# 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 a two-parameter family of generalized $(\alpha_1, \alpha_2)$-Jensen-Shannon (JS) divergences; and $(ii)$ the sensitivity of the transfer learner output $W$ to each individual sample of the data set $Z_i$ via the mutual information $I(W; Z_i)$. For $\alpha_1 \in (0, 1)$, the $(\alpha_1, \alpha_2)$-JS divergence can be bounded even when the support of $P_Z$ is not included in that of $P'_Z$. This contrasts the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P_Z||P'_Z)$-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 $\phi$-divergence based bound of Wu *et al.* [1]. We also obtain new upper bounds on the average transfer excess risk in terms of the $(\alpha_1, \alpha_2)$-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* 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*. 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. We define the transfer learner as a stochastic mapping $P_{W|Z^M}$ from the input training set to the output space of model parameters. The difference between the generalization loss and the training loss, $\Delta L(w|Z^M) = L_g(w) - L_t(w|Z^M)$, 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.

Existing works on transfer learning [3]–[6] have largely focused on obtaining *high-probability*, probably approximately correct (PAC), bounds on the transfer generalization gap. These bounds have the general form: With probability at least $1 - \delta$, with $\delta \in (0, 1)$, over the training set $Z^M$, the bound $|\Delta L(w|Z^M)| \leq \epsilon$ holds uniformly for all $w \in \mathcal{W}$. The upper bound $\epsilon$ has been expressed as a function of a distance measure $d(\mathcal{S}, \mathcal{T})$ that quantifies the distributional shift between source ($\mathcal{S}$) and target ($\mathcal{T}$) domains. Specifically, the main goal of these studies has been to define appropriate distance measures $d(\mathcal{S}, \mathcal{T})$ that can be estimated from finite data with reasonable accuracy. For example, Ben et al. in [3] and [4] introduce the $d_{\mathcal{A}}$ distance and $\mathscr{H}\Delta\mathscr{H}$-divergence respectively for the 0-1 loss, while Mansour et al. [5] proposed a *discrepancy distance* that holds for any loss functions. These measures depend on the structural properties of the model class $\mathcal{W}$ through the model complexity measures such as Vapnik-Chervonenkis (VC) dimension and Rademacher complexity. Similar high probability bounds have also been studied for the optimality gap, i.e., $\mathbb{E}_{P_{W|Z^m}}[L_g(w)] - \min_{w \in \mathcal{W}} L_g(w)$.

In contrast to these prior works, this paper focuses on obtaining information-theoretic bounds on the average transfer generalization gap, $\mathbb{E}_{P_{Z^M} P_{W|Z^M}}[\Delta L(W|Z^M)]$, where the average is with respect to the training data and the transfer learner. These bounds are fundamentally different from the existing high-probability bounds, and thus they are not directly comparable. Unlike the high-probability bounds which ignore the properties of the training algorithm, the information-theoretic bounds describe the generalization capability of arbitrary transfer learners via their sensitivity to the input training set.

Our work is related to the recent study in [1] on information-theoretic bounds for transfer learning. The resulting bound

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).The authors thank Prof. Tan (NUS) for useful discussions.

captures the impact of the domain shift via the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P_Z||P'_Z)$ 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 $\infty$ 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.

### A. Contributions

In this work, we mitigate the above drawback of KL divergence based bounds on average transfer generalization gap, by using a two-parameter $(\alpha_1, \alpha_2)$-family of Jensen-Shannon (JS) divergences with $\alpha_1, \alpha_2 \in [0, 1]$ to capture the domain shift. This family includes as special cases the conventional JS divergence with $\alpha_1 = \alpha_2 = 0.5$, as well as Nielsen's symmetric $\alpha$-skew and asymmetric $\alpha$-skew JS divergences [7], which corresponds to the choices $\alpha_2 = 0.5$ and $\alpha_1 = \alpha_2$ respectively. 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 $(\alpha_1, \alpha_2)$-JS divergence 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 $(\alpha_1, \alpha_2)$-JS divergence is bounded for $\alpha_1 \in (0, 1)$ [8, Thm. 1], and gives non-vacuous bounds even when $P_Z \not\ll P'_Z$. Moreover, the obtained bound holds for unbounded loss functions with bounded cumulant generating function (CGF). In contrast, the $\phi$-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 $L_\infty$-norm.

