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

Title:
On the Generalization for Transfer Learning: An Information-Theoretic Analysis

Date:
2024-01-01

Citation:
Wu, X., Manton, J. H., Aickelin, U. & Zhu, J. (2024). On the Generalization for Transfer Learning: An Information-Theoretic Analysis. IEEE Transactions on Information Theory, 70 (10), pp.7089-7124. https://doi.org/10.1109/TIT.2024.3441574.

Persistent Link:
https://hdl.handle.net/11343/348144

On the Generalization for Transfer Learning: An Information-Theoretic Analysis

Xuetong Wu, Jonathan H. Manton, *Fellow, IEEE,* Uwe Aickelin, *Fellow, IEEE,* Jingge Zhu, *Member, IEEE*

***Abstract*—Transfer learning, or domain adaptation, is concerned with machine learning problems in which training and testing data come from possibly different probability distributions. In this work, we give an information-theoretic analysis of the generalization error and excess risk of transfer learning algorithms. Our results suggest, perhaps as expected, that the Kullback-Leibler (KL) divergence $D(\mu\|\mu')$ plays an important role in the characterizations where μ and μ' denote the distribution of the training data and the testing data, respectively. Specifically, we provide generalization error and excess risk upper bounds for learning algorithms where data from both distributions are available in the training phase. Recognizing that the bounds could be sub-optimal in general, we provide improved excess risk upper bounds for a certain class of algorithms, including the empirical risk minimization (ERM) algorithm, by making stronger assumptions through the *central condition*. To demonstrate the usefulness of the bounds, we further extend the analysis to the Gibbs algorithm and the noisy stochastic gradient descent method. We then generalize the mutual information bound with other divergences such as ϕ-divergence and Wasserstein distance, which may lead to tighter bounds and can handle the case when μ is not absolutely continuous with respect to μ'. Several numerical results are provided to demonstrate our theoretical findings. Lastly, to address the problem that the bounds are often not directly applicable in practice due to the absence of the distributional knowledge of the data, we develop an algorithm (called InfoBoost) that dynamically adjusts the importance weights for both source and target data based on certain information measures. The empirical results show the effectiveness of the proposed algorithm.**

***Index Terms*—Transfer learning, generalization error, KL divergence, mutual information, ϕ-divergence.**

Xuetong Wu, Jonathan H. Manton and Jingge Zhu are with the Department of Electrical and Electronic Engineering, University of Melbourne, VIC, 3010, Australia, E-mail: xuetongw1@student.unimelb.edu.au; {jmanton, jingle.zhu}@unimelb.edu.au;

Uwe Aickelin is with the School of Computing and Information Systems, University of Melbourne, VIC, 3010, Australia, E-mail: uwe.aickelin@unimelb.edu.au.

I. Introduction

A learning algorithm is viewed as a stochastic mapping, which takes training data as its input and produces a hypothesis as the output. The output hypothesis will then be used on data not seen before (testing data). 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 labelled (or unlabelled) data from a distribution μ, namely the source domain, on which our learning algorithm trains, but wish to use the trained hypothesis for data coming from a different distribution μ', namely the target domain, from which we have limited data for training. Generalization error, a crucial measure of learning performance, is defined as the difference between the empirical loss and the population loss for a given hypothesis and indicates if the hypothesis suffers from overfitting or underfitting for the target domain of interest. Conventionally, many bounding techniques are proposed under different conditions and assumptions for traditional machine learning methods. For example, Vapnik and Chervonenkis [2] proposed VC-dimension which describes the richness of a hypothesis class for generalization ability. The notion of "algorithmic stability" was introduced in [3] and [4] for bounding the generalization error, by examining if a single training sample has a significant effect on the expected loss. PAC-Bayes bounds are a class of algorithm-dependent bounds first introduced by McAllester [5]. Xu and Mannor [6] develop another notion, namely the robustness for the generalization.

Unlike the stability, the robustness conveys geometric intuition and it can be extended to non-standard setups such as Markov chain or quantile loss, facilitating new bounds on generalization.

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. [7] defined the generalization error for transfer learning problems and gave its VC dimension-style bounds for classification problems with the proposed $\mathcal{H}$-divergence. Blitzer et al. [8] studied transfer learning problems with a similar setup and obtained upper bounds in terms of Rademacher complexity. Long et al. [9] developed a more general framework for transfer learning and the error is bounded by the distribution difference and adaptability of the hypothesis output. Dai et al. [10] and Eaton et al. [11] proposed two boosting-based transfer learning algorithms that emphasize the significance of source instances in transfer boosting, and it can be shown that the error bounds are increasingly smaller with iterations increasing. Zhang et al. [12] proposed an extension of theories in [7] to multiclass classification in transfer learning with a novel domain divergence called margin disparity discrepancy. 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. Traditional bounds used in transfer learning theory, such as VC dimension or Rademacher complexity, typically do not take the learning algorithm into account. They primarily focus on characterizing the complexity of the whole hypothesis class or making assumptions about the algorithms and loss functions. Consequently, the results often provide an overly pessimistic view of the learning problem, especially when the data or the algorithm has some underlying structure that could be exploited to simplify the learning task. On the other hand, some of the works focus specifically on certain transfer learning algorithms and loss functions [7, 10], and the resulting bounds on the generalization error are not universally applicable. Moreover, most bounds mentioned above are only concerned with the hypothesis or the algorithm solely, which fails to account for the complex interplay between data distribution, model hypothesis, and learning algorithm in determining the generalization error.

To characterize the intrinsic nature of a learning algorithm, some recent works have shown that the generalization error can be upper bounded using information-theoretic quantities. In particular, Russo and Zou [13] study the connection between mutual information and generalization error in the context of adaptive learning. The authors show 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 [14]. This paper exploits the information-theoretic framework in the transfer learning settings and derives the upper bounds for the generalization error. To summarize our main contributions, we highlight the following points.

1. We give an information-theoretic upper bound on the generalization error and the excess risk of transfer learning algorithms where training and testing data come from different distributions. Specifically, the upper bound involves the mutual information $I(W;Z)$ where W denotes the output hypothesis and Z denotes the training instance, and an additional term $D(\mu||\mu')$ that captures the effect of domain adaptation where μ, μ' denotes the distributions of the source and the target domains, respectively. Such a result can be easily extended to multi-source transfer learning problems. To show the usefulness of the bounds, we further specialize the upper bounds on three specific algorithms: empirical risk minimization (ERM), the noisy stochastic gradient descent, and the Gibbs algorithm.
2. We observe that the mutual information upper bounds derived from existing methods are in general sub-optimal in terms of the convergence rate. To arrive at the correct learning rate, we further tighten the bounds for specific algorithms such as ERM and regularized ERM using the proposed (η, c)-central condition. We demonstrate through a few examples that, compared with the mutual information bounds previously derived, the new bound improves the convergence rate of the generalization error from $O(\sqrt{1/n})$ to $O(1/n)$ up to the domain divergence. We could further arrive at intermediate rates under the relaxed (v, c)-central conditions.

3. We extend the results by using other types of divergences such as ϕ-divergence and Wasserstein distance, which can be tighter than the mutual information bound under mild conditions. Such an extension also allows us to handle more general learning scenarios where the mutual information bound may be vacuous, e.g. when μ is not absolutely continuous with respect to μ'.
4. Finally, we give a few examples of some simple transfer learning problems to validate our proposed bounds. However, in practical scenarios, these bounds are often not directly applicable due to the lack of knowledge of the data distributions. To address this, we propose a boosting-type algorithm called `InfoBoost` in which the importance weights for source and target data are adjusted adaptively in accordance with information measures. We also conduct several experiments on real datasets, and the empirical results show that, in most cases, our algorithm outperforms the state-of-the-art benchmarks.

The outline of this paper is structured as follows. We provided an in-depth review of the literature on the information-theoretic analysis for machine learning and transfer learning in Section II. Then, we formally formulate the transfer learning problem and give the main results in Section III. We then specialize the bounds on the noisy iterative algorithms and the Gibbs algorithm following a similar intuition in Section IV. In Section V, for scenarios where the mutual information-based bounds are vacuous, we extend the results to other divergences such as the ϕ-divergence and the Wasserstein distance. Some examples are illustrated in Section VI to show the effectiveness of the bounds. Additionally, from a practical perspective, we propose an intuitive algorithm for transfer learning problems in Section VI that potentially works in real-world scenarios, inspired by the mutual information bounds. Section VII concludes the paper with some remarks.

II. Literature Review

In this section, we take a closer look at the related works that have shaped our understanding of machine learning and transfer learning. We will begin with an exploration of the information-theoretic analysis for traditional machine learning. Considering that the current information-theoretic analysis has a couple of primary limitations, such as the bounds being ineffective for deterministic algorithms and typically showing a slower rate of generalization error w.r.t. the data size, we will also delve into related works that improve the learning rate using various novel approaches.

A. Information-theoretic analysis for machine learning

Generally speaking, the *generalization error* measures how well a learned model performs on previously unseen data, which is usually characterized by the gap between the training loss and testing loss. The generalization error of a learning algorithm lies in the core analysis of the statistical learning theory, the estimation of which becomes remarkably crucial in machine learning problems. Conventionally, many bounding techniques are proposed with different notions as aforementioned [15, 16, 5, 6, 14, 17]. However, most bounds mentioned above are only concerned with the hypothesis or the algorithm solely. For example, VC-dimension methods care about the worst-case bound, which depends only on the hypothesis space. The stability methods only specify the properties of learning algorithms but do not require additional assumptions on hypothesis space.

Information theory has been demonstrated to not only offer theoretical insights into the generalization error but also steer the intuition behind specific learning algorithms as it is a useful tool that extracts the statistical characterizations of the input data. Recently, Russo and Zou [13] and Xu and Raginsky [18] studied a general statistical learning problem, and the authors give theoretical bounds for striking the balance between the data fit and generalization by controlling the mutual information between the output hypothesis and input data. In contrast to conventional generalization error bounds, such a result hinges on the data distribution, algorithms, and the learned hypothesis. Moreover, the dependence between the input data and the output hypothesis can be regarded as a measure that prompts the hypothesis class and algorithmic stability, hence recovering many other existing results such as VC dimension, algorithmic stability, and differential privacy. Similarly, Bassily et al. [19] studied learning algorithms that only use a small amount of information from input samples by focusing on the capacity of the algorithm space. Furthermore, the mutual information bounds are improved in [20], and the authors proposed a tighter

version with the mutual information between the output hypothesis and single data instance for the generalization error. In [21], several information-theoretic measures from an algorithmic stability perspective are used to upper bound the generalization error. In contrast to the mutual information, the generalization error bounds based on the Wasserstein distance are proposed in [22] and have been extended in [23] with the total variation distance by exploiting the geometric nature of the Kantorovich-Rubinstein duality theorem.

B. Improvements on information-theoretic bounds

However, there are several drawbacks to the information-theoretic bounds for typical machine learning problems. The first is that, for some deterministic algorithms, the mutual information quantities may be infinite and the bound will become vacuous. To tackle the infinity issue, instead of measuring information with the whole dataset, Bu et al. [20] propose the bound based on the mutual information between the single instance and the hypothesis, which is finite even for deterministic algorithms. Negrea et al. [24] propose generalization error bounds based on a subset of the whole dataset chosen uniformly at random. Steinke and Zakynthinou [25] and Haghifam et al. [26] improve the results with the conditional mutual information that is always finite by introducing a set of discrete (binary) random variables. Rodríguez Gálvez et al. [27] have further tightened the bound using the random subset techniques under the Wasserstein distance. Another drawback is that these bounds are usually hard to estimate if the hypothesis and the data are of high dimensions. To make the quantity estimable, Harutyunyan et al. [28] derive a novel generalization error bound that measures information with the predictions instead of the hypothesis produced by the training samples, which is significantly easier to estimate. The third drawback is that the bounds are usually sub-optimal, in terms of the convergence rate w.r.t. the sample size. In most of the relevant works, the convergence rate of the expected generalization error in traditional statistical learning problems is in the form of $O(\sqrt{\lambda/n})$ where λ is some information-theoretic quantities such as the mutual information between the data sample and the learned hypothesis. However, such a learning rate is typically considered to be "slow", compared to a "fast rate" of $O(1/n)$ in many learning scenarios. Fast rate conditions are less investigated under the information-theoretic framework. Only a few works are dedicated to it, e.g., Grünwald et al. [29] applies the conditional mutual information [25] for the fast rate characterization under the PAC-Bayes framework and the results rely on prior knowledge of the hypothesis space. Wu et al. [1] proposed the (η, c)-central condition for fast rate generalization error with the mutual information between the hypothesis and single data instance. Bu et al. [30] characterize the exact generalization error of the transfer learning under the Gibbs algorithm using the symmetric KL divergence and arrive at the fast rate under mild conditions.

C. Transfer learning bounds and comparisons

The transfer learning problem is of particular interest in real-world applications, as in many cases we often have easy access to a substantial amount of labelled (or unlabelled) data from one distribution, on which our learning algorithm trains, but wish to use the learned hypothesis for data coming from a different distribution, from which we only have limited data for training. In practice, there will be perturbations or shifts in the distributions of the training and testing data, or obtaining the training data for some tasks can be very expensive and difficult such as robotics [31, 32], medical images [33, 34, 35] and rare language translation [36, 37]. Popular empirical risk minimization (ERM)–based methods usually minimize a convex combination of source and target data. The performance of the ERM has been initiated and investigated in works such as [8, 38, 7, 9]. These studies generally offer different bounds on the generalization error, contingent on specific domain divergences between source and target distributions. For example, the high-probability bounds for generalization error based on the error distance are presented in works such as Theorem 2 in [8], Theorem 3 in [7], and Theorem 2 in [9]. Other works, such as Theorem 8 in [38] and Theorem 3.7 in [12], place their results within the context of the $\mathcal{A}$-discrepancy and disparity discrepancy with the Rademacher complexity of the hypothesis space. However, such discrepancies are typically linked to the hypothesis space only as they assess the largest discrepancy between two domains among all possible hypotheses given certain loss functions. It is also worth noting that these bounds are usually worst-case bounds as they work for any hypothesis in the hypothesis space. In specific learning regimes, such as in [10], the researchers investigate the boosting type

transfer learning algorithm, introducing a learning bound in Theorem 3 that hinges on the iteration number and target sample error. However, this bound is closely tied to the chosen hypothesis space and does not capture the value of the source data. Kuzborskij and Orabona [39] consider the scenario in that only the source hypothesis induced from the source data is available. The authors conduct an algorithmic stability analysis on a class of hypothesis transfer learning problems with the regularized least squares algorithm, and the result suggests the relatedness of the source and target domains (e.g., how the source hypothesis performs on the target domains) determines the effectiveness of the transfer. Germain et al. [40] take the first attempt at the transfer learning problem under the PAC-Bayesian framework and provide a novel pseudo-distance on domain distributions, which leverages the idea from [7, 38] by changing the pointwise disagreement to an averaging disagreement to fit into the PAC-Bayesian analysis.

In our work, we introduce an information-theoretic framework for the ERM setup that presents multiple advantages over previous results. Specifically, our suggested bounds delve deeply into the dependence among data distribution, output hypothesis, and the algorithm itself. Such a bound may, in fact, be tighter than traditional bounds because the relationship between the dataset and the hypothesis can be seen as a metric that encourages algorithmic stability. Our bound also contains the KL divergence between the source and target domains, effectively capturing the domain divergence. Moreover, whereas the majority of previous bounds exhibit sublinear convergence w.r.t. the sample size up to the domain divergence term, our bound can provide the correct linear convergence rate for some algorithms.

III. 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 absolutely continuous with respect to μ'. 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 with size n. More precisely, for a fixed number $\beta \in [0, 1)$, we assume that βn is an integer and the samples $S' = \{Z'_1, \ldots, Z'_{\beta n}\}$ are drawn IID from the target distribution μ', and the samples $S = \{Z_{\beta n+1}, \ldots, Z_n\}$ are drawn IID from the source distribution μ.

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*

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

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 the focus of this paper. We call $(\mu', \mu, \ell, \mathcal{W}, \mathcal{A})$ as a transfer learning tuple.

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 the 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' defined as

$$\hat{L}_\alpha(w, S, S') := \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \ell(w, Z'_i) + \frac{1-\alpha}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \ell(w, Z_i) \tag{3}$$

for some weight parameter $\alpha \in [0, 1]$ to be determined. We will use $\hat{L}_\alpha(w)$ interchangeably for $\hat{L}_\alpha(w, S, S')$ to simplify the notation, and then we define the ERM solution by

$$w_{\mathsf{ERM}} := \operatorname{argmin}_w \hat{L}_\alpha(w, S, S'). \tag{4}$$

Accordingly, we define the optimal hypothesis with respect to the distribution μ' as

$$w^* = \operatorname{argmin}_{w \in \mathcal{W}} L_{\mu'}(w). \tag{5}$$

We are interested in two quantities for any general machine learning algorithms, including the ERM algorithm. The first one is the *generalization error* defined as

$$\text{gen}(w, S, S') := L_{\mu'}(w) - \hat{L}_\alpha(w, S, S'), \quad (6)$$

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

$$R_{\mu'}(w) = L_{\mu'}(w) - L_{\mu'}(w^*), \quad (7)$$

which is the difference between the population risk of w compared to that of the optimal hypothesis w^*. Notice that the excess risk of the ERM solution is related to the generalization error via the following upper bound:

$$\begin{aligned}
&L_{\mu'}(w_{\mathsf{ERM}}) - L_{\mu'}(w^*) \\
&= L_{\mu'}(w_{\mathsf{ERM}}) - \hat{L}_\alpha(w_{\mathsf{ERM}}, S, S') + \hat{L}_\alpha(w_{\mathsf{ERM}}, S, S') \\
&- \hat{L}_\alpha(w^*, S, S') + \hat{L}_\alpha(w^*, S, S') \\
&- L_\alpha(w^*) + L_\alpha(w^*) - L_{\mu'}(w^*) \\
&\le \text{gen}(w_{\mathsf{ERM}}, S, S') + \hat{L}_\alpha(w^*, S, S') - L_\alpha(w^*) \\
&\quad + (1-\alpha)(L_\mu(w^*) - L_{\mu'}(w^*)), \quad (8)
\end{aligned}$$

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

$$\begin{aligned}
L_\alpha(w) &:= (1-\alpha)L_\mu(w) + \alpha L_{\mu'}(w) \\
&= (1-\alpha)\mathbb{E}_{Z\sim\mu}\left[\ell(w, Z)\right] + \alpha\mathbb{E}_{Z'\sim\mu'}\left[\ell(w, Z')\right]. \quad (9)
\end{aligned}$$

B. Upper bound on the generalization error

We consider the hypothesis W as a random variable induced by the random samples S, S' with some algorithm $\mathcal{A}$, characterized by a conditional distribution $P_{W|SS'}$. We will first study the expectation of the generalization error

$$\begin{aligned}
&\mathbb{E}_{WSS'}\left[\text{gen}(W, S, S')\right] \\
&= \mathbb{E}_{WSS'}\left[L_{\mu'}(W) - \hat{L}_\alpha(W, S, S')\right], \quad (10)
\end{aligned}$$

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

$$\begin{aligned}
&P_{WSS'}(w, S, S') \\
&= P_{W|SS'}(w|S, S') \prod_{i=1}^{\beta n} \mu'(z'_i) \prod_{i=\beta n+1}^{n} \mu(z_i). \quad (11)
\end{aligned}$$

Furthermore, we use P_W to denote the marginal distribution of W induced by the joint distribution $P_{WSS'}$. Following the characterization used in [20], 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 hypothesis W induced by a certain algorithm $P_{W|SS'}$, as well as the KL-divergence between the source and target distributions. As pointed out in [20], using mutual information between the hypothesis and individual samples $I(W; Z_i)$, in general, gives a tighter upper bound than using $I(W; S)$.

Theorem 1 (Generalization error of generic algorithms)**.** *Assume that the hypothesis W is distributed over P_W induced by some algorithm, and the cumulant generating function of the random variable $\ell(W, Z) - \mathbb{E}\left[\ell(W, Z)\right]$ 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 (10) is upper bounded as*

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

where we define

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

The proof can be found in Appendix A. Notice that the bound above is not specific to the ERM algorithm but applicable to any hypothesis generated by a learning algorithm that has a bounded cumulant generating function. Comparing the derived results with other transfer learning bounds such as Theorem 2 in [8], and Theorem 3 in [7] that are applied in worst-case scenarios for any hypothesis, our proposed bound suggests that the generalization error inherently

depends on the mutual information $I(W;Z_i)$ data distribution, the output hypothesis, and the algorithm itself, which may lead to a tighter characterization as it also takes the algorithm $P_{W|SS'}$ into account. From a stability point of view, good algorithms (ERM, for example) should ensure that the mutual information $I(W;Z_i)$ vanishes as $n \to \infty$, which has similar insights as [16] from the point view that a single instance should not affect the output hypothesis much. 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. The KL divergence is only dependent on the data distributions and irrelevant to the loss function and hypothesis. This is in contrast to other metrics that may depend on the hypothesis space and the prediction functions such as discrepancy distance [38] and $\mathcal{H}\Delta\mathcal{H}$-divergence[7].

Remark 1. *It is natural to consider the problem of minimizing the upper bound with respect to the parameter α as it mediates the balance between performance on the source and target domains. This is, however, a non-trivial problem as the output hypothesis W implicitly involves α, and optimization of bound w.r.t. α is challenging. In principle, α should depend on the source and target data sizes and distribution differences between the two domains. Specifically, when the target domain data is abundant and the source and target domains are significantly different, α can be biased towards the target domain (i.e., α should be closer to 1). This is because the model can rely more on the target domain data for better learning performance. Conversely, when the target domain data is limited and the source and target domains are closely related, α should be biased towards the source domain (i.e., α should be closer to 0). This is because the model can benefit more from leveraging knowledge from the source domain to avoid overfitting the limited target data.*

The optimal value of α often requires empirical validation. Notice that 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 approximate as β as suggested in [41, 7] that this choice is shown to achieve a tighter bound empirically. Overall, we suggest that non-asymptotically, α should approach 1 with the target sample size βn increasing, say, $\alpha = 1 - O(\frac{1}{\beta n})$. In real practice, techniques such as cross-validation or grid search could also be used to tune α by evaluating model performance across a range of values.

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 the 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*

$$\mathbb{E}_{WS}\left[\text{gen}(W,S)\right] \le \frac{1}{n}\sum_{i=1}^{n} \psi_{-}^{*-1}(I(W;Z_i) + D(\mu||\mu')),$$

$$-\mathbb{E}_{WS}\left[\text{gen}(W,S)\right] \le \frac{1}{n}\sum_{i=1}^{n} \psi_{+}^{*-1}(I(W;Z_i) + D(\mu||\mu')).$$

The proof of this result is given in Appendix B. If the loss function $\ell(W,Z)$ is r^2-subgaussian, namely

$$\log \mathbb{E}\left[e^{\lambda(\ell(W,Z)-\mathbb{E}[\ell(W,Z))]}\right] \le \frac{r^2\lambda^2}{2}$$

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)**.** *Let P_W be the marginal distribution induced by S, S' and $P_{W|SS'}$ for some algorithm. If $\ell(W,Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu'$, then the expectation of the generalization error is upper bounded as*

$$|\mathbb{E}_{WSS'}\left[\text{gen}(W,S,S')\right]| \le \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W;Z_i')} + \frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W;Z_i)+D(\mu||\mu'))}. \tag{12}$$

If $\beta = 0$, for any hypothesis W induced by S and a learning algorithm $P_{W|S}$, we have the upper bound

$$|\mathbb{E}_{WS}\left[\text{gen}(W,S)\right]| \leq \frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{I(W;Z_i)+D(\mu||\mu')}. \tag{13}$$

The above result follows directly from Theorem 1 and Corollary 1 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 2. *Using the chain rule of mutual information and the fact that Z_i's are IID, we can relax the upper bound in (13) as*

$$\mathbb{E}_{WS}\left[\text{gen}(W,S)\right] \leq \sqrt{2r^2\left(\frac{I(W;S)}{n}+D(\mu||\mu')\right)}, \tag{14}$$

which recovers the result in [18] if $\mu = \mu'$. Moreover, we see that the effect of the "domain shift" is simply captured by the KL divergence between the source and the target distribution.

Remark 3. *Even though we focus on the supervised learning setups for transfer learning with the convexly combined empirical loss. Such a framework can be easily extended to various different transfer learning setups such as multi-source transfer learning problems, pre-trained hypothesis setups, and unsupervised setups. To maintain the paper focus and prevent excessive use of notations, we have placed all detailed results and discussions in Appendix C, D and E, respectively.*

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 (7). To do this, we first define a distance quantity between the two divergent distributions as

$$d_{\mathcal{W}}(\mu,\mu') = \sup_{w\in\mathcal{W}}|L_\mu(w) - L_{\mu'}(w)|. \tag{15}$$

The following theorem gives an upper bound on the excess risk.

Theorem 2 (Excess risk of ERM)**.** *Let P_W be the marginal distribution induced by S, S' and $P_{W|SS'}$ for the ERM algorithm, assume the loss function $\ell(W,Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu'$. Then the following inequality holds.*

$$\begin{aligned}\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq{}& \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i')}\\ &+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))}\\ &+(1-\alpha)d_{\mathcal{W}}(\mu,\mu').\end{aligned} \tag{16}$$

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

$$\begin{aligned}\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq{}& \frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))}\\ &+d_{\mathcal{W}}(\mu,\mu').\end{aligned} \tag{17}$$

The proof of this theorem is given in Appendix F. Note that $d_{\mathcal{W}}(\mu,\mu')$ is normally known as the integral probability metric, which is challenging to evaluate. Sriperumbudur et al. [42] investigated the data-dependent estimation to compute the quantity using the Kantorovich metric, Dudley metric, and kernel distance, respectively. Another evaluation method is proposed in [7] to resolve the issue for classification problems. We verify our bound for the transfer learning with a toy example studied in [20]. In Section VI-B, we also verify the bounds on the logistic regression transfer problem where the hypothesis cannot be explicitly calculated.

