Characterizing and Understanding the Generalization Error of Transfer Learning with Gibbs Algorithm

Yuheng Bu[⋆†] **Gholamali Aminian**[⋆‡] **Laura Toni**[‡] **Gregory W. Wornell**[†] **Miguel R. D. Rodrigues**[‡]
[†] Massachusetts Institute of Technology [‡] University College London

Abstract

We provide an information-theoretic analysis of the generalization ability of Gibbs-based transfer learning algorithms by focusing on two popular empirical risk minimization (ERM) approaches for transfer learning, α-weighted-ERM and two-stage-ERM. Our key result is an exact characterization of the generalization behavior using the conditional symmetrized Kullback-Leibler (KL) information between the output hypothesis and the target training samples given the source training samples. Our results can also be applied to provide novel distribution-free generalization error upper bounds on these two aforementioned Gibbs algorithms. Our approach is versatile, as it also characterizes the generalization errors and excess risks of these two Gibbs algorithms in the asymptotic regime, where they converge to the α-weighted-ERM and two-stage-ERM, respectively. Based on our theoretical results, we show that the benefits of transfer learning can be viewed as a bias-variance trade-off, with the bias induced by the source distribution and the variance induced by the lack of target samples. We believe this viewpoint can guide the choice of transfer learning algorithms in practice.

1 INTRODUCTION

A common assumption in supervised learning is that both the training and test data samples are generated from the same distribution. However, this assumption does not always hold in many applications, as we often have easy access to samples generated from a source distribution, and we want to use the hypothesis trained using source training samples, which can be readily available, on a different target task, from which only limited data is available. Transfer learning and domain adaptation methods are developed to tackle this challenge, and the state-of-the-art transfer learning algorithms based on pre-trained models and fine-tuning have led to significant improvements in various applications such as computer vision, natural language processing, and more (Li et al., 2012; Long et al., 2015; Yosinski et al., 2014; Raffel et al., 2019).

Many works have attempted to explain the empirical success of transfer learning from different perspectives. The first theoretical analysis for domain adaptation is proposed by Ben-David et al. (2007) for binary classification, where the authors provide a VC-dimension-based excess risk bound for the zero-one loss in terms of $d_{\mathcal{A}}$-distance as a measure of discrepancy between source and target tasks. A new notion of discrepancy measure for transfer learning called transfer-exponent under the covariate-shift assumption is proposed in Hanneke and Kpotufe (2019). A minimax lower bound on the generalization error for transfer learning in neural networks is derived in Kalan and Fabian (2020). Recently, an Empirical Risk Minimization (ERM) algorithm via representation learning is proposed in Tripuraneni et al. (2020), and an upper bound on the excess risk of the new task is provided in terms of Gaussian complexity. Wang et al. (2019) provides an upper bound on excess risk based on instance weighting. Using KL divergence as a measure of similarity between the source and target data-generating distribution, an information-theoretic generalization error upper bound for transfer learning is proposed in Wu et al. (2020).

However, these upper bounds on excess risk and generalization error may not entirely capture the generalization ability of a transfer learning algorithm. One immediate concern is the tightness issue, as the proposed bounds (Wang et al., 2019) can be loose or even vacuous when evaluated in practice. More importantly, the current definitions of discrepancy metric do not

⋆ Equal contribution. Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

fully characterize all the aspects that could influence the performance of a transfer learning approach, e.g., most discrepancy measures are either algorithm independent (KL divergence in Wu et al. (2020)), or only depend on the hypothesis class ($d_{\mathcal{A}}$-distance in Ben-David et al. (2007)), or apply only under specific assumptions (the transfer exponent under covariate shift assumption in Hanneke and Kpotufe (2019)). Therefore, our understanding of transfer learning algorithms is still limited.

To overcome these limitations, we study transfer learning approaches using two Gibbs algorithms, i.e., α-weighted Gibbs algorithm and two-stage Gibbs algorithm, which can be viewed as randomized versions of two ERM-based transfer learning algorithms, i.e., α-weighted-ERM (Ben-David et al., 2010; Zhang et al., 2012) and two-stage-ERM (Tripuraneni et al., 2020; Donahue et al., 2014) by information-theoretic tools.

Our main contributions are as follows:

- We derive exact characterizations of the generalization errors for α-weighted Gibbs algorithm and two-stage Gibbs algorithm in terms of the conditional symmetrized KL information. We also provide novel distribution-free upper bounds, which quantify how the number of samples from the source and target will influence the generalization error of these transfer learning algorithms.

- We further demonstrate how to use our method to characterize the asymptotic behavior of the generalization error for these two Gibbs algorithms under large inverse temperature, where the α-weighted Gibbs algorithm and two-stage Gibbs algorithm converge to the α-weighted-ERM and two-stage-ERM, respectively.

- Finally, by studying the excess risk of the α-weighted-ERM and two-stage-ERM algorithms in the asymptotic regime, we show that transfer learning algorithms admit a bias-variance trade-off viewpoint, whereby the choice of a transfer learning algorithm should depend on both the bias induced by the source distribution and the variance caused by the limited target samples.

Notations: Throughout the paper, upper-case letters denote random variables (e.g., Z), lower-case letters denote the realizations of random variables (e.g., z), and calligraphic letters denote spaces (e.g., $\mathcal{Z}$). All the logarithms are the natural ones, and all information measure units are in nats. $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

2 PROBLEM FORMULATION

Let $D_s = \{Z_i^s\}_{i=1}^n$ and $D_t = \{Z_j^t\}_{j=1}^m$ be the source and target training sets, respectively, where Z_i^s and Z_j^t are defined on the same alphabet $\mathcal{Z}$. Note that D_s and D_t are independent, but neither D_s nor D_t is required to be i.i.d generated from the data-generating source or target distributions P_Z^s or P_Z^t. We denote the joint distribution of all source training samples as P_{D_s} and that of the target training samples as P_{D_t}. We denote the hypotheses by $w \in \mathcal{W}$, where $\mathcal{W}$ is a hypothesis class. The performance of any hypotheses is measured by a non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$, and we can define the empirical risk and the population risk of a source task for a given source dataset d_s as

$$L_E(w, d_s) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i^s), \tag{1}$$

$$L_P(w, P_{D_s}) \triangleq \mathbb{E}_{P_{D_s}}[L_E(w, D_s)], \tag{2}$$

and the empirical risk and the population risk of the target task for a given target dataset d_t,

$$L_E(w, d_t) \triangleq \frac{1}{m} \sum_{j=1}^{m} \ell(w, z_j^t), \tag{3}$$

$$L_P(w, P_{D_t}) \triangleq \mathbb{E}_{P_{D_t}}[L_E(w, D_t)]. \tag{4}$$

A transfer learning algorithm can be modeled as a randomized mapping from the source and target training sets (D_s, D_t) onto a hypothesis $W \in \mathcal{W}$ according to the conditional distribution $P_{W|D_s,D_t}$. Thus, the expected transfer generalization error quantifying the degree of over-fitting on the target training data can be written as

$$\overline{\text{gen}}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t}) \triangleq \mathbb{E}_{P_{W,D_s,D_t}}[L_P(W, P_{D_t}) - L_E(W, D_t)], \tag{5}$$

where the expectation is taken over the joint distribution $P_{W,D_s,D_t} = P_{W|D_s,D_t} \otimes P_{D_s} \otimes P_{D_t}$.

2.1 Transfer Learning Algorithms

We focus on the following two transfer learning approaches: α-weighted-ERM and two-stage-ERM.

α-Weighted-ERM Transfer Learning: We denote the hypothesis by $w_\alpha \in \mathcal{W}$ as the output of α-weighted-ERM learning algorithm. The hypothesis w_α is typically obtained by minimizing a convex combination of the source and target task empirical risks as in Ben-David et al. (2010), i.e.,

$$L_E(w_\alpha, d_s, d_t) = (1-\alpha) L_E(w_\alpha, d_s) + \alpha L_E(w_\alpha, d_t), \tag{6}$$

for $0 \leq \alpha \leq 1$.

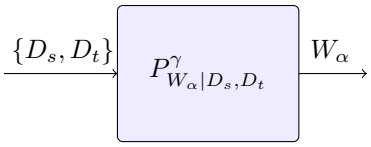

Figure 1: α-weighted Gibbs Algorithm

Two-stage-ERM Transfer Learning: Suppose that the hypothesis $w \in \mathcal{W}$ can be written as $w = (w_\phi, w_c)$, where $w_\phi \in \mathcal{W}_\phi$ is the shared hypothesis (parameter) across both source and target tasks, and $w_c^s \in \mathcal{W}_c$ and $w_c^t \in \mathcal{W}_c$ denote some task-specific hypothesis (parameter) for the source and target tasks, respectively. In practice, w_ϕ aggregates the parameters of the first few layers of a neural network for both tasks (e.g., shared featurizer), and w_c^s, w_c^t collect the remaining parameters for the source and target tasks, respectively. The performance of the pair (w_ϕ, w_c) is measured by a non-negative loss function $\ell : \mathcal{W}_c \times \mathcal{W}_\phi \times \mathcal{Z} \to \mathbb{R}_0^+$. We consider the following two-stage-ERM transfer learning algorithm inspired by Tripuraneni et al. (2020).

First Stage: The algorithm first learns the shared hypothesis w_ϕ and the source-specific hypothesis w_c^s by minimizing the following empirical risk function defined on the source dataset at Stage 1:

$$L_E^{S1}(w_\phi, w_c^s, d_s) \triangleq \frac{1}{n}\sum_{i=1}^{n} \ell(w_\phi, w_c^s, z_i^s). \tag{7}$$

Second Stage: The algorithm fixes the shared hypothesis w_ϕ and learns the target-specific hypothesis w_c^t by minimizing the following empirical risk function defined on the target dataset at Stage 2:

$$L_E^{S2}(w_\phi, w_c^t, d_t) = \frac{1}{m}\sum_{j=1}^{m} \ell(w_\phi, w_c^t, z_j^t). \tag{8}$$

2.2 Transfer Learning with Gibbs algorithms

Intending to understand the generalization behavior of transfer learning techniques, we now consider the Gibbs counterpart of the aforementioned ERM-based transfer learning algorithms. In particular, the $(\gamma, \pi(w), f(w,d))$-Gibbs distribution, which was first proposed by Gibbs (1902) in statistical mechanics, is defined as:

$$P_{W|D}^{\gamma}(w|d) \triangleq \frac{\pi(w)e^{-\gamma f(w,d)}}{V(d,\gamma)}, \quad \gamma \geq 0, \tag{9}$$

where γ is the inverse temperature, $\pi(w)$ is an arbitrary prior distribution on W, $f(w,d)$ is energy function, and $V(d,\gamma) \triangleq \int \pi(w)e^{-\gamma f(w,d)}dw$ is the partition function.