Our work is motivated by the recent study [9] that employs the conventional JS divergence, 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 $w$. Moving beyond [9], 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 $(\alpha_1, \alpha_2)$-JS divergence-based bounds on average transfer generalization gap to EWRM, and obtain new upper bounds on the *average optimality gap* for EWRM. This is unlike prior work [1], [4], which obtain high probability bounds on the optimality gap. We show via an example that by choosing the parameters $\alpha_1, \alpha_2$, the $(\alpha_1, \alpha_2)$-JS divergence can better capture the relative impact of source and target data sets on the performance of EWRM, yielding tighter bounds than with the conventional JS divergence.

## 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 $\beta M$, for some fixed $\beta \in (0, 1]$, independent and identically distributed (i.i.d.) samples $Z^{\beta M} = (Z_1, \ldots, Z_{\beta M}) \sim P_Z^{\beta M}$[1] 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 $l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$. 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 evaluate 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 [4], [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).

[1]We use $P_X^N$ to denote the $N$-fold product distribution induced by $P_X$.

## III. $\alpha$-JS Divergence-Based Bounds on Average Transfer Generalization Gap

In this section, we obtain bounds on the average transfer generalization gap $\Delta L^{\text{avg}} := \mathbb{E}_{P_{Z^M}P_{W|Z^M}}[\Delta L(W|Z^M)]$, 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.

*Assumption 3.1:* The loss function $l(W,Z)$ is $\sigma^2$-sub-Gaussian[2] under $(W,Z) \sim P_W R_Z^{\alpha_1}$, where $P_W$ is the marginal of the joint distribution $P_{W|Z^M}P_{Z^M}$ and

$$R_Z^{\alpha_1}(z) = \alpha_1 P_Z(z) + (1-\alpha_1)P_Z'(z), \tag{5}$$

for some $\alpha_1 \in [0,1]$, 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 $R_Z^{\alpha_1}$ for $\alpha_1 \in [0,1]$.

To derive bounds on the average transfer generalization gap, we consider the following family of $(\alpha_1,\alpha_2)$-JS divergences,

$$D_{\text{JS}}^{\alpha_1,\alpha_2}(P_Z'||P_Z) = \alpha_2 D_{\text{KL}}(P_Z'||R_Z^{\alpha_1}) + (1-\alpha_2)D_{\text{KL}}(P_Z||R_Z^{\alpha_1})), \tag{6}$$

where $\alpha_1,\alpha_2 \in [0,1]$. We refer to Section I for connections with existing JS divergences. Towards obtaining $(\alpha_1,\alpha_2)$-JS-divergence-based bounds, we decompose the transfer generalization gap (4) as

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

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 for $(\beta,\alpha_2) \in (0,1)$, the following upper bound on the average transfer generalization gap holds for any algorithm $P_{W|Z^M}$,

$$\Delta L^{\text{avg}} \leq \frac{\gamma\sigma\sqrt{2\hat{\alpha}_2}}{\beta M}\sum_{i=1}^{\beta M}\sqrt{D_{\text{JS}}^{\alpha_1,\alpha_2}(P_Z'||P_Z) + (1-\alpha_2)I(W;Z_i)} + \frac{2(1-\gamma)\sigma}{(1-\beta)M}\sum_{i=\beta M+1}^{M}\sqrt{2D_{\text{KL}}(P_Z'||R_Z^{\alpha_1}) + I(W;Z_i)}, \tag{8}$$

where $\hat{\alpha}_2 = 1/\alpha_2 + 1/(1-\alpha_2)$.

*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 samples 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 $(\alpha_1,\alpha_2)$-JS-divergence $D_{\text{JS}}^{\alpha_1,\alpha_2}(P_Z'||P_Z)$. 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 $D_{\text{KL}}(P_Z'||R_Z^{\alpha_1})$, which quantify the distance between the target distribution $P_Z'$ and the mixture distribution $R_Z^{\alpha_1}$.

We note that the KL divergence term $D_{\text{KL}}(P_Z'||R_Z^{\alpha_1})$ arises here since the sub-Gaussianity of the loss function $l(W,Z)$ is assumed under $(W,Z) \sim P_W R_Z^{\alpha_1}$ (Assum. 3.1). We also note that, for $\alpha_1 < 1$, we have $\text{supp}(P_Z') \subseteq \text{supp}(R_Z^{\alpha_1})$ with $\text{supp}(\cdot)$ denoting the support of '$\cdot$', and hence the KL divergence $D_{\text{KL}}(P_Z'||R_Z^{\alpha_1})$ is well-defined. Moreover, for fixed $\gamma$, $\beta$ and $M$, the bound in (8) can be tightened by optimizing over the choice of $\alpha_1$ and $\alpha_2$. For instance, for the extreme case when $\gamma = 0$, the bound in (8) is minimized by choosing $\alpha_1 = \gamma = 0$.