Example 1 (Estimating the mean of Gaussian)**.** *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 $i = 1, 2, \cdots, n$. 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 for any $\lambda > 0$:

$$\log \mathbb{E} e^{\lambda(\ell(w,z_i)-\mathbb{E}[\ell(w,z_i)])} \leq \sigma_\ell^4 \lambda^2 + \frac{2\lambda^2\sigma_\ell^2(m-m')^2}{1+2\lambda\sigma_\ell^2}.$$

Then it can be seen that the loss function is $\sqrt{2\sigma_\ell^4 + 4\sigma_\ell^2(m-m')^2}$-subgaussian under the distribution $P_W \otimes \mu'$. Let $\sigma_{\ell'}^2 = 2\sigma_\ell^4 + 4\sigma_\ell^2(m-m')^2$, we reach at

$$\mathbb{E}_{WS}\left[\text{gen}(W_{\mathsf{ERM}}, S)\right] \leq \sqrt{\sigma_{\ell'}^2 \log \frac{n}{n-1} + 2\sigma_{\ell'}^2 D(\mu\|\mu')},$$

where $D(\mu||\mu') = \frac{(m-m')^2}{2\sigma^2}$. Then the excess risk is upper bounded by,

$$\mathbb{E}_W[L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*)] \leq \sqrt{\sigma_{\ell'}^2 \log \frac{n}{n-1} + 2\sigma_{\ell'}^2 D(\mu\|\mu')} + d_{\mathcal{W}}\left(\mu, \mu'\right).$$

In this case, the generalization error and the excess risk of W_{ERM} can be calculated exactly to be

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

$$\mathbb{E}_W[L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*)] = \frac{\sigma^2}{n} + 2\sigma^2 D(\mu||\mu').$$

The derived excess risk bound approaches $\sqrt{4\sigma^4 D(\mu\|\mu') + 16\sigma^4 D(\mu\|\mu')^2} + d_{\mathcal{W}}(\mu, \mu')$ as $n \to \infty$ with a decay rate of $O(1/\sqrt{n})$, which does not capture the bound asymptotically well as the true value should be $\sqrt{4\sigma^4 D(\mu\|\mu')^2}$. Moreover, the hypothesis space-dependent quantity $d_{\mathcal{W}}(\mu, \mu')$ will be infinite if w is unbounded, resulting in a vacuous bound. To further tighten the bound, we propose various "easiness" conditions on the excess risk, which is shown to capture the true behavior up to a scaling factor in Section III-D.

D. Fast rate upper bound on the excess risk of ERM

As can be seen from previous sections, the convergence rate of the excess risk is in the form of $O\left(\sqrt{\frac{\lambda}{\beta n}} + \sqrt{\frac{\lambda'}{(1-\beta)n} + D(\mu\|\mu')} + d_{\mathcal{W}}(\mu, \mu')\right)$ where λ and λ' is some information-theoretic related quantities such as the mutual information, $D(\mu\|\mu')$ and $d_{\mathcal{W}}(\mu, \mu')$ are the domain divergences between the source and target domains. However, such a learning rate is in general suboptimal.

In this section, we give an alternative analysis for the excess risk under various "easiness" conditions following the idea from [29, 43]. With the new technique, we can show that the excess risk is characterized by the mutual information between the hypothesis and data instances and the rate will be of the form $O\left(\frac{\eta}{\beta n} + \frac{\eta'}{(1-\beta)n} + D(\mu\|\mu')\right)$ where η and η' are some information-theoretic related quantities different from λ and λ' for specific learning algorithms such as empirical risk minimization and the hypothesis dependent term $d_{\mathcal{W}}(\mu, \mu')$ vanishes in the new bound. While the results presented in this section offer a more refined analysis for certain scenarios, it is important to note that they do not entirely supersede the 'slow-rate' results. For example, in the case of the Gaussian problem, we could show that the fast rate bound in Theorem 3 does indeed provide a strictly better bound on the excess risk than the slow rate result in Theorem 2. However, in a more general context, the two bounds on excess risk are not directly comparable, as they pertain to different underlying assumptions and scenarios. To simplify the notation, we also define the *empirical* excess risk w.r.t. w^* for some $w \in \mathcal{W}$ given the dataset S as

$$\hat{R}(w, S) := \hat{L}(w, S) - \hat{L}(w^*, S). \tag{18}$$

The empirical excess risk combined with both source and target is defined as

$$\hat{R}_\alpha(w, S, S') := \hat{L}_\alpha(w, S, S') - \hat{L}_\alpha(w^*, S, S'). \tag{19}$$

We also define the unexpected excess risk:

$$r(w, z_i) := \ell(w, z_i) - \ell(w^*, z_i) \tag{20}$$

for single instance z_i as well. We further define the excess risk as

$$R_{\mu'}(w_{\mathsf{ERM}}) := L_{\mu'}(w_{\mathsf{ERM}}) - L_{\mu'}(w^*)]. \tag{21}$$

Then the expected excess risk over W_{ERM} can be bounded by the following inequality:

$$\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq \mathbb{E}_{WSS'}[\mathcal{E}(W_{\mathsf{ERM}}, S, S')], \tag{22}$$

where the empirical excess risk gap is defined as

$$\mathcal{E}(w, S, S') = \alpha\left(R_{\mu'}(w) - \hat{R}(w, S')\right) + (1-\alpha)\left(R_{\mu'}(w) - \hat{R}(w, S)\right). \tag{23}$$

Here we have used the fact $\hat{L}_\alpha(W_{\mathsf{ERM}}, S, S') - \hat{L}_\alpha(w^*, S, S') \leq 0$ by the definition of W_{ERM}. Recently, numerous studies have focused on formulating fast information-theoretic bounds [29, 43], aiming to eliminate the square root present in the upper bounds proposed in the previous sections. This is mainly achieved by changing the assumptions from the loss function $\ell(w, z)$ to the unexpected excess risk $r(w, z)$. In this context, referencing the optimal hypothesis w^* becomes crucial to further narrow down the concentration condition. This essential requirement is recognized as the central condition [44, 29, 43]. In our work, we utilize this condition and make subtle modifications to cater to our goals in transfer learning as follows.

Definition 1 ((η, c)-Central Condition)**.** *Let $\eta > 0$ and $0 < c \leq 1$. We say that a transfer learning problem satisfies the expected (η, c)-central condition under the target distribution μ' if the following inequality holds for the optimal hypothesis w^*:*

$$\log \mathbb{E}_{P_W \otimes \mu'} \left[e^{-\eta(\ell(W,Z) - \ell(w^*,Z))} \right] \leq - c\eta \mathbb{E}_{P_W \otimes \mu'} \left[\ell(W, Z) - \ell(w^*, Z) \right]. \tag{24}$$

where P_W is the marginal distribution of the output hypothesis.

This condition is similar to the central condition [44, 45], where some assumptions are made on the small lower tail for the excess risk function with the exponential concavity, implying good concentration properties of the excess risk. Compared to the η-central condition defined in [44, Def. 3.1] which can be retrieved by setting $c = 0$, the R.H.S. of (24) will be negative and has tighter control of the tail behavior for some $c > 0$. We also point out that such a condition is indeed the key assumption for improving the rate by removing the square root, which also coincides with some well-known conditions that lead to a fast rate such as the Bernstein condition [46, 47, 48, 49] and the central condition with the witness condition [44, 45]. Next, we provide several instances where the (η, c)-central condition is satisfied. While some of these examples are discussed in [43], we revisit them here for the sake of completeness.

Example 2. *If $r(W, Z)$ is (ν^2, α)-sub-exponential under the distribution $P_W \otimes \mu'$, then the learning tuple satisfies $(\min(\frac{1}{\alpha}, \frac{\nu^2}{\mathbb{E}_{P_W \otimes \mu'}[r(W,Z)]}), \frac{1}{2})$-central condition.*

Example 3. *If $r(W, Z)$ is (ν^2, α)-sub-Gamma under the distribution $P_W \otimes \mu'$, then the learning tuple satisfies $(\frac{\mathbb{E}_{P_W \otimes \mu'}[r(W,Z)]}{\nu^2 + \alpha \mathbb{E}_{P_W \otimes \mu'}[r(W,Z)]}, \frac{1}{2})$-central condition.*

Example 4. *Let $\gamma \in [0, 1]$ and $B \geq 1$. Let P_W is induced by $P_{WSS'}$. We assume that the **Bernstein condition** holds under the target distribution $P_W \otimes \mu'$. Namely, the following inequality holds for the optimal hypothesis w^*:*

$$\mathbb{E}_{P_W \otimes \mu'} \left[\left(\ell\left(W, Z'\right) - \ell\left(w^*, Z'\right) \right)^2 \right] \leq B \left(\mathbb{E}_{P_W \otimes \mu'} \left[\ell\left(W, Z'\right) - \ell\left(w^*; Z'\right) \right] \right)^\gamma.$$

In case, and if $\gamma = 1$ and $r(w, z_i)$ is bounded by $-b$ with some $b > 0$ for all w and z_i, then the learning tuple also satisfies $(\min(\frac{1}{b}, \frac{1}{2B(e-2)}), \frac{1}{2})$-central condition.

The Bernstein condition is commonly identified as a way to describe the 'easiness' of a learning problem. The typical Bernstein condition necessitates that the inequality is satisfied for all $w \in \mathcal{W}$. However, in our case, we only require that the inequality is satisfied in expectation over P_W. In particular, consider the source-only case, $\gamma = 1$ corresponds to the easiest and the learning rate will be $O(\frac{1}{n} + cD(\mu\|\mu'))$ if $I(W; Z_i)$ is converging with the rate of $O(\frac{1}{n})$ for some leading constant c. For the bounded loss, the Bernstein condition will automatically hold with $\gamma = 0$ and it will recover the results in Corollary 1 with the rate of $O(\sqrt{\frac{1}{n} + cD(\mu\|\mu')})$.

Example 5. *The second condition is the central condition with the witness condition [44, 45], which also implies the (η, c)-central condition. We say the learning tuple $(\mu, \mu', \ell, \mathcal{W}, \mathcal{A})$ satisfies the η-central condition [44, 45] if for the optimal hypothesis w^*, the following inequality holds,*

$$\mathbb{E}_{P_W \otimes \mu'} \left[e^{-\eta(\ell(W,Z) - \ell(w^*,Z))} \right] \leq 1.$$

We also say the learning tuple $(\mu, \mu', \ell, \mathcal{W}, \mathcal{A})$ satisfies the (u, c)-witness condition [45] if for constants $u > 0$ and $c \in (0, 1]$, the following inequality holds.

$$\mathbb{E}_{P_W \otimes \mu'}[(\ell(W, Z) - \ell(w^*, Z)) \cdot \mathbf{1}_{\{\ell(W,Z) - \ell(w^*,Z) \leq u\}}] \geq c \mathbb{E}_{P_W \otimes \mu'} \left[\ell(W, Z) - \ell(w^*, Z) \right],$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. Then we have the following statement: If the learning tuple satisfies both η-central condition and (u, c)-witness condition, then the learning tuple also satisfies the $(\eta', \frac{c - \frac{c\eta'}{\eta}}{\eta' u + 1})$-central condition for any $0 < \eta' < \eta$.

The standard η-central condition [44, 50, 45] is essential for deriving fast-rate bounds of generalization error. Some typical examples include exponential concave loss functions (including log-loss) with $\eta = 1$ [50, 51] and bounded loss functions with Massart noise condition with various η [44]. The witness condition [45, Def. 12] is applied to exclude cases where learnability is inherently impossible. This condition ensures that we exclude the poorly performing hypothesis w with almost zero probability (even though it can still affect the expected loss), which we will never observe empirically. One trivial example is, if the excess risk is upper bounded by some constant b, we may always take $u = b$ and $c = 1$ so that a witness condition is satisfied.

With the definitions in place, we derive the fast rate bounds for the excess risk in transfer learning under the (η, c)-central condition as follows.

Theorem 3 (Fast Rate with (η, c)-Central Condition)**.** *Assume the learning problem with the ERM algorithm satisfies the expected (η, c)-central condition under the target distribution μ'. Then for any $0 < \eta' \leq \eta$, the expected excess risk can be upper bounded by,*

$$\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq \frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n} I(W_{\mathsf{ERM}}; Z_i') + \frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{ERM}}; Z_i) + D(\mu\|\mu')\right). \tag{25}$$

More generally, for any algorithm $\mathcal{A}$ and any W induced by the algorithm, if the expected (η, c)-central condition holds, we have that for any $0 < \eta' \leq \eta$,

$$\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq \frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n} I(W; Z_i') + \frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W; Z_i) + D(\mu\|\mu')\right) + \frac{1}{c}\left[\alpha\mathbb{E}_{WS'}[\hat{R}(W, S')] + (1-\alpha)\mathbb{E}_{WS}[\hat{R}(W, S)\right]. \tag{26}$$

The proof can be found in Appendix G. Now we compare (25) with the bound in Theorem 2 which we reproduce below,

$$\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{n}\sqrt{I\left(W_{\mathsf{ERM}}; Z_i'\right)} + \frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{\left(I\left(W_{\mathsf{ERM}}; Z_i\right) + D\left(\mu\|\mu'\right)\right)} + (1-\alpha)d_{\mathcal{W}}\left(\mu, \mu'\right).$$

In the new bound, the square root term is removed and we may achieve a faster rate for converging to the domain divergence $D(\mu\|\mu')$. Furthermore, the new bound does not contain the hypothesis space-dependent divergence term $d_{\mathcal{W}}\left(\mu, \mu'\right)$, which might be very large or unbounded for certain distributions and hypothesis space. In the following Gaussian mean estimation example, we verify that the new bound is tighter than the previous bound and captures the true behaviour of the excess risk.

Example 6 (Continuing from Example 1)**.** *We continue to examine the bound in Theorem 3 that achieves the correct rate of convergence in the Gaussian mean estimation, which satisfies the (η, c)-central condition for certain η and c. To this end, for a large sample size n, we check,*

$$\log \mathbb{E}_{P_W\otimes\mu'}\left[e^{-\eta r(W,Z)}\right] = \log\sqrt{\frac{n}{n + 2\eta\sigma^2(1-2\eta\sigma^2)}} + (2\eta^2\sigma^2 - \eta)(m-m')^2 \leq -c\eta\left(\frac{\sigma^2}{n} + (m-m')^2\right).$$

From the above inequality, this learning problem satisfy the (η, c)-central condition for any $0 < \eta < \frac{1}{2\sigma^2}$ and any

$$c \leq -\frac{1}{\eta}\lim_{n\to\infty}\frac{\frac{1}{2}\log\frac{n}{n+2\eta\sigma^2(1-2\eta\sigma^2)} + (2\eta^2\sigma^2 - \eta)(m-m')^2}{\frac{\sigma^2}{n} + (m-m')^2} = 1 - 2\eta\sigma^2,$$

by the quotient law of limits, where the choice of c is independent of the sample size and thus does not affect the convergence rate. Therefore, take $\eta = \frac{1}{4\sigma^2}$ and $c = \frac{1}{2}$, the excess risk bound in (26) for ERM under the source only case becomes

$$\begin{aligned}\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] &\leq \frac{1}{c\eta' n}\sum_{i=1}^{n}\left(I(W; Z_i) + D(\mu\|\mu')\right) + \frac{1}{c}\mathbb{E}_{P_{WS}}[\hat{R}(W, S)] \\ &= 4\sigma^2\log\frac{n}{n-1} + 4(m-m')^2 - \frac{2\sigma^2}{n} - 2(m-m')^2 \\ &= 4\sigma^2\log\frac{n}{n-1} - \frac{2\sigma^2}{n} + 2(m-m')^2 \\ &\asymp \frac{2\sigma^2}{n} + 2(m-m')^2,\end{aligned} \tag{27}$$

for large n. While the true excess risk can be calculated by

$$\begin{aligned}\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] &= (m-m')^2 + \frac{\sigma^2}{m} + \sigma^2 - \sigma^2 \\ &= (m-m')^2 + \frac{\sigma^2}{n}.\end{aligned}$$

The new bound is tight in the sense that it captures the true excess risk up to a scaling factor. However, if we apply (17) and the bound becomes,

$$\begin{aligned}\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq &\sqrt{\sigma_{\ell'}^2 \log\frac{n}{n-1} + 2\sigma_{\ell'}^2 D(\mu\|\mu')} \\ &+ d_{\mathcal{W}}(\mu,\mu'),\end{aligned}$$

where $\sigma_{\ell'}^2 = 2\sigma_\ell^4 + 4\sigma_\ell^2(m-m')^2$ and $d_{\mathcal{W}}(\mu,\mu') = \sup_{w\in\mathcal{W}} \left|(w-m)^2 - (w-m')\right|$. Then this bound approaches

$$\sqrt{4(m-m')^4 + 2\sigma^2(m-m')^2} + d_{\mathcal{W}}(\mu,\mu') \quad (28)$$

with the rate of $\sqrt{1/n}$, which is apparently worse than (27).

Remark 4 (Justification of the tightness)**.** *In the following, we examine the tightness of the bound and show why $r(w,z)$ is a more sensible choice than $\ell(w,z)$. For simplicity, we consider the source-only case with the Gaussian mean estimation problem where in the proof we used the Donsker-Varadhan representation for the KL divergence between the $D(P_{WZ_i}\|P_W\otimes\mu')$ for each Z_i:*

$$\begin{aligned}D(P_{WZ_i}\|P_W\otimes\mu') = &\sup_{f:\mathcal{W}\otimes\mathcal{Z}\to\mathbb{R}} \mathbb{E}_{WZ_i}[f(W,Z_i)] \\ &- \log\left(\mathbb{E}_{P_W\otimes\mu'}\left[e^{f(W,Z_i)}\right]\right). \quad (29)\end{aligned}$$

It is known that under mild conditions [52], the optimal function where the equality is achieved for Eq. (29) is chosen by $f'(dP_{WZ_i}/(d(P_W\otimes\mu')$ where $f(t) = t\log t$. We will calculate this optimizer explicitly and show that the choice of $r(w,z_i)$ is actually tight. To this end, we firstly calculate the densities of P_W and $P_{W|Z_i}$ as: $dP_W = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}\exp(\frac{-(W-\mu)^2 n}{2\sigma^2})$, $dP_{W|Z_i} = \frac{n}{\sqrt{2\pi\sigma^2(n-1)}}\exp(-\frac{(W-\frac{n-1}{n}\mu-\frac{1}{n}Z_i)^2 n^2}{2\sigma^2(n-1)})$. Then we can calculate the optimizer as:

$$\begin{aligned}f'(dP_{WZ_i}/d(P_W\otimes\mu')) &= \log\frac{dP_{W|Z_i}}{dP_W} + \log\frac{d\mu}{d\mu'} + 1 \\ &= \underbrace{\frac{1}{2}\log\frac{n}{n-1} - \frac{(w-z_i)^2-(\mu-z_i)^2}{2\sigma^2} - \frac{(w-z_i)^2}{2\sigma^2(n-1)}}_{\log\frac{dP_{W|Z_i}}{dP_W}} \\ &\quad + \underbrace{\frac{(\mu'-z_i)^2-(\mu-z_i)^2}{2\sigma^2}}_{\log\frac{d\mu}{d\mu'}} + 1 \\ &= \frac{1}{2}\log\frac{n}{n-1} - \frac{(w-z_i)^2-(\mu'-z_i)^2}{2\sigma^2} - \frac{(w-z_i)^2}{2\sigma^2(n-1)} + 1\end{aligned}$$

for fixed w and z_i. The above function can be written as:

$$\begin{aligned}f'(dP_{W|Z_i}/dP_W) = &-\frac{r(w,z_i)}{2\sigma^2} - \frac{\ell(w,z_i)}{2\sigma^2(n-1)} \\ &+ \frac{1}{2}\log\frac{n}{n-1} + 1.\end{aligned}$$

The unexpected excess risk $r(w,z_i)$ clearly appears in the optimizer with some scaling factor and shifting constant (up to a $O(\frac{1}{n})$ difference), which, however, will not affect the convergence. To rigorously show this, we state the following result.

Lemma 1. *The choice of the function $-\frac{r(w,z_i)}{2\sigma^2}$ satisfies the following inequality:*

$$\begin{aligned}&\frac{n-1}{n}\left(I(W;Z_i) + D(\mu\|\mu')\right) \\ &\leq \mathbb{E}_{WZ_i}[-\frac{r(w,z_i)}{2\sigma^2}] - \log\mathbb{E}_{P_W\otimes\mu'}[e^{-\frac{r(w,z_i)}{2\sigma^2}}] \\ &\leq I(W;Z_i) + D(\mu\|\mu'),\end{aligned}$$

The provided lemma confirms that for the Gaussian example, the variational representation is actually tight. However, when using the loss function $\ell(w,z_i)$ without referencing w^*, the mutual information bound might not be tight as the equality in the variational representations may not be achieved (up to $O(\frac{1}{n} + D(\mu\|\mu'))$). Moreover, we posit that selecting $-r(w,z_i)$ also results in a tight upper bound on the generalization error. This is further supported by our demonstration of a corresponding lower bound for the generalization error in the following.

Lemma 2 (Matching Lower Bound)**.** *Consider the Gaussian mean estimation problem with $\beta = 0$. With a large n, the following inequality holds for ERM:*

$$\begin{aligned}&\mathbb{E}_{WS}[\mathrm{gen}(W,S)] \geq \\ &2\sigma^2\frac{n-1}{n^2}\sum_{i=1}^{n} I(W;Z_i) + D(\mu\|\mu').\end{aligned}$$

From our result, it could be seen that the sample-wise mutual information is present in both the upper and lower bounds. When considering the generalization error, the rates of convergence for both bounds align, albeit with varying leading constants. In the

context of the Gaussian mean example, the excess risk upper bound is precise, given that both the empirical excess risk and generalization error are of order $O\left(\frac{1}{n} + D(\mu\|\mu')\right)$*. Yet, for a broader range of learning problems, the lower bound for the excess risk primarily hinges on the data distributions and may not be calculated easily and explicitly.*

Moreover, the learning bound in Theorem 3 can be applied to the regularized ERM algorithm as:

$$w_{\mathsf{RERM}} = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\, \hat{L}_\alpha(w, S, S') + \frac{\lambda}{n} g(w),$$

where $g : \mathcal{W} \to \mathbb{R}$ denotes the regularizer function and λ is some penalizing coefficient. We define $\hat{R}_{\text{reg}}(w, S, S') = \hat{R}_\alpha(w, S, S') + \frac{\lambda}{n}(g(w) - g(w^*))$, then we have the following lemma.

Lemma 3. *We assume conditions in Theorem 3 hold for the regularized ERM and also assume* $|g(w_1) - g(w_2)| \le B$ *for any* w_1 *and* w_2 *in* $\mathcal{W}$ *with some* $B > 0$*. Then for* W_{RERM}*:*

$$\begin{aligned}
&\mathbb{E}_W\left[L_{\mu'}(W_{\mathsf{RERM}}) - L_{\mu'}(w^*)\right] \le \\
&\frac{1}{c}\mathbb{E}_{P_{WSS'}}\left[\hat{R}_{\text{reg}}\left(W_{\mathsf{RERM}}, S, S'\right)\right] + \frac{\lambda B}{cn} \\
&+ \frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n} I(W_{\mathsf{RERM}}; Z_i') \\
&+ \frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}}; Z_i) + D(\mu\|\mu')\right).
\end{aligned}$$

The proof of this result is given in Appendix H. As $\hat{R}_{\text{reg}}(w, S, S')$ will be negative for w_{RERM}, the regularized ERM algorithm can lead to the fast rate up to the domain divergence (by ignoring the multiplicative constant) if both $I(W_{\mathsf{RERM}}; Z_i')$ and $I(W_{\mathsf{RERM}}; Z_i)$ are of $O(1/n)$.

From Theorem 3, we can achieve the fast rate if the mutual information between the hypothesis and data example is converging up to the domain divergence with the rate $O(1/n)$, and one may ask whether we could arrive at an intermediate rate between $O(1/\sqrt{n})$ and $O(1/n)$. To further relax the (η, c)-central condition, we can also derive the intermediate rate with the order of $O(n^{-\alpha})$ for $\alpha \in [\frac{1}{2}, 1]$. Similar to the v-central condition, which is a weaker condition of the η-central condition [44, 45], we propose the (v, c)-central condition first and derive the intermediate rate results in Theorem 4.

Definition 2 ((v, c)-Central Condition)**.** *Let* $v : [0, \infty) \to [0, \infty)$ *is a bounded and non-decreasing function satisfying* $v(\epsilon) > 0$ *for all* $\epsilon > 0$*. We say that* $(\mu, \mu', \ell, \mathcal{W}, \mathcal{A})$ *satisfies the* (v, c)*-central condition if for all* $\epsilon \ge 0$*, it holds that*

$$\begin{aligned}
&\log \mathbb{E}_{P_W\otimes\mu'}\left[e^{-v(\epsilon)(\ell(W,Z)-\ell(w^*,Z))}\right] \le \\
&-cv(\epsilon)\mathbb{E}_{P_W\otimes\mu'}\left[\ell(W, Z) - \ell(w^*, Z)\right] + v(\epsilon)\epsilon. \quad (30)
\end{aligned}$$

Example 7 (Sub-Gaussian Condition)**.** *We assume that the* σ^2*-subgaussian holds under the target distribution* $P_W \otimes \mu'$ *for a certain* σ^2*. Then the learning tuple also satisfies* $(\min(v, c)$*-central condition where* $v(\epsilon) = \frac{2\epsilon}{\sigma^2}$ *and* $c = 1$*.*

Example 8 (Bernstein Condition)**.** *Let* $\gamma \in (0, 1]$ *and* $B \ge 1$*. We assume that the* ***Bernstein condition*** *holds under the target distribution* $P_W \otimes \mu'$ *for given* γ *and* B*. Additionally, if* $r(w, z_i)$ *is bounded by* $-b$ *with some* $b > 0$ *for all* w *and* z_i*, the learning tuple also satisfies* $(\min(v, c)$*-central condition where* $v(\epsilon) = \frac{\epsilon^{1-\gamma}}{2Bc(1-\gamma)^{1-\gamma}}$ *and* $c = \min\{\frac{1}{2}, \gamma\}$*.*

Then we can derive the following results for the intermediate rate result.