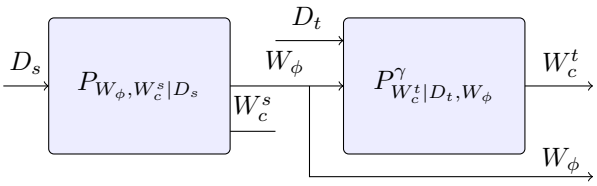

Figure 2: Two-stage Gibbs Algorithm

The $(\gamma, \pi(w), L_E(w,d_t))$-Gibbs distribution can be viewed as a randomized version of an ERM algorithm using only target samples if we specify the energy function $f(w,d) = L_E(w,d_t)$. Moreover, as the inverse temperature $\gamma \to \infty$, the prior distribution $\pi(w)$ becomes negligible, and the Gibbs algorithm converges to the standard supervised-ERM algorithm.

Similarly, we can immediately define the following α-weighted Gibbs algorithm and two-stage Gibbs algorithm, which can be viewed as randomized versions of α-weighted-ERM and two-stage-ERM, respectively.

α-weighted Gibbs algorithm generalizes the α-weighted-ERM via a $(\gamma, \pi(w_\alpha), L_E(w_\alpha, d_s, d_t))$-Gibbs algorithm (see, Figure 1)

$$P_{W_\alpha|D_s,D_t}^{\gamma}(w_\alpha|d_s,d_t) = \frac{\pi(w_\alpha)e^{-\gamma L_E(w_\alpha,d_s,d_t)}}{V_\alpha(d_s,d_t,\gamma)}. \tag{10}$$

The expected transfer generalization error of the α-weighted Gibbs algorithm is denoted as

$$\overline{\text{gen}}_\alpha(P_{D_s}, P_{D_t}) \triangleq \overline{\text{gen}}(P_{W_\alpha|D_s,D_t}^{\gamma}, P_{D_s}, P_{D_t}). \tag{11}$$

Two-stage Gibbs algorithm generalizes the two-stage-ERM via a $(\gamma, \pi(w_c^t), L_E^{S2}(w_\phi, w_c^t, d_t))$-Gibbs algorithm algorithm

$$P_{W_c^t|D_t,W_\phi}^{\gamma}(w_c^t|d_t,w_\phi) = \frac{\pi(w_c^t)e^{-\gamma L_E^{S_2}(w_\phi,w_c^t,d_t)}}{V_\beta(w_\phi,d_t,\gamma)} \tag{12}$$

in the second stage, where the shared hypothesis w_ϕ is the output of the learning algorithm $P_{W_\phi,W_c^s|D_s}$ at the first stage. As shown in Figure 2, the two-stage Gibbs algorithm is constructed by concatenating two randomized mappings $P_{W_c^t|D_t,W_\phi}^{\gamma}$ and $P_{W_\phi,W_c^s|D_s}$.

The population risk for the target task is defined as:

$$L_P(w_\phi, w_c^t, P_{D_t}) = \mathbb{E}_{P_{D_t}}[L_E^{S2}(w_\phi, w_c^t, D_t)], \tag{13}$$

and the expected transfer generalization error under two-stage Gibbs algorithm can be denoted as

$$\overline{\text{gen}}_\beta(P_{D_s}, P_{D_t}) \triangleq \tag{14}$$
$$\mathbb{E}_{P_{W_\phi,W_c^t,D_t}}[L_P(W_\phi, W_c^t, P_{D_t}) - L_E^{S2}(W_\phi, W_c^t, D_t)],$$

where the expectation is taken over the joint distribution $P_{W_\phi,W_c^t,D_t} = P_{W_c^t|D_t,W_\phi}^{\gamma} \otimes P_{W_\phi} \otimes P_{D_t}$.

2.3 Information Measures

We will characterize the aforementioned generalization errors using various information measures. If P and Q are probability measures over space $\mathcal{X}$, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by $D(P\|Q) \triangleq \int_{\mathcal{X}} \log\left(\frac{dP}{dQ}\right) dP$. If Q is absolutely continuous with respect to P, the symmetrized KL divergence (a.k.a., Jeffrey's divergence (Jeffreys, 1946)) is

$$D_{\mathrm{SKL}}(P\|Q) \triangleq D(P\|Q) + D(Q\|P). \tag{15}$$

The mutual information between two random variables X and Y is the KL divergence between the joint distribution and product-of-marginal distribution

$$I(X;Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y)$$

or equivalently, the conditional KL divergence between $P_{Y|X}$ and P_Y averaged over P_X, $D(P_{Y|X}\|P_Y|P_X) \triangleq \int_{\mathcal{X}} D(P_{Y|X=x}\|P_Y) dP_X(x)$. By swapping the role of $P_{X,Y}$ and $P_X \otimes P_Y$ in mutual information, we get the lautum information introduced by Palomar and Verdú (2008),

$$L(X;Y) \triangleq D(P_X \otimes P_Y\|P_{X,Y}).$$

Finally, the symmetrized KL information between X and Y is given by Aminian et al. (2015):

$$I_{\mathrm{SKL}}(X;Y) \triangleq D_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y) = I(X;Y) + L(X;Y). \tag{16}$$

The conditional mutual information between two random variables X and Y conditioned on Z is the KL divergence between $P_{X,Y|Z}$ and $P_{X|Z} \otimes P_{Y|Z}$ averaged over P_Z,

$$I(X;Y|Z) \triangleq \int_{\mathcal{Z}} D(P_{X,Y|Z=z}\|P_{Y|Z=z} \otimes P_{X|Z=z}) dP_Z(z).$$

Similarly, we can also define the conditional lautum information $L(X;Y|Z)$, and the conditional symmetrized KL information is given by

$$I_{\mathrm{SKL}}(X;Y|Z) \triangleq I(X;Y|Z) + L(X;Y|Z). \tag{17}$$

3 RELATED WORK

Other Interpretations for Gibbs Algorithm: Besides viewing the Gibbs algorithm as a randomized ERM, there are additional interpretations for considering Gibbs algorithm in transfer learning.

SGLD: The Stochastic Gradient Langevin Dynamics (SGLD), which can be viewed as noisy version of Stochastic Gradient Descent (SGD), is defined as:

$$W_{k+1} = W_k - \eta \nabla L_E(W_k, d_t) + \sqrt{\frac{2\beta}{\gamma}}\zeta_k, \quad k = 0, 1, \cdots,$$

where ζ_k is a standard Gaussian random vector and $\eta > 0$ is the step size. In Raginsky et al. (2017), it is proved that under some conditions on the loss function, the conditional distribution $P_{W_k|D_t}$ induced by SGLD algorithm is close to $(\gamma, \pi(W_0), L_E(w_k, d_t))$-Gibbs distribution in 2-Wasserstein distance for sufficiently large k.

Information Risk Minimization: The Gibbs algorithm also arises within the information risk minimization framework, where one adopts a conditional KL divergence regularizer to reduce over-fitting. In particular, it is shown in (Xu and Raginsky, 2017; Zhang, 2006; Zhang et al., 2006) that the solution to the regularized ERM problem

$$P^{\star}_{W|D_t} = \arg\inf_{P_{W|D_t}} \Big(\mathbb{E}_{P_{W,D_t}}[L_E(W, D_t)] + \frac{1}{\gamma} D(P_{W|D_t}\|\pi(W)|P_{D_t})\Big),$$

corresponds to the $(\gamma, \pi(w), L_E(w, d_t))$-Gibbs distribution. The inverse temperature γ controls the regularization term and balances between over-fitting and generalization.

Supervised Learning with the Gibbs Algorithm: An exact characterization of the generalization error of the Gibbs algorithm in terms of symmetrized KL information is provided by Aminian et al. (2021). The authors also provide a generalization error upper bound with the rate of $\mathcal{O}(\alpha/n)$ under the sub-Gaussian assumption. An information-theoretic upper bound with similar rate $\mathcal{O}(\alpha/n)$ is provided in Raginsky et al. (2016) for the Gibbs algorithm with bounded loss function, and PAC-Bayesian bounds using a variational approximation of Gibbs posteriors are studied in Alquier et al. (2016). Asadi and Abbe (2020); Kuzborskij et al. (2019) both focus on bounding the excess risk of the Gibbs algorithm.

Other Analyses of Transfer Learning: In hypothesis transfer learning problem (Kuzborskij and Orabona, 2013), where we only have access to the learned source hypotheses instead of the source training data, an upper bound on the leave-one-out error measured by square loss is provided. An extension of hypothesis transfer learning is studied in Kuzborskij and Orabona (2017), involving an algorithm combining the hypotheses from multiple sources based on regularized ERM principle. There are also works focusing on the theoretical aspects of domain adaptation, see (Ben-David et al., 2007, 2010; Mansour et al., 2009a,b; Germain et al., 2016; David et al., 2010), which are also related to our problem. Note that in domain adaptation, there is no labeled target data and only unlabeled target samples are available. Actually, having access to

target labeled data would improve the performance of the learning algorithm for target task (Mansour et al., 2021; Wang et al., 2019).

Note that we provide an *exact* characterization of the generalization error for these Gibbs algorithms in transfer learning scenarios, which differs from this body of research.

