{2\sigma^2 D^{\alpha}_{\mathrm{JS}}(P'_Z||P_Z)}. \tag{17}$$

Proof: See Appendix C. ■

IV. Example

In this section, we consider the problem of estimating the mean of a discrete random variable Z taking values in set $\mathcal{Z} = \{0, 1, 2\}$. The source domain is defined by data distributed as $Z \sim P_Z$, with $P_Z(0) = p_s$ and $P_Z(1) = 1 - p_s$, and the target-domain data is distributed as $Z \sim P'_Z$, with $P'_Z(1) = p_t$ and $P'_Z(2) = 1 - p_t$. The transfer learner infers an estimate $w \in \mathcal{W}$ of the mean of the random variable Z. The loss function $l(w, z) = (w - z)^2$ measures the quadratic error between the estimate w and a test input z. For a training data set Z^M, the EWRM transfer learner in (3) outputs the estimate

$$W^{\mathrm{EWRM}} = \frac{\gamma}{\beta M}\sum_{i=1}^{\beta M} Z_i + \frac{(1-\gamma)}{(1-\beta)M}\sum_{i=\beta M+1}^{M} Z_i. \tag{18}$$

The average transfer generalization gap evaluates to

$$\begin{aligned} &\mathbb{E}_{P_{Z^M}}[\Delta L(W^{\mathrm{EWRM}}|Z^M)] = 2\mathbb{E}_{P_{Z^M}}[(W^{\mathrm{EWRM}})^2] \\ &\quad - 2\mu_t\mathbb{E}_{P_{Z^M}}[W^{\mathrm{EWRM}}] + \gamma(\nu_t + \mu_t^2 - \nu_s - \mu_s^2), \end{aligned} \tag{19}$$

where μ_t and ν_t are the mean and variance respectively of the random variable $Z \sim P'_Z$; while μ_s, and ν_s are the mean and variance respectively of the random variable $Z \sim P_Z$. The averages in (19) can be computed explicitly as $\mathbb{E}_{P_{Z^M}}[W^{\mathrm{EWRM}}] = \gamma\mu_s + (1 - \gamma)\mu_t$, and $\mathbb{E}_{P_{Z^M}}[(W^{\mathrm{EWRM}})^2] = \gamma^2\nu_s/(\beta M) + (1-\gamma)^2\nu_t/((1-\beta)M) + (\mathbb{E}_{P_{Z^M}}[W^{\mathrm{EWRM}}])^2$.

Since the support of the target-domain data distribution does not include the support of the source-domain data distribution, the KL divergence evaluates to $D(P_Z||P'_Z) = \infty$. In contrast, the α-JS divergence can be evaluated as $D^{\alpha}_{\mathrm{JS}}(P'_Z||P_Z) =$

$$0.5\bigg(-p_s\log(\alpha) + (1 - p_s)\log\frac{1 - p_s}{\alpha(1 - p_s) + (1 - \alpha)p_t}$$
$$+ p_t\log\frac{p_t}{\alpha(1 - p_s) + (1 - \alpha)p_t} - (1 - p_t)\log(1 - \alpha)\bigg). \tag{20}$$

Furthermore, using (18) and the alphabet $\mathcal{Z} \in \{0, 1, 2\}$, we can, without loss of generality, consider the model parameter space $\mathcal{W}$ limited to the interval $[0, 2]$. Therefore, the loss function $l(w, z)$ is bounded in the interval $[0, 4]$, and hence it is 4-sub-Gaussian.