Theorem 4. *Assume the learning tuple* $(\mu, \mu', \ell, \mathcal{W}, \mathcal{A})$ *satisfies the* (v, c)*-central condition up to* ϵ *for some function* v *as defined in Def. 2 and* $0 < c < 1$*. Then it holds that for any* $\epsilon \ge 0$ *and any* $0 < \eta' \le v(\epsilon)$*,*

$$\begin{aligned}
&\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \le \frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_\alpha\left(W, S, S'\right)] \\
&+ \frac{\alpha}{c\beta n}\sum_{i=1}^{\beta n}\left(\frac{I(W; Z_i')}{\eta'} + \epsilon\right) \\
&+ \frac{1-\alpha}{c(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\frac{I(W; Z_i) + D(\mu\|\mu')}{\eta'} + \epsilon\right).
\end{aligned} \quad (31)$$

In particular, if $v(\epsilon) = \epsilon^{1-\gamma}$ for some $\gamma \in (0, 1]$, then the generalization error is bounded by,

$$\begin{aligned}
&\mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \le \\
&\frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_\alpha\left(W, S, S'\right)] + \frac{2\alpha}{c\beta n}\sum_{i=1}^{\beta n} I(W; Z_i')^{\frac{1}{2-\gamma}} \\
&+ \frac{2(1-\alpha)}{c(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W; Z_i) + D(\mu\|\mu')\right)^{\frac{1}{2-\gamma}}.
\end{aligned}$$

The proof can be found in Appendix I. Thus, the expected generalization is found to have an order of

$I(W;Z_i)^{\frac{1}{2-\gamma}}$, which corresponds to the typical results under Bernstein's condition [48, 49, 29].

IV. Applications and Extensions

A. Generalization error of stochastic noisy iterative algorithms

The upper bound obtained in the 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 is obtained 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 is 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 iterative algorithm:

$$W(t) = W(t-1) - \eta_t \nabla \hat{L}_\alpha(W(t-1), S, S') + n(t), \tag{32}$$

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 $n(t) \sim \mathcal{N}(0, \sigma_t^2 I_d)$.

To obtain a generalization error bound for the above algorithm aligning with the theorem derived, we make the following assumptions.

Assumption 1. *We assume the loss function $\ell(w,z)$ is r^2-subgaussian under the distribution μ' for any $w \in \mathcal{W}$.*

Assumption 2. *The gradient is bounded, e.g., $\|\nabla \ell(w(t), z_i)\|_2 \leq K_S$, for all $z_i \in S$, and $\|\nabla \ell(w(t), z_i)\|_2 \leq K_T$, for all $z_i \in S'$ with $K_S, K_T > 0$, $\forall t \geq 1$. Then it follows that $\left\| \nabla(\hat{L}_\alpha(w(t), S, S')) \right\|_2 \leq (1-\alpha)K_S + \alpha K_T \triangleq K_{ST}$.*

Remark 5. *The bounded gradient assumption is a common assumption made in many analyses of machine learning and optimization algorithms, particularly in the context of gradient descent and its variants [53, 54]. In simpler terms, it ensures that the function does not have any abrupt or infinitely steep changes, and this assumption can be easily satisfied in many learning setups. For instance, functions like the absolute loss $f(x) = |x|$ and the linear loss $f(x) = mx + c$ are inherently Lipschitz continuous, thus adhering to the bounded gradient criteria. Similarly, polynomials of limited degrees, such as the quadratic function $f(x) = x^2$, satisfy this condition within a closed interval $x \in [a, b]$. Furthermore, the sigmoid and hyperbolic tangent functions, two popular activation functions used in neural networks, also comply with this assumption, having bounded constants of $K = 0.25$ and $K = 1$, respectively.*

Now we will apply the bound in Corollary 2 by further characterizing the mutual information $I(W;Z_i)$ with the relevant optimization parameters.

Theorem 5 (Generalization error of stochastic noisy iterative algorithm)**.** *Suppose that Assumptions 1 and 2 hold and $W(T)$ is obtained from (32) at Tth iteration. Then the generalization error is upper bounded by*

$$\mathbb{E}_{WSS'}\left[\mathrm{gen}\left(W(T), S, S'\right)\right] \leq \alpha\sqrt{\frac{2r^2}{\beta n}\hat{I}(S)} + (1-\alpha)\sqrt{2r^2\left(\frac{\hat{I}(S)}{(1-\beta)n} + D\left(\mu\|\mu'\right)\right)}, \tag{33}$$

where we define

$$\hat{I}(S) := \frac{d}{2}\sum_{t=1}^{T}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right) - \sum_{t=1}^{T} h(n_t). \tag{34}$$

The proof is given in Appendix J. 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.

Remark 6. *For fixed learning parameters $T, \sigma_t^2, K_{ST}, \eta_t$ not depending on the sample size n, if we increase the sample size n, the bound decays up to $(1-\alpha)\sqrt{2r^2 D(\mu\|\mu')}$. This shows that as the sample size increases, the mutual information diminishes as each individual instance*

has less influence on the gradient descent updates. Consequently, the decreased $\hat{I}(S)$ will result in a reduced generalization error.

Remark 7. *If we fix the sample size instead but increase the iteration number T, the bounds would go to infinity with a fast-growing rate if η_t and σ_t^2 are not wisely chosen. To determine proper choices of the optimization parameters to achieve a low generalization error, we consider the case $d = 1$ and $h(n_t) = d\log 2\pi e\sigma_t^2$ where $n_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$ for simplicity, and we can then further upper bound $\hat{I}(S)$ as:*

$$\hat{I}(S) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\frac{\eta_t^2 K_{ST}^2}{\sigma_t^2}\right) \leq \frac{K_{ST}^2}{2}\sum_{t=1}^{T}\frac{\eta_t^2}{\sigma_t^2}$$

with the inequality that $\log(1+x) \leq x$ for $x \geq 0$. To have better control of the mutual information, we may need to control the rate of the summation of $\frac{\eta_t^2}{\sigma_t^2}$. One typical choice of the noisy stochastic gradient descent is $\eta_t = \frac{1}{t}$ and the noise is set as $\sigma_t = \sqrt{\eta_t}$, then $\hat{I}(S) \leq O(\log(T))$, which scales logarithmically and coincides with the results in [55, 56]. Recently, in [57], it is shown that the mutual information bounds fail to give vanishing bounds w.r.t. T for both generalization error and excess risk (see (2) and (3) for example). Regarding this, while the mutual information bounds may not provide optimal bounds, we still include these results to showcase the practicality of mutual information bounds in analyzing the noisy gradient descent by knowing the optimization parameters. However, it should be noted that more refined techniques with surrogate algorithms may be required for a tighter characterization. From the ratio, it could also be seen that the mutual information is controlled by the step size η and the noise variance σ_t^2. Given a larger step size η, the data sample will have a greater influence on the hypothesis. On the contrary, if the noise variance is dominating, altering the step size might have less impact on the hypothesis.

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 Proposition 3 in [58] with the assumption of a strongly convex loss function, which guarantees the convergence of the hypothesis. We now provide an excess risk upper bound when the loss function is strongly convex. Recall that the excess risk of $W(T)$ is defined as

$$R_{\mu'}\left(W(T)\right) = L_{\mu'}\left(W(T)\right) - L_{\mu'}\left(w^*\right). \tag{35}$$

Following the result in Theorem 2, we present the upper bound for the excess risk if the following two assumptions hold.

Assumption 3. *$\ell(w,z)$ is ν-strongly convex, namely*

$$\ell(w_1,z) \geq \ell(w_2,z) + \nabla\ell(w_2)(w_1-w_2) + \frac{\nu}{2}\|w_1-w_2\|^2 \tag{36}$$

for some $\nu > 0$ and any $w_1, w_2 \in \mathcal{W}$.

Remark 8. *The Assumption 3 is a fundamental condition in the context of optimization. For algorithms like gradient descent, strong convexity can guarantee faster convergence rates and can be introduced through regularization techniques like L_2 regularization, which can prevent overfitting in machine learning models [59, 60]. For example, in the context of linear regression, the loss function is given by the mean squared error. When we add an L_2 regularization term (i.e., ridge regression), the loss function satisfies the strongly convex properties. Similarly, the cross-entropy loss in the logistic regression with an added L_2 regularization could also satisfy the strong convex condition[61]. Another type of loss - the exponential loss used in the boosting algorithm - is also strongly convex, especially in the context of AdaBoost [62].*

Assumption 4. *The loss function $\ell(w,z)$ has $\mathcal{L}$-Lipschitz-continuous gradient such that*

$$|\nabla\ell(w_1,z) - \nabla\ell(w_2,z)| \leq \mathcal{L}|w_1 - w_2| \tag{37}$$

for any $w_1, w_2 \in \mathcal{W}$ with respect to any $z \in \mathcal{Z}$.

Remark 9. *The Lipschitz-continuous gradient condition (sometimes referring to $\mathcal{L}$-smooth condition) ensures that the gradient does not change too abruptly and has been widely applied in many optimization problems, particularly in the context of stochastic gradient descent (SGD) and its variants [63, 64]. This condition can be used to bound the updates, ensuring that the gradient change does not become too large and destabilize the learning process. Some typical examples include the mean squared loss, the Huber loss [65], and the log-cosh loss [66]. However, it is important to highlight that not all loss functions in machine learning exhibit smoothness. For instance, the least absolute loss and hinge loss are examples of non-smooth loss functions.*

Corollary 3 (Excess risk of strongly convex loss function)**.** *Suppose Theorem 5 holds and the loss function $\ell(w,z)$ satisfies Assumptions 3 and 4. Define $\kappa = \frac{\nu}{\mathcal{L}}$, setting $\eta = \frac{1}{\mathcal{L}}$, and W is arbitrarily initialized with $W(0)$. Then the excess risk can be bounded as follows:*

$$\mathbb{E}_W\left[R_{\mu'}\left(W(T)\right)\right] \leq (1-\alpha)d_{\mathcal{W}}(\mu,\mu') + \alpha\sqrt{\frac{2r^2}{\beta n}\hat{I}(S)}$$
$$+ (1-\alpha)\sqrt{2r^2\left(\frac{\hat{I}(S)}{(1-\beta)n} + D\left(\mu\|\mu'\right)\right)}$$
$$+ K_{ST}(1-\kappa)^T\mathbb{E}[\|W(0) - W_{\mathsf{ERM}}\|]$$
$$+ K_{ST}\sum_{t=1}^{T}(1-\kappa)^{T-t}\mathbb{E}\left[\|n(t)\|\right]. \tag{38}$$

Remark 10. *From the above bound, we could also optimize κ and η for a tighter bound. Let us consider the same setup that $n(t)$ is Gaussian distributed and $d = 1$. First, we set $\eta_t = \frac{1}{t}$ and $\sigma_t = \sqrt{\eta_t}$ to make the generalization error tight. Now we assume that the hypothesis space is bounded, i.e., $\|w_1 - w_2\| \leq W_B$ for any $w_1, w_2 \in \mathcal{W}$. Then the third term in R.H.S. will decay exponentially when T goes to infinity, which means the initialization does not affect the generalization error with a large number of iterations. While the fourth term in the R.H.S. can be upper bounded by:*

$$K_{ST}\sum_{t=1}^{T}(1-\kappa)^{T-t}\mathbb{E}\left[\|n(t)\|\right] = \frac{K_{ST}}{\sqrt{2\pi}}\sum_{t=1}^{T}\frac{(1-\kappa)^{T-t}}{\sqrt{t}}$$
$$\leq \frac{K_{ST}}{\sqrt{2\pi}}\frac{1}{1-(1-\kappa)}.$$

Then the final bound would have the form of

$$\mathbb{E}_W\left[R_{\mu'}\right] \leq (1-\alpha)d_{\mathcal{W}}(\mu,\mu')$$
$$+ O\left(\alpha\sqrt{\frac{\log(T)}{\beta n}} + (1-\alpha)\sqrt{\frac{\log(T)}{(1-\beta)n} + cD(\mu\|\mu')}\right.$$
$$\left. + (1-\kappa)^T + \frac{K_{ST}}{\sqrt{2\pi\kappa}}\right),$$

where c is some leading constant. From the bound, we can see that selecting κ as a constant would not affect the rate of the bound, but the choice of η could be crucial as it controls the variance level of the noise.

The proof is provided in Appendix K. Notice that the summation of the terms $\|n(t)\|$ needs to be finite. Hence this upper bound is effective when $n(t)$ is sampled from a bounded random variable (for example, truncated Gaussian or uniform random variable), but not for the case $n(t)$ is Gaussian distributed, where $\|n(t)\|$ is not bounded. We give a toy example in Section VI-A for the Bernoulli transfer to show the effectiveness of the bounds.

B. Generalization error on Gibbs algorithm

Theorem 2 shows that the generalization error can be upper-bounded in terms of the mutual information between the input data and output hypothesis and the KL divergence between the source and target domains. Since the KL divergence $D(\mu\|\mu')$ is usually uncontrollable as μ and μ' are unknown in real settings, it is natural to consider an algorithm that minimizes the empirical risk regularized by $I(W,Z_i)$ as

$$P^*_{W|S,S'} = \operatorname*{argmin}_{P_{W|S,S'}}\left(\mathbb{E}_{WSS'}[\hat{L}_\alpha(W,S,S')] + \frac{1}{k}\sum_{i=1}^{n}I(W;Z_i)\right). \tag{39}$$

With the chain rule, we have

$$\sum_{i=1}^{\beta n}I(W;Z'_i) + \sum_{i=\beta n+1}^{n}I(W;Z_i) \leq \sum_{i=1}^{\beta n}I(W;Z'_i|(Z')^{i-1})$$
$$+ \sum_{i=\beta n+1}^{n}I(W;Z_i|S',Z_{\beta n}^{i-1}) = I(W;S,S'), \tag{40}$$

with the definition $(Z')^{i-1} = \{Z'_1, Z'_2, \cdots, Z'_{i-1}\}$ and $Z_{\beta n}^{i-1} = \{Z_{\beta n+1}, \cdots, Z_{i-1}\}$. Then we aim to minimize the relaxed Gibbs algorithm as

$$P^*_{W|S,S'} = \operatorname*{argmin}_{P_{W|S,S'}}\left(\mathbb{E}_{WSS'}[\hat{L}_\alpha(W,S,S')] + \frac{1}{k}I(W;S,S')\right). \tag{41}$$

We can also relax the above optimization problem by replacing $I(W;S,S')$ with an upper bound

$$D\left(P_{W|S,S'}\|Q|P_{S,S'}\right) = I(W;S,S') + D\left(P_W\|Q\right)$$

where Q is an arbitrary distribution on W. We can also rewrite Q as

$$D\left(P_{W|S,S'}\|Q|P_{S,S'}\right) =$$
$$\int_{\mathcal{Z}^n}D\left(P_{W|S=s,S'=s'}\|Q\right)d^{\otimes\beta n}(\mu')d^{\otimes(1-\beta)n}(\mu),$$

which does not depend on source distribution μ or target distribution μ'. Thus we can relax (41) and arrive at the following surrogate solution as

$$P^*_{W|S,S'} = \underset{P_{W|S,S'}}{\operatorname{argmin}} \Big(\mathbb{E}_{WSS'}[\hat{L}_\alpha(W,S,S')] + \frac{1}{k} D\left(P_{W|S,S'} \| Q | P_{S,S'}\right) \Big). \tag{42}$$

Theorem 6. *The solution to the optimization problem (42) is the Gibbs algorithm, which satisfies*

$$P^*_{W|S',S}(\mathrm{d}w) = \frac{e^{-k\hat{L}_\alpha(W,S,S')} Q(\mathrm{d}w)}{\mathbb{E}_Q\left[e^{-k\hat{L}_\alpha(W,S,S')}\right]} \tag{43}$$

for each $(S,S') \in \mathcal{Z}^n$.

The proof can be found in Appendix L. For fixed α, we denote the hypothesis that achieves the combined minimum population risk among $\mathcal{W}$ by $w^*_{st}(\alpha)$ such that

$$w^*_{st}(\alpha) = \underset{w\in\mathcal{W}}{\operatorname{argmin}}\, \alpha L_{\mu'}(w) + (1-\alpha) L_\mu(w). \tag{44}$$

In particular, we define $w^*_s = w^*_{st}(0)$ and we also have that $w^* = w^*_{st}(1)$. We further denote w_G as the output hypothesis of the Gibbs algorithm. We have

$$\begin{aligned}
&\mathbb{E}_W[L_{\mu'}(W_G)] = \mathbb{E}_{WSS'}[L_{\mu'}(W_G) - \hat{L}_\alpha(W_G,S,S')] \\
&+ \mathbb{E}_{WSS'}[\hat{L}_\alpha(W_G,S,S')] \\
&\le \mathbb{E}_{WSS'}[\mathrm{gen}(W_G,S,S')] + \mathbb{E}_{WSS'}[\hat{L}_\alpha(W_G,S,S')] \\
&+ \frac{1}{k} D(P^*_{W_G|S,S'} \| Q | S,S').
\end{aligned} \tag{45}$$

As a direct application of Corollary 2, we then specialize the upper bound on the population risk for the Gibbs algorithm by further upper bounding the generalization error on the R.H.S. of (45). For any $\ell \in [0,1]$, we reach the following corollary for countable hypothesis space.

Corollary 4. *Suppose $\mathcal{W}$ is countable. Let W_G denote the output of the Gibbs algorithm applied on dataset S,S'. For some $\alpha \in [0,1]$ and $\ell(w,z) \in [0,1]$ for any w and z, the generalization error is upper bounded by:*

$$\begin{aligned}
\left|\mathbb{E}_{WSS'}\left[\mathrm{gen}(W_G,S,S')\right]\right| &\le \frac{\alpha^2 k}{4\beta n} + \frac{(1-\alpha)^2 k}{4(1-\beta)n} \\
&+ (1-\alpha)\sqrt{\frac{D(\mu\|\mu')}{2}},
\end{aligned} \tag{46}$$

and the population risk of the Gibbs algorithm satisfies:

$$\begin{aligned}
\mathbb{E}_W\left[L_{\mu'}(W_G)\right] &\le L_\alpha(w^*_{st}(\alpha)) + \frac{1}{k}\log\frac{1}{Q\left(w^*_{st}(\alpha)\right)} \\
&+ \frac{\alpha^2 k}{4\beta n} + \frac{(1-\alpha)^2 k}{4(1-\beta)n} + (1-\alpha)\sqrt{\frac{D(\mu\|\mu')}{2}}.
\end{aligned} \tag{47}$$

Remark 11. *From (46), we can see that the generalization error bound converges up to the domain divergence with $O(\frac{1}{n})$, which coincides with the rate in [30] for the exact generalization error characterization of the $\alpha-$weighted ERM algorithm. The only difference lies in the definition of the generalization error: we define the generalization error as the gap between the combination of the empirical risks in both the source and target domains and the population risk in the target domain. While in [30], the generalization error is defined as the gap between the empirical risk in the target domain only and its population risk. Thus, there is an additional domain divergence term in our upper bound. Apart from that, in both cases, the Gibbs algorithm is able to achieve a fast convergence rate.*

If we set $\alpha = 1$, $\beta = 1$ (only use the target data), we will retrieve the results for conventional machine learning, derived in [18, Corollary 2] as follows

$$\mathbb{E}_W\left[L_{\mu'}(W_G)\right] \le L_{\mu'}(w^*) + \frac{1}{k}\log\frac{1}{Q\left(w^*\right)} + \frac{k}{4n}. \tag{48}$$

We consider the special case $\beta = 0$, $\alpha = 0$ (only using the source data), then we have

$$\begin{aligned}
\mathbb{E}_W\left[L_{\mu'}(W_G)\right] \le& L_\mu(w^*_s) + \frac{1}{k}\log\frac{1}{Q\left(w^*_s\right)} \\
&+ \frac{k}{4n} + \sqrt{\frac{D(\mu\|\mu')}{2}}.
\end{aligned} \tag{49}$$

In this case, it is observed that $D(\mu\|\mu')$ would not vanish. Even when n increases, the excess risk also depends on the population risk of w^*_s w.r.t. the source distribution μ.

When $\mathcal{W}$ is uncountable (e.g., $\mathcal{W} = \mathbb{R}^d$), the term $\frac{1}{k} D(P_{W_G|S,S'} \| Q | S,S')$ will be bounded using the approximation by a Gaussian distribution for W_G given S and S'. As a result, we could bound the population risk using the following corollary.

Corollary 5. *Suppose $\mathcal{W} = \mathbb{R}^d$ and $\ell(\cdot,z)$ is ρ-Lipschitz for all $z \in \mathcal{Z}$. Let W_G denote the output of the Gibbs algorithm induced by datasets S and S'. For $\ell \in [0,1]$ and some $\alpha \in [0,1]$, the population risk of W_G satisfies*

$$\mathbb{E}_W\left[L_{\mu'}(W_G)\right] \le L_\alpha(w^*_{st}(\alpha)) + \frac{\alpha^2 k}{4\beta n} + \frac{(1-\alpha)^2 k}{4(1-\beta)n}$$

$$+(1-\alpha)\sqrt{\frac{D(\mu\|\mu')}{2}}$$
$$+\inf_{a>0}\left(a\rho\sqrt{d}+\frac{1}{k}D\left(\mathcal{N}\left(w_{st}^*(\alpha),a^2\mathbf{I}_d\right)\|Q\right)\right). \tag{50}$$

V. Bounding with other divergences

The usage of KL divergence to quantify distribution shifts has been popular in many previous works [67, 68, 69]. However, an important observation made by Hanneke and Kpotufe [70] demonstrates certain limitations of this approach - specifically in the context of transfer learning - where in Theorem 2, the result is not effective for a class of supervised machine learning problems if μ is not absolutely continuous with respect to μ' as the KL divergence goes to infinity and the bound becomes vacuous. In Example 1 of [70], the authors also illustrate why KL divergence in the hypothesis space is not the right measure for this purpose, and the issue is wisely bypassed by changing the hypothesis distribution from $P(w)$ to a Bernoulli distribution $P(w=w^*)$. The transfer component - the divergence proposed in their work - can easily handle this case, providing insights into the data values in transfer learning. Now we give two examples.

Example 9. *Let the sample Z be a pair (X,Y) where X denotes features and Y denotes the corresponding label, and we assume that Y is determined by X, i.e., $Y=f(X)$ for some deterministic function f. In this case, the distribution μ of Z can be factored as $\mu(z)=\mu(x,y)=P_X(x)P_{Y|X}(y|x)=P_X(x)\mathbf{1}_{y=f(x)}$ where P_X is the distribution of X and $\mathbf{1}$ denotes the indicator function. Let target distribution μ' factor as $\mu'(x,y)=P'_X(x)\mathbf{1}_{y=f'(x)}$ for some distribution P'_X and function f'. Notice that in this case, μ is not absolutely continuous with respect to μ' unless $f=f'$. Indeed, for some (x,y) we have $\mathbf{1}_{y=f'(x)}=0$ while $\mathbf{1}_{y=f(x)}=1$ unless f agrees with f' almost everywhere. Therefore, the KL divergence is $D(\mu\|\mu')=\infty$, and the upper bound becomes vacuous unless $f=f'$ (however P_X and P'_X could still be different, hence the problem is not necessarily trivial in this case).*

Example 10. *Let the sample Z be a discrete random variable over the set $\{1,2,3\}$. Assume that for the source domain $\mu(Z=1)=\mu(Z=2)=\mu(Z=3)=\frac{1}{3}$ while $\mu'(Z=1)=\mu'(Z=2)=\frac{1}{2}$ and $\mu'(Z=3)=0$. In this case, the KL divergence $D(\mu\|\mu')=\infty$ as μ is not absolutely continuous w.r.t. μ' when $Z=3$. However, if we consider the total variation [71] between μ and μ' (where the detailed definition is given in the later context), it can be calculated that $TV(\mu,\mu')=\sum_{z\in\mathcal{Z}}\frac{1}{2}|\mu(z)-\mu'(z)|=\frac{1}{3}$, which is finite instead.*

We mitigate these issues tied to KL divergence by introducing bounds using other divergences, such as the Wasserstein distance and ϕ-divergence, as we show in Example 10. These alternative measures provide us with a more general and robust framework to quantify the distribution shift, which may also give a tighter characterization.

A. ϕ-divergence bounds

To develop an appropriate upper bound to handle the case where the KL divergence is infinite, we may extend the results by using other types of divergence following the work of [72], which do not impose the absolute continuity restriction. To this end, we first introduce a more general divergence between two distributions that can handle such a case, namely, the ϕ-divergence.

Definition 3 (ϕ-divergence)**.** *Given two measures μ, ν and a convex functional ϕ, we define the ϕ divergence by:*

$$D_\phi(\nu\|\mu)=\mathbb{E}_\mu\left[\phi(\frac{d\nu}{d\mu})\right] \tag{51}$$

where $d\nu/d\mu$ is the Radon-Nikodym derivative.