4 GENERALIZATION ERROR OF TRANSFER LEARNING ALGORITHM

We now offer exact characterizations of the expected transfer generalization errors in terms of symmetrized KL information for the α-weighted and two-stage Gibbs algorithms, respectively. Then, combining the exact characterization of expected transfer generalization error for Gibbs algorithms with a conditional mutual information-based generalization error upper bound, we derive novel distribution-free upper bounds for these two Gibbs algorithms. Finally, we provide another exact characterization of the generalization errors in terms of symmetrized KL divergence, which is shown to be useful in the asymptotic analysis.

4.1 Exact Characterization of Generalization Error Using Conditional Symmetrized KL Information

One of our main results, which characterizes the exact expected transfer generalization error of the α-weighted Gibbs algorithm with prior distribution $\pi(w_\alpha)$, is as follows:

Theorem 1 (Proved in Appendix A)**.** *For the α-weighted Gibbs algorithm, $0<\alpha<1$ and $\gamma>0$,*

$$P^{\gamma}_{W_\alpha|D_s,D_t}(w_\alpha|d_s,d_t)=\frac{\pi(w_\alpha)e^{-\gamma L_E(w_\alpha,d_s,d_t)}}{V_\alpha(d_s,d_t,\gamma)}, \quad (18)$$

the expected transfer generalization error is given by

$$\overline{gen}_\alpha(P_{D_s},P_{D_t})=\frac{I_{\mathrm{SKL}}(W_\alpha;D_t|D_s)}{\gamma\alpha}. \quad (19)$$

We also provide an exact characterization of the expected transfer generalization error for two-stage Gibbs algorithm using conditional symmetrized KL information.

Theorem 2 (Proved in Appendix A)**.** *The expected transfer generalization error of the two-stage Gibbs algorithm in (12) is given by*

$$\overline{gen}_\beta(P_{D_s},P_{D_t})=\frac{I_{\mathrm{SKL}}(W_c^t;D_t|W_\phi)}{\gamma}. \quad (20)$$

To the best of our knowledge, these results are the first exact characterizations of the expected transfer generalization error for the α-weighted and two-stage Gibbs algorithm. Note that both Theorem 1 and Theorem 2 only assume that the loss function is non-negative and the training set of source and target are independent, and they hold even for non-i.i.d source and target training samples.

The expected transfer generalization errors are non-negative, i.e., $\overline{\mathrm{gen}}_\alpha(P_{D_s},P_{D_t})\geq 0$ and $\overline{\mathrm{gen}}_\beta(P_{D_s},P_{D_t})\geq 0$, which follows by the non-negativity of the conditional symmetrized KL information.

4.2 Example: Mean Estimation

We now consider a simple mean estimation problem, where the symmetrized KL information can be computed exactly, to demonstrate the usefulness of our Theorems. All details are provided in Appendix B.

Consider the problem of learning the mean $\boldsymbol{\mu}_t\in\mathbb{R}^d$ of the target task using n i.i.d. source samples $D_s=\{Z_i^s\}_{i=1}^n$ and m i.i.d. target samples $D_t=\{Z_j^t\}_{j=1}^m$. We assume that the samples from the source and target tasks satisfying $\mathbb{E}[Z^s]=\boldsymbol{\mu}_s$, $\mathrm{cov}[Z^s]=\sigma_s^2 I_d$ and $\mathbb{E}[Z^t]=\boldsymbol{\mu}_t$, $\mathrm{cov}[Z^t]=\sigma_t^2 I_d$, respectively. We adopt the mean-squared loss $\ell(\boldsymbol{w},\boldsymbol{z})=\|\boldsymbol{z}-\boldsymbol{w}\|_2^2$, and assume a Gaussian prior for the mean $\pi(\boldsymbol{w})=\mathcal{N}(\boldsymbol{\mu}_0,\sigma_0^2 I_d)$.

For the α-weighted Gibbs algorithm, if we set the inverse-temperature $\gamma=\frac{m+n}{2\sigma^2}$ and $\alpha=\frac{m}{m+n}$, then the $(\frac{m+n}{2\sigma^2},\mathcal{N}(\boldsymbol{\mu}_0,\sigma_0^2 I_d),L_E(\boldsymbol{w}_\alpha,d_s,d_t))$-Gibbs algorithm is given by the following posterior (Murphy, 2007),

$$P^{\gamma}_{W_\alpha|D_t,D_s}(\boldsymbol{w}_\alpha|D_s,D_t)\sim\mathcal{N}\Big(\boldsymbol{m}_\alpha,\sigma_1^2 I_d\Big), \quad (21)$$

with $\boldsymbol{m}_\alpha=\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0+\frac{\sigma_1^2}{\sigma^2}\big(\sum_{i=1}^n Z_i^s+\sum_{j=1}^m Z_j^t\big)$, and $\sigma_1^2=\frac{\sigma_0^2\sigma^2}{(m+n)\sigma_0^2+\sigma^2}$. Since $P^{\gamma}_{W_\alpha|D_s,D_t}$ is Gaussian, the conditional symmetrized KL information does not depend on the distribution P_{Z^t} as long as $\mathrm{cov}[Z^t]=\sigma_t^2 I_d$, i.e.,

$$I_{\mathrm{SKL}}(W_\alpha;D_t|D_s)=\frac{md\sigma_0^2\sigma_t^2}{((m+n)\sigma_0^2+\sigma^2)\sigma^2}. \quad (22)$$

From Theorem 1, the expected transfer generalization error of this algorithm can be computed exactly as:

$$\begin{aligned}\overline{\mathrm{gen}}_\alpha(P_{D_s},P_{D_t})&=\frac{I_{\mathrm{SKL}}(W_\alpha;D_t|D_s)}{\gamma\alpha}\\&=\frac{2d\sigma_0^2\sigma_t^2}{(m+n)(\sigma_0^2+\frac{1}{2\gamma})}.\end{aligned} \quad (23)$$

For the two-stage Gibbs algorithm, we learn the first d_ϕ components $\boldsymbol{\mu}_\phi\in\mathbb{R}^{d_\phi}$ using source samples. Then,

we set inverse temperature $\gamma = \frac{m}{2\sigma^2}$ and use the $(\frac{m}{2\sigma^2}, \mathcal{N}(\boldsymbol{\mu}_{0,c}, \sigma_0^2 I_{d_c}), L_E^{S2}(\boldsymbol{\mu}_\phi, \boldsymbol{w}_c^t, d_t))$-Gibbs algorithm to learn the remaining $d_c = d - d_\phi$ components. Following similar steps, by Theorem 2, we have

$$\overline{\text{gen}}_\beta(P_{D_s}, P_{D_t}) = \frac{I_{\text{SKL}}(W_c^t; D_t|W_\phi)}{\gamma} = \frac{2d_c\sigma_0^2\sigma_t^2}{m(\sigma_0^2 + \frac{1}{2\gamma})}. \quad (24)$$

Remark 1 (Comparison with Supervised Learning). *It is shown in Aminian et al. (2021) that the generalization error of a supervised Gibbs algorithm for this mean estimation example is*

$$\overline{gen}(P_{W|D_t}^\gamma, P_{D_t}) = \frac{2d\sigma_0^2\sigma_t^2}{m(\sigma_0^2 + \frac{1}{2\gamma})}, \quad (25)$$

where $P_{W|D_t}^\gamma$ is $(\frac{m}{2\sigma^2}, \mathcal{N}(\boldsymbol{\mu}_0, \sigma_0^2 I_d), L_E(w, d_t))$-Gibbs algorithm. Comparing to the supervised Gibbs algorithm, the α-weighted Gibbs algorithm reduces the generalization error to $\mathcal{O}(\frac{d}{m+n})$ by fitting n source samples and m target samples simultaneously, and the two-stage Gibbs algorithm achieves the rate of $\mathcal{O}(\frac{d_c}{m})$ by only learning $\boldsymbol{w}_c^t \in \mathbb{R}^{d_c}$ from the target samples D_t.

Remark 2 (Effect of Source samples). *As shown in (23) and (24), the transfer generalization errors of this mean estimation problem do not depend on the distribution of sources samples D_s. The reason is that the effect of sources samples is cancelled out in generalization error by subtracting the empirical risk from the population risk. Although different source distributions do not change generalization errors, they will influence the population risks and excess risks, and more detailed discussion is provided in Appendix B.*

4.3 Distribution-free Upper Bounds

To understand the behavior of expected transfer generalization error, we also provide distribution-free upper bounds in this subsection. These bounds quantify how the generalization errors of the α-weighted and two-stage Gibbs algorithms depend on the number of target (source) samples m (n), and can be applied when directly computing symmetrized KL information is hard.

We first provide a conditional mutual information based upper bound on the expected transfer generalization error for any general learning algorithm $P_{W|D_s,D_t}$ under i.i.d and σ-sub-Gaussian assumptions.

Theorem 3 (Proved in Appendix C). *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W, Z)$ is σ-sub-Gaussian[1] under the distribution $P_Z^t \otimes P_W$. Then the following upper bound holds*

$$|\overline{gen}(P_{W|D_s,D_t}, P_{D_s,D_t})| \leq \sqrt{\frac{2\sigma^2}{m} I(W; D_t|D_s)}. \quad (26)$$

The following distribution-free upper bound on the expected transfer generalization error for α-weighted Gibbs algorithm can be obtained by combining the upper bound in Theorem 3 with the exact characterization in Theorem 1.