In Figure 1, we compare the the average transfer generalization gap (19) with the α-JS-divergence bound of (9) with $\alpha = 0.5$ and the ϕ-divergence based bound in [1, Cor. 3] with $\phi(x) = |x - 1|$ for the case when $\beta = 1$ (i.e., only source-domain data set available for training) as a function of increasing values of M. For fixed P_Z with $p_s = 0.48$, we

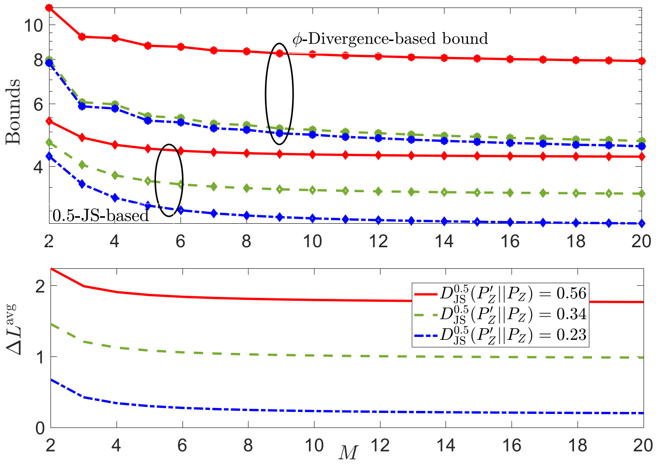

Fig. 1: Average transfer generalization gap (19) (bottom) and the 0.5-JS-based bound in (9) and ϕ-divergence based bound in [1, Cor. 3] (top) as a function of M (when $\beta = 1$) for varying JS divergence between P_Z' and a fixed P_Z with $p_s = 0.48$.

vary the 0.5-JS-divergence by varying p_t. As predicted by our bound, the transfer generalization gap decreases with increase in the number of source-data samples M available for training. However, there exists a non-vanishing generalization gap even at high M, which is a direct consequence of the domain shift. Moreover, a larger 0.5-JS-divergence between P_Z and P_Z' is predictive of a larger average transfer generalization gap. Finally, we show that 0.5-JS-divergence based bounds outperform the ϕ-divergence based bound in [1, Cor. 3] when $\beta = 1$ at varying JS distances. However, we note that when $\beta < 1$, the α-divergence based bounds need not necessarily outperform the ϕ-divergence based bounds.

REFERENCES

[1] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-Theoretic Analysis for Transfer Learning," *arXiv preprint arXiv:2005.08697*, 2020.

[2] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[3] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of Representations for Domain Adaptation," in *Advances in Neural Information Processing Systems*, 2007, pp. 137–144.

[4] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain Adaptation: Learning Bounds and Algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[5] C. Zhang, L. Zhang, and J. Ye, "Generalization Bounds for Domain Adaptation," in *Advances in Neural Information Processing Systems*, 2012, pp. 3320–3328.

[6] 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. 1-2, pp. 151–175, 2010.

[7] A. Xu and M. Raginsky, "Information-Theoretic Analysis of Generalization Capability of Learning Algorithms," in *Proc. of Adv. in Neural Inf. Processing Sys. (NIPS)*, Dec. 2017, pp. 2524–2533.

[8] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening Mutual Information Based Bounds on Generalization Error," in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, July 2019, pp. 587–591.

[9] F. Nielsen, "A family of statistical symmetric divergences based on jensen's inequality," *arXiv preprint arXiv:1009.4004*, 2010.

[10] C. Shui, Q. Chen, J. Wen, F. Zhou, C. Gagné, and B. Wang, "Beyond $\mathcal{H}$-divergence: Domain adaptation theory with jensen-shannon divergence," *arXiv preprint arXiv:2007.15567*, 2020.

[11] F. Hellström and G. Durisi, "Generalization Bounds via Information Density and Conditional Information Density," *arXiv preprint arXiv:2005.08044*, 2020.

[12] Y. Polyanskiy and Y. Wu, "Lecture Notes on Information Theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, no. 2012-2016, p. 7, 2014.