To tackle the absolute continuity issue between μ and μ', we shall choose $\phi(x)=\frac{1}{2}|x-1|$ and arrive at the bounds with the total variation distance, which is always bounded by $[0,1]$. Following [72], we suppose that the loss function $\ell(w,z)$ is L_∞-norm upper bounded by σ where the L_∞-norm of a random variable is defined as

$$\|X\|_\infty=\inf\{M:P(X>M)=0\}.$$

Then we have the following corollary.

Corollary 6. *(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 μ'. Then the following inequality holds.*

$$\mathbb{E}_{WSS'}\left[\text{gen}(W_{\mathsf{ERM}},S,S')\right]\leq\frac{2\alpha\|\sigma\|_\infty}{\beta n}\sum_{i=1}^{\beta n}I_\phi(W_{\mathsf{ERM}};Z'_i)$$

$$+\frac{2(1-\alpha)\|\sigma\|_\infty}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I_\phi(W_{\mathsf{ERM}};Z_i)+TV(\mu,\mu')\right), \tag{52}$$

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* $P_{W_{\mathsf{ERM}}}\otimes P_{Z_i}$ *(similar to* Z_i'*) with* $D_\phi(P||Q)=\frac{1}{2}\int|dP-dQ|$ *and* $TV(\mu,\mu')=D_\phi(\mu||\mu')$ *denotes the total variation distance between the distribution* μ *and* μ'*.*

The proof can be found in Appendix N. Note that with this bound, the divergence $TV(\mu\|\mu')$ is always bounded by $[0,1]$ for any μ and μ'. More generally, the generalization error can be upper bounded using different ϕ-divergences, and the key to unifying different divergences is the Legendre-Fenchel duality that we used in Theorem 1. For example, by choosing $\phi(x)=x\log x$ and the function ψ for bounding the moment generating function, we will end up with the KL-divergence-based bound. Likewise, if we choose $\phi(x)=\frac{x^2}{2}$ and the function ψ for bounding the variance, we will arrive at the χ^2-divergence based bound as shown in [73], such an extension allows the result to hold for a wider family of distributions (e.g., the sub-exponential random variables) where the mutual information bound may be invalid.

B. Generalization error with Wasserstein distance

Another interesting metric, called the Wasserstein distance, has a very close connection with the KL divergence via the transportation-cost inequality [74]. Such a distance has several advantages over the KL divergence. As we show later, the Wasserstein distance can handle the deterministic algorithm where the mutual information is infinite. Furthermore, under mild conditions, the Wasserstein distance-based bound is naturally tighter for transfer learning, compared to the mutual information bound as shown in Corollary 2. To show our results, we first give some definitions of the probability measures. Let $(\mathcal{Z},d)$ be a metric space and $p\in[1,+\infty)$, we define $\mathcal{P}_p(\mathcal{Z})$ as the set of probability measures μ on $\mathcal{Z}$ satisfying $(\mathbb{E}_{Z\sim\mu}[d(Z,z_0)^p])^{\frac{1}{p}}<\infty$ for some $z_0\in\mathcal{Z}$. Then we define the p-Wasserstein distance as follows.

Definition 4 (Wasserstein distance)**.** *Assume* $\mu,\nu\in\mathcal{P}_p(\mathcal{Z})$*. The* p*-Wasserstein distance between* μ *and* ν *is defined as*

$$\mathbb{W}_p(\mu,\nu)=\inf_{\pi\in\Pi(\mu,\nu)}\left(\mathbb{E}_\pi\left[d(Z,Z')^p\right]\right)^{1/p}. \tag{53}$$

where $\Pi(\mu,\nu)$ *denotes the coupling of* μ *and* ν*, i.e., the set of all the joint probability measures* $\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z})$ *with marginals equal to* μ *and* ν*.*

With the definition in place, we give the generalization error bound of the ERM algorithm using the Wasserstein distance in the following theorem.

Theorem 7 (Generalization error bound of ERM with Wasserstein distance)**.** *Let* P_W *be the marginal distribution induced by* S,S' *and* $P_{W|SS'}$ *with the ERM algorithm. Assume that for any* $w\in\mathcal{W}$*, the loss function* $\ell(w,Z)$ *is* $\mathcal{L}$*-Lipschitz for any* $W\in\mathcal{W}$*,* $Z\in\mathcal{Z}$*. Then the following inequality holds.*

$$\mathbb{E}_W\left[\mathrm{gen}(W_{\mathsf{ERM}},S,S')\right]\le\frac{\alpha\mathcal{L}}{\beta n}\sum_{i=1}^{\beta n}\mathbb{E}_{\mu'}[\mathbb{W}_1(P_W,P_{W|Z_i'})]$$
$$+\frac{(1-\alpha)\mathcal{L}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\mathbb{E}_\mu[\mathbb{W}_1(P_W,P_{W|Z_i})]+\mathbb{W}_1(\mu,\mu')\right). \tag{54}$$

Remark 12. *Under the Wasserstein distance, the domain divergence is captured by the first order Wasserstein distance* $\mathbb{W}_1(\mu,\mu')$*, which also resolves the absolutely continuous issue in mutual information bound. This bound requires that the loss function is* $\mathcal{L}$*-Lipschitz with respect to any hypothesis* W *and data instance* Z *while the mutual information bound is derived under the subgaussian assumption. It is also worth noting that the bound is based on the term* $\mathbb{W}_1(P_W,P_{W|Z_i})$*. Intuitively, this term measures how distribution diverges with a given single instance* Z_i*. In other words, it measures how one instance can affect the distribution of* W *given a specific algorithm. This intuition is also related to stability in the algorithmic perspective, similar to* $I(W,Z_i)$ *in the mutual information bound.*

The proof can be found in Appendix O. Based on the result of the generalization error, we can derive the excess risk upper bound using the Wasserstein distance.

Theorem 8 (Excess risk bound of ERM with Wasserstein distance)**.** *Assume the conditions in Theorem 7 hold and assume the loss function* $\ell(w,z)$ *is bounded by* $[0,1]$ *for any* w *and* z*. Then the following inequality holds:*

$$\mathbb{E}_W\left[R_{\mu'}(W_{\mathsf{ERM}})\right]\le(1-\alpha)d_{\mathcal{W}}(\mu,\mu')$$
$$+\frac{(1-\alpha)\mathcal{L}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\mathbb{E}_\mu[\mathbb{W}_1(P_W,P_{W|z_i})]+\mathbb{W}_1(\mu,\mu')\right)$$

$$+\frac{\alpha\mathcal{L}}{\beta n}\sum_{i=1}^{\beta n}\mathbb{E}_{\mu'}[\mathbb{W}_1(P_W,P_{W|z_i'})]. \tag{55}$$

We can show that this bound is tighter than the mutual information bound under mild conditions. Specifically, we consider the case that only the source domain is available ($\alpha=\beta=0$) with some n. From Theorem 2, we can easily bound the **expected excess risk** for ERM algorithm and we denote the bound by $\mathbb{B}_{\text{Info}}$, which is defined as:

$$\mathbb{B}_{\text{Info}} := \frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))} + d_{\mathcal{W}}(\mu,\mu'). \tag{56}$$

From Theorem 8, we can also derive the bound for the expected excess risk with the ERM algorithm, and we denote the Wasserstein type bound by $\mathbb{B}_{\text{Wass}}$ as

$$\mathbb{B}_{\text{Wass}} := \frac{\mathcal{L}}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\mu}[\mathbb{W}_1(P_W,P_{W|z_i})]+\mathbb{W}_1(\mu,\mu')\right) + d_{\mathcal{W}}(\mu,\mu'). \tag{57}$$

Next, we will prove that the $\mathbb{B}_{\text{Wass}}$ is, in general, tighter than $\mathbb{B}_{\text{Info}}$ under the mild assumption. To this end, we first introduce the transportation cost inequality (TCI) following the definition 3.4.2 from [74].

Definition 5 (Transportation Cost Inequality)**.** *We say that a probability measure μ on $(\mathcal{X},d)$ satisfies an L^p transportation cost inequality with constant $c>0$, or a $\mathrm{T}_p(c)$ inequality for short, if for every probability measure $\nu\ll\mu$ we have*

$$\mathbb{W}_p(\nu,\mu)\le\sqrt{2cD(\nu\|\mu)}. \tag{58}$$

This is a typical definition in many learning setups and we will give several examples below.

Example 11. *Under the Hamming distance, it can be proved that the first-order Wasserstein distance is equivalent to the total variation [74], then a well-known application of this inequality is the Pinsker's inequality where $p=1$ and $c=\frac{1}{4}$. This inequality provides an upper bound on the Wasserstein distance in terms of the divergence.*

Example 12. *It is proved that the inequality (58) holds if and only if for some c and every 1-Lipschitz function f, it satisfies the subgaussian property under μ, e.g., for any $t\in\mathbb{R}$,*

$$\mathbb{E}_{\mu}[e^{tf}]\le e^{\frac{ct^2}{2}}.$$

Readers can refer to Theorem 1 in [75] and Theorem $\diamond$ in [73] for more details.

From example 12, it could be shown that if we set $\nu=P_{WZ_i}$ and $\mu=P_W\otimes P_{Z_i}$ in (58), the Wasserstein distance between ν and μ will be tighter than the mutual information measure. We provide a rigorous argument in the following proposition.

Proposition 1. *Consider the case where $\beta=0$ for simplicity (e.g., with the source data only), let P_W be the marginal distribution induced by some algorithm and the source sample distribution $\mu^{\otimes n}$. We assume that the induced conditional distribution $P_{W|z_i}$ is absolutely continuous w.r.t. P_W for any $z_i\in\mathcal{Z}$, and both P_W and μ' satisfies the $T_1(\frac{r^2}{2\mathcal{L}^2})$ transportation cost inequality. Then the following inequality holds:*

$$\mathbb{B}_{\text{Wass}}\le\mathbb{B}_{\text{Info}}.$$

The proof can be found in Appendix P. Essentially, if both μ' and the resulting hypothesis P_W exhibit subgaussian characteristics, which include having light tails and preserving strong concentration for any Lipschitz functions, then the Wasserstein distance could be a better metric than the mutual information when evaluating the generalization error.

In this section, we demonstrate how various distribution metrics, including the total variation, ϕ-divergence, and the Wasserstein distance, can be employed to derive different information-theoretic bounds. Metrics like the Wasserstein distance are particularly noteworthy as they not only address the absolute continuity concerns but have also proven to provide tighter bounds than those based on mutual information. However, a caveat is that evaluating these bounds requires the knowledge of data and hypothesis distributions, and estimating these distributions can be very challenging in practical scenarios. As an initial approach to this concern, the subsequent section introduces a heuristic algorithm that capitalizes on the properties and insights drawn from the aforementioned bounds.

VI. Examples and Algorithms

In this section, we provide three examples to illustrate the upper bounds we obtained in previous sections. First we provide an example of calculating the learning bounds on the Bernoulli transfer problem given the optimization parameters with the stochastic gradient descent algorithm. Then we present the

logistic regression transfer learning problem with the mutual information bounds, and lastly we evaluate the proposed InfoBoost algorithm in several real-world transfer learning scenarios.

A. Transfer with stochastic gradient descent

In this example, we assume that data samples $(Z'_1, \ldots, Z'_{\beta n}) \sim Ber(p)$ and $(Z_{\beta n+1}, \ldots, Z_n) \sim Ber(p')$ where $Ber(p)$ denotes the Bernoulli distribution with probability p. For any $z_i \in (S, S')$, the loss function is defined as binary cross-entropy

$$\ell(w, z_i) = -(z_i \log(w) + (1 - z_i) \log(1 - w)).$$

In this case, there is no closed-form solution to the ERM algorithm so we apply the algorithm in (32) to obtain an hypothesis $W(T)$ where we choose $n(t) \sim N(0, \sigma_t^2)$.

The numerical results are shown in Figure 1, where the calculation for the upper bounds and detailed experimental setups can be found in Appendix Q. In the first row, we compare different number of samples while in second row we compare the bounds when β varies, and the last row investigates the effect of initialization value $W(0)$.

From the results, it is obvious that both the excess risk and generalization error are upper bounded by our developed upper bounds. We notice that that the tightness of the bound varies for different parameters of the algorithm. For example, if the initial value $W(0)$ is close to W_{ERM}, the upper bounds are tighter in this case. Also one observes that our bounds are in general becomes tighter if the number of samples n increases. The results confirms that the bounds captures the dependence of the input data and output hypothesis, as well as the stochasticity of the learning algorithm.

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 2. 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

$$\begin{aligned}\ell(w, z_i) := &- (y_i \log(\sigma(w^T x_i)) \\ &+ (1 - y_i) \log(1 - \sigma(w^T x_i))),\end{aligned}$$

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

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 [7, 41]. We give the empirical estimation for r^2 within the according hypothesis space such that

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

To evaluate the mutual information $I(W_{\mathsf{ERM}}, Z_i)$ efficiently, we follow the work [76] by repeatedly generating W_{ERM} and Z_i. As $\mu \ll \mu'$, we decompose $D(\mu(X, Y) \| \mu'(X', Y')) = D(\mu(X) \| \mu'(X)) + D(\mu(Y|X) \| \mu'(Y|X)|X)$ 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.

Theorem 9. *(Generalization error of ERM with Rademacher complexity) [41, Theorem 6.2] Assume that for any $w \in \mathcal{W}$, the loss function $\ell(w, Z)$ is bounded between $[a, b]$ for any w and z. Then for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ (over the randomness of samples and the learning algorithm)*

$$\begin{aligned}\mathrm{gen}(w_{\mathsf{ERM}}) \leq\ & (1 - \alpha) d_{\mathcal{W}}(\mu, \mu') + 2\alpha \mathbb{E}_{\sigma \otimes \mu} \left[\sup_{w \in \mathcal{W}} \sigma \ell(w, Z) \right] \\ & + \frac{2(1 - \alpha)}{\beta n} \mathbb{E}_{\sigma} \left[\sup_{w \in \mathcal{W}} \sum_{i=1}^{\beta n} \sigma_i \ell(w, z_i) \right] + 3\alpha \sqrt{\frac{(b - a) \ln(4/\delta)}{2\beta n}} \\ & + (1 - \alpha) \sqrt{\frac{(b - a)^2}{2} \ln(\frac{2}{\delta}) \left(\frac{\alpha^2}{\beta n} + \frac{(1 - \alpha)^2}{(1 - \beta) n} \right)},\end{aligned}$$

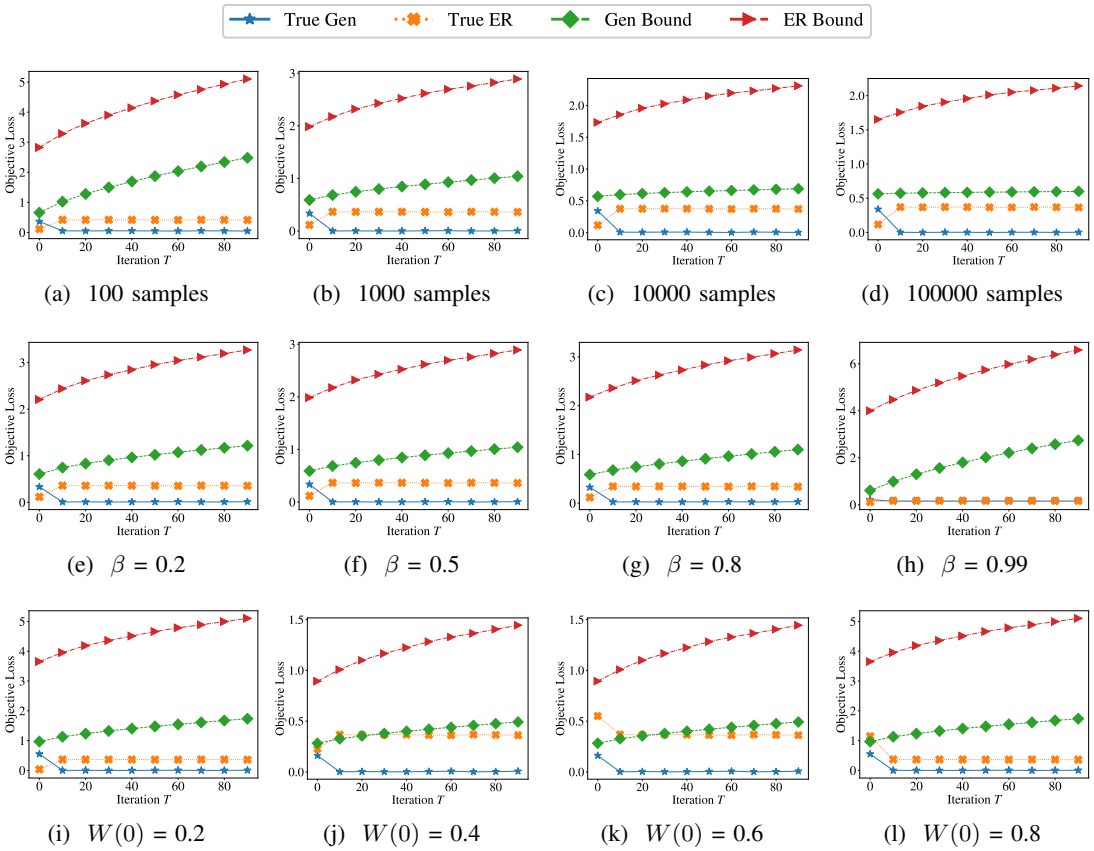


Fig. 1. Comparisons for testing results of true generalization error(blue), generalization error bound (green), true excess risk (orange) and excess risk bound(red). We set a series of parameters $\alpha = 0.5, p' = w^* = 0.1, p = 0.9, w_{\text{ERM}} = 0.5, T = 100$, $K_{ST}(0) = 10, \eta(0) = 0.1, \sigma_t = \sqrt{\theta\eta(t)/t}, \theta = 0.001, W(0) = 0.3$ and $\delta = 0.01$ to be fixed for all experiments. For comparison tests, we set $\beta = 0.5, W(0) = 0.3$ for the first row, $n = 1000, W(0) = 0.3$ for the second row, and $\beta = 0.5, n = 1000$ for the last row, respectively.

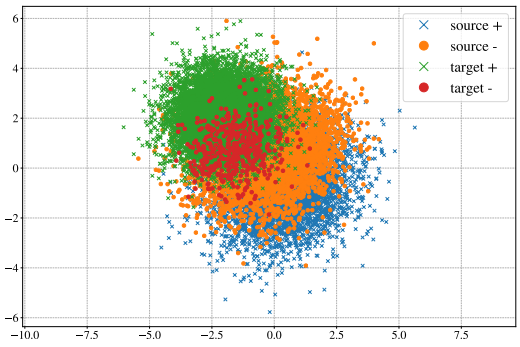


Fig. 2. The source data x_i are sampled from the **truncated** Gaussian distribution $\mathcal{N}_{tc} \sim (\mathbf{0}, 2\mathbf{I})$ while the target data are sampled from the **truncated** Gaussian distribution $\mathcal{N}_{tc} \sim ((-2, 2), \mathbf{I})$. The according label $y \in \{0, 1\}$, is 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.

where σ (and σ_i) are randomly selected from {-1,+1} with equal probability.

The comparisons of generalization error bound and excess risk bound are shown in figure 3. It is obvious that the true losses are bounded by our developed upper bounds. The result also suggests that our bound is tighter than the 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 behavior of generalization error while the Rademacher complexity bound fails to do so. It is noteworthy that both bounds converge as n increases. The result confirms that the bounds capture the dependence of the input data and output hypothesis, as well as the stochasticity of the algorithm.

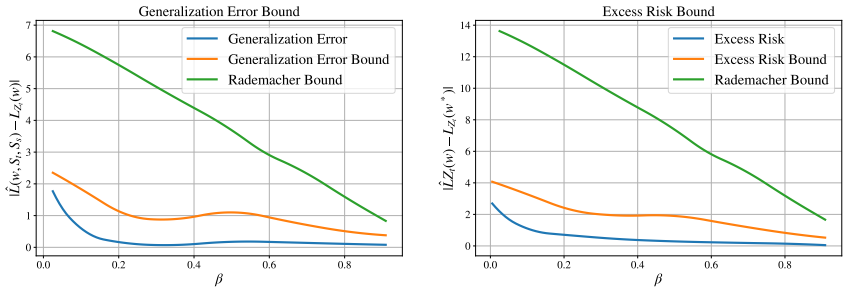


Fig. 3. Comparisons for generalization error and excess risk where we fix $n_s = 10000$ and vary n_t by setting $\alpha = \beta$.

C. Fast Rate Logistic Regression

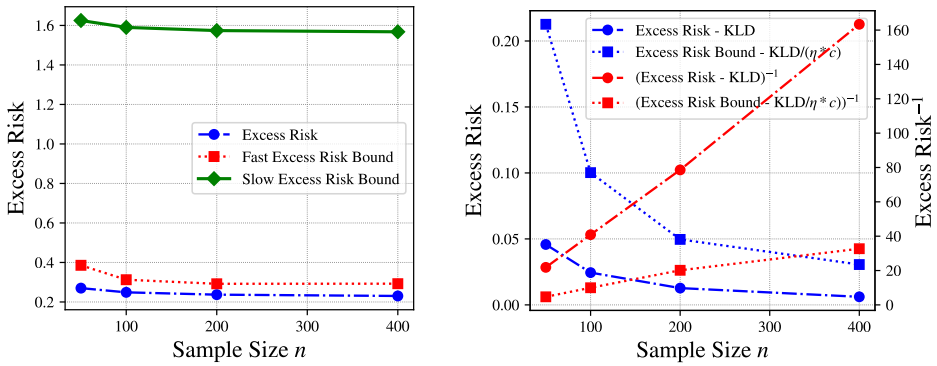
In this section, we apply our fast-rate bound in the logistic regression problem in a 2-dimensional space to further evaluate the effectiveness of the bounds in Equation (25). Again, we assume each $w \in \mathbb{R}^2$ and $z_i = (x_i, y_i) \in \mathbb{R}^2 \times \{0, 1\}$ as a similar setup in the previous section. Now we assume the source-only scenario where $\alpha = \beta = 0$, and each x_i is drawn from a standard multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{I}_2)$ and Let $w_s = (0.5, 0.5)$ and $w_t = (-0.5, -0.5)$ for a different setup for generalizing the experimental results. We also restrict hypothesis space as $\mathcal{W} = \{w : \|w\|_2 < 2\}$ where W_{ERM} falls in this area with high probability. Since the hypothesis is bounded and under the log-loss, then the learning problem will satisfy the central and witness condition [44, 45]. Therefore, it will satisfy the (η, c)-central condition. To estimate η, c, and the mutual information $I(W_{\mathsf{ERM}}, Z_i)$ efficiently, we repeatedly generate samples of W_{ERM} and Z_i and use their empirical density for estimation. Specifically, we vary the sample size n in the range $[50, 400]$ and for each value of n, we repeat the logistic regression algorithm 2000 times to generate a set of W_{ERM} samples. By setting $\eta = 0.8$ as an example, we can empirically estimate the CGF and the expected excess risk using the data sample and a set of ERM hypotheses, which results in an estimate of $c \approx 0.195$. Importantly, our experiments indicate that once η is fixed, the choice of c remains independent of the sample size n, providing empirical support for the (η, c)-central condition. For the mutual information estimation, we used a similar method as illustrated in the previous section by decomposing the mutual information into marginal and conditional divergences. To demonstrate the usefulness of the results, we also compare the bounds among the true excess risk, the slow rate excess risk in Equation (17) and the fast rate excess risk in Equation (25). The comparisons are shown in Figure 4. From the figure, it can be seen that the fast excess risk bound in Equation (25) is even tighter than the slow excess risk bound in Equation (17). Moreover, both the excess risk and its fast rate bound exhibit linear convergence scaling as $O(\frac{1}{n})$, up to the domain divergence (with some leading constant). Importantly, this simple toy example showed that the bounds presented in Theorem 3 are tight, accurately reflecting the true behaviours at the same decay rate.

D. Algorithms and Real Dataset

Many information-theoretic bounds are challenging to apply directly to real-world machine-learning tasks primarily because they require knowledge of the data and the hypothesis distributions, and estimating these bounds can be very difficult. Recognizing this gap, we introduce a heuristic boosting algorithm named `InfoBoost` where this innovative approach leverages the concepts behind the proposed bounds.

With the standard empirical risk minimization algorithm, both the source and target instances are equally weighted within their domains as we define the empirical risk by Eq (3). However, in many real-world scenarios, the weight of each instance can be different, especially for the transfer learning problem. To select the source data that are useful for the prediction on the target domain, we can re-weight each source instance so that the source data will have a similar performance as the target data. To interpret, we start from the following fact on change of measure for the expected risk if μ' is absolutely continuous w.r.t. μ:

$$\mathbb{E}_{\mu'}[\ell(W, Z)] = \mathbb{E}_{\mu}\left[\frac{\mu'(Z)}{\mu(Z)}\ell(W, Z)\right]. \tag{59}$$



(a) Excess Risk (b) Convergence

Fig. 4. We represent the true expected generalization error in (a) along with its bounds in Theorem 2 and Theorem 3. Here we vary n from 50 to 400. To show the convergence up to the domain divergence, we also plot the quantity $\mathbb{E}_W[R_{\mu'}(W)] - D(\mu\|\mu')$ and the fast rate bound $\frac{1}{c\eta n}\sum_{i=1}^{n} I(W;Z_i)$, along with their reciprocals to show the rate w.r.t. sample size n. All results are derived by 2000 experimental repeats.