Theorem 4 (Proved in Appendix D). *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W, Z)$ is σ_α-sub-Gaussian under the distribution $P_Z^t \otimes P_{W_\alpha}$. If we further assume $C_\alpha \leq \frac{L(W_\alpha; D_t|D_s)}{I(W_\alpha; D_t|D_s)}$ for some $C_\alpha \geq 0$, then for the α-weighted Gibbs algorithm and $0 < \alpha < 1$,*

$$\overline{gen}_\alpha(P_{D_s}, P_{D_t}) \leq \frac{2\sigma_\alpha^2\gamma\alpha}{(1 + C_\alpha)m}. \quad (27)$$

Remark 3. *Let $\alpha = \frac{m}{n+m}$, then we have*

$$\overline{gen}_\alpha(P_{D_s}, P_{D_t}) \leq \frac{2\sigma_\alpha^2\gamma}{(1 + C_\alpha)(n + m)}, \quad (28)$$

which is lower than the distribution-free upper bound for supervised learning under $(\gamma, \pi(w), L_E(w, d_t))$-Gibbs algorithm $P_{W|D_t}^\gamma$ provided in (Aminian et al., 2021, Theorem 2), i.e., $\overline{gen}_\alpha(P_{W|D_t}^\gamma, P_{D_t}) \leq \frac{2\sigma^2\gamma}{(1+C_E)m}$, if $C_E = C_\alpha$ and $\sigma^2 = \sigma_\alpha^2$.

Using similar approach, we can obtain a distribution-free upper bound on the expected transfer generalization error for the two-stage Gibbs algorithm.

Theorem 5 (Proved in Appendix D). *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W_c, w_\phi, Z)$ is σ_β-sub-Gaussian under distribution $P_Z^t \otimes P_{W_c^t|W_\phi = w_\phi}$ for all $w_\phi \in \mathcal{W}_\phi$. If we further assume $C_\beta \leq \frac{L(W_c^t; D_t|W_\phi)}{I(W_c^t; D_t|W_\phi)}$ for some $C_\beta \geq 0$, then for the two-stage Gibbs algorithm in (12), we have*

$$\overline{gen}_\beta(P_{D_s}, P_{D_t}) \leq \frac{2\sigma_\beta^2\gamma}{(1 + C_\beta)m}. \quad (29)$$

Remark 4 (Choice of C_β and C_α). *Setting $C_\alpha = 0$ in Theorem 4 and $C_\beta = 0$ in Theorem 5 is always valid since the lautum information is always positive whenever the mutual information is positive.*

4.4 Exact Characterization of Generalization Error Using Conditional Symmetrized KL Divergence

In this section, we provide exact characterizations of expected transfer generalization errors for α-weighted

[1] A random variable X is σ-sub-Gaussian if $\log \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \leq \frac{\sigma^2\lambda^2}{2}$, $\forall \lambda \in \mathbb{R}$.

and two-stage Gibbs algorithms using conditional symmetrized KL divergence by considering the Gibbs algorithm with the population risks as energy functions in (9). Such a result is very useful in the asymptotic analysis Section 5.1.

Theorem 6 (Proved in Appendix E). *The expected transfer generalization error of the α-weighted Gibbs algorithm in (10) is given by:*

$$\overline{gen}_{\alpha}(P_{D_s},P_{D_t}) = \frac{D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s,D_t}\|P^{\gamma,L_\alpha(w_\alpha,d_s,P_{D_t})}_{W_\alpha|D_s}|P_{D_s}P_{D_t})}{\gamma\alpha}, \tag{30}$$

where $P^{\gamma,L_\alpha(w_\alpha,d_s,P_{D_t})}_{W_\alpha|D_s}$ is $(\gamma,\pi(w_\alpha),L_\alpha(w_\alpha,d_s,P_{D_t}))$-Gibbs algorithm with $L_\alpha(w,d_s,P_{D_t}) \triangleq \alpha L_P(w_\alpha,P_{D_t})+(1-\alpha)L_E(w_\alpha,d_s)$.

A similar result can be obtained for the two-stage Gibbs algorithm.

Theorem 7 (Proved in Appendix E). *The expected transfer generalization error of the two-stage Gibbs algorithm in (12) is given by:*

$$\overline{gen}_{\beta}(P_{D_s},P_{D_t}) = \frac{D_{\mathrm{SKL}}(P^{\gamma}_{W^t_c|D_t,W_\phi}\|P^{\gamma,L_P(w_\phi,w^t_c,P_{D_t})}_{W^t_c|W_\phi}|P_{D_t}P_{W_\phi})}{\gamma}, \tag{31}$$

where $P^{\gamma,L_P(w_\phi,w^t_c,P_{D_t})}_{W^t_c|W_\phi}$ is $(\gamma,\pi(w^t_c),L_P(w_\phi,w^t_c,P_{D_t}))$-Gibbs algorithm.

More discussions about the connection between the results obtained using symmetrized KL information and those of symmetrized KL divergence are provided in Appendix E.

5 ASYMPTOTIC BEHAVIOR OF GENERALIZATION ERROR AND EXCESS RISK

In this section, we first consider the asymptotic behavior of the generalization error for the two Gibbs algorithms as the inverse temperature $\gamma \to \infty$. Note that, in this regime, both Gibbs algorithms converge to the corresponding ERM algorithms, and the distribution-free upper bounds obtained in the previous section become vacuous. Then, we show that such results can be applied to characterize the excess risks of the two ERM algorithms as $m,n \to \infty$, leading up to useful intuition about how to select different transfer learning approaches.

5.1 Generalization Error

α-weighted-ERM: We assume that there exists a unique $\hat{W}_\alpha(D_s,D_t)$ and a unique $\hat{W}_\alpha(D_s)$ that minimizes the risk $L_E(w,D_s,D_t)$ and $L_\alpha(w,D_s,P_{D_t})$, respectively, i.e.,

$$\hat{W}_\alpha(D_s,D_t) = \arg\min_{w\in\mathcal{W}} L_E(w,D_s,D_t), \tag{32}$$

$$\hat{W}_\alpha(D_s) = \arg\min_{w\in\mathcal{W}} L_\alpha(w,D_s,P_{D_t}). \tag{33}$$

It is shown in Hwang (1980) that if the following Hessian matrices

$$H^*(D_s,D_t) \triangleq \nabla^2_w L_E(w,D_s,D_t)\big|_{w=\hat{W}_\alpha(D_s,D_t)}, \tag{34}$$

$$H^*(D_s) \triangleq \nabla^2_w L_\alpha(w,D_s,P_{D_t})\big|_{w=\hat{W}_\alpha(D_s)} \tag{35}$$

are not singular, then, as $\gamma\to\infty$

$$P^{\gamma}_{W_\alpha|D_s,D_t} \to \mathcal{N}(\hat{W}_\alpha(D_s,D_t),\frac{1}{\gamma}H^*(D_s,D_t)^{-1}),$$
$$P^{\gamma,L_\alpha(w_\alpha,d_s,P_{D_t})}_{W_\alpha|D_s} \to \mathcal{N}(\hat{W}_\alpha(D_s),\frac{1}{\gamma}H^*(D_s)^{-1}). \tag{36}$$

Thus, the conditional symmetrized KL divergence in Theorem 6 can be evaluated directly using Gaussian approximations.

Proposition 1 (Proved in Appendix F.1). *If the Hessian matrices $H^*(D_s,D_t) = H^*(D_s) = H^*$ are independent of D_s and D_t, then the generalization error of the α-weighted-ERM algorithm is*

$$\overline{gen}_{\alpha}(P_{D_t},P_{D_s}) = \frac{\mathbb{E}_{P_{D_s}\otimes P_{D_t}}[\|\hat{W}_\alpha(D_s,D_t)-\hat{W}_\alpha(D_s)\|^2_{H^*}]}{\alpha},$$

where $\|W\|^2_H \triangleq W^\top HW$.

Remark 5. *The assumptions that the two Hessian matrices $H^*(D_s,D_t)$ and $H^*(D_s)$ coincide and are independent of D_s and D_t, are only needed to simplify the expression in Proposition 1. Note that these assumptions are satisfied in the asymptotic regime where $m,n\to\infty$ in standard maximum likelihood estimates (MLE) setting discussed below.*

We can use Proposition 1 to obtain the generalization error of (MLE) in the asymptotic regime $m,n\to\infty$. More specifically, suppose that we have m and n i.i.d. samples generated from the target distribution P^t_Z and source distribution P^s_Z, respectively. We want to fit the training data with a parametric distribution family $\{f(z|\boldsymbol{w}_\alpha)\}$ using the α-weighted-ERM algorithm, where $\boldsymbol{w}_\alpha \in \mathcal{W}\subset\mathbb{R}^d$ denotes the parameter of the model. Here, the true data-generating distribution may not belong to the parametric family, i.e., $P^s_Z, P^t_Z \notin \{f(\cdot|\boldsymbol{w}_\alpha)|\boldsymbol{w}_\alpha\in\mathcal{W}\}$.

If we use the log-loss $\ell(\boldsymbol{w}_\alpha,z) = -\log f(z|\boldsymbol{w}_\alpha)$ in the α-weighted Gibbs algorithm, and set $\alpha = \frac{m}{m+n}$, as $\gamma\to\infty$, it converges to the α-weighted-ERM algo-

rithm, which is equivalent to the following MLE, i.e.,

$$\hat{W}_\alpha(D_s, D_t) = \arg\max_{\boldsymbol{w}_\alpha \in \mathcal{W}} \sum_{i=1}^{n} \log f(Z_i^s|\boldsymbol{w}_\alpha) + \sum_{j=1}^{m} \log f(Z_j^t|\boldsymbol{w}_\alpha). \tag{37}$$

If we further let $m, n \to \infty$, under regularization conditions for MLE (details in Appendix F.2) which guarantee that $\hat{W}_\alpha(D_s, D_t)$ and $\hat{W}_\alpha(D_s)$ are unique, we can show that

$$\hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s) \to \mathcal{N}\big(0, \frac{m\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\bar{\mathcal{I}}(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}}{(m+n)^2}\big),$$

where

$$\boldsymbol{w}_\alpha^* \triangleq \arg\min_{\boldsymbol{w}\in\mathcal{W}} nD(P_Z^s \| f(\cdot|\boldsymbol{w})) + mD(P_Z^t \| f(\cdot|\boldsymbol{w})),$$

$\bar{J}(\boldsymbol{w}_\alpha^*)$ is the weighted expectation of the Hessian matrix, and $\bar{\mathcal{I}}(\boldsymbol{w}_\alpha^*)$ is the weighted Fisher information matrix. Detailed definitions of $\bar{J}$ and $\bar{\mathcal{I}}$ and proofs are provided in Appendix F.3.