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 $\beta = 1$, $\Delta L^{\text{avg}} \leq$

$$\frac{\sigma\sqrt{2\hat{\alpha}_2}}{M}\sum_{i=1}^{M}\sqrt{D_{\text{JS}}^{\alpha_1,\alpha_2}(P_Z'||P_Z) + (1-\alpha_2)I(W;Z_i)}. \tag{9}$$

The bound in (8) can be proven to hold also under the following assumption, similar to the one considered in [10].

*Assumption 3.2:* The loss function $l(w,Z)$ is $\sigma^2-$sub-Gaussian under $Z \sim R_Z^{\alpha_1}$ 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 $\mathbb{E}_{R_Z^{\alpha_1}}[\exp(\lambda(l(w,Z)-\mathbb{E}_{R_Z^{\alpha_1}}[l(w,Z)])-\lambda^2\sigma^2/2)] \leq 1$ over $W \sim P_W$. As discussed in [11], 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 $(\alpha_1,\alpha_2)$-JS-divergence-based bounds on average transfer generalization gap can be generalized to loss functions $l(W,Z)$ whose CGF is upper bounded by a function $\Psi(\lambda)$ for $\lambda \in [b_-,b_+]$ under $(W,Z) \sim P_W R_Z^{\alpha_1}$. We refer to [12] for details. This class of functions include the sub-Gaussian loss $l(W,Z)$ with $\Psi(\lambda) = \Psi(-\lambda) = \lambda^2\sigma^2/2$ and $b_+ = b_- = \infty$, and the sub-gamma loss $l(W,Z)$ with variance parameter $\sigma$ 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

$$w^* = \arg\min_{w\in\mathcal{W}} L_g(w) \tag{10}$$

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

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

$$\Delta L_g^* = \mathbb{E}_{P_{Z^M}}[L_g(W^{\text{EWRM}})] - L_g(w^*), \tag{11}$$

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

To obtain an upper bound on the average excess risk $\Delta L_g^*$, we use the decomposition

$$\Delta L_g^* = \underbrace{\mathbb{E}_{P_{Z^M}}[L_g(W^{\text{EWRM}})] - \mathbb{E}_{P_{Z^M}}[L_t(W^{\text{EWRM}}|Z^M)]}_{A} + \underbrace{\mathbb{E}_{P_{Z^M}}[L_t(W^{\text{EWRM}}|Z^M)] - L_g(w^*)}_{B}. \tag{12}$$

Term A in (12) 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

$$\begin{aligned} B &\leq \mathbb{E}_{P_{Z^M}}[L_t(w^*|Z^M)] - L_g(w^*) \\ &= \gamma\Big[\mathbb{E}_{P_Z}[l(w^*, Z)] - \mathbb{E}_{P'_Z}[l(w^*, Z)]\Big], \end{aligned} \tag{13}$$

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 $\text{UB}(W^{\text{EWRM}}) =$

$$\frac{\gamma\sigma\sqrt{2\hat{\alpha}_2}}{\beta M}\sum_{i=1}^{\beta M}\sqrt{D_{\text{JS}}^{\alpha_1,\alpha_2}(P'_Z||P_Z) + (1-\alpha_2)I(W^{\text{EWRM}};Z_i)} + \frac{2(1-\gamma)\sigma}{(1-\beta)M}\sum_{i=\beta M+1}^{M}\sqrt{2D_{\text{KL}}(P'_Z||R_Z^{\alpha_1}) + I(W^{\text{EWRM}};Z_i)},$$

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

*Theorem 3.3:* Under Assumption 3.2, the following bound holds for $\beta \in (0, 1]$

$$\Delta L_g^* \leq \text{UB}(W^{\text{EWRM}}) + \gamma\sqrt{2\sigma^2\hat{\alpha}_2 D_{\text{JS}}^{\alpha_1,\alpha_2}(P'_Z||P_Z)}, \tag{14}$$

where $\hat{\alpha}_2 = 1/\alpha_2 + 1/(1-\alpha_2)$.
We refer to [12] for proof.