Ideally, if we can find the distribution density ratio $\frac{\mu'(z_i)}{\mu(z_i)}$ for each instance in the source domain, the expected risk induced by the source is obviously an unbiased estimate of the expected risk under the target distribution. However, this quantity is non-trivial to estimate. For the supervised learning problem where $Z_i = (X_i, Y_i)$, under the co-variate assumption such that $P(Y|X)$ remains unchanged across different domains, this ratio can be estimated through statistical methods such as correcting sample selection bias [77], kernel mean matching [78] and direct density ratio estimation [79]. On the other hand, the iterative algorithm based on model aggregation (e.g., boosting) has been proposed to adaptively adjust the instance weights in transfer learning problems. To name a few, Dai et al. [10] take the first step in the application of boosting to transfer learning where the weights for the target data are adjusted using Adaboost [62] while for the source the weights are gradually decreasing following the weighted majority algorithm. Such a strategy is known as the TradaBoost algorithm. Eaton et al. [11] further develop the TransferBoost algorithm where the source instances are re-weighted according to whether the source domain can improve or hurt the prediction performance on the target. More recently, Wang et al. [80] propose the GapBoost algorithm by taking the hypothesis performance gap between two domains into consideration.

Inspired by the information-theoretic bounds and the boosting algorithm, we heuristically develop an information-theoretic-based boosting algorithm where the re-weighting scheme will involve mutual information $I(W;Z_i)$ and the domain divergence $D(\mu\|\mu')$, which appear in our derived upper bounds. From the excess risk bound (17), we can see that the learning performance is controlled by the mutual information terms $I(W;Z_i)$, the domain divergence $D(\mu\|\mu')$, the summation of the weights in the target domains controlled by α. Instead of using the same weights for all instances in the same domain, we assign different weights to different instances to show their importance and then introduce the novel boosting type algorithm based on the mutual information bound, namely, the `InfoBoost` algorithm. For simplicity, we consider the binary classification task where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and present the algorithm in Alg 1. The multi-label task and regression boosting algorithms can be extended following similar steps as shown in [62].

The procedures of the algorithm are sketched as follows:

- Given both source and target training data, firstly we initialize the weights $\gamma_1(i), i = 1, \cdots, n$ uniformly (arbitrary initialization is also acceptable). The subscript for all variables in the algorithm indicates the timestamp.
- We will then update weights iteratively in each round. At each timestamp t, we first learn a base hypothesis denoted by $h_t : \mathcal{X} \to \mathcal{Y}$ with the weights γ_t using both source and target data. To introduce the concept of information stability for a single instance z_i, we need to train the hypothesis without i-th sample and compare it with h_t, which we denote by h_t^{-i}.
- To capture the domain divergence,

Algorithm 1: InfoBoost Algorithm

Input : Source Sample S, Target Sample S', Iteration T, Loss function $\ell(h, x, y)$

Initialize weights $\gamma_1(i) = \frac{1}{n}$ for all i;
Choose hyper-parameters Γ, ζ, η and iteration T;
for $t = 1, \cdots, T$ **do**
  Learn a base hypothesis h_t using $S \cup S'$;
  Learn h_t^{-i} using $S \cup S'$;
  Learn domain hypothesis $h_{S'}$ and h_S using S' and S separately;
  Learn a domain discriminator h_{dis} for the domain features;
  $\epsilon_t = \sum_{i=1}^{n\beta} \gamma_t(i)\mathbf{1}_{h_t(x_i)\neq y_i} + \sum_{i=\beta n+1}^{n} \gamma_t(i)\mathbf{1}_{h_t(x_i)\neq y_i}$;
  $\alpha_t = \log \frac{1-\epsilon_t}{\epsilon_t}$;
  for $i = 1, \cdots, \beta n$ **do**
    Incur stability loss $d_i(h_t, h_t^{-i})$ according to (60);
    $\gamma_{t+1}(i) = \gamma_t(i) \cdot e^{\alpha_t \ell(h_t, x_i, y_i) - \eta d_i(h_t, h_t^{-i})}$;
  end
  for $i = \beta n + 1, \cdots, n$ **do**
    Incur stability loss $d_i(h_t, h_t^{-i})$ according to (60);
    Incur domain discrimination loss $\ell(h_{\text{dis}}(x_i))$ according to (61);
    Incur labelling divergence loss $d(h_{S'}(x_i, y_i), h_S(x_i, y_i))$ according to (62);
    $\gamma_{t+1}(i) = \gamma_t(i) \cdot e^{\alpha_t \ell(h_t, x_i, y_i) - \eta d_i(h_t, h_t^{-i}) - \zeta(\ell(h_{\text{dis}}(x_i)) + d_i(h_S, h_{S'}))}$;
  end
  Let $Z_t = \sum_{i=1}^{n} \gamma_t(i)$;
  Normalize $\gamma_t(i) = \frac{\gamma_t(i)}{Z_t}$ for $i = 1, 2, \cdots, n$;
end
Output: $f(x) = \mathbf{1}_{\sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t}$.

we decompose the KL divergence as $D(\mu(X,Y)\|\mu'(X,Y)) = D(\mu(X), \mu'(X)) + D(\mu(Y|X)\|\mu'(Y|X)|X)$ into two terms. Instead of directly estimating the KL divergence, we may use surrogate quantities learned from the data to approximate these two KL divergences. In practice, we train a domain discriminator $h_{\text{dis}} : \mathcal{X} \to [0, 1]$ to capture the feature divergence $D(\mu(X), \mu'(X))$, which takes the features X as the input and the probability of being the target domain as the output. In other words, if $h_{\text{dis}}(x)$ is closer to 1, then x is more likely to be drawn from the target domain, and the divergence should be small. In addition, we train two domain-specific classifiers $h_{S'} : \mathcal{X} \to \mathcal{Y}$ and $h_S : \mathcal{X} \to \mathcal{Y}$ using S' and S respectively to capture the condition divergence term $D(\mu(Y|X)\|\mu'(Y|X)|X)$, where the differences of their predictions on the same input can be heuristically interpreted as the labelling divergence.

- Having outlined the basic concept, let us explain in detail how the algorithm works. We firstly evaluate h_t over S and S' to obtain the error rate ϵ_t and α_t similarly as in [62]. For each target data, we use the exponentially updating rule following the boosting strategy in [62]. In addition to that, we also examine the difference between h_t and h_t^{-i}, which can be heuristically interpreted as the information stability loss by:

$$d_i(h_t, h_t^{-i}) = |h_t(x_i) - h_t^{-i}(x_i)|. \tag{60}$$

A smaller d_i indicates that the sample z_i has less effect on learning the hypothesis. Nonetheless, learning h_t^{-i} may be very time-consuming due to the large training sample size n. Instead, by evenly splitting the training sample into K folds where K is much smaller than the sample size n, we will train a set of base classifier $h_k, k = 1, 2, \cdots, K$ for each fold from the rest $K - 1$ folds. That is, we will use the same hypothesis h_k for all data in k-th fold to approximate h_t^{-i}. We choose $K = 20$ for the subsequent experiments. For each source data, we will further take the domain divergence into account, which is decomposed into two parts as mentioned earlier. The first is the divergence of the feature of X induced by the domain discriminator h_{dis}:

$$\ell(h_{\text{dis}}(x_i)) = 1 - h_{\text{dis}}(x_i). \tag{61}$$

Here, if $\ell(h_{\text{dis}}(x_i))$ is large, then i-th sample is more different from the target data and we shall assign a smaller weight to this instance. Regarding the labelling divergence $D(\mu(Y|X)\|\mu'(Y|X)|X)$, we will utilize the hypothesis h_S and $h_{S'}$ trained from the source and

target domains, which can be seen as a mapping from X to Y, and define the following quantity to measure their differences:

$$d_i(h_S, h_{S'}) = |\ell(h_S, x_i, y_i) - \ell(h_{S'}, x_i, y_i)|. \tag{62}$$

A smaller labelling divergence implies that h_S and $h_{S'}$ are more similar. In summary, we use $d_i(h_t, h_t^{-i})$ as a proxy for the mutual information $I(W; Z_i)$ and $\text{div}(x_i, y_i) = \ell(h_{\text{dis}}(x_i)) + d_i(h_S, h_{S'})$ as a proxy for the domain divergence $D(\mu \| \mu')$. We also introduce the hyper-parameters ζ and η to allow more flexible control of these two quantities, and an appropriate choice of which could also prevent us from focusing too much on the particular data.

- After T iterations, we aggregate all base classifiers h_t with the corresponding weights α_t for some new input x and output the prediction.

We now evaluate the proposed InfoBoost algorithm on several typical transfer learning tasks to show its effectiveness. The datasets used are listed as follows.

- Office-Caltech-10 [81]: This dataset contains four subsets, and each domain has a set of office photos with the same 10 classes. In particular, the four subsets are **W**ebcam (W for short), **D**SLR (D for short), **A**mazon (A for short) and **C**altech dataset (C for short). We use each subset as a domain. Consequently, we get four domains (A, C, D and W), leading to 12 transfer learning problems. Since each domain shares the same 10 classes, we therefore constructed 5 binary classification tasks and reported the average error in each transfer learning problem. We use the SURF features as described in [81] encoded with a visual dictionary of 800 dimensions.
- 20 Newsgroups [1]: the dataset contains approximately 20000 reviews from 7 major categories that can be split into 20 subcategories. The source and target domains were picked from the same major categories but different subcategories in each transfer learning task, in the same way as in [80].
- MNIST and USPS: These two datasets contain black and white hand-written digits from 0 to 9 where MNIST [2] has approximately 70000 images and USPS [3] has approximately 10000 images in total. Since each domain shares the same digits from 0 to 9 (but a different writing style), we therefore constructed 5 binary classification tasks and reported the classification error for each task. We use all samples from USPS datasets but only use 10000 images from MNIST.

The benchmarks we compare with are the typical boosting algorithms for transfer learning with the same base classifier, e.g., the logistic regression model with very small regularization. We list all the competitors as follows.

- **AdaBoost$_{\mathcal{T}}$ [62]:** The first baseline method is the AdaBoost algorithm with the target training data only, and the initial weights are assigned uniformly.
- **AdaBoost$_{\mathcal{S}\&\mathcal{T}}$:** We also directly apply the AdaBoost algorithm with both source and target data, and the initial weights are assigned uniformly over all instances.
- **TrAdaBoost [10]:** TrAdaBoost algorithm is the firstly proposed boosting method for transfer learning where at each iteration less weights are assigned to the source and we adaptively focus more on the target data.
- **TransferBoost [11]:** TransferBoost is another boosting algorithm that selects the useful source data by examining whether the source domain improves the learning performance in the target domain. We use the same way as described in [11] to choose the hyper-parameters α_t and β_t.
- **GapBoost [80]:** The GapBoost algorithm minimizes the proposed performance gap between the source and target domains by training auxiliary classifiers on both source and target domains. The gap of the predictions of the auxiliary classifiers is regarded as the performance gap. Such a quantity is taken into consideration when updating instance weights. We use the same hyper-parameters as described in the experiments section in [80].

Performance Comparisons In all comparisons, η and ζ are both set to 1 in the InfoBoost algorithm. We used the entire source data set for different tasks. As each dataset has a different target sample size, we selected 10 target instances in the training phase

[1] http://qwone.com/~jason/20Newsgroups/
[2] http://yann.lecun.com/exdb/mnist/
[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass.html#usps

TABLE I
ACCURACY IN % WITH 10 TARGET DATA FOR SURF OFFICE-CALTECH DATASET

	Adaboost$_{\mathcal{T}}$	Adaboost$_{\mathcal{S}\&\mathcal{T}}$	Tradaboost	Transfer Boost	GapBoost	InfoBoost
$A \to W$	64.69	75.47	75.88	76.42	77.78	**78.30**
$A \to D$	71.60	75.83	76.16	75.42	76.89	**78.44**
$A \to C$	57.19	76.88	**77.24**	74.97	75.08	74.84
$D \to A$	54.94	56.98	57.12	57.68	57.87	**58.04**
$D \to W$	48.91	51.84	51.54	52.05	52.31	**52.62**
$D \to C$	54.57	60.97	60.88	61.01	**61.11**	61.05
$W \to A$	54.76	62.90	62.75	63.04	62.86	**63.28**
$W \to D$	59.41	61.32	**61.44**	60.66	60.96	61.20
$W \to C$	55.91	63.07	63.01	63.13	63.07	**63.27**
$C \to A$	57.68	76.34	**78.05**	76.51	75.14	77.87
$C \to D$	68.24	68.61	74.19	70.93	70.08	**75.80**
$C \to W$	65.28	75.01	74.85	76.39	74.16	**80.49**
Average	59.43	67.10	67.76	67.35	67.36	**68.77**

TABLE II
ACCURACY IN % WITH 10% TRAINING TARGET DATA FOR 20 NEWSGROUPS DATA

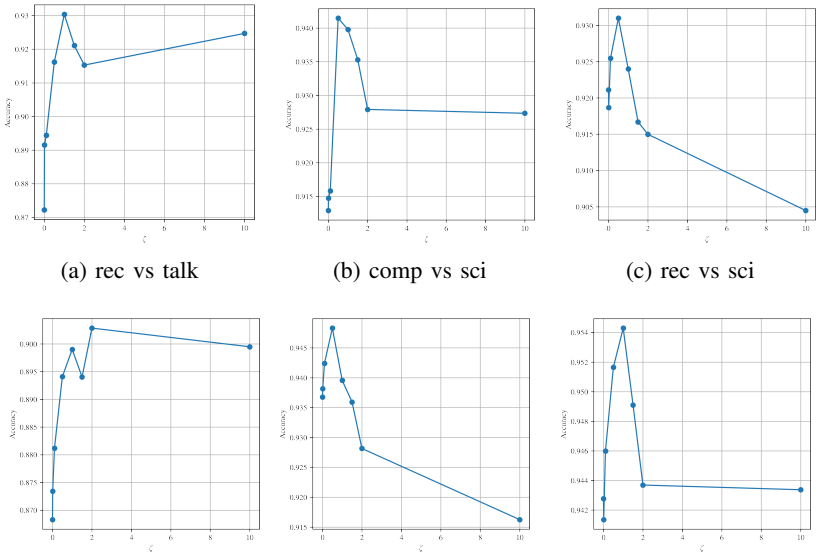
Tasks	AdaBoost$_{\mathcal{T}}$	AdaBoost$_{\mathcal{S}\&\mathcal{T}}$	TradaBoost	TransferBoost	GapBoost	InfoBoost
rec vs talk	90.30 ± 3.14	87.42 ± 2.21	71.97 ± 5.89	89.42 ± 2.99	91.77 ± 3.51	$\mathbf{93.11 \pm 1.71}$
comp vs sci	92.87 ± 1.03	84.66 ± 1.53	92.71 ± 1.78	93.35 ± 1.05	93.55 ± 2.54	$\mathbf{94.15 \pm 1.04}$
rec vs sci	90.17 ± 1.02	86.68 ± 1.06	90.49 ± 1.36	$\mathbf{93.62 \pm 0.69}$	90.70 ± 2.14	92.58 ± 0.89
talk vs sci	88.66 ± 2.74	71.72 ± 3.68	81.39 ± 4.98	87.78 ± 3.25	90.14 ± 2.55	$\mathbf{90.40 \pm 1.78}$
comp vs rec	91.44 ± 1.35	87.90 ± 1.80	91.52 ± 1.12	$\mathbf{94.95 \pm 0.60}$	92.62 ± 2.95	94.00 ± 1.58
comp vs talk	93.86 ± 1.46	89.57 ± 1.32	75.78 ± 1.14	94.83 ± 0.78	94.90 ± 1.02	$\mathbf{95.08 \pm 0.89}$
Average	91.22	84.66	83.98	92.33	92.28	**93.22**

TABLE III
ACCURACY IN % WITH 1% TRAINING TARGET DATA FOR MNIST AND USPS DATASETS

Tasks	AdaBoost$_{\mathcal{T}}$	AdaBoost$_{\mathcal{S}\&\mathcal{T}}$	TradaBoost	TransferBoost	GapBoost	InfoBoost
U to M$_{1\text{ vs }7}$	61.52 ± 4.15	57.73 ± 1.92	52.67 ± 1.08	63.47 ± 2.34	63.58 ± 1.99	$\mathbf{64.51 \pm 2.11}$
U to M$_{2\text{ vs }3}$	59.69 ± 3.96	56.49 ± 1.87	52.18 ± 3.87	63.29 ± 1.71	$\mathbf{63.37 \pm 1.87}$	62.83 ± 1.57
U to M$_{5\text{ vs }6}$	61.05 ± 2.87	57.22 ± 1.24	55.15 ± 1.44	61.54 ± 1.74	61.46 ± 1.70	$\mathbf{62.66 \pm 1.63}$
U to M$_{0\text{ vs }8}$	59.78 ± 3.75	57.61 ± 2.57	51.48 ± 1.29	64.50 ± 1.88	64.77 ± 1.88	$\mathbf{65.00 \pm 2.02}$
U to M$_{4\text{ vs }9}$	57.97 ± 3.12	58.03 ± 4.24	57.15 ± 4.69	61.87 ± 1.64	$\mathbf{61.91 \pm 1.71}$	61.51 ± 2.01
M to U$_{1\text{ vs }7}$	87.42 ± 4.94	90.44 ± 12.54	72.64 ± 13.98	92.23 ± 3.66	91.26 ± 3.50	$\mathbf{92.43 \pm 3.75}$
M to U$_{2\text{ vs }3}$	74.38 ± 5.33	74.32 ± 11.68	76.80 ± 7.34	$\mathbf{77.01} \pm 7.08$	76.94 ± 5.83	76.29 ± 6.34
M to U$_{5\text{ vs }6}$	71.87 ± 7.14	59.11 ± 8.62	69.53 ± 3.79	69.42 ± 7.92	75.33 ± 5.83	$\mathbf{75.63 \pm 5.60}$
M to U$_{0\text{ vs }8}$	74.47 ± 3.85	74.84 ± 4.87	71.83 ± 4.57	77.89 ± 4.08	$\mathbf{78.43 \pm 4.22}$	77.96 ± 4.43
M to U$_{4\text{ vs }9}$	87.88 ± 4.04	88.80 ± 5.93	79.24 ± 1.43	91.07 ± 4.03	91.21 ± 3.78	$\mathbf{91.33 \pm 3.94}$
Average	69.60	67.46	63.87	72.23	72.83	**73.02**

for the Office-Caltech dataset, 10% of the target data for the 20 Newsgroup dataset for training, and 1% target instances for training on the handwritten digit datasets. We used generalized linear regression (e.g., logistic regression) for various boosting algorithms as our base classifier. All performance comparisons are listed on Table I, II, and III. From the comparisons, we can see that the InfoBoost algorithm outperforms other competitors in most cases for all three different transfer learning problems, showing the benefits of taking the domain divergence and the mutual information into account.

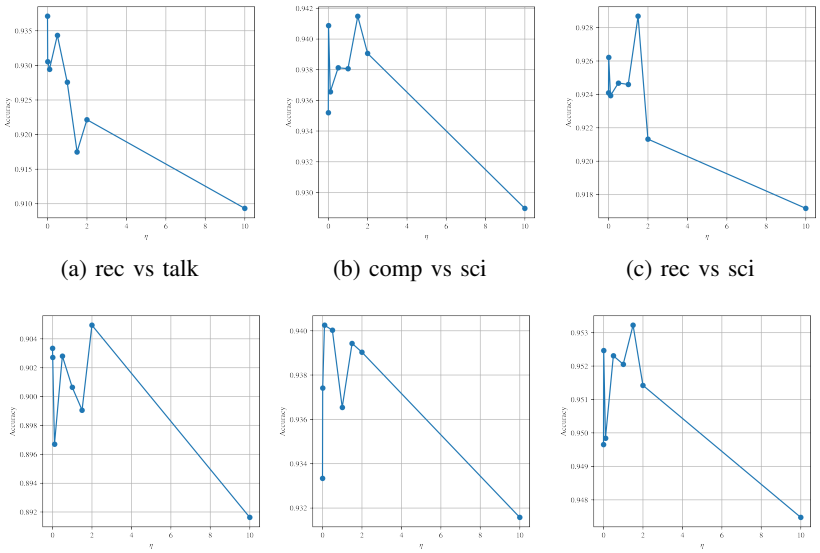
Hyperparameter Sensitivity We carefully examine the effects of ζ and η by fixing one variable and varying another. The results are shown in Figure 5 and 6, respectively. One the one hand, we can see from Figure 5, $\zeta = 1$ achieves higher accuracy compared to other choices. This is understandable because in the case of too small ζ, the domain divergence is not well taken into account in the source data updating step and the algorithm may be over-fitted. If ζ is too large, then the domain divergence will overwhelm the



(a) rec vs talk (b) comp vs sci (c) rec vs sci

(d) talk vs sci (e) comp vs rec (f) comp vs talk

Fig. 5. Effect of ζ varying from 0 to 10 when fixing $\eta = 1$. The results are averaged over 20 experiments.

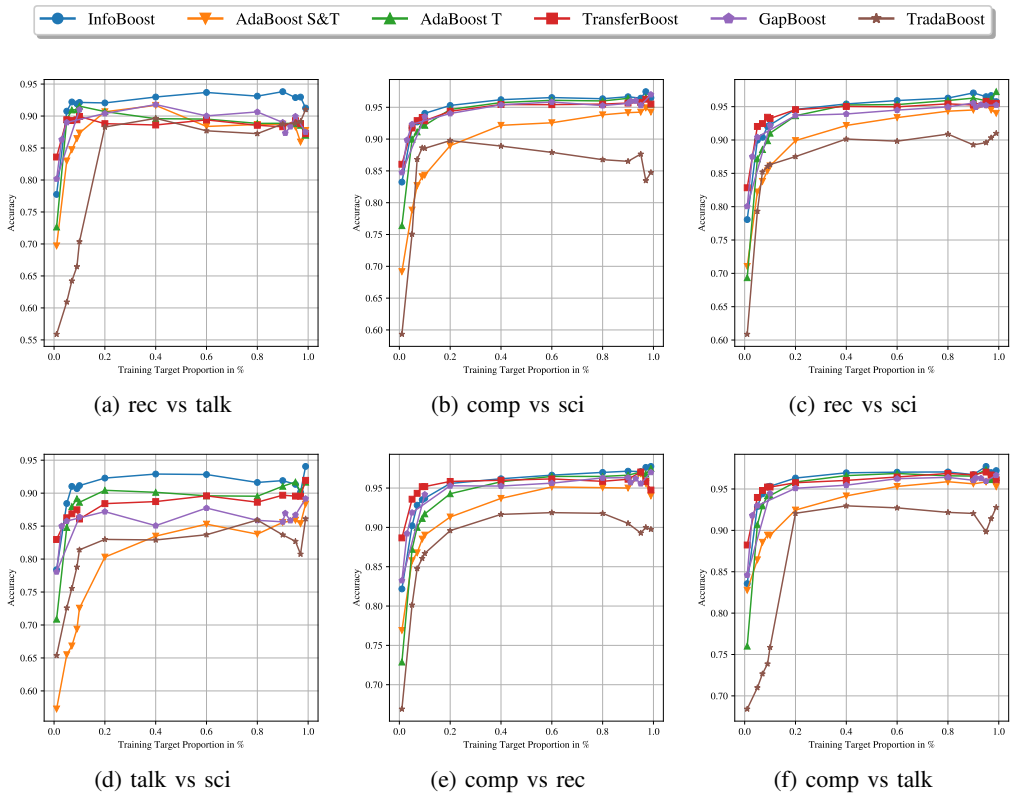


(a) rec vs talk (b) comp vs sci (c) rec vs sci

(d) talk vs sci (e) comp vs rec (f) comp vs talk

Fig. 6. Effect of η varying from 0 to 10 when fixing $\zeta = 1$. The results are averaged over 20 experiments.

effect of the empirical risk, which will easily lead to under-fitting. Therefore, a proper selection of ζ is needed to yield satisfactory performance. On the other hand, if we fix $\zeta = 1$ and vary η from 0 to 10, from Figure 6 we can then see that a reasonable choice of η (e.g., less and equal than 2) can lead to a good performance in most cases. Even with a very small η, the accuracy is still comparable to the best result. But the performance will be degraded if we choose a large η (e.g., $\eta = 10$); this is because after several iterations, we mainly select the samples that have less effect on the prediction and a large η overwhelms both empirical risk and domain divergence, which may lead to under-fitting in a similar manner. Although the accuracy rate dropped, the drop was not as great as the change of ζ. Therefore, the effect of η on the model performance is less than that of ζ. To achieve the optimal choice of the hyper-parameters,



(a) rec vs talk (b) comp vs sci (c) rec vs sci

(d) talk vs sci (e) comp vs rec (f) comp vs talk

Fig. 7. Accuracy with different proportions of target training sizes for different tasks. We utilize all the source data and the results are averaged over 20 experiments.

we recommend setting both ζ and η to 1 around.

We also examine the effect of the size of the target sample. Specifically, we split the target dataset into the training and testing sets by varying the proportion of the training size from 0.01 to 0.99 for 20 Newsgroup dataset as an example, and the results are shown in Figure 7. From the comparisons, when the target training sample size is sufficient, the InfoBoost algorithm will have the best performance compared to other competitors. However, when there is a lack of a target training sample, the InfoBoost algorithm might not have enough data to precisely capture the domain divergence and the mutual information, and the performance will be degraded with a small target training sample size.