In addition, the Hessian matrix $H^*(D_s, D_t) \to \bar{J}(\boldsymbol{w}_\alpha^*)$ as $m, n \to \infty$, which is independent of the samples D_s, D_t. Thus, Proposition 1 gives

$$\overline{\text{gen}}_\alpha(P_{D_t}, P_{D_s}) = \frac{\text{tr}(\bar{\mathcal{I}}(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1})}{n+m}, \tag{38}$$

which scales as $\mathcal{O}(\frac{d}{m+n})$.

Two-stage-ERM: We assume that there exists one unique $\hat{W}_c^t(D_t, W_\phi)$ which minimize the empirical risk of stage 2,

$$\hat{W}_c^t(D_t, W_\phi) \triangleq \arg\min_{w_c\in\mathcal{W}_c} L_E^{S2}(W_\phi, w_c, D_t), \tag{39}$$

and there is one unique $\hat{W}_c^t(W_\phi)$ which minimize the population risk given a fixed W_ϕ,

$$\hat{W}_c^t(W_\phi) \triangleq \arg\min_{w_c\in\mathcal{W}_c} L_P(W_\phi, w_c, P_{D_t}). \tag{40}$$

Similarly, if the following Hessian matrices

$$H_c^*(D_t,W_\phi) \triangleq \nabla_{w_c}^2 L_E^{S2}(W_\phi,w_c,D_t)\big|_{w_c=\hat{W}_c^t(D_t,W_\phi)} \tag{41}$$

$$H_c^*(W_\phi) \triangleq \nabla_{w_c}^2 L_P(W_\phi, w_c, P_{D_t})\big|_{w_c=\hat{W}_c^t(W_\phi)} \tag{42}$$

are not singular, we can obtain the following result by evaluating the conditional symmetrized KL divergence in Proposition 7 using similar Gaussian approximation as in (36).

Proposition 2 (Proved in Appendix F.1)**.** *If Hessian matrices $H_c^*(D_t, W_\phi) = H_c^*(W_\phi) = H_c^*$ are independent of D_s, D_t, then the generalization error of the two-stage-ERM algorithm is*

$$\overline{gen}_\beta(P_{D_t}, P_{D_s}) = \mathbb{E}_{P_{D_s,D_t,W_\phi}}[\|\hat{W}_c^t(D_t, W_\phi) - \hat{W}_c^t(W_\phi)\|_{H_c^*}^2].$$

Consider a similar MLE setting as we did for the α-weighted-ERM algorithm, except that now we want to fit data with a parametric distribution family $\{f(z_j^t|\boldsymbol{w}_\phi, \boldsymbol{w}_c^t)\}_{j=1}^m$ using the two-stage-ERM algorithm, where $\boldsymbol{w}_\phi \in \mathcal{W}_\phi \subset \mathbb{R}^{d_\phi}$, $\boldsymbol{w}_c^t \in \mathcal{W}_c \subset \mathbb{R}^{d_c}$ denote the shared and specific parameters, respectively.

If we use the log-loss $\ell(\boldsymbol{w}_\phi, \boldsymbol{w}_c^t, z) = -\log f(z|\boldsymbol{w}_\phi, \boldsymbol{w}_c^t)$ in the two-stage Gibbs algorithm, as $\gamma \to \infty$, it converges to the following two-stage MLE approach,

$$[\hat{W}_\phi(D_s), \hat{W}_c^s(D_s)] \triangleq \arg\max_{[\boldsymbol{w}_\phi,\boldsymbol{w}_c]\in\mathcal{W}} \sum_{i=1}^{n} \log f(Z_i^s|\boldsymbol{w}_\phi, \boldsymbol{w}_c),$$

$$\hat{W}_c^t(D_t, \hat{W}_\phi) \triangleq \arg\max_{\boldsymbol{w}_c\in\mathcal{W}_c} \sum_{j=1}^{m} \log f(Z_j^t|\hat{W}_\phi, \boldsymbol{w}_c).$$

As $m, n \to \infty$, under similar regularization conditions (details in Appendix F.2) which guarantee the uniqueness of these estimates, we can show that

$$\hat{W}_c^t(D_t, \hat{W}_\phi) - \hat{W}_c^t(\hat{W}_\phi) \to \mathcal{N}\big(0, \frac{J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}}{m}\big),$$

where

$$[\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{s*}] \triangleq \arg\min_{[\boldsymbol{w}_\phi,\boldsymbol{w}_c]\in\mathcal{W}} D(P_{Z^s} \| f(\cdot|\boldsymbol{w}_\phi, \boldsymbol{w}_c)), \tag{43}$$

$$\boldsymbol{w}_c^{st*} \triangleq \arg\min_{\boldsymbol{w}_c\in\mathcal{W}_c} D(P_{Z^t} \| f(\cdot|\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c)), \tag{44}$$

and $J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})$, $\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})$ stands for the expected Hessian matrix and Fisher information matrix over $\boldsymbol{w}_c$ under target distribution, respectively. Detailed proofs are provided in Appendix F.3. As the Hessian matrix $H_c^*(D_t, W_\phi) = H_c^*(W_\phi) \to J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})$ as $m, n \to \infty$, by Proposition 2, we have

$$\overline{\text{gen}}_\beta(P_{D_t}, P_{D_s}) = \mathcal{O}(\frac{d_c}{m}). \tag{45}$$

5.2 Excess Risk in MLE Setting

We further consider the excess risks of the α-weighted-ERM algorithm and the two-stage-ERM algorithm in the aforementioned MLE setting when $m, n \to \infty$, and show that such analyses provide some intuitions in selecting different transfer learning algorithms. All the details are provided in Appendix F.4.

The excess risk (Mohri et al., 2018) is defined as the difference between the population risk achieved by the learning algorithm and that achieved by the optimal hypothesis given the knowledge of the true target distribution P_{Z_t}, i.e.,

$$\mathcal{E}_r(P_W) \triangleq \mathbb{E}_{P_{W,D_s,D_t}}[L_P(W, P_{D_t})] - L_P(\boldsymbol{w}_t^*, P_{D_t}),$$
$$\text{with} \quad \boldsymbol{w}_t^* \triangleq \arg\min_{\boldsymbol{w}\in\mathcal{W}} L_P(\boldsymbol{w}, P_{D_t}), \tag{46}$$

Table 1: Comparison of different algorithms under MLE setting.

	Standard ERM	**α-weighted-ERM**	**Two-stage-ERM**
Excess risk bias	0	$\|\boldsymbol{w}_\alpha^* - \boldsymbol{w}_t^*\|^2_{J_t(\boldsymbol{w}_t^*)}$	$\left\|[\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}] - [\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*}]\right\|^2_{J_t(\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*})}$
Excess risk variance	$\mathcal{O}(\frac{d}{m})$	$\mathcal{O}(\frac{d}{m+n})$	$\mathcal{O}(\frac{d}{n} + \frac{d_c}{m})$
Generalization error	$\mathcal{O}(\frac{d}{m})$	$\mathcal{O}(\frac{d}{m+n})$	$\mathcal{O}(\frac{d_c}{m})$

where $\boldsymbol{w}_t^* = \arg\min_{\boldsymbol{w}\in\mathcal{W}} D(P_Z^t \| f(\cdot|w))$ holds in the MLE setting considered here.

α-weighted-ERM: In general, a proper transfer learning algorithm should have small excess risk $\mathcal{E}_r$, which justifies the following approximation of the excess risk

$$\mathcal{E}_r(P_{\hat{W}_\alpha(D_s,D_t)}) \approx \frac{1}{2}\mathbb{E}_{P_{D_s,D_t}}\left[\left\|\hat{W}_\alpha(D_s,D_t) - \boldsymbol{w}_t^*\right\|^2_{J_t(\boldsymbol{w}_t^*)}\right]$$
$$= \frac{1}{2}\|\boldsymbol{w}_\alpha^* - \boldsymbol{w}_t^*\|^2_{J_t(\boldsymbol{w}_t^*)} + \frac{\mathrm{tr}\big(J_t(\boldsymbol{w}_t^*)\mathrm{Cov}(\hat{W}_\alpha(D_s,D_t))\big)}{2}.$$

As we can see from the above expression, the excess risk can be decomposed into squared bias and variance terms. The bias is caused by learning from the mixture of the source and target distributions instead of just the target distribution P_Z^t. In addition, it can be shown that $\mathrm{tr}(J_t(\boldsymbol{w}_t^*)\mathrm{Cov}(\hat{W}_\alpha(D_s,D_t))) = \mathcal{O}(\frac{d}{m+n})$, which has the same order as the generalization error in (38).

Two-stage-ERM: In the two-stage algorithm, $\boldsymbol{w}^{t*}$ can be written as $\boldsymbol{w}^{t*} = [\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*}]$, and using similar approximation, we have

$$\mathcal{E}_r(P_{\hat{W}_\phi(D_s),\hat{W}_c^t(D_t,\hat{W}_\phi)})$$
$$\approx \frac{1}{2}\left\|[\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}] - [\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*}]\right\|^2_{J_t(\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*})} \quad (47)$$
$$+ \frac{\mathrm{tr}\big(J_t(\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*})\mathrm{Cov}(\hat{W}_\phi(D_s), \hat{W}_c^t(D_t,\hat{W}_\phi))\big)}{2}.$$

Here the bias is caused by sharing the parameter $\boldsymbol{w}_\phi^{s*}$ learned from the source distribution. If $\boldsymbol{w}_\phi^{t*} = \boldsymbol{w}_\phi^{s*}$, then $\boldsymbol{w}_c^{st*} = \boldsymbol{w}_c^{t*}$ and the bias is zero. It can be shown that the variance term scales as $O(\frac{d}{n} + \frac{d_c}{m})$. When $n \gg m$, it reduces to $\mathcal{O}(\frac{d_c}{m})$, which is the same as the generalization error in (45).