## 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

$$W^{\text{EWRM}} = \frac{\gamma}{\beta M}\sum_{i=1}^{\beta M} Z_i + \frac{(1-\gamma)}{(1-\beta)M}\sum_{i=\beta M+1}^{M} Z_i. \tag{15}$$

The average transfer generalization gap evaluates to

$$\begin{aligned} &\mathbb{E}_{P_{Z^M}}[\Delta L(W^{\text{EWRM}}|Z^M)] \\ &= 2\bar{\nu} - 2\mu_t\bar{\mu} + \gamma(\nu_t + \mu_t^2 - \nu_s - \mu_s^2), \end{aligned} \tag{16}$$

where $\mu_t$ and $\nu_t$ are the mean and variance respectively of the random variable $Z \sim P'_Z$; while $\mu_s$, and $\nu_s$ are the mean and variance respectively of the random variable $Z \sim P_Z$. The averages in (16) can be computed explicitly as $\bar{\mu} = \gamma\mu_s + (1-\gamma)\mu_t$ and $\bar{\nu} = \gamma^2\nu_s/(\beta M) + (1-\gamma)^2\nu_t/((1-\beta)M) + (\bar{\mu})^2$.

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 $(\alpha_1, \alpha_2)$-JS divergence can be evaluated in closed form. Furthermore, using (15) 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.

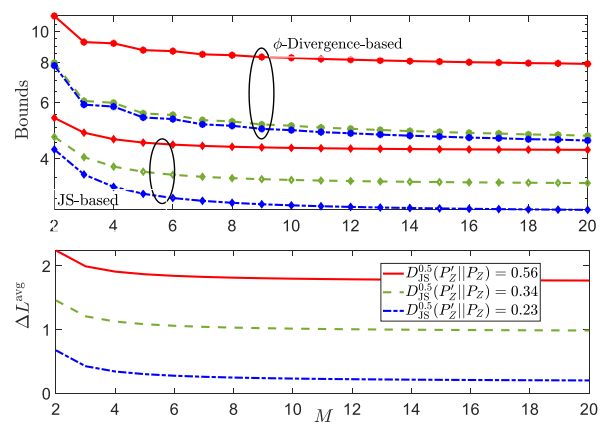


Fig. 1: Average transfer generalization gap (16) (bottom) and the $(0.5, 0.5)$-JS-based bound in (9) and $\phi$-divergence based bound in [1, Cor. 3] (top) as a function of $M$ (when $\beta = 1$) for varying JS divergence between $P'_Z$ and a fixed $P_Z$ with $p_s = 0.48$.

In Figure 1, we compare the the average transfer generalization gap (16) with the conventional JS-divergence bound of (9) for $\alpha_1 = \alpha_2 = 0.5$ and the $\phi$-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 vary the 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 JS-divergence between $P_Z$ and $P'_Z$ is predictive of a larger average transfer generalization gap. Finally, we show that JS-divergence based bounds outperform the $\phi$-divergence based bound in [1, Cor. 3] when $\beta = 1$ at varying JS distances.

We now study the advantage of considering the general family of $(\alpha_1, \alpha_2)$-JS divergence over the JS divergence.

Since the loss function is bounded, Assumption 3.1 holds for mixture distribution $R_Z^{\alpha_1}$ for any $\alpha_1 \in [0,1]$. Consequently, the bound in (8) can be tightened by optimizing over $\alpha_1$ and $\alpha_2$. In Figure 2, we evaluate the tightness of the bound in

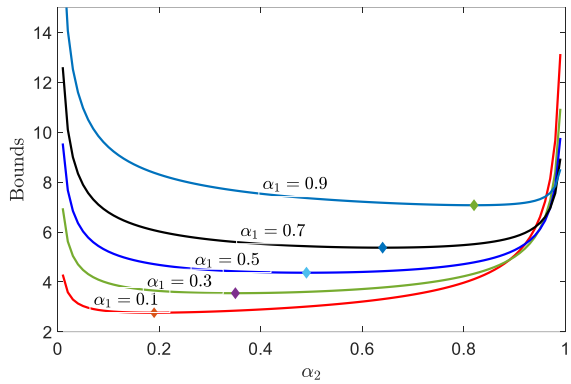


Fig. 2: The $(\alpha_1, \alpha_2)$-JS divergence based bound in (8) as a function of $\alpha_2$ for varying $\alpha_1$ ($M = 30$, $\beta = 2/3$, $\gamma = 0.3$) . (8) as a function of $\alpha_2$ for varying values of $\alpha_1$. As can be seen, the choice ($\alpha_1 = 0.1$, $\alpha_2 \approx 0.2$) yields the tightest bound. Therefore, the optimizing choice of $(\alpha_1, \alpha_2)$ does not result in existing JS divergences which assume $\alpha_2 = 0.5$ (symmetric skew JS divergence) or $\alpha_2 = \alpha_1$ (asymmetric skew JS divergence).