Further Discussion Different from the works [11] and [80] that utilize the domain performance gap, we propose a new boosting algorithm inspired by the mutual information and the domain divergence for transfer learning. We empirically verify the effectiveness of our proposed algorithm and particularly examine the sensitivity of the hyper-parameters. Since the information-theoretic quantities are usually hard to estimate, we use the surrogate quantities (e.g., $d_i(h_t, h_t^{-i})$, $\mathrm{div}(x_i, y_i)$ that heuristically represents the mutual information and the domain divergence. Instead, one could also improve the algorithm with other related quantities by information conditioning and processing techniques, such as the conditional mutual information [25] and f-information proposed in [28], which might be easier to estimate from the data, which is left as our future work.

VII. Concluding Remarks

In this work, we developed a set of upper bounds on the generalization error and the excess risk for general transfer learning algorithms under an information-theoretic framework. The derived bounds are particularly useful for various widely used algorithms in machine learning such as ERM, regularized ERM, Gibbs algorithm, and stochastic gradient descent algorithms. Moreover, we extend the results with different divergences other than the KL divergence, such as the ϕ-divergence and the Wasserstein distance, which can give a tighter bound and handle more general scenarios where the KL divergence may be vacuous. We also tighten the bounds for the ERM and regularized ERM

algorithms, and give the fast rate characterization of the transfer learning problems. Finally, we propose the InfoBoost algorithm that in each re-weighting iteration, the quantities of the mutual information bounds are utilized and the empirical verification shows the effectiveness of our algorithm. However, in some cases, it is hard to estimate the information-theoretic bound in practice. In particular, the density estimation from the source and target data might be inaccurate only given a few data samples in a high dimensional data space. One can thus relax the condition by either assuming the source and target distributions are lying in a restricted space and the divergence term will not be too deviated even with a small sample size or by finding a completely different divergence that is easier to estimate from the data, which is left as our future work.

In this paper we only studied the upper bound of transfer learning, ensuring the performance for some commonly used algorithms. However, since it is an upper bound, we are not sure when the introduction of source data hurts or improves the performance on the target domain. It is often known as the "negative transfer" if the source data hurts. Therefore, under the information-theoretic framework, another possible future direction is to develop the lower bound for transfer learning algorithms to rigorously determine whether the source data will be useful or not, which might help mitigate the effect of the negative transfer.

ACKNOWLEDGEMENT

The preliminary version of this work is presented at the ISIT2020 conference, we greatly appreciate useful feedback and comments from all reviewers. This research is supported by Melbourne Research Scholarships (MRS), and Australian Defence Science and Technology Group (DSTG) under the scheme The Artificial Intelligence for Decision Making Initiative 2022 and in part by Australian Research Council under project DE210101497.

REFERENCES

[1] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-theoretic analysis for transfer learning," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2819–2824.

[2] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications*, 1971.

[3] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," in *Proceedings of the tenth annual conference on Computational learning theory*, 1997, pp. 152–162.

[4] L. Devroye and T. Wagner, "Distribution-free inequalities for the deleted and holdout error estimates," *IEEE Transactions on Information Theory*, vol. 25, no. 2, pp. 202–207, 1979.

[5] D. A. McAllester, "Some pac-bayesian theorems," *Machine Learning*, vol. 37, no. 3, pp. 355–363, 1999.

[6] H. Xu and S. Mannor, "Robustness and generalization," *Machine learning*, vol. 86, no. 3, pp. 391–423, 2012.

[7] 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.

[8] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning bounds for domain adaptation," *Advances in neural information processing systems*, vol. 20, 2007.

[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] 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.

[11] E. Eaton *et al.*, "Selective transfer between learning tasks using task-based boosting," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

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

[13] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Artificial Intelligence and Statistics*. PMLR, 2016, pp. 1232–1240.

[14] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[15] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 1999.

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

[17] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, 2015, pp. 117–126.

[18] 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.

[19] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," in *Algorithmic Learning Theory*. PMLR, 2018, pp. 25–55.

[20] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 587–591.

[21] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, "Information-theoretic analysis of stability and bias of learning algorithms," in *2016 IEEE Information Theory Workshop (ITW)*. IEEE, 2016, pp. 26–30.

[22] A. T. Lopez and V. Jog, "Generalization error bounds using wasserstein distances," in *2018 IEEE Information Theory Workshop (ITW)*. IEEE, 2018, pp. 1–5.

[23] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.

[24] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for sgld via data-dependent estimates," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*. PMLR, 2020, pp. 3437–3452.

[26] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9925–9935, 2020.

[27] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 109–19 121, 2021.

[28] H. Harutyunyan, M. Raginsky, G. V. Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," *arXiv preprint arXiv:2110.01584*, 2021.

[29] P. Grünwald, T. Steinke, and L. Zakynthinou, "Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes," *arXiv preprint arXiv:2106.09683*, 2021.

[30] Y. Bu, G. Aminian, L. Toni, G. W. Wornell, and M. Rodrigues, "Characterizing and understanding the generalization error of transfer learning with gibbs algorithm," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8673–8699.

[31] U. Côté-Allard, C. L. Fall, A. Campeau-Lecours, C. Gosselin, F. Laviolette, and B. Gosselin, "Transfer learning for semg hand gestures recognition using convolutional neural networks," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 1663–1668.

[32] J. Van Baar, A. Sullivan, R. Cordorel, D. Jha, D. Romeres, and D. Nikovski, "Sim-to-real transfer learning using robustified controllers in robotic tasks involving complex dynamics," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6001–6007.

[33] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, "Machine learning for medical imaging," *Radiographics*, vol. 37, no. 2, p. 505, 2017.

[34] V. Cheplygina, I. P. Pena, J. H. Pedersen, D. A. Lynch, L. Sørensen, and M. De Bruijne, "Transfer learning for multicenter classification of chronic obstructive pulmonary disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1486–1496, 2017.

[35] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using imagenet," *Computers in biology and medicine*, vol. 128, p. 104115, 2021.

[36] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[37] T. Kocmi and O. Bojar, "Trivial transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1809.00357*, 2018.

[38] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[39] I. Kuzborskij and F. Orabona, "Stability and hypothesis transfer learning," in *International Conference on Machine Learning*. PMLR, 2013, pp. 942–950.

[40] P. Germain, A. Habrard, F. Laviolette, and E. Morvant, "A pac-bayesian approach for domain adaptation with specialization to linear classifiers," in *International conference on machine learning*. PMLR, 2013, pp. 738–746.

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

[42] 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.

[43] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Fast rate generalization error bounds: Variations on a theme," *arXiv preprint arXiv:2205.03131*, 2022.

[44] T. Van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson, "Fast rates in statistical and online learning," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1793–1861, 2015.

[45] P. D. Grünwald and N. A. Mehta, "Fast rates for general unbounded loss functions: From erm to generalized bayes." *J. Mach. Learn. Res.*, vol. 21, pp. 56–1, 2020.

[46] P. L. Bartlett and S. Mendelson, "Empirical minimization," *Probability theory and related fields*, vol. 135, no. 3, pp. 311–334, 2006.

[47] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[48] S. Hanneke, "Refined error bounds for several learning algorithms," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 4667–4721, 2016.

[49] Z. Mhammedi, P. D. Grunwald, and B. Guedj, "Pac-bayes un-expected bernstein inequality," *arXiv preprint arXiv:1905.13367*, 2019.

[50] N. Mehta, "Fast rates with high probability in exp-concave statistical learning," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1085–1093.

[51] J. Zhu, "Semi-supervised learning: The case when unlabeled data is equally useful," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 709–718.

[52] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time, i," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.

[53] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," *Advances in neural information processing systems*, vol. 24, 2011.

[54] C. Malherbe and N. Vayatis, "Global optimization of lipschitz functions," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2314–2323.

[55] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 681–688.

[56] A. Pensia, V. Jog, and P.-L. Loh, "Generalization error bounds for noisy, iterative algorithms," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 546–550.

[57] M. Haghifam, B. Rodríguez-Gálvez, R. Thobaben, M. Skoglund, D. M. Roy, and G. K. Dziugaite, "Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization," in *International Conference on Algorithmic Learning Theory*. PMLR, 2023, pp. 663–706.

[58] 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.

[59] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[60] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," *arXiv preprint arXiv:1109.5647*, 2011.

[61] F. Salehi, E. Abbasi, and B. Hassibi, "The impact of regularization on high-dimensional logistic regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[62] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of online learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[63] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild: A lock-free approach to parallelizing stochastic gradient descent," *Advances in neural information processing systems*, vol. 24, 2011.

[64] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč, "Sarah: A novel method for machine learning problems using stochastic recursive gradient," in *International conference on machine learning*. PMLR, 2017, pp. 2613–2621.

[65] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 492–518.

[66] R. A. Saleh and A. Saleh, "Statistical properties of the log-cosh loss function used in machine learning," *arXiv preprint arXiv:2208.04564*, 2022.

[67] S. Gupta and D. Rothenhäusler, "The s-value: Evaluating stability with respect to distributional shifts," *arXiv preprint arXiv:2105.03067*, 2021.

[68] A. T. Nguyen, T. Tran, Y. Gal, P. H. Torr, and A. G. Baydin, "Kl guided domain adaptation," *arXiv preprint arXiv:2106.07780*, 2021.

[69] G. Aminian, M. Abroshan, M. M. Khalili, L. Toni, and M. Rodrigues, "An information-theoretical approach to semi-supervised learning under covariate-shift," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 7433–7449.

[70] S. Hanneke and S. Kpotufe, "On the value of target data in transfer learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[71] R. M. Dudley, "Distances of probability measures and random variables," in *Selected Works of RM Dudley*. Springer, 2010, pp. 28–37.

[72] 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)*. IEEE, 2017, pp. 1475–1479.

[73] A. R. Esposito and M. Gastpar, "From generalisation error to transportation-cost inequalities and back," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 294–299.

[74] M. Raginsky, I. Sason *et al.*, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[75] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic sobolev inequalities," *Journal of Functional Analysis*, vol. 163, no. 1, pp. 1–28, 1999.

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

[77] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola, "Correcting sample selection bias by unlabeled data," *Advances in neural information processing systems*, vol. 19, 2006.

[78] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf, "Covariate shift by kernel mean matching," *Dataset shift in machine learning*, vol. 3, no. 4, p. 5, 2009.

[79] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[80] B. Wang, J. Mendez, M. Cai, and E. Eaton, "Transfer learning via minimizing the performance gap between domains," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[81] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2066–2073.

[82] S. Boucheron, G. Lugosi, and P. Massart, *Con-*

centration inequalities: A nonasymptotic theory of independence. OUP Oxford, Feb. 2013.

[83] B. Liu, Y. Cai, Y. Guo, and X. Chen, "Transtailor: Pruning the pre-trained model for improved transfer learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 10, 2021, pp. 8627–8634.

[84] K. You, Y. Liu, J. Wang, and M. Long, "Logme: Practical assessment of pre-trained models for transfer learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 133–12 143.

[85] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 2521–2530.

[86] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2018.

[87] S. Niu, M. Liu, Y. Liu, J. Wang, and H. Song, "Distant domain transfer learning for medical imaging," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3784–3793, 2021.

[88] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information Theory*, vol. 37, no. 6, pp. 1501–1518, 1991.

Appendix

Appendix A
Proof of Theorem 1

Proof. First, we rewrite the expectation of the generalization error for certain algorithm $P_{W|SS'}$ as

$$\begin{aligned}
&\mathbb{E}_{WSS'}\left[L_{\mu'}(W)-\hat{L}_{\alpha}(W)\right]=\mathbb{E}_{WSS'}\left[L_{\mu'}(W)-(1-\alpha)\hat{L}(W,S)-\alpha\hat{L}(W,S')\right]\\
&=\mathbb{E}_{WSS'}\left[(1-\alpha)L_{\mu'}(W)+\alpha L_{\mu'}(W)-\frac{1}{n}\sum_{i=\beta n+1}^{n}\frac{1-\alpha}{1-\beta}\ell(W,Z_i)-\frac{1}{n}\sum_{i=1}^{\beta n}\frac{\alpha}{\beta}\ell(W,Z_i')\right]\\
&=\frac{1}{n}\mathbb{E}_{WSS'}\left[\sum_{i=1}^{\beta n}\frac{\alpha}{\beta}(L_{\mu'}(W)-\ell(W,Z_i'))+\sum_{i=\beta n+1}^{n}\frac{1-\alpha}{1-\beta}(L_{\mu'}(W)-\ell(W,Z_i))\right]\\
&=\frac{1}{n}\frac{\alpha}{\beta}\sum_{i=1}^{\beta n}\mathbb{E}_{WZ_i}\left[(L_{\mu'}(W)-\ell(W,Z_i'))\right]+\frac{1}{n}\frac{1-\alpha}{1-\beta}\sum_{i=\beta n+1}^{n}\mathbb{E}_{WZ_i}\left[L_{\mu'}(W)-\ell(W,Z_i))\right],
\end{aligned}$$

where the joint distribution $P_{WSS'}(w,s,s')$ on (W,S,S') is given by $P_S(s)P_{S'}(s')P_{W|SS'}(w|s,s')$. Recall that the variational representation of the KL divergence between two distributions P and Q defined over $\mathcal{X}$ is given as: (see, e.g. [82])

$$D(P||Q)=\sup_{f}\left\{\mathbb{E}_P\left[f(X)\right]-\log\mathbb{E}_Q\left[e^{f(x)}\right]\right\}, \tag{63}$$

where the supremum is taken over all measurable functions such that $\mathbb{E}_Q\left[e^{f(x)}\right]$ exists. For each $i=1,\ldots,n$, we define the joint distribution $P_{WZ_i}(w,z_i)$ ($P_{WZ_i'}(w,z_i')$) between an individual sample Z_i (Z_i') and the hypothesis W as induced by $P_{WSS'}(w,z^n)$ by marginalizing all samples other than z_i, and let P_W be the marginal distribution on W induced from $P_{WSS'}$.

We first show the first inequality in the Theorem. For any $i=1,\ldots,\beta n$, let $P=P_{WZ_i'}$, $Q=P_W\otimes\mu'$ in (63), and define $f:=\lambda\ell(W,Z_i')$ for some λ. The representation in (63) implies that

$$\mathbb{E}_{WZ_i'}\left[\lambda\ell(W,Z_i')\right]\le D(P_{WZ_i'}||P_W\otimes\mu')+\log\mathbb{E}\left[e^{\lambda\ell(W,Z_i')}\right],$$

where the expectation on the R.H.S. is taken w.r.t. the distribution $P_W\otimes\mu'$. By the assumption that

$$\log\mathbb{E}\left[e^{\lambda(\ell(W,Z_i')-\mathbb{E}[\ell(W,Z_i')])}\right]\le\psi(\lambda)$$

for some $\lambda\in[b_-,0]$ under the distribution $P_W\otimes\mu'$, we have

$$\mathbb{E}_{WZ_i'}\left[\lambda(\ell(W,Z_i')-\mathbb{E}_{WZ_i'\sim P_W\otimes\mu'}\left[\ell(W,Z_i')\right])\right]\le D(P_{WZ_i'}||P_W\otimes\mu')+\psi(\lambda),$$

which is equivalent to

$$\begin{aligned}
\mathbb{E}_{WZ_i'}\left[L_{\mu'}(W)-\ell(W,Z_i')\right]&\le-\frac{1}{\lambda}\left(D(P_{WZ_i'}||P_W\otimes\mu')+\psi(\lambda))\right)\\
&=-\frac{1}{\lambda}\left(I(W;Z_i')+D(P_{Z_i'}||\mu')+\psi(\lambda))\right)\\
&=-\frac{1}{\lambda}\left(I(W;Z_i')+\psi(\lambda))\right),
\end{aligned}$$

as $P_{Z_i'}=\mu'$ for $i=1,\ldots,\beta n$. The best upper bound is obtained by minimizing the R.H.S., giving

$$\mathbb{E}_{WZ_i'}\left[L_{\mu'}(W)-\ell(W,Z_i')\right]\le\min_{\lambda\in[0,-b_-]}\frac{1}{\lambda}(I(W;Z_i')+\psi(-\lambda))=\psi^{*-1}(I(W;Z_i')). \tag{64}$$

For $i = \beta n + 1, \ldots, n$ in the source domain, using the same argument we can show that

$$\mathbb{E}_{WZ_i}\left[L_{\mu'}(W) - \ell(W, Z_i)\right] \leq \psi^{*-1}(I(W; Z_i) + D(\mu||\mu')) \tag{65}$$

Summing over i using the upper bounds in (64) and (65), we obtain the first inequality in the theorem.

The second inequality is shown in the same way by using the fact that the cumulant generating function is upper bounded by $\psi(\lambda)$ in $[0, b_+)$. □

APPENDIX B
PROOF OF COROLLARY 1

Proof. In the case when $\beta = 0$, for any hypothesis W induced by the sample S and the learning algorithm $P_{W|S}$, the generalization error is

$$\begin{aligned}\mathbb{E}_{WS}\left[L_{\mu'}(W) - \hat{L}(W, S)\right] &= \mathbb{E}_{WS}\left[L_{\mu'}(W) - \frac{1}{n}\sum_{i=1}^{n}\ell(W, Z_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{WZ_i}\left[L_{\mu'}(W) - \ell(W, Z_i)\right].\end{aligned}$$

From here, we can use the same argument as in the proof of Theorem 1 to upper and lower bound the term $\mathbb{E}_{WZ_i}\left[L_{\mu'}(W) - \ell(W, Z_i)\right]$ to arrive at the result. Notice that here, we do not assume that W is the solution of the ERM algorithm. □

APPENDIX C
MULTI-SOURCE TRANSFER LEARNING PROBLEM

In this subsection, we consider the learning regime that we have more than one source domain. Specifically, assume that the target domain consists of n_t samples drawn IID from μ', e.g., $S' = \{Z'_1, Z'_2, \cdots, Z'_{n_t}\}$. We also have k source domains $S_1^k = (S_1, S_2, \cdots, S_k)$ and each source domain S_i consists of n_i samples IID drawn from μ_i, e.g., $S_i = \{Z_{i,1}, Z_{i,2}, \cdots, Z_{i,n_i}\}$, for $i = 1, \cdots, k$. Then we assign the weight α_i to the source domain i and the weight α_t for the target domain where $\sum_{i=1}^{k}\alpha_i + \alpha_t = 1$. We then define the empirical loss by

$$\hat{L}_\alpha(w, S_1^k, S') = \sum_{i=1}^{k}\alpha_i\hat{L}(w, S_i) + \alpha_t\hat{L}(w, S'). \tag{66}$$

and the corresponding combined expected risk is defined by

$$L_\alpha(w) = \sum_{i=1}^{k}\alpha_i L_{\mu_i}(w, S_i) + \alpha_t L_{\mu'}(w). \tag{67}$$

Then the generalization error for the multi-source transfer learning is defined as

$$\text{gen}(w, S_1^k, S') = L_{\mu'}(w) - \hat{L}_\alpha(w, S_1^k, S'). \tag{68}$$

We can easily extend the result of the generalization error from the single source domain to multiple source domains.

Corollary 7 (Generalization error for multi-source transfer learning)**.** *Let P_W be the marginal distribution induced by S_1^k, S' and $P_{W|S_1^k S'}$ for some algorithm (not necessarily the ERM solution). If $\ell(W, Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu'$, then the expectation of the generalization error is upper bounded as*

$$|\mathbb{E}_{WS_1^k S'}\left[\text{gen}(W, S_1^k, S')\right]| \leq \frac{\alpha_t\sqrt{2r^2}}{n_t}\sum_{i=1}^{n_t}\sqrt{I(W; Z'_i)} + \sum_{i=1}^{k}\frac{(1-\alpha_i)\sqrt{2r^2}}{n_i}\sum_{j=1}^{n_i}\sqrt{(I(W; Z_{i,j}) + D(\mu_i||\mu'))}. \tag{69}$$

For ERM solution, we can decompose the excess risk in the following way:

$$\begin{aligned} L_{\mu'}(w_{\mathsf{ERM}}) - L_{\mu'}(w^*) &= L_{\mu'}(w_{\mathsf{ERM}}) - \hat{L}_\alpha(w_{\mathsf{ERM}}, S_1^k, S') + \hat{L}_\alpha(w_{\mathsf{ERM}}, S_1^k, S') - \hat{L}_\alpha(w^*, S_1^k, S') \\ &\quad + \hat{L}_\alpha(w^*, S_1^k, S') - L_\alpha(w^*) + L_\alpha(w^*) - L_{\mu'}(w^*) \\ &\leq \mathrm{gen}(w_{\mathsf{ERM}}, S, S') + \hat{L}_\alpha(w^*, S_1^k, S') - L_\alpha(w^*) \\ &\quad + \sum_{i=1}^k \alpha_i (L_\mu(w^*) - L_{\mu'}(w^*)). \end{aligned} \tag{70}$$

Therefore, we can give the upper bounds on the excess risk.

Corollary 8 (Excess risk for multi-source transfer learning)**.** *Let P_W be the marginal distribution induced by S_1^k, S' and $P_{W|S_1^k S'}$ for the ERM algorithm, assume the loss function $\ell(W, Z)$ is r^2-subgaussian under the distribution $P_W \otimes \mu'$. Then the following inequality holds.*

$$\begin{aligned} \mathbb{E}_W[R_{\mu'}(W_{\mathsf{ERM}})] \leq & \frac{\alpha_t \sqrt{2r^2}}{n_t} \sum_{i=1}^{n_t} \sqrt{I(W_{\mathsf{ERM}}; Z_i')} + \sum_{i=1}^k \frac{\alpha_i \sqrt{2r^2}}{n_i} \sum_{j=1}^{n_i} \sqrt{(I(W_{\mathsf{ERM}}; Z_{i,j}) + D(\mu_i || \mu'))} \\ & + \sum_{i=1}^k \alpha_i d_{\mathcal{W}}(\mu_i, \mu'). \end{aligned} \tag{71}$$

Remark 13. *To minimize the upper bound, we will be optimizing α_t and α_i for target and source domains, which is not trivial as these weights are also implicitly embedded in the mutual information $I(W_{\mathsf{ERM}}; Z_i)$. Intuitively, less weights need to be assigned to those source domains which have large KL divergence $D(\mu_i \| \mu')$. To this end, one can apply the InfoBoost algorithm for optimizing α.*

APPENDIX D
TRANSFER LEARNING WITH PRE-TRAINED HYPOTHESIS

In transfer learning, it is quite common to first pre-train on the source domain and subsequently fine-tune on the target dataset. This methodology has been extensively adopted and affirmed across various situations, especially in the setups of deep learning [83, 84]. Using our method, we can further apply this bounding technique to the pre-trained hypothesis.

First, let us introduce some new notations for this particular setup: we denote the target dataset by $S' = \left\{Z_1', Z_2', \cdots, Z_{n_t}'\right\}$ and the source dataset by $S = \{Z_1, Z_2, \cdots, Z_{n_s}\}$. For deep learning model pre-training, it is common for the source and target to have some shared layers, represented by w_C. Initially, we train a model $w_{\text{source}} = (w_C, w_S)$ using the source data, where w_S denotes the domain-specific model parameters. Subsequently, we use w_C as a fixed layer and fine-tune the model $w_{\text{target}} = (w_C, w_T)$ for the target domain using the target data. Let $\ell(w_C, w_{\text{domain}}, Z)$ be the loss function where w_{domain} denotes the domain specific model parameters. In this case, the learning procedures are described as follows:

- We initially train w_{source} by ERM with the source data:

$$w_C, w_S = \operatorname*{argmin}_{w_C, w_S} \frac{1}{n_s} \sum_{i=1}^{n_s} \ell(w_C, w_S, Z_i') \tag{72}$$

- Given the trained w_C, we next train w_T by ERM with the target data:

$$w_T = \operatorname*{argmin}_{w_T} \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(w_C, w_T, Z_i') \tag{73}$$

- The output w_T and w_C will be tested with the data from the target distribution.

With the above definitions, we can then define the generalization error on the target domain as:

$$\mathrm{gen}(w_T, w_C, S') = \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(w_C, w_T, Z_i') - \mathbb{E}_{Z_i'}\left[\ell(w_C, w_T, Z_i')\right] \tag{74}$$

Now we can state the theorem as follows.

Theorem 10. *Let (w_C, w_S) be the hypothesis learned from the source data S. For any w_C, we assume that the loss function is r^2-subgaussian under the distribution $P_{W_T|w_C} \otimes \mu'$ where W_T is the hypothesis learned from the target data S' by fixing w_C. Then the generalization error can be bounded as:*

$$\mathbb{E}_{W_T W_C S'}\left[gen(W_T, W_C, S')\right] \leq \frac{1}{n_t}\sum_{i=1}^{n_t}\sqrt{2r^2 I(W_T; Z_i'|W_C)}. \tag{75}$$

Proof. The proof of the above theorem simply follows with re-arranging the Donsker-Varadhan variational representation for the KL divergence $D(P_{W_T Z_i'|w_C}\|P_{W_T\otimes\mu'|w_C})$, and for any w_C we will have that for each Z_i':

$$\left|\mathbb{E}_{W_T Z_i'|w_C}\left[\ell(W_T, w_C, Z_i')\right]\right| - \mathbb{E}_{W_T\otimes\mu'|w_C}\left[\ell(W_T, w_C, Z_i')\right] \leq \sqrt{2r^2 D(P_{W_T Z_i'|w_C}\|P_{W_T\otimes\mu'|w_C})}. \tag{76}$$

Then we have:

$$\begin{aligned}\left|\mathbb{E}_{W_T W_C S'}\left[\mathrm{gen}(W_T, W_C, S')\right]\right| \leq& \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbb{E}_{W_C}\left[\sqrt{2r^2 I(W_T; Z_i'|w_C)}\right] &(77)\\ \leq& \frac{1}{n_t}\sum_{i=1}^{n_t}\left[\sqrt{2r^2 I(W_T; Z_i'|W_C)}\right], &(78)\end{aligned}$$

and the second inequality follows from Jensen's inequality. Then it is also natural to derive the excess risk bound:

$$\begin{aligned}\mathbb{E}_{W_T W_C}\left[R_{\mu'}(W_T, W_C)\right] \leq& \frac{1}{n_t}\sum_{i=1}^{n_t}\sqrt{2r^2 I(W_T; Z_i'|W_C)} &(79)\\ &+ \frac{1}{n_t}\sum_{i=1}^{n_t}\left(\mathbb{E}_{W_C W_T\otimes Z_i'}\left[\ell(W_C, W_T, Z_i')\right] - \arg\min_{w_T, w_C}\mathbb{E}_{Z_i'}\left[\ell(w_C, w_T, Z_i')\right]\right) &(80)\end{aligned}$$

since $w^* = \arg\min_{w_T, w_C}\mathbb{E}_{Z_i'}\left[\ell(w_C, w_T, Z_i')\right]$. □

The above theorem has several implications:

- If not using w_C, we then don't use any source data, and we can recover the typical results in [85].
- If we totally depend on w_C, the generalization error will be zero, but the excess risk may be large due to the second difference term.
- The insights from this bound is that the pre-trained model w_C plays an important role in the generalization error: a good source-induced hypothesis would ensure that w_C would lead to low $\arg\min_{w_T}\mathbb{E}_{Z_i'}\left[\ell(w_C, w_T, Z_i')\right]$ and a good fine-tuning would ensure that the conditional mutual information $I(W_T; Z_i'|w_C)$ is small given w_C.