In Table 1, we summarize the excess risk, and generalization error results for the two transfer learning algorithms studied in the paper and those of the standard supervised learning under MLE setting (Van der Vaart, 2000) as $m, n \to \infty$. The improvement of the excess risk for transfer learning algorithms comes from trading the variance induced by the lack of target samples with the bias introduced by the source distribution, which suggests that the choice of learning algorithm should depend on both source distribution and the number of samples m, n.

The bias term in the excess risks can be interpreted as another notion of discrepancy measure, which is algorithm-dependent, as $\boldsymbol{w}_\alpha^*$ and $\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{s*}$ are defined as the optimal parameters under different algorithms given the knowledge of both source and target distributions. Sometimes, these bias terms are more useful in choosing an algorithm than the discrepancy measure used in the literature. For example, consider the mean estimation example in Section 4.2, if we set $\boldsymbol{\mu}_s = \boldsymbol{\mu}_t$, $\sigma_s^2 \ll \sigma_t^2$, and let $m, n \to \infty$, then the bias term for both α-weighted-ERM and two-stage-ERM should be zero, and transfer learning algorithms are preferred over the standard ERM. However, the KL divergence between the source and target distribution, which is proposed as a discrepancy measure in Wu et al. (2020), would be large.

The generalization error can be interpreted as the variance of the excess risk when $n \gg m$, and the analysis provided in the paper could help us to find a good balance in the bias and variance trade-off. Our results can be also used in transfer learning algorithm selection in the MLE setting by comparing the sum of the corresponding generalization error term and empirical risk achieved by each algorithm, which generalizes the standard Akaike information criterion (AIC) (Akaike, 1998) used in the supervised learning to transfer learning setting.

6 CONCLUSION

We provide an exact characterization of the generalization error for two Gibbs-based transfer learning algorithms, i.e., α-weighted Gibbs algorithm and two-stage-ERM Gibbs algorithm, using conditional symmetrized KL information and divergence. Based on our results, we show that the benefits of transfer learning can be viewed as a bias-variance trade-off, and importantly, the term relating to the bias points to a new discrepancy measure that merits further investigation.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback, which helped us to improve the paper greatly. Yuheng Bu is supported, in part, by NSF under Grant CCF-1717610 and by the MIT-IBM Watson AI Lab. Gholamali Aminian is supported by the Royal Society Newton International Fellowship, grant no. NIF\R1 \192656.

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Gholamali Aminian, Hamidreza Arjmandi, Amin Gohari, Masoumeh Nasiri-Kenari, and Urbashi Mitra. Capacity of diffusion-based molecular communication networks over LTI-Poisson channels. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 1(2):188–201, 2015.

Gholamali Aminian, Yuheng Bu, Laura Toni, Miguel Rodrigues, and Gregory Wornell. An exact characterization of the generalization error for the Gibbs algorithm. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Amir R Asadi and Emmanuel Abbe. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research*, 21(139):1–32, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137, 2007.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

Shai Ben David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 129–136. JMLR Workshop and Conference Proceedings, 2010.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new PAC-Bayesian perspective on domain adaptation. In *International conference on machine learning*, pages 859–868. PMLR, 2016.

Josiah Willard Gibbs. Elementary principles of statistical mechanics. *Compare*, 289:314, 1902.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32:9871–9881, 2019.

Chii-Ruey Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

MM Kalan and Z Fabian. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Neural Information Processing Systems (NeuRIPS 2020)*, 2020.

Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

Ilja Kuzborskij and Francesco Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2):171–195, 2017.

Ilja Kuzborskij, Nicolò Cesa-Bianchi, and Csaba Szepesvári. Distribution-dependent analysis of Gibbs-erm principle. In *Conference on Learning Theory*, pages 2028–2054. PMLR, 2019.

Wei Li, Rui Zhao, and Xiaogang Wang. Human reidentification with transferred metric learning. In *Asian conference on computer vision*, pages 31–44. Springer, 2012.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *Conference on Learning Theory, (COLT)*, 2009a.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Multiple source adaptation and the Rényi divergence. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 367–374, 2009b.

Yishay Mansour, Mehryar Mohri, Jae Ro, Ananda Theertha Suresh, and Ke Wu. A theory of multiple-source adaptation with limited target labeled data. In *International Conference on Artificial Intelligence and Statistics*, pages 2332–2340. PMLR, 2021.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Kevin P Murphy. Conjugate Bayesian analysis of the gaussian distribution. *def*, $1(2\sigma 2)$:16, 2007.

Daniel P Palomar and Sergio Verdú. Lautum information. *IEEE transactions on information theory*, 54 (3):964–975, 2008.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33, 2020.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton. Transfer learning via minimizing the performance gap between domains. *Advances in Neural Information Processing Systems*, 32:10645–10655, 2019.

Xuetong Wu, Jonathan H Manton, Uwe Aickelin, and Jingge Zhu. Information-theoretic analysis for transfer learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2819–2824. IEEE, 2020.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

Chao Zhang, Lei Zhang, and Jieping Ye. Generalization bounds for domain adaptation. *Advances in neural information processing systems*, 4:3320, 2012.

Huiming Zhang and Song Xi Chen. Concentration inequalities for statistical inference. *arXiv preprint arXiv:2011.02258*, 2020.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Tong Zhang et al. From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34(5): 2180–2210, 2006.

Supplementary Material: Characterizing and Understanding the Generalization Error of Transfer Learning with Gibbs Algorithm

A Exact Characterization of Generalization Error Based on Symmetrized KL Information

A.1 α-weighted Gibbs Algorithm

Theorem 1. ***(restated)*** *For the α-weighted Gibbs algorithm, $0 < \alpha < 1$ and $\gamma > 0$,*

$$P^{\gamma}_{W_\alpha|D_s,D_t}(w_\alpha|d_s,d_t) = \frac{\pi(w_\alpha)e^{-\gamma L_E(w_\alpha,d_s,d_t)}}{V_\alpha(d_s,d_t,\gamma)},$$

the expected transfer generalization error is given by

$$\overline{gen}_\alpha(P_{D_s},P_{D_t}) = \frac{I_{\mathrm{SKL}}(W_\alpha;D_t|D_s)}{\gamma\alpha}.$$

Proof. By the definition of conditional symmetrized KL information, we have

$$\begin{aligned} I_{\mathrm{SKL}}(W_\alpha;D_t|D_s) &= \mathbb{E}_{P_{D_s}}\Big[\mathbb{E}_{P_{W_\alpha,D_t|D_s}}\Big[\log\Big(\frac{P^{\gamma}_{W_\alpha|D_s,D_t}}{P_{W_\alpha|D_s}}\Big)\Big] + \mathbb{E}_{P_{W_\alpha|D_s}P_{D_t|D_s}}\Big[\log\Big(\frac{P_{W_\alpha|D_s}}{P^{\gamma}_{W_\alpha|D_s,D_t}}\Big)\Big]\Big] \\ &= \mathbb{E}_{P_{D_s}}\big[\mathbb{E}_{P_{W_\alpha,D_t|D_s}}[\log(P^{\gamma}_{W_\alpha|D_s,D_t})] - \mathbb{E}_{P_{W_\alpha|D_s}P_{D_t|D_s}}[\log(P^{\gamma}_{W_\alpha|D_s,D_t})]\big]. \end{aligned} \tag{48}$$

Combining with fact that D_s and D_t are independent, and plug in the posterior of α-weighted Gibbs algorithm, we have

$$\begin{aligned} I_{\mathrm{SKL}}(W_\alpha;D_t|D_s) &= \mathbb{E}_{P_{D_s}}[\gamma\mathbb{E}_{P_{W_\alpha,D_t|D_s}}[L_E(W_\alpha,D_s,D_t)] - \gamma\mathbb{E}_{P_{W_\alpha|D_s}P_{D_t}}[L_E(W_\alpha,D_s,D_t)]] \\ &= \gamma\mathbb{E}_{P_{D_s}}[\mathbb{E}_{P_{W_\alpha,D_t|D_s}}[(1-\alpha)L_E(w_\alpha,d_s)+\alpha L_E(w_\alpha,d_t)]] \\ &\quad - \gamma\mathbb{E}_{P_{D_s}}[\mathbb{E}_{P_{W_\alpha|D_s}P_{D_t}}[(1-\alpha)L_E(w_\alpha,d_s)+\alpha L_E(w_\alpha,d_t)]] \\ &= \gamma\alpha\big[\mathbb{E}_{P_{W_\alpha,D_t,D_s}}[L_E(w_\alpha,d_t)] - [\mathbb{E}_{P_{W_\alpha,D_s}P_{D_t}}[L_E(w_\alpha,d_t)]\big] \\ &= \gamma\alpha\overline{\mathrm{gen}}_\alpha(P_{D_s},P_{D_t}). \end{aligned} \tag{49}$$

□

Due to the symmetry of the α-weighted Gibbs algorithm, if we use $\overline{\mathrm{gen}}_\alpha(P_{D_t},P_{D_s})$ to denote the generalization error of treating P_{D_t} as source task and D_s as the target, we can obtain that $\overline{\mathrm{gen}}_\alpha(P_{D_t},P_{D_s}) = \frac{I_{\mathrm{SKL}}(W_\alpha;D_s|D_t)}{\gamma\alpha}$.

It is also worthwhile to mention that the α-weighted expected generalization error of both source and target tasks can be characterized in terms of symmetrized KL information as shown in the following Proposition.