APPENDIX A: PROOF OF THEOREM 3.1

To obtain an upper bound on $\mathit{\Delta} L^{\mathrm{avg}}$, we use the decomposition (7) and separately bound the two differences. The average of the first difference in (7) can be equivalently written as $\mathbb{E}_{P_{Z^M,W}}[L_g(W) - L_t(W|Z^{\beta M})] =$

$$\frac{1}{\beta M}\sum_{i=1}^{\beta M}\Big[\mathbb{E}_{P_W P_Z'}[l(W,Z)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)]\Big] \tag{21}$$

and similarly $\mathbb{E}_{P_{Z^M,W}}[L_g(W) - L_t(W|Z^{(1-\beta)M})] =$

$$\frac{\sum_{i=\beta M+1}^{M}\Big[\mathbb{E}_{P_W P_Z'}[l(W,Z)] - \mathbb{E}_{P_{Z_i}' P_{W|Z_i}}[l(W,Z_i)]\Big]}{(1-\beta)M}. \tag{22}$$

We first bound the difference $\mathbb{E}_{P_W P_Z'}[l(W,Z)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)]$ in (21). Towards this, we use the the exponential inequalities in Lemma A.1 obtained based on the change of measure approach adopted in [11]. Fix $\lambda = \lambda_1 > 0$ in (25) and $\lambda = -\lambda_1$ in (26), and applying Jensen's inequality yield the following inequalities

$$\mathbb{E}_{P_W P_{Z_i}'}[l(W,Z_i)] - \mathbb{E}_{P_W R_Z^\alpha}[l(W,Z)] \leq \frac{\lambda_1\sigma^2}{2} + \frac{D_{\mathrm{KL}}(P_Z'||R_Z^\alpha)}{\lambda_1} \tag{23}$$

$$\mathbb{E}_{P_W R_Z^\alpha}[l(W,Z)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)] \leq \frac{\lambda_1\sigma^2}{2} + \frac{1}{\lambda_1}\Big(D_{\mathrm{KL}}(P_Z||R_Z^\alpha) + I(W;Z_i)\Big). \tag{24}$$

Adding (23) and (24) and choosing $\lambda_1 = \sqrt{2D_{\mathrm{JS}}^\alpha(P_Z'||P_Z) + I(W;Z_i)}/\sigma$ gives that $\mathbb{E}_{P_W P_{Z_i}'}[l(W,Z_i)] -$ $\mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)] \leq 2\sqrt{\sigma^2\Big(2D_{\mathrm{JS}}^\alpha(P_Z'||P_Z) + I(W;Z_i)\Big)}$. Similarly, we can bound the difference $\mathbb{E}_{P_W P_Z'}[l(W,Z)] - \mathbb{E}_{P_{Z_i}' P_{W|Z_i}}[l(W,Z_i)]$ in (22) by fixing $\lambda = -\lambda_1$ in (27) and applying Jensen's inequality. Adding the resultant inequality and (23), and optimizing over $\lambda_1 > 0$ gives the corresponding bound.

Lemma A.1: Under Assumption 3.1, the following inequalities hold for all $\lambda \in \mathbb{R}$ when $i = 1, \ldots, \beta M$,

$$\mathbb{E}_{P_W P_{Z_i}'}\Bigg[\exp\Bigg(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R_Z^\alpha}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{dP_{Z_i}'(Z_i)}{dR_{Z_i}^\alpha(Z_i)}\Bigg)\Bigg] \leq 1, \tag{25}$$

$$\mathbb{E}_{P_{Z_i}P_{W|Z_i}}\Bigg[\exp\Bigg(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R_Z^\alpha}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{dP_{Z_i}(Z_i)}{dR_{Z_i}^\alpha(Z_i)} - i(W;Z_i)\Bigg)\Bigg] \leq 1, \tag{26}$$

where $i(W;Z_i) = \log(dP_{W,Z_i}(W,Z_i)/(dP_W P_{Z_i}(W,Z_i))$ is the information density between random variables W and Z_i.