While the outcome of the above theorem appears straightforward and intuitive, it offers only minimal direction regarding the training of w_C. At a glance, one might expect that a smaller domain divergence would lead w_C closer to w_C^*. However, the manner in which the source influences the common hypothesis w_C remains somewhat implicit.

APPENDIX E
TRANSFER LEARNING WITH UNLABELED SOURCE DATA

Unsupervised transfer learning has been extensively explored in numerous literature, especially in the medical area where the data may be difficult to get [86, 87]. In many real-world situations, there are often abundant unlabeled source data but a limited amount of labeled target data accessible for training. In this section, we will examine the generalization error within the domain of unlabeled source data, and the

framework can be readily adapted to encompass both labeled source and unlabeled target data regions. Let $S' = (x'_1, x'_2, \cdots, z'_{\beta n}) \in (\mathcal{X} \times \mathcal{Y})^{\beta n}$ and each $Z'_i = (x'_i, y'_i)$ is a feature-label pair. For the source data, let $S = (X_{\beta n+1}, X_{\beta n+2}, \cdots, X_n) \in \mathcal{X}^{(1-\beta)n}$ as each instance X_i is unlabeled, where $\mathcal{X}$ and $\mathcal{Y}$ are the feature and label spaces. Assume each source instance is i.i.d. drawn from μ_X, and each target pair is i.i.d. drawn from μ'_{XY}. For the loss function, we make the definitions for the supervised and unsupervised metrics (also leveraged from [69]):

- Supervised loss function: A loss function $\ell : \mathcal{W} \times \mathcal{X} \times Y \to \mathbb{R}$ that measures the performance of the prediction.
- Unsupervised loss function: A loss function $\ell_u : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$ that measures the performance on the unlabeled data.

Then we can define the empirical risk similarly as (3):

$$\hat{L}_\alpha(w, S, S') = \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \ell(w, x'_i, y'_i) + \frac{1-\alpha}{(1-\beta)n} \ell_u(w, x_i).$$

The excess risk is defined as:

$$L_{\mu'}(w) = \mathbb{E}_{\mu'_{XY}} \left[\ell(w, X', Y') \right].$$

Then we define the generalization error as:

$$\text{gen}(w, S, S') = \hat{L}_\alpha(w, S, S') - L_{\mu'}(w).$$

Now we state the results as follows:

Theorem 11. *Assume that the supervised loss functions $\ell(w, x, y)$ is r_l^2-subgaussian under the $\mu'_X \otimes \mu'_{Y|X}$ for all $w \in W$ and $\ell_u(w, x)$ is r_u^2-subgaussian under marginal distribution μ'_X for all $w \in W$. The following upper bound holds on the expected generalization error:*

$$\begin{aligned} |gen(w, S, S')| \leq & \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \sqrt{2r_l^2 I(W; X'_i, Y'_i)} + \frac{(1-\alpha)}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \sqrt{2r_u^2 I(W; X'_i) + D(\mu_X \| \mu'_X)} \\ & + (1-\alpha) \mathbb{E}_{W \otimes \mu'_X} [\ell_u(W, X) - \mathbb{E}_{\mu'_{Y|X}} [\ell(W, X, Y)]]. \end{aligned}$$

Proof. The proof simply follows the decomposition of $\text{gen}(w, S, S')$ as:

$$\begin{aligned} \text{gen}(w, S, S') &= \hat{L}_\alpha(w, S, S') - L_{\mu'}(w) \\ &= \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \left(\ell(w, x'_i, y'_i) - \mathbb{E}_{\mu'_{XY}} [\ell(w, x'_i, y'_i)] \right) && (81) \\ &\quad + \frac{1-\alpha}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \left(\ell_u(w, x_i) - \mathbb{E}_{\mu'_X} [\ell_u(w, X_i)] \right) && (82) \\ &\quad + (1-\alpha) \left(\mathbb{E}_{\mu'_X} [\ell_u(w, X)] - \mathbb{E}_{\mu'_{XY}} [\ell(w, x'_i, y'_i)] \right) && (83) \end{aligned}$$

From the assumption, we follow similar procedures as in Corollary 2 as:

$$\mathbb{E}_{W X'_i Y'_i} [\ell(W, X'_i, Y'_i)] - \mathbb{E}_{W \otimes \mu'_{XY}} [\ell(W, X'_i, Y'_i)] \leq \sqrt{2r_l^2 I(W; X'_i, Y'_i)},$$

and

$$\mathbb{E}_{W X_i} [\ell_u(W, X_i)] - \mathbb{E}_{W \otimes \mu'_X} [\ell_u(W, X_i)] \leq \sqrt{2r_u^2 I(W; X_i) + D(\mu_X \| \mu'_X)}.$$

By taking the expectation over the generalization error, we get:

$$|\mathbb{E}_{WSS'} \left[\text{gen}(w, S, S') \right]| \leq \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \sqrt{2r_l^2 I(W; X'_i, Y'_i)} + \frac{(1-\alpha)}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \sqrt{2r_u^2 I(W; X'_i) + D(\mu_X \| \mu'_X)}$$

$$+ (1-\alpha)\mathbb{E}_{W\otimes\mu'_X}[\ell_u(W,X) - \mathbb{E}_{\mu'_{Y|X}}[\ell(W,X,Y)]].$$

□

The above results are very similar to Corollary 1 and Theorem 1 for the supervised settings. However, there are still some new implications:

- Since we do not have the labels for the source domain, the second term in the R.H.S. only contains the domain divergence of the source feature. However, we do have an extra term $|\mathbb{E}_{W\otimes\mu'_X}[\ell_u(W,X) - \mathbb{E}_{\mu'_{Y|X}}[\ell(W,X,Y)]]|$ that captures the conditional distribution shift.
- The effect of the conditional distribution shift is reflected in the choice of the unsupervised loss function ℓ_u, for any w and x, if $\ell_u(w,x)$ can approach $\mathbb{E}_{\mu'_{Y|X}}[\ell(w,x,Y)]$, then we would have a better transfer.
- This now is directly related to the pseudo-labeling techniques for choosing the ℓ_u, say if we rewrite $\ell_u(w,x) = \mathbb{E}_{\hat{\mu}_{Y|x}}[\ell(w,x,Y)]$, then we can directly view the unsupervised loss function as the average loss of $\ell(w,x,Y)$ under the approximated distribution $\hat{\mu}_{Y|x}$, which could be derived by many empirical methods such as deep learning and statistical transfer methods.
- Overall, if we want to achieve a small generalization error with the unlabeled target data, we may require that the algorithm would yield low mutual information, the divergence between μ_X and μ'_X should also be small, and the approximated conditional distribution $\hat{\mu}_{Y|x}$ that is used in pseudo-labeling should be close to the true conditional distribution $\mu'_{Y|x}$.

APPENDIX F
PROOF OF THEOREM 2

The proof is built up on (8):

$$L_{\mu'}(w_{\mathsf{ERM}}) - L_{\mu'}(w^*) \leq \mathrm{gen}(w_{\mathsf{ERM}}, S, S') + \hat{L}_\alpha(w^*) - L_\alpha(w^*) + (1-\alpha)(L_\mu(w^*) - L_{\mu'}(w^*)). \tag{84}$$

The Corollary 2 provides an upper bound on the generalization error $\mathrm{gen}(w_{\mathsf{ERM}}, S, S')$, and the claim follows immediately as the expectation of $\hat{L}_\alpha(w^*) - L_\alpha(w^*)$ is zero.

APPENDIX G
PROOF OF THEOREM 3

Proof. We will build up on the decomposition of the expected excess risk:

$$\begin{aligned}\mathbb{E}_W[L_{\mu'}(W) - L_{\mu'}(w^*)] = \mathbb{E}_W[L_{\mu'}(W) - L_{\mu'}(w^*)] - \mathbb{E}_{WSS'}[\hat{L}_\alpha(W) - L_\alpha(w^*)] \\ + \mathbb{E}_{WSS'}[\hat{L}_\alpha(W) - \hat{L}_\alpha(w^*)],\end{aligned} \tag{85}$$

with the fact that $\mathbb{E}_{SS'}[\hat{L}_\alpha(w^*,S,S')] = L_\alpha(w^*)$. For the target domain, we rewrite the excess risk for any w by:

$$R_{\mu'}(w) = \mathbb{E}_{\mu'}[\ell(w,Z)] - \mathbb{E}_{\mu'}[\ell(w^*,Z)] = \mathbb{E}_{S'}[\hat{R}(w,S')]. \tag{86}$$

Then the expected excess risk in (85) can be written as,

$$\mathbb{E}_W[L_{\mu'}(W) - L_{\mu'}(w^*)] = \mathbb{E}_{WSS'}[\mathcal{E}(W,S,S')] + \mathbb{E}_{WSS'}[\hat{R}_\alpha(W,S,S')]. \tag{87}$$

We will bound the first term by taking the expectation w.r.t. W learned from S and S' as follows.

$$\mathbb{E}_{WSS'}[\mathcal{E}(W,S,S')] = \alpha\left(\mathbb{E}_{P_W\otimes\mu'^{\otimes\beta n}}[\hat{R}(W,S')] - \mathbb{E}_{WS'}[\hat{R}(W,S')]\right) \tag{88}$$

$$+ (1-\alpha)\left(\mathbb{E}_{P_W\otimes\mu'^{\otimes(1-\beta)n}}[\hat{R}(W,S)] - \mathbb{E}_{WS}[\hat{R}(W,S)]\right) \tag{89}$$

$$= \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\mathbb{E}_{P_W\otimes\mu'}[r(W,Z'_i)] - \mathbb{E}_{WZ'_i}[r(W,Z'_i)] \tag{90}$$

$$+\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i)]-\mathbb{E}_{WZ_i}[r(W,Z_i)]. \tag{91}$$

Again, we use the variational representation of the KL divergence between two distributions P and Q defined over $\mathcal{X}$ is given as

$$D(P||Q)=\sup_f\left\{\mathbb{E}_P\left[f(X)\right]-\log\mathbb{E}_Q\left[e^{f(x)}\right]\right\}. \tag{92}$$

We firstly examine the summation of the target data portion in (90) for $i=1,2,\cdots,\beta n$. Under the expected (η,c)-central condition, for any $0<\eta'\leq\eta$, let $f(w,z_i')=-\eta' r(w,z_i')$, we have,

$$\begin{aligned}
D(P_{WZ_i'}\|P_W\otimes P_{Z_i'}) &\geq \mathbb{E}_{P_{WZ_i'}}\left[-\eta' r(W,Z_i')\right]-\log\mathbb{E}_{P_W\otimes\mu'}[e^{-\eta'(r(W,Z_i'))}]\\
&=\mathbb{E}_{P_{WZ_i'}}\left[-\eta' r(W,Z_i')\right]-\log\mathbb{E}_{P_W\otimes\mu'}[e^{-\eta'(r(W,Z_i')-\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')])}]+\mathbb{E}_{P_W\otimes\mu'}\left[\eta' r(W,Z_i')\right]\\
&=\eta'\left(\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]\right)-\log\mathbb{E}_{P_W\otimes\mu'}[e^{\eta'(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}].
\end{aligned} \tag{93}$$

Next we will upper bound the second term $\log\mathbb{E}_{P_W\otimes\mu'}[e^{\eta'(\mathbb{E}_{P_W\otimes\mu}[r(W,Z_i')]-r(W,Z_i'))}]$ in R.H.S. using the expected (η,c)-central condition. From the (η,c)-central condition, we have,

$$\log\mathbb{E}_{P_W\otimes\mu'}[e^{\eta(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}]\leq(1-c)\eta\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]. \tag{94}$$

Since $\eta'\leq\eta$, Jensen's inequality yields:

$$\begin{aligned}
\log\mathbb{E}_{P_W\otimes\mu'}[e^{\eta'(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}] &= \log\mathbb{E}_{P_W\otimes\mu'}[e^{\frac{\eta'}{\eta}\eta(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}] &(95)\\
&\leq \log\left(\mathbb{E}_{P_W\otimes\mu'}[e^{\eta(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}]\right)^{\frac{\eta'}{\eta}} &(96)\\
&\leq \frac{\eta'}{\eta}(1-c)\eta\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')] &(97)\\
&= \eta'(1-c)\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]. &(98)
\end{aligned}$$

Substitute (98) into (93), we arrive at,

$$I(W;Z_i')\geq\eta'\left(\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]\right)-(1-c)\eta'\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]. \tag{99}$$

Divide η' on both sides, we have that

$$\frac{I(W;Z_i')}{\eta'}\geq\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]-(1-c)\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]. \tag{100}$$

Rearrange the equation and yields,

$$c\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]\leq\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]+\frac{I(W;Z_i')}{\eta'}. \tag{101}$$

Therefore,

$$\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]\leq(\frac{1}{c}-1)\left(\mathbb{E}_{P_{WZ_i'}}[r(w,Z_i')]\right)+\frac{I(W;Z_i')}{c\eta'}. \tag{102}$$

Using the similar arguments for the source domain portion in (91), we have that for any $i=\beta n+1,\cdots,n$,

$$\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i)\right]-\mathbb{E}_{P_{WZ_i}}\left[r(W,Z_i)\right]\leq(\frac{1}{c}-1)\left(\mathbb{E}_{P_{WZ_i}}[r(w,Z_i)]\right)+\frac{I(W;Z_i)+D(\mu\|\mu')}{c\eta'}. \tag{103}$$

Summing up every term for Z_i, we end up with:

$$\mathbb{E}_{WSS'}\left[\mathcal{E}(W,S,S')\right]\leq\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W;Z_i')+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W;Z_i)+D(\mu\|\mu')\right)$$

$$+(\frac{1}{c}-1)\left(\mathbb{E}_{WSS'}[\alpha\hat{R}(W,S')+(1-\alpha)\hat{R}(W,S)]\right). \tag{104}$$

Since $\frac{1}{c}-1>0$ and $\mathbb{E}_{WSS'}[\alpha\hat{R}(W,S')+(1-\alpha)\hat{R}(W,S)]$ will be negative for W_{ERM}, we completes the proof for ERM by:

$$\mathbb{E}_W\left[R_{\mu'}(W_{\mathsf{ERM}})\right]=\mathbb{E}_{WSS'}\left[\mathcal{E}(W_{\mathsf{ERM}},S,S')\right]+\mathbb{E}_{WSS'}[\hat{R}_\alpha(W_{\mathsf{ERM}},S,S')] \tag{105}$$

$$\leq\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{ERM}};Z_i')+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{ERM}};Z_i)+D(\mu\|\mu')\right). \tag{106}$$

If (η,c)-central condition holds for general $\hat{W}$, following the same analysis, we arrive at,

$$\mathbb{E}_{\hat{W}SS'}\left[\mathcal{E}(\hat{W},S,S')\right]\leq\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(\hat{W};Z_i')+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(\hat{W};Z_i)+D(\mu\|\mu')\right)$$

$$+(\frac{1}{c}-1)\mathbb{E}_{\hat{W}SS'}\left[\hat{R}_\alpha(\hat{W},S,S')\right]. \tag{107}$$

Therefore, we complete the proof by,

$$\mathbb{E}_{\hat{W}}\left[R_{\mu'}(\hat{W})\right]=\mathbb{E}_{\hat{W}SS'}\left[\mathcal{E}(\hat{W},S,S')\right]+\mathbb{E}_{\hat{W}SS'}\left[\hat{R}_\alpha(\hat{W},S,S')\right] \tag{108}$$

$$\leq\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(\hat{W};Z_i')+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(\hat{W};Z_i)+D(\mu\|\mu')\right)+\frac{1}{c}\mathbb{E}_{\hat{W}SS'}\left[\hat{R}_\alpha(\hat{W},S,S')\right]. \tag{109}$$

□

APPENDIX H
PROOF OF LEMMA 3

Proof. We firstly define the combined regularized loss as,

$$\hat{L}_{\text{reg}}(w,S,S'):=\hat{L}_\alpha(w,S,S')+\frac{\lambda}{n}g(w). \tag{110}$$

Based on Theorem 3, we can bound the excess risk for W_{RERM} by,

$$\mathbb{E}_W[R_{\mu'}(W_{\mathsf{RERM}})]\leq\frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_\alpha\left(W_{\mathsf{RERM}},S,S'\right)]+\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{RERM}};Z_i')$$

$$+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}};Z_i)+D(\mu\|\mu')\right) \tag{111}$$

$$=\frac{1}{c}\mathbb{E}_{WSS'}[\hat{L}_\alpha\left(W_{\mathsf{RERM}},S,S'\right)-\hat{L}_\alpha\left(w^*,S,S'\right)]+\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{RERM}};Z_i')$$

$$+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}};Z_i)+D(\mu\|\mu')\right) \tag{112}$$

$$\overset{(a)}{\leq}\frac{1}{c}\left(\mathbb{E}_{WSS'}[\hat{L}_{\text{reg}}\left(W_{\mathsf{RERM}},S,S'\right)]-\mathbb{E}_{SS'}[\hat{L}_{\text{reg}}\left(w^*,S,S'\right)]\right)+\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{RERM}};Z_i')$$

$$+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}};Z_i)+D(\mu\|\mu')\right)+\frac{\lambda B}{cn} \tag{113}$$

$$=\frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_{\text{reg}}\left(W_{\mathsf{RERM}},S,S'\right)]+\frac{\lambda B}{cn}+\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{RERM}};Z_i')$$
$$+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}};Z_i)+D(\mu\|\mu')\right) \tag{114}$$
$$\overset{(b)}{\leq}\frac{\lambda B}{cn}+\frac{1}{c\eta'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}I(W_{\mathsf{RERM}};Z_i')+\frac{1}{c\eta'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W_{\mathsf{RERM}};Z_i)+D(\mu\|\mu')\right), \tag{115}$$

where (a) follows since $|g(w^*)-g(W_{\mathsf{RERM}}))|\leq B$ the expected empirical risk is negative for W_{ERM} and (b) holds due to that W_{RERM} is the minimizer of the regularized loss. □

APPENDIX I
PROOF OF LEMMA 4

Proof. Firstly we examine the bounds for target instances z_i' for $i=1,2,\cdots,\beta n$. We will build upon (93). With the (v,c)-central condition, for any $\epsilon\geq 0$ and any $0<\eta'\leq v(\epsilon)$, the Jensen's inequality yields:

$$\log\mathbb{E}_{P_W\otimes\mu'}[e^{\eta'(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}]=\log\mathbb{E}_{P_W\otimes\mu'}[e^{\frac{\eta'}{v(\epsilon)}v(\epsilon)(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}] \tag{116}$$
$$\leq\log\left(\mathbb{E}_{P_W\otimes\mu'}[e^{v(\epsilon)(\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-r(W,Z_i'))}]\right)^{\frac{\eta'}{v(\epsilon)}} \tag{117}$$
$$\leq\frac{\eta'}{v(\epsilon)}\left((1-c)v(\epsilon)\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]+v(\epsilon)\epsilon\right) \tag{118}$$
$$=\eta'(1-c)\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]+\eta'\epsilon. \tag{119}$$

Substitute (119) into (93), we arrive at,

$$I(W;Z_i')\geq\eta'\left(\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]\right)-(1-c)\eta'\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-\eta'\epsilon. \tag{120}$$

Divide η' on both sides, we arrive at,

$$\frac{I(W;Z_i')}{\eta'}\geq\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]-(1-c)\mathbb{E}_{P_W\otimes\mu'}[r(W,Z_i')]-\epsilon. \tag{121}$$

Rearrange the equation and yields,

$$c\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]\leq\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]+\frac{I(W;Z_i')}{\eta'}+\epsilon. \tag{122}$$

Therefore,

$$\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i')\right]-\mathbb{E}_{P_{WZ_i'}}\left[r(W,Z_i')\right]\leq(\frac{1}{c}-1)\left(\mathbb{E}_{P_{WZ_i'}}[r(w,Z_i')]\right)+\frac{I(W;Z_i')}{c\eta'}+\frac{\epsilon}{c}. \tag{123}$$

Regarding the source instances, following the same procedures we have the following inequality for $z_i, i=\beta n+1,\cdots,n$:

$$\mathbb{E}_{P_W\otimes\mu'}\left[r(W,Z_i)\right]-\mathbb{E}_{P_{WZ_i}}\left[r(W,Z_i)\right]\leq(\frac{1}{c}-1)\left(\mathbb{E}_{P_{WZ_i}}[r(w,Z_i)]\right)+\frac{I(W;Z_i)+D(\mu\|\mu')}{c\eta'}+\frac{\epsilon}{c}. \tag{124}$$

Summing up every term for Z_i and dividing the summation by n, we end up with,

$$\mathbb{E}_W[L_{\mu'}(W)-L_{\mu'}(w^*)]\leq\frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_\alpha\left(W,S,S'\right)]+\frac{\alpha}{c\beta n}\sum_{i=1}^{\beta n}\left(\frac{I(W;Z_i)}{\eta'}+\epsilon\right)$$
$$+\frac{1-\alpha}{c(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\frac{I(W;Z_i)+D(\mu\|\mu')}{\eta'}+\epsilon\right). \tag{125}$$

In particular, if $v(\epsilon) = \epsilon^{1-\beta}$ for some $\beta \in [0,1]$, then by choosing $\eta' = v(\epsilon)$ and $\frac{I(W;Z_i)}{c\eta'} + \frac{\epsilon}{c}$ is optimized when $\epsilon = I(W;Z_i)^{\frac{1}{2-\beta}}$ for the target instances and $\epsilon = (I(W;Z_i) + D(\mu\|\mu'))^{\frac{1}{2-\beta}}$ for the source instances. Then the bound becomes,

$$\begin{aligned}\mathbb{E}_W[L_{\mu'}(W) - L_{\mu'}(w^*)] \leq & \frac{1}{c}\mathbb{E}_{WSS'}[\hat{R}_\alpha\left(W,S,S'\right)] + \frac{\alpha}{c\beta n}\sum_{i=1}^{\beta n} I(W_{\mathsf{ERM}};Z_i)^{\frac{1}{2-\beta}} \\ & + \frac{1-\alpha}{c(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I(W;Z_i) + D(\mu\|\mu')\right)^{\frac{1}{2-\beta}}, \end{aligned} \tag{126}$$

which completes the proof. □

APPENDIX J
PROOF OF THEOREM 5

Proof. The following lemma is used to prove the theorem.