## Appendix A: Proof of Theorem 3.1

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

$$\frac{1}{\beta M}\sum_{i=1}^{\beta M}\left[\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)]\right] \quad (17)$$

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

$$\frac{\sum_{i=\beta M+1}^{M}\left[\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P'_{Z_i}P_{W|Z_i}}[l(W,Z_i)]\right]}{(1-\beta)M}. \quad (18)$$

We first bound the difference $\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)]$ in (17). Towards this, we use the exponential inequalities in Lemma A.1 (proof included in [12]) obtained based on the change of measure approach adopted in [13]. Fix $\lambda = \lambda_1/\alpha_2$ for some $\lambda_1 > 0$ in (21) and $\lambda = -\lambda_1/(1-\alpha_2)$ in (22), and apply Jensen's inequality to get the following inequalities

$$\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P_W R_Z^{\alpha_1}}[l(W,Z)] \le \frac{\lambda_1\sigma^2}{2\alpha_2} + \frac{\alpha_2 D_{\text{KL}}(P'_Z||R_Z^{\alpha_1})}{\lambda_1} \quad (19)$$

$$\mathbb{E}_{P_W R_Z^{\alpha_1}}[l(W,Z)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)] \le \frac{\lambda_1\sigma^2}{2(1-\alpha_2)} + \frac{1-\alpha_2}{\lambda_1}\left(D_{\text{KL}}(P_Z||R_Z^{\alpha_1}) + I(W;Z_i)\right). \quad (20)$$

Adding (19) and (20) optimizing over $\lambda_1 > 0$ gives the required bound on $\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P_{Z_i}P_{W|Z_i}}[l(W,Z_i)]$. Similarly, we can bound the difference $\mathbb{E}_{P_W P'_{Z_i}}[l(W,Z_i)] - \mathbb{E}_{P'_{Z_i}P_{W|Z_i}}[l(W,Z_i)]$ in (18) by fixing $\lambda = \lambda_1 > 0$ in (21) and $\lambda = -\lambda_1$ in (23). Applying Jensen's inequality on both bounds, adding the resultant inequalities 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 $i = 1, \ldots, \beta M$,

$$\mathbb{E}_{P_W P'_{Z_i}}\left[\exp\left(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R_Z^{\alpha_1}}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{P'_{Z_i}(Z_i)}{R_{Z_i}^{\alpha_1}(Z_i)}\right)\right] \le 1, \quad (21)$$

$$\mathbb{E}_{P_{Z_i}P_{W|Z_i}}\left[\exp\left(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R_Z^{\alpha_1}}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{P_{Z_i}(Z_i)}{R_{Z_i}^{\alpha_1}(Z_i)} - \imath(W,Z_i)\right)\right] \le 1, \quad (22)$$

where $\imath(W,Z_i) = \log(P_{W,Z_i}(W,Z_i)/(P_W P_{Z_i}(W,Z_i))$ is the information density between random variables $W$ and $Z_i$. For $i = \beta M + 1, \ldots, M$, the inequality (21) holds along with the following inequality

$$\mathbb{E}_{P'_{Z_i}P_{W|Z_i}}\left[\exp\left(\lambda(l(W,Z_i) - \mathbb{E}_{P_W R_Z^{\alpha_1}}[l(W,Z)]) - \frac{\lambda^2\sigma^2}{2} - \log\frac{P'_{Z_i}(Z_i)}{R_{Z_i}^{\alpha_1}(Z_i)} - \imath(W,Z_i)\right)\right] \le 1. \quad (23)$$

## 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] 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.

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

[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] F. Nielsen, "On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid," *Entropy*, vol. 22, no. 2, p. 221, 2020.

[8] T. Yamano, "Some bounds for skewed $\alpha$-Jensen-Shannon divergence," *Results in Applied Mathematics*, vol. 3, p. 100064, 2019.

[9] 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.

[10] 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.

[11] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization rrror," in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, July 2019, pp. 587–591.

[12] S. T. Jose and O. Simeone, "Information-theoretic bounds on transfer generalization gap based on Jensen-Shannon divergence," *arXiv preprint: arXiv: 2010.09484*.

[13] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *arXiv preprint arXiv:2005.08044*, 2020.