Proposition 3. *For $(\gamma,\pi(w_\alpha),L_E(w_\alpha,d_s,d_t))$-Gibbs algorithm and $0<\alpha<1$, we have*

$$\alpha\overline{gen}_\alpha(P_{D_s},P_{D_t}) + (1-\alpha)\overline{gen}_\alpha(P_{D_t},P_{D_s}) = \frac{I_{\mathrm{SKL}}(W_\alpha;D_t,D_s)}{\gamma}. \tag{50}$$

Proof. The symmetrized KL information can be written as

$$I_{\mathrm{SKL}}(W_\alpha;D_t,D_s) = \mathbb{E}_{P_{W_\alpha,D_t,D_s}}\big[\log(P^{\gamma}_{W_\alpha|D_s,D_t})\big] - \mathbb{E}_{P_{W_\alpha}P_{D_t,D_s}}\big[\log(P^{\gamma}_{W_\alpha|D_s,D_t})\big]. \tag{51}$$

Plug in the posterior of α-weighted Gibbs algorithm,

$$\begin{aligned}
&I_{\mathrm{SKL}}(W_\alpha; D_t, D_s)\\
&= \mathbb{E}_{P_{W_\alpha,D_t,D_s}}\big[-\gamma L_E(w_\alpha,d_s,d_t)\big] + \mathbb{E}_{P_{W_\alpha}P_{D_t,D_s}}\big[\gamma L_E(w_\alpha,d_s,d_t)\big]\\
&= -\gamma\mathbb{E}_{P_{W_\alpha,D_t,D_s}}[\alpha L_E(w_\alpha,d_t)+(1-\alpha)L_E(w_\alpha,d_s)] + \gamma\mathbb{E}_{P_{W_\alpha}P_{D_t,D_s}}[\alpha L_E(w_\alpha,d_t)+(1-\alpha)L_E(w_\alpha,d_s)] \qquad (52)\\
&= \alpha\gamma\overline{\mathrm{gen}}_\alpha(P_{D_s},P_{D_t}) + (1-\alpha)\gamma\overline{\mathrm{gen}}_\alpha(P_{D_t},P_{D_s}). \qquad \square
\end{aligned}$$

Note that the Proposition 3 holds even for dependent source D_s and target D_t samples.

A.2 Two-stage Gibbs Algrotihm

Theorem 2. ***(restated)*** *The expected transfer generalization error of the two-stage Gibbs algorithm,*

$$P^{\gamma}_{W_c^t|D_t,W_\phi}(w_c^t|d_t,w_\phi) = \frac{\pi(w_c^t)e^{-\gamma L_E^{S_2}(w_\phi,w_c^t,d_t)}}{V_\beta(w_\phi,d_t,\gamma)},$$

is given by

$$\overline{gen}_\beta(P_{D_s},P_{D_t}) = \frac{I_{\mathrm{SKL}}(D_t;W_c^t|W_\phi)}{\gamma}.$$

Proof. In the second stage we freeze the share parameters W_ϕ, and we will update the specific target task parameter. Thus,

$$\begin{aligned}
&I_{\mathrm{SKL}}(W_c^t; D_t|W_\phi)\\
&= \mathbb{E}_{P_{W_\phi}}\big[\mathbb{E}_{P_{W_c^t,D_t|W_\phi}}[\log(P_{W_c^t|D_t,W_\phi})] - \mathbb{E}_{P_{W_c^t|W_\phi}P_{D_t|W_\phi}}[\log(P_{W_c^t|D_s,W_\phi})]\big]\\
&= \gamma\left(\mathbb{E}_{P_{W_\phi}}\big[\mathbb{E}_{P_{W_c^t|W_\phi}P_{D_t|W_\phi}}[L_E^{S2}(W_\phi,W_c^t,D_t)] - \mathbb{E}_{P_{W_c^s,D_t|W_\phi}}[L_E^{S2}(W_\phi,W_c^t,D_t)]\big]\right) \qquad (53)\\
&= \gamma\overline{\mathrm{gen}}_\beta(P_{D_s},P_{D_t}). \qquad \square
\end{aligned}$$

B Example: Mean Estimation

B.1 Symmetrized KL Divergence

The following lemma from Palomar and Verdú (2008) characterizes the mutual and lautum information for the Gaussian channel.

Lemma 1. *(Palomar and Verdú, 2008, Theorem 14) Consider the following model*

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{N}_{\mathrm{G}}, \qquad (54)$$

where $\boldsymbol{X} \in \mathbb{R}^{d_X}$ denotes the input random vector with zero mean (not necessarily Gaussian), $\boldsymbol{A} \in \mathbb{R}^{d_Y\times d_X}$ denotes the linear transformation undergone by the input, $\boldsymbol{Y} \in \mathbb{R}^{d_Y}$ is the output vector, and $\boldsymbol{N}_{\mathrm{G}} \in \mathbb{R}^{d_Y}$ is a Gaussian noise vector independent of $\boldsymbol{X}$. The input and the noise covariance matrices are given by $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{N_{\mathrm{G}}}$. Then, we have

$$I(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_{N_{\mathrm{G}}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) - D\big(P_{\boldsymbol{Y}}\|P_{N_{\mathrm{G}}}\big), \qquad (55)$$

$$L(\boldsymbol{X};\boldsymbol{Y}) = \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_{N_{\mathrm{G}}}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) + D\big(P_{\boldsymbol{Y}}\|P_{N_{\mathrm{G}}}\big). \qquad (56)$$

In the α-weighted Gibbs algorithm, the output W_α can be written as

$$W_\alpha = \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{\sigma_1^2}{\sigma^2}\Big(\sum_{i=1}^n Z_i^s + \sum_{j=1}^m Z_j^t\Big) + N = \frac{\sigma_1^2}{\sigma^2}\sum_{j=1}^m (Z_j^t - \boldsymbol{\mu}_t) + \frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{m\sigma_1^2}{\sigma^2}\boldsymbol{\mu}_t + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n Z_i^s + N, \qquad (57)$$

where $N \sim \mathcal{N}(0, \sigma_1^2 I_d)$, and $\sigma_1^2 = \frac{\sigma_0^2\sigma^2}{(m+n)\sigma_0^2+\sigma^2}$. For fixed sources training sample d_s, we can set $P_{N_G} \sim \mathcal{N}(\frac{\sigma_1^2}{\sigma_0^2}\boldsymbol{\mu}_0 + \frac{m\sigma_1^2}{\sigma^2}\boldsymbol{\mu}_t + \frac{\sigma_1^2}{\sigma^2}\sum_{i=1}^n z_i^s, \sigma_1^2 I_d)$ and $\boldsymbol{\Sigma} = \sigma_t^2 I_{nd}$ in Lemma 1 gives

$$I_{\mathrm{SKL}}(W_\alpha; D_t|D_s = d_s) = \mathrm{tr}\big(\boldsymbol{\Sigma}_{N_G}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) = \mathrm{tr}\big(\frac{\sigma_t^2}{\sigma_1^2}\boldsymbol{A}\boldsymbol{A}^\top\big). \tag{58}$$

Noticing that $\boldsymbol{A}\boldsymbol{A}^\top = \frac{m\sigma_1^4}{\sigma^4} I_d$ and taking expectation over P_S, we have

$$I_{\mathrm{SKL}}(W_\alpha; D_t|D_s) = \frac{md\sigma_0^2\sigma_t^2}{((m+n)\sigma_0^2+\sigma^2)\sigma^2}. \tag{59}$$

For the two-stage Gibbs algorithm, the output W_c^t can be written as

$$W_c^t = \frac{\sigma_c^2}{\sigma_0^2}\boldsymbol{\mu}_{0,c} + \frac{\sigma_c^2}{\sigma^2}\sum_{j=1}^m Z_{j,c}^t + N_c = \frac{\sigma_c^2}{\sigma^2}\sum_{j=1}^m (Z_{j,c}^t - \boldsymbol{\mu}_{t,c}) + \frac{\sigma_c^2}{\sigma_0^2}\boldsymbol{\mu}_{0,c} + \frac{n\sigma_c^2}{\sigma^2}\boldsymbol{\mu}_{t,c} + N_c, \tag{60}$$

where $N_c \sim \mathcal{N}(0, \sigma_c^2 I_{d_c})$, $\sigma_c^2 = \frac{\sigma_0^2\sigma^2}{m\sigma_0^2+\sigma^2}$, and subscript c stands for the task-specific component of the parameters. Since W_c^t is independent of the source samples, setting $P_{N_G} \sim \mathcal{N}(\frac{\sigma_c^2}{\sigma_0^2}\boldsymbol{\mu}_{0,c} + \frac{n\sigma_c^2}{\sigma^2}\boldsymbol{\mu}_{t,c}, \sigma_c^2 I_{d_c})$ and $\boldsymbol{\Sigma} = \sigma_t^2 I_{nd_c}$ in Lemma 1 gives

$$I_{\mathrm{SKL}}(W_c^t; D_t|W_\phi) = \mathrm{tr}\big(\boldsymbol{\Sigma}_{N_G}^{-1}\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\big) = \mathrm{tr}\big(\frac{\sigma_t^2}{\sigma_c^2}\boldsymbol{A}\boldsymbol{A}^\top\big) = \frac{md_c\sigma_0^2\sigma_t^2}{(m\sigma_0^2+\sigma^2)\sigma^2}, \tag{61}$$

where the last step follows due to the fact $\boldsymbol{A}\boldsymbol{A}^\top = \frac{m\sigma_c^4}{\sigma^4} I_{d_c}$ in this case.

B.2 Effect of Source samples

As shown in (23) and (24), the transfer generalization errors of this mean estimation problem only depend on the number of samples of D_s, and do not depend on the distribution P_{D_s}. In this subsection, we will show that, though different sources samples (distribution) do not change generalization error, they will influence the population risks and excess risks.