Lemma 4. *For all t, if the noise $n(t) \sim \mathcal{N}(0, \sigma_t I_d)$, we have*

$$I\left(W(t);S|W(t-1)\right) \leq \frac{d}{2}\log\left(1 + \frac{\eta_t^2 K_{ST}^2}{d\sigma_t^2}\right),$$

$$I\left(W(t);S'|W(t-1)\right) \leq \frac{d}{2}\log\left(1 + \frac{\eta_t^2 K_{ST}^2}{d\sigma_t^2}\right).$$

Proof of Lemma 4. From the definition of mutual entropy

$$I\left(W(t);S|W(t-1)\right) = h\left(W(t)|W(t-1)\right) - h\left(w(t)|W(t-1),S\right).$$

Each term can be bounded separately. First, we have the update rule on $W(t)$:

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

Note that

$$h\left(W(t) - W(t-1)|W(t-1)\right) = h\left(W(t)|W(t-1)\right),$$

since the subtraction term does not affect the entropy of a random variable. Also the perturbation $n(t)$ is independent of the gradient term, thus we can compute the upper bound of the expected squared norm of $w(t) - w(t-1)$:

$$\begin{aligned}\mathbb{E}\left[\|W(t) - W(t-1)\|_2^2\right] &= \mathbb{E}\left[\left\|\eta_t(\alpha\nabla(\hat{L}_\alpha(W(t-1),S') + (1-\alpha)\nabla(\hat{L}_\alpha(W(t-1),S))\right\|_2^2 + \|n_t\|_2^2\right] \\ &\leq \eta_t^2(\alpha K_S + (1-\alpha)K_T)^2 + d\sigma_t^2 \\ &\leq \eta_t^2 K_{ST}^2 + d\sigma_t^2\end{aligned}$$

where in the expression above, we used the assumption that $n_t \sim N(0, \sigma_t^2 I_d)$. Among all random variables X with a fixed expectation bound $\mathbb{E}\|X\|_2^2 < A$, then the norm distribution $Y \sim N(0, \sqrt{\frac{A}{d}} I_d)$ has the largest entropy given by:

$$h(Y) = d\log\left(\sqrt{2\pi e\sigma_Y^2}\right) = \frac{d}{2}\log\left(\frac{2\pi e A}{d}\right)$$

which indicates that:

$$h\left(W(t)|W(t-1)\right) \leq \frac{d}{2}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right)$$

By entropy power inequality [88], we have:

$$\begin{aligned} h\left(W(t)|W(t-1),S\right) =& h\left(W(t-1)+\eta_t\nabla\hat{L}_\alpha(W(t-1),S,S')+n_t|W(t-1),S\right)\\ =& h\left(n_t+\eta_t\alpha\nabla\hat{L}_\alpha(W(t-1),S')|W(t-1),S\right)\\ \geq& \frac{1}{2}\log(e^{2h(n_t)}+e^{2h\left(\eta_t\alpha\nabla\hat{L}_\alpha(W(t-1),S')|W(t-1),S\right)})\\ \geq& h\left(n_t\right). \end{aligned}$$

This leads to the following desired bound for the mutual entropy $I(W(t);S|W(t-1))$:

$$h\left(W(t)|W(t-1)\right)-h\left(W(t)|S,W(t-1)\right)\leq\frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-h(n_t)$$

Similarly, we can achieve the upper bound for the mutual entropy $I(W(t);S'|W(t-1))$:

$$h\left(W(t)|W(t-1)\right)-h\left(W(t)|S',W(t-1)\right)\leq\frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-h(n_t)$$

Therefore, consider the mutual information $I(W(t);S'|W(t-1))$ and $I(W(t);S|W(t-1))$ with Gaussian noise $n(t)$, e.g., $h(t)=\frac{d}{2}\log 2\pi e\sigma_t^2$, we can write

$$\begin{aligned} I(W(t);S'|W(t-1)) &= h\left(W(t)|W(t-1)\right)-h\left(W(t)|S',W(t-1)\right)\\ &\leq \frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-\frac{d}{2}\log 2\pi e\sigma_t^2\\ &= \frac{d}{2}\log\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d\sigma_t^2}\\ &= \frac{d}{2}\log\left(1+\frac{\eta_t^2K_{ST}^2}{d\sigma_t^2}\right). \end{aligned}$$

Similarly, we have:

$$I(W(t);S|W(t-1))\leq\frac{d}{2}\log\left(1+\frac{\eta_t^2K_{ST}^2}{d\sigma_t^2}\right).$$

□

With Lemma 4, we reach the following bound by Jensen's inequality:

$$\begin{aligned} \mathbb{E}_{WSS'}\left[\mathrm{gen}\left(W(T),S,S'\right)\right] &\leq \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I\left(W(T);Z_i'\right)}+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{\left(I\left(W(T);Z_i\right)+D\left(\mu\|\mu'\right)\right)}\\ &\leq \alpha\sqrt{\frac{2r^2}{\beta n}I(W(T);S')}+(1-\alpha)\sqrt{2r^2\left(\frac{I(W(T);S)}{(1-\beta)n}+D\left(\mu\|\mu'\right)\right)}. \end{aligned} \tag{127}$$

Let $W^T=(W(1),W(2),W(3),\cdots,W(T))$, with the characteristic of the gradient descent algorithm, we can show that

$$h(W(t)|W^{(t-1)},S)=h(W(t)|W(t-1),S), \tag{128}$$

which follows from the Markov chain that $S\to W(1)\to W(2)\cdots\to W(T)$. Using lemma 4, both the mutual information $I(W(T);S)$ and $I(W(T);S')$ are bounded as:

$$I(W(T);S)\leq I(W^T;S)$$

$$
\begin{aligned}
=&I(W(1);S|W(0)) + I(W(2);S|W(1)) + I(W(3);S|W(2),W(1)) \\
&+ I(W(4);S|(W(3),W(2),W(1))) + \cdots + I(W(T);S|W^{T-1}) \\
=&\sum_{t=1}^{T} I(W(t);S|W(t-1)) \\
\leq&\frac{d}{2}\sum_{t=1}^{T}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right) - \sum_{t=1}^{T} h(n(t)),
\end{aligned}
$$

where the first inequality follows from the Markov chain $S \to W^T$. □

APPENDIX K
PROOF OF COROLLARY 3

We leverage the following proposition that

Proposition 2. *Under the given assumptions, we define $\kappa = \frac{\nu}{\mathcal{L}} \in (0,1)$, setting $\eta = \frac{1}{\mathcal{L}}$, for all $T \geq 1$, we have:*

$$
\begin{aligned}
\hat{L}_\alpha(W(T),S,S') - \hat{L}_\alpha(w_{\mathsf{ERM}},S,S') &\leq K_{ST}\left\|W(T) - w_{\mathsf{ERM}}\right\| \\
&\leq K_{ST}(1-\kappa)^T(\|W(0) - w_{\mathsf{ERM}}\| + \hat{A}_T),
\end{aligned}
$$

where we define $\hat{A}_T$

$$
\hat{A}_T := \sum_{t=1}^{T}(1-\kappa)^{-t}\left\|n(t)\right\|.
$$

We firstly claim that $\hat{L}_\alpha$ is K_{ST}-Lipschitz continuous with K_{ST} bounded gradient, then the proof follows the proposition 3 in the work [58].

Proof. (of Corollary 3) We firstly decompose the excess risk $L_{\mu'}(W(T)) - L_{\mu'}(w^*)$ into five fractions as follows.

$$
\begin{aligned}
L_{\mu'}(W(T)) - L_{\mu'}(w^*) =&L_{\mu'}(W(T)) - \hat{L}_\alpha(W(T),S,S') + \hat{L}_\alpha(W(T),S,S') - \hat{L}_\alpha(W_{\mathsf{ERM}},S,S') \\
&+ \hat{L}_\alpha(W_{\mathsf{ERM}},S,S') - \hat{L}_\alpha(w^*,S,S') + \hat{L}_\alpha(w^*,S,S') - L_\alpha(w^*) + L_\alpha(w^*) - L_{\mu'}(w^*).
\end{aligned}
$$

Following corollary 2, we have

$$
\begin{aligned}
\mathbb{E}\left[L_{\mu'}(W(T)) - \hat{L}_\alpha(W(T),S,S')\right] \leq&\frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I\left(W(T);Z_i'\right)} \\
&+ \frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{\left(I\left(W(T);Z_i\right) + D\left(\mu\|\mu'\right)\right)}. \qquad (129)
\end{aligned}
$$

Then use proposition 2, for any $W(T)$, we reach

$$
\mathbb{E}[\hat{L}_\alpha(W(T),S,S') - \hat{L}_\alpha(W_{\mathsf{ERM}},S,S')] \leq K_{ST}(1-\kappa)^T(\mathbb{E}[\|W(0) - W_{\mathsf{ERM}}\|] + \sum_{t=1}^{T}(1-\kappa)^{-t}\mathbb{E}[\|n(t)\|]). \qquad (130)
$$

The remaining term $\hat{L}_\alpha(w^*,S,S') - L_\alpha(w^*) + L_\alpha(w^*) - \hat{L}_{\mu'}(w^*)$ can be bounded with Theorem 2 for some $w^* \in \mathcal{W}$ that

$$
\mathbb{E}\left[\hat{L}_\alpha(w^*) - L_\alpha(w^*) + L_\alpha(w^*) - \hat{L}_{\mu'}(w^*)\right] \leq (1-\alpha)d_{\mathcal{W}}(\mu,\mu'). \qquad (131)
$$

With the property $\hat{L}_\alpha(W_{\mathsf{ERM}}) - \hat{L}_\alpha(w^*) < 0$, we combine the inequality (129), (130) and (131) and claim the result. □

APPENDIX L
PROOF OF THEOREM 6

Proof. We know that

$$\begin{aligned}
&\inf_{P_{W|S,S'}} \left(\mathbb{E}\left[\hat{L}_\alpha(W,S,S')\right] + \frac{1}{k} D\left(P_{W|S,S'} \| Q | P_{S,S'}\right) \right) \\
&= \inf_{P_{W|S,S'}} \int_{\mathcal{Z}^{\beta n}} (\mu')^{\otimes \beta n}(\mathrm{d}s) \int_{\mathcal{Z}^{(1-\beta)n}} \mu^{\otimes (1-\beta)n}(\mathrm{d}s') \left(\mathbb{E}\left[\hat{L}_\alpha(W,S,S') | S'=s', S=s\right] + \frac{1}{k} D\left(P_{W|S'=s',S=s} \| Q\right) \right) \\
&= \int_{\mathcal{Z}^{\beta n}} (\mu')^{\otimes \beta n}(\mathrm{d}s') \int_{\mathcal{Z}^{(1-\beta)n}} \mu^{\otimes (1-\beta)n}(\mathrm{d}s) \inf_{P_{W|S,S'}} \left(\mathbb{E}\left[\hat{L}_\alpha(W,S,S') | S'=s', S=s\right] + \frac{1}{k} D\left(P_{W|S'=s',S=s} \| Q\right) \right),
\end{aligned}$$

which leads to the well-known Gibbs algorithm such that

$$\begin{aligned}
&\inf_{P_{W|S,S'}} \left(\mathbb{E}\left[\hat{L}_\alpha(W) | S'=s', S=s\right] + \frac{1}{k} D\left(P_{W|S'=s',S=s} \| Q\right) \right) \\
&= \inf_{P_{W|S,S'}} \int P_{W|S,S'}(w) \left(k\hat{L}_\alpha(w,s,s') + \log \frac{P_{W|S,S'}(w)}{Q(w)} \right) dw \\
&= \inf_{P_{W|S,S'}} \int P_{W|S,S'}(w) \log \frac{P_{W|S,S'}(w)}{Q(w) e^{-k\hat{L}_\alpha(w,s,s')} / \mathbb{E}_Q[e^{-k\hat{L}_\alpha(W,s,s')}]} dw.
\end{aligned}$$

Hence,

$$P^*_{W|S'=s',S=s}(\mathrm{d}w) = \frac{e^{-k\hat{L}_\alpha(w,s,s')} Q(\mathrm{d}w)}{\mathbb{E}_Q\left[e^{-k\hat{L}_\alpha(W,s,s')}\right]} \quad \text{for each } s \in \mathcal{Z}^n. \tag{132}$$

□

APPENDIX M
PROOF OF COROLLARY 4

Proof. Since the bounded loss function $\ell(w,z) \in [0,1]$ also satisfies r^2-subgaussian with $r^2 = \frac{1}{4}$, we have

$$\begin{aligned}
|\mathbb{E}_{WSS'}\left[\mathrm{gen}(W_G,S,S')\right]| &\le \frac{\alpha\sqrt{2r^2}}{\beta n} \sum_{i=1}^{\beta n} \sqrt{I(W_G;Z_i')} + \frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \sqrt{(I(W_G;Z_i) + D(\mu||\mu'))} \\
&\le \frac{\alpha}{\beta n} \sum_{i=1}^{\beta n} \sqrt{\frac{I(W_G;Z_i'|S'_{-i},S)}{2}} + \frac{(1-\alpha)}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \sqrt{\frac{I(W_G;Z_i|S',S_{-i}) + D(\mu||\mu')}{2}},
\end{aligned}$$

where we denote S_{-i} by deleting the i-th element in S. The second inequality uses the fact that Z_i and S^{-i} are independent. Then using the Hoeffding inequality with the fact that the loss is bounded by $[0,1]$, we have,

$$\begin{aligned}
I(W_G;Z_i'|S'_{-i},S) &\le \frac{\alpha^2 k^2}{8\beta^2 n^2}, \text{ for } i = 1,\cdots,\beta n \\
I(W_G;Z_i|S_{-i},S') &\le \frac{(1-\alpha)^2 k^2}{8(1-\beta)^2 n^2}, \text{ for } i = \beta n+1,\cdots,n.
\end{aligned}$$

Finally, we arrive at,

$$\begin{aligned}
|\mathbb{E}_{WSS'}\left[\mathrm{gen}(W_G,S,S')\right]| &\le \frac{\alpha^2 k}{4\beta n} + \frac{(1-\alpha)}{(1-\beta)n} \sum_{i=\beta n+1}^{n} \sqrt{\frac{(1-\alpha)^2 k^2}{16(1-\beta)^2 n^2} + \frac{D(\mu\|\mu')}{2}} &(133)\\
&\le \frac{\alpha^2 k}{4\beta n} + \frac{(1-\alpha)^2 k}{4(1-\beta)n} + (1-\alpha)\sqrt{\frac{D(\mu\|\mu')}{2}}, &(134)
\end{aligned}$$

where we use $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for both non-negative x and y in the last inequality. With the fact that $D(P^*_{W_G|S,S'}\|Q|S,S')$ is positive for any Q,

$$\begin{aligned}\mathbb{E}_W[L_{\mu'}(W_G)] &= \mathbb{E}_{WSS'}[L_{\mu'}(W_G) - \hat{L}_\alpha(W_G,S,S')] + \mathbb{E}_{WSS'}[\hat{L}_\alpha(W_G,S,S')] \\ &\leq \mathbb{E}_{WSS'}[\text{gen}(W_G,S,S')] + \mathbb{E}_{WSS'}[\hat{L}_\alpha(W_G,S,S')] + \frac{1}{k}D(P^*_{W_G|S,S'}\|Q|S,S') \\ &\leq \mathbb{E}_{WSS'}[\text{gen}(W_G,S,S')] + \mathbb{E}_{WSS'}[\hat{L}_\alpha(w^*_{st}(\alpha),S,S')] + \frac{1}{k}D(\delta_{w^*_{st}(\alpha)}\|Q|S,S') \end{aligned} \tag{135}$$

as $P^*_{W_G|S,S'}$ is the minimizer of the regularized empirical risk in Equation (42). Since $\mathcal{W}$ is finite, we have that $D(\delta_{w^*_{st}(\alpha)}\|Q|S,S') = -\log\frac{1}{Q(w^*_{st}(\alpha))}$ and $\mathbb{E}_{WSS'}[\hat{L}_\alpha(w^*_{st}(\alpha),S,S')] = L_\alpha(w^*_{st}(\alpha))$, which completes the proof. □

APPENDIX N
PROOF OF COROLLARY 6

Proof. Suppose $\ell(W,Z_i)$ is L_∞-norm upper bounded by σ, the L_∞-norm of a random variable is defined as

$$\|X\|_\infty = \inf\{M : P(X > M) = 0\},$$

then followed by [72, Theorem 3], we have

$$|\mathbb{E}_P[\ell(W,Z_i)] - \mathbb{E}_Q[\ell(W,Z_i)]| \leq 2\|\sigma\|_\infty D_\phi(P\|Q), \tag{136}$$

where $D_\phi(P\|Q) = \frac{1}{2}\int|dP - dQ|$ is referred to as the ϕ-divergence with $\phi(x) = \frac{1}{2}|x-1|$. If $Z'_i \sim \mu'$, $D_\phi(P\|Q) = D_\phi(P_{WZ'}||P_W \otimes \mu') := I_\phi(Z'_i;W)$. If $Z_i \sim \mu$, we have

$$\begin{aligned} D_\phi(P\|Q) &= \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|dP_{W,Z_i} - dP_W d\mu'\right| && (137)\\ &= \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|dP_{W,Z_i} - dP_W d\mu + dP_W d\mu - dP_W d\mu'\right| && (138)\\ &\leq \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|dP_{W,Z_i} - dP_W d\mu\right| + \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|dP_W d\mu - dP_W d\mu'\right| && (139)\\ &= I_\phi(W;Z_i) + TV(\mu,\mu'), && (140)\end{aligned}$$

where $TV(\mu,\mu') = D_\phi(\mu||\mu')$ denotes the total variation distance between the distribution μ and μ'. By this, we can extend the mutual information measure to ϕ-divergence. □

APPENDIX O
PROOF OF THEOREM 7

Proof. We firstly look at the generalization error in terms of $P_{WSS'}$ and $P_W \otimes P_S \otimes P_{S'}$ as,

$$\begin{aligned}&\mathbb{E}_{WSS'}\left[\text{gen}(W_{\mathsf{ERM}},S,S')\right] = \mathbb{E}_{WSS'}\left[L_{\mu'}(W_{\mathsf{ERM}}) - \hat{L}_\alpha(W_{\mathsf{ERM}},S,S')\right] \\ &= \int_{\mathcal{W}\times\mathcal{Z}}\alpha\ell(W_{\mathsf{ERM}},Z') + (1-\alpha)\ell(W_{\mathsf{ERM}},Z)dP_W d\mu' - \int_{\mathcal{W}\times\mathcal{S}\times\mathcal{S}'}\hat{L}_\alpha(W_{\mathsf{ERM}},S,S')dP_{WSS'} \\ &= \int_{\mathcal{W}\times\mathcal{Z}}\alpha\ell(W_{\mathsf{ERM}},Z')dP_W d\mu' - \int_{\mathcal{W}\times\mathcal{S}'}\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\ell(W_{\mathsf{ERM}},Z'_i)dP_{WS'} \\ &\quad + \int_{\mathcal{W}\times\mathcal{Z}}(1-\alpha)\ell(W_{\mathsf{ERM}},Z)dP_W d\mu' - \int_{\mathcal{W}\times\mathcal{S}'}\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\ell(W_{\mathsf{ERM}},Z_i)dP_{WS} \\ &= \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\left(\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z'_i)dP_W d\mu' - \int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z'_i)dP_{WZ'_i}\right)\end{aligned}$$

$$+\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z)dP_W d\mu'-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_{WZ_i}\right)$$

$$=\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\left(\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i')dP_W d\mu'-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i')dP_{WZ_i'}\right) \tag{141}$$

$$+\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z)dP_W d\mu'-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_W d\mu\right) \tag{142}$$

$$+\frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_W d\mu-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_{WZ_i}\right) \tag{143}$$

Theorem 12 (Kantorovich-Rubinstein duality Theorem)**.** *Let $(\mathcal{X},d)$ be a metric space and let μ, v denote two Radon probability measures contained in $\mathcal{P}_d(\mathcal{X})$. Then*

$$\mathbb{W}_1(\mu,v)=\sup\left\{\int_{\mathcal{X}}f\mathrm{d}\mu-\int_{\mathcal{X}}fdv:f\in\mathrm{Lip}_1(\mathcal{X},d)\right\}, \tag{144}$$

where $\mathrm{Lip}_1(\mathcal{X},d)$ denotes the collection of all 1-Lipschitz continuous functions on $\mathcal{X}$.

Using the theorem above, we can bound the generalization error using the Wasserstein distance. By assuming that the loss function is $\mathcal{L}$-Lipschitz for any $Z\in\mathcal{Z}$ and $W\in\mathcal{W}$, then we can bound (141), (142), (143) using the following inequalities:

$$\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i')dP_W d\mu'-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i')dP_{WZ_i'}\leq\mathcal{L}\mathbb{E}_{\mu'}\left[\mathbb{W}_1(P_W,P_{W|Z_i'})\right], \tag{145}$$

$$\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z)dP_W d\mu'-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_W d\mu\leq\mathcal{L}\mathbb{W}_1(\mu',\mu), \tag{146}$$

$$\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_W d\mu-\int_{\mathcal{W}\times\mathcal{Z}}\ell(W_{\mathsf{ERM}},Z_i)dP_{WZ_i}\leq\mathcal{L}\mathbb{E}_{\mu}\left[\mathbb{W}_1(P_W,P_{W|Z_i})\right], \tag{147}$$

respectively, which completest the proof. □

APPENDIX P
PROOF OF PROPOSITION 1

Proof. In the following, we denote the random variable W by the ERM solution W_{ERM} for simplicity. With the fact that $\sqrt{x+y}\geq\sqrt{\frac{x}{2}}+\sqrt{\frac{y}{2}}$ for both $x,y\geq 0$, we can further lower bound the mutual information based quantity $\mathbb{B}_{\text{Info}}$ by,

$$\mathbb{B}_{\text{Info}}=\frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))}\geq\frac{\sqrt{r^2}}{n}\sum_{i=1}^{n}\left(\sqrt{I(W_{\mathsf{ERM}};Z_i)}+\sqrt{D(\mu||\mu')}\right).$$

With the assumption that P_W and μ' satisfy the $T_1(\frac{r^2}{2\mathcal{L}^2})$ transport cost inequality, we have that for $P_{W|z_i}\ll P_W$ with any z_i and $\mu\ll\mu'$,

$$\mathbb{W}_1(P_{W|z_i},P_W)\leq\sqrt{\frac{r^2}{\mathcal{L}^2}D(P_{W|z_i}\|P_W)},$$

$$\mathbb{W}_1(\mu,\mu')\leq\sqrt{\frac{r^2}{\mathcal{L}^2}D(\mu\|\mu')}.$$

By Jensen's inequality, we can show that for all i,

$$\mathbb{E}_{Z_i}\left[\mathcal{L}\mathbb{W}_1(P_{W|Z_i},P_W)\right]\leq\mathbb{E}_{Z_i}\left[\sqrt{r^2D(P_{W|Z_i}\|P_W)}\right]$$

$$\leq \sqrt{r^2 \mathbb{E}_{Z_i}\left[D(P_{W|Z_i}\|P_W)\right]}$$
$$= \sqrt{r^2 I(W;Z_i)},$$

where $I(W;Z_i) = D(P_{WZ_i}\|P_W \otimes P_{Z_i}) = \mathbb{E}_{Z_i}[D(P_{W|Z_i}\|P_W)]$. Similarly, we can also prove that,

$$\mathcal{L}\mathbb{W}_1(\mu,\mu') \leq \sqrt{2r^2 D(\mu\|\mu')},$$

which completes the proof and it naturally shows that the Wasserstein distance based bound is tighter than the mutual information based bound. □

APPENDIX Q
SETUP OF BERNOULLI ADAPTATION

As we use the binary cross-entropy loss function:

$$\ell(w,z_i) = -(z_i \log(w) + (1-z_i)\log(1-w)), \tag{148}$$

the corresponding gradient is given by

$$\nabla\ell(w,z_i) = \frac{1-z_i}{1-w} - \frac{z_i}{w}. \tag{149}$$

Then the population risk is then defined as

$$L_{\mu'}(w(T)) = -(p'\log(w(T)) + (1-p')\log(1-w(T))), \tag{150}$$

and the corresponding empirical risk is defined to be

$$\hat{L}_\alpha\left(w(T),S,S'\right) = \frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\ell\left(w(T),z_i'\right) + \frac{1-\alpha}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\ell\left(w(T),z_i\right). \tag{151}$$

We notice that $\|\nabla\ell(w,z_i)\|$ and $\ell(w,z_i)$ are not bounded in our case if w approaches 0 or 1, however, in the simulation, within finite T iterations, we set the relative iterative parameters to be the maximum value among all iterations such that

$$K_T(T) = \max_{\substack{t=1,2,\cdots,T,\\ i=1,2,\cdots,\beta n}}\left\|\nabla\ell(w(t);z_i)\right\|, \tag{152}$$

$$K_S(T) = \max_{\substack{t=1,2,\cdots,T,\\ i=\beta n+1,\cdots,n}}\left\|\nabla\ell(w(t);z_i)\right\|, \tag{153}$$

$$K_{ST}(T) = \alpha K_T(T) + (1-\alpha)K_S(T), \tag{154}$$

$$\eta(T) = \frac{1}{K_{ST}(T)}, \tag{155}$$

$$r(T) = \max_{t=1,2,\cdots,T}\frac{|\log\frac{w(t)}{1-w(t)}|}{\sqrt{2}}. \tag{156}$$

The generalization error bound after T iterations is given by

$$|\mathbb{E}_{WSS'}\left[\text{gen}(w(T),S,S')\right]| \leq \alpha\sqrt{\frac{2r^2(T)}{\beta n}\sum_{t=1}^{T}\frac{1}{2}\log\left(1+\frac{\eta(T)^2K_{ST}^2(T)}{\sigma_t^2}\right)}$$
$$+ (1-\alpha)\sqrt{2r^2(T)\left(\frac{\sum_{t=1}^{T}\frac{1}{2}\log\left(1+\frac{\eta(T)^2K_{ST}^2(T)}{\sigma_t^2}\right)}{(1-\beta)n} + D\left(\mu\|\mu'\right)\right)}. \tag{157}$$

By setting $\kappa = 0$ loosely, within the finite iterations such that $\sum_{i=1}^{T} \|n(i)\| < \infty$, the according excess risk bound for this case is then expressed as

$$\begin{aligned}\mathbb{E}\left[L_{\mu'}(w(T) - L_{\mu'}(w^*)\right] \leq & \alpha \sqrt{\frac{2r^2(T)}{\beta n} \sum_{t=1}^{T} \frac{1}{2} \log\left(1 + \frac{\eta^2(T) K_{ST}^2(T)}{\sigma_t^2}\right)} \\ & + (1-\alpha) \sqrt{2r^2(T) \left(\frac{\sum_{t=1}^{T} \frac{1}{2} \log\left(1 + \frac{\eta^2(T) K_{ST}^2(T)}{\sigma_t^2}\right)}{(1-\beta)n} + D\left(\mu \| \mu'\right) \right)} \\ & + (1-\alpha) \sup_{w \in W^T} |\hat{L}(w, S) - \hat{L}(w, S')| \\ & + K_{ST}(T) \|w(0) - W_{\mathsf{ERM}}\| + K_{ST}(T) \sum_{t=1}^{T} \|n(t)\|, \end{aligned} \tag{158}$$

where $D(\mu||\mu')$ is calculated by

$$D(\mu||\mu') = \sum_{i=0}^{1} \mu(i) \log \frac{\mu(i)}{\mu'(i)} = (1-p) \log \frac{1-p}{1-p'} + p \log \frac{p}{p'}. \tag{159}$$