For $i = \beta M + 1, \ldots, M$, the inequality (25) holds along with the following inequality

$$\mathbb{E}_{P'_{Z_i} P_{W|Z_i}}\bigg[\exp\bigg(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R^\alpha_Z}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{dP'_{Z_i}(Z_i)}{dR^\alpha_{Z_i}(Z_i)} - i(W;Z_i)\bigg)\bigg] \leq 1. \tag{27}$$

APPENDIX B
PROOF OF LEMMA A.1

The derivation of the exponential inequalities leverage the change of measure approach adopted in [11]. For $i = 1, \ldots, M$, Assumption 3.1 gives that

$$\mathbb{E}_{P_W R^\alpha_{Z_i}}\bigg[\exp\bigg(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R^\alpha_Z}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2}\bigg)\bigg] \leq 1. \tag{28}$$

For $i = 1, \ldots, \beta M$, the inequality (28) implies the following inequality,

$$\mathbb{E}_{P_W R^\alpha_{Z_i}}\bigg[\mathbb{I}_{\mathcal{E}_1}\exp\bigg(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R^\alpha_Z}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2}\bigg)\bigg] \leq 1, \tag{29}$$

where $\mathcal{E}_1 = \mathrm{supp}(P'_{Z_i})$, and $\mathbb{I}_{\mathcal{E}}$ denotes the indicator function which takes value 1 when the event $\mathcal{E}$ is true, and is zero otherwise. Now, performing a change of measure from $R^\alpha_{Z_i}$ to P'_{Z_i} as in [12, Prop. 17] yields (25). To get to inequality (26), we similarly first perform change of measure from $R^\alpha_{Z_i}$ to P_{Z_i} on (28), and then perform another change of measure from $P_{Z_i}P_W$ to $P_{Z_i}P_{W|Z_i}$.

For $i = \beta M + 1, \ldots, M$, (25) can be verified to hold as before. To get to inequality (27), we perform a change of measure on (25) from $P'_{Z_i}P_W$ to $P'_{Z_i}P_{W|Z_i}$.

APPENDIX C
PROOF OF THEOREM 3.5

To obtain the required bound in (17), we show that the term $\mathbb{E}_{P_Z}[l(w^*,Z)] - \mathbb{E}_{P'_Z}[l(w^*,Z)]$ can be bounded by $\sqrt{8\sigma^2 D^\alpha_{\mathrm{JS}}(P'_Z||P_Z)}$. Towards this, note that the following exponential inequality holds under Assumption 3.2 with $w = w^*$

$$\mathbb{E}_{R^\alpha_Z}\bigg[\exp\bigg(\lambda(l(w^*,Z) - \mathbb{E}_{R^\alpha_Z}[l(w^*,Z)]) - \frac{\lambda^2\sigma^2}{2}\bigg)\bigg] \leq 1. \tag{30}$$

Denote $\delta l(w^*) := l(w^*,Z) - \mathbb{E}_{R^\alpha_Z}[l(w^*,Z]$. Now, performing change of measure from R^α_Z to P'_Z, and from R^α_Z to P_Z on (30) respectively as in Lemma A.1, yields the following inequalities

$$\mathbb{E}_{P'_Z}\bigg[\exp\bigg(\lambda\delta l(w^*) - \log\frac{dP'_Z(Z)}{dR^\alpha_Z(Z)} - \frac{\lambda^2\sigma^2}{2}\bigg)\bigg] \leq 1, \tag{31}$$

$$\mathbb{E}_{P_Z}\bigg[\exp\bigg(\lambda\delta l(w^*) - \log\frac{dP_Z(Z)}{dR^\alpha_Z(Z)} - \frac{\lambda^2\sigma^2}{2}\bigg)\bigg] \leq 1. \tag{32}$$

Take $\lambda = \lambda_1 > 0$ in (32) and $\lambda = -\lambda_1$ in (31). Now apply Jensen's inequality on (31) and (32), and add the resulting inequalities to get the following inequality

$$\mathbb{E}_{P_Z}[l(w^*,Z)] - L_g(w^*) \leq \lambda_1\sigma^2 + \frac{1}{\lambda_1}2D^\alpha_{\mathrm{JS}}(P'_Z||P_Z). \tag{33}$$

Now, letting $\lambda_1 = \sqrt{2D^\alpha_{\mathrm{JS}}(P'_Z||P_Z)}/\sigma$ yields the required bound.