In this mean estimation example, the population risk of any W can be decomposed into

$$\begin{aligned} L_P(W, P_{D_t}) &= \mathbb{E}_{Z_t}[\|W - Z_t\|_2^2] = \mathbb{E}_{Z_t}[\|W - \mathbb{E}[W] + \mathbb{E}[W] - \boldsymbol{\mu}_t + \boldsymbol{\mu}_t - Z_t\|_2^2] \\ &= \|\mathbb{E}[W] - \boldsymbol{\mu}_t\|_2^2 + \mathrm{tr}(\mathrm{Cov}[W]) + d\sigma_t^2, \end{aligned} \tag{62}$$

where the first term, $\|\mathbb{E}[W] - \boldsymbol{\mu}_t\|_2^2$, is the squared bias, and the second term, $\mathrm{tr}(\mathrm{Cov}[W])$, is the variance. It is easy to verify that the optimal $\boldsymbol{w}^* = \arg\min L_P(W, P_{D_t})$ is just the target mean $\boldsymbol{\mu}_t$, and $L_P(\boldsymbol{w}^*, P_{D_t}) = d\sigma_t^2$, then the excess risk defined in (46) can be written as,

$$\mathcal{E}_r(P_W) = \|\mathbb{E}[W] - \boldsymbol{\mu}_t\|_2^2 + \mathrm{tr}(\mathrm{Cov}[W]). \tag{63}$$

For the α-weighted Gibbs algorithm in (60), it can be shown that

$$\mathrm{Bias} = \mathbb{E}[W_\alpha] - \boldsymbol{\mu}_t = \frac{\sigma^2(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_t) + n\sigma_0^2(\boldsymbol{\mu}_s - \boldsymbol{\mu}_t)}{(m+n)\sigma_0^2 + \sigma^2}, \tag{64}$$

$$\mathrm{tr}(\mathrm{Cov}[W_\alpha]) = \frac{d\sigma_1^4}{\sigma^4}(n\sigma_s^2 + m\sigma_t^2) + d\sigma_1^2. \tag{65}$$

The Bias term will be zero if $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_s = \boldsymbol{\mu}_t$. Thus, the excess risk of α-weighted Gibbs algorithm will be minimized when $\boldsymbol{\mu}_s = \boldsymbol{\mu}_t$ and $\sigma_s^2 = 0$, which is equivalent to the case that the target mean $\boldsymbol{\mu}_t$ is known.

For the two-stage Gibbs algorithm, if we learn the first d_ϕ components $\boldsymbol{\mu}_\phi \in \mathbb{R}^{d_\phi}$ using the $(\frac{n}{2\sigma^2}, \mathcal{N}(\boldsymbol{\mu}_{1,\phi}, \sigma_0^2 I_{d_\phi}), L_E^{S1}(\boldsymbol{w}_\phi, \boldsymbol{w}_c^s, d_s))$-Gibbs algorithm, and use the $(\frac{m}{2\sigma^2}, \mathcal{N}(\boldsymbol{\mu}_{2,c}, \sigma_0^2 I_{d_c}), L_E^{S2}(\boldsymbol{\mu}_\phi, \boldsymbol{w}_c^t, d_t))$-Gibbs

algorithm to learn the remain d_c components in the second stage, it can be shown that

$$\text{Bias}_\phi = \mathbb{E}[W_\phi] - \boldsymbol{\mu}_{t,\phi} = \frac{\sigma^2(\boldsymbol{\mu}_{1,\phi} - \boldsymbol{\mu}_{t,\phi}) + n\sigma_0^2(\boldsymbol{\mu}_{s,\phi} - \boldsymbol{\mu}_{t,\phi})}{n\sigma_0^2 + \sigma^2}, \tag{66}$$

$$\text{Bias}_c = \mathbb{E}[W_c^t] - \boldsymbol{\mu}_{t,c} = \frac{\sigma^2(\boldsymbol{\mu}_{2,c} - \boldsymbol{\mu}_{t,c})}{m\sigma_0^2 + \sigma^2}, \tag{67}$$

$$\text{tr}(\text{Cov}[W_\phi]) = \frac{nd_\phi\sigma_\phi^4\sigma_s^2}{\sigma^4} + d_\phi\sigma_\phi^2, \tag{68}$$

$$\text{tr}(\text{Cov}[W_c^t]) = \frac{md_c\sigma_c^4\sigma_t^2}{\sigma^4} + d_c\sigma_c^2, \tag{69}$$

with $\sigma_\phi^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$ and $\sigma_c^2 = \frac{\sigma_0^2\sigma^2}{m\sigma_0^2+\sigma^2}$. The excess risk of the two-stage Gibbs algorithm will be minimized when $\boldsymbol{\mu}_{s,\phi} = \boldsymbol{\mu}_{t,\phi}$ and $\sigma_s^2 = 0$, i.e., the optimal shared parameter $\boldsymbol{\mu}_{t,\phi}$ is known.

C Expected Transfer Generalization Error Upper Bound for General Learning Algorithm

C.1 Preliminaries

We first provide some preliminaries for our proofs in this section by introducing the notion of cumulant generating function, which characterizes different tail behaviors of random variables.

Definition 1. *The cumulant generating function (CGF) of a random variable X is defined as*

$$\Lambda_X(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]. \tag{70}$$

Assuming $\Lambda_X(\lambda)$ exists, it can be verified that $\Lambda_X(0) = \Lambda_X'(0) = 0$, and that it is convex.

Definition 2. *For a convex function ψ defined on the interval $[0, b)$, where $0 < b \leq \infty$, its Legendre dual $\psi^\star$ is defined as*

$$\psi^\star(x) \triangleq \sup_{\lambda\in[0,b)} \big(\lambda x - \psi(\lambda)\big). \tag{71}$$

The following lemma characterizes a useful property of the Legendre dual and its inverse function.

Lemma 2. *(Boucheron et al., 2013, Lemma 2.4) Assume that $\psi(0) = \psi'(0) = 0$. Then $\psi^\star(x)$ defined above is a non-negative convex and non-decreasing function on $[0, \infty)$ with $\psi^\star(0) = 0$. Moreover, its inverse function $\psi^{\star-1}(y) = \inf\{x \geq 0 : \psi^\star(x) \geq y\}$ is concave, and can be written as*

$$\psi^{\star-1}(y) = \inf_{\lambda\in[0,b)} \Big(\frac{y + \psi(\lambda)}{\lambda}\Big), \quad b > 0. \tag{72}$$

We consider the distributions with the following tail behaviors in the appendices:

- **Sub-Gaussian:** A random variable X is σ-sub-Gaussian, if $\psi(\lambda) = \frac{\sigma^2\lambda^2}{2}$ is an upper bound on $\Lambda_X(\lambda)$, for $\lambda \in \mathbb{R}$. Then by Lemma 2,
$$\psi^{\star-1}(y) = \sqrt{2\sigma^2 y}.$$
- **Sub-Exponential:** A random variable X is (σ_e^2, b)-sub-Exponential, if $\psi(\lambda) = \frac{\sigma_e^2\lambda^2}{2}$ is an upper bound on $\Lambda_X(\lambda)$, for $0 \leq |\lambda| \leq \frac{1}{b}$ and $b > 0$. Using Lemma 2, we have
$$\psi^{\star-1}(y) = \begin{cases} \sqrt{2\sigma_e^2 y}, & \text{if } y \leq \frac{\sigma_e^2}{2b}; \\ by + \frac{\sigma_e^2}{2b}, & \text{otherwise.} \end{cases}$$
- **Sub-Gamma:** A random variable X is $\Gamma(\sigma_s^2, c_s)$-sub-Gamma (Zhang and Chen, 2020), if $\psi(\lambda) = \frac{\lambda^2\sigma_s^2}{2(1-c_s|\lambda|)}$ is an upper bound on $\Lambda_X(\lambda)$, for $0 < |\lambda| < \frac{1}{c_s}$ and $c_s > 0$. Using Lemma 2, we have
$$\psi^{\star-1}(y) = \sqrt{2\sigma_s^2 y} + c_s y.$$

C.2 Proof of Theorem 3

We prove a more general form of Theorem 3 as follows:

Theorem 8. *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t and the loss function $\ell(W,Z)$ satisfies $\Lambda_{\ell(W,Z)}(\lambda) \leq \psi(-\lambda)$, for $\lambda \in (-b,0)$ and $\Lambda_{\ell(W,Z)}(\lambda) \leq \psi(\lambda)$, for $\lambda \in (0,b)$ and $b > 0$ under the distribution $P_Z^t \otimes P_W$. The following upper bound holds:*

$$|\overline{gen}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| \leq \psi^{\star -1}(\frac{I(W;D_t|D_s)}{m}). \tag{73}$$

Proof. The generalization error can be written as

$$|\overline{\text{gen}}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| \leq \frac{1}{m}\sum_{i=1}^{m} |\mathbb{E}_{P_{W,Z_i^t}}[\ell(W,Z_i^t)] - \mathbb{E}_{P_W \otimes P_Z^t}[\ell(W,Z^t)]|. \tag{74}$$

Using the Donsker–Varadhan variational representation (Boucheron et al., 2013), for all $\lambda \in (-b,+b)$,

$$D(P_{W,Z_i^t|d_s} \| P_{W|d_s} \otimes P_Z^t) \geq \mathbb{E}_{P_{W,Z_i^t|d_s}}[\lambda \ell(W,Z_i^t)] - \log(\mathbb{E}_{P_{W|d_s} \otimes P_Z^t}[e^{\lambda \ell(W,Z^t)}]). \tag{75}$$

Taking expectation respect to D_s over both sides, then we have

$$\begin{aligned} I(W;Z_i^t|D_s) &\geq \mathbb{E}_{P_{W,Z_i^t}}[\lambda \ell(W,Z_i^t)] - \mathbb{E}_{P_{D_s}}[\log(\mathbb{E}_{P_{W|D_s} \otimes P_Z^t}[e^{\lambda \ell(W,Z^t)}])] \\ &\geq \mathbb{E}_{P_{W,Z_i^t}}[\lambda \ell(W,Z_i^t)] - \log(\mathbb{E}_{P_W \otimes P_Z^t}[e^{\lambda \ell(W,Z^t)}]) \\ &\geq \lambda(\mathbb{E}_{P_{W,Z_i^t}}[\ell(W,Z_i^t)] - \mathbb{E}_{P_W \otimes P_Z^t}[\ell(W,Z^t)]) - \psi(\lambda). \end{aligned} \tag{76}$$

Using similar approach as in (Bu et al., 2020, Theorem 1),

$$|\mathbb{E}_{P_{W,Z_i^t}}[\ell(W,Z_i^t)] - \mathbb{E}_{P_W \otimes P_Z^t}[\ell(W,Z^t)]| \leq \psi^{\star -1}(I(W;Z_i^t|D_s)). \tag{77}$$

Now by combining (74) and (77), we have:

$$\begin{aligned} |\overline{\text{gen}}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| &\leq \frac{1}{m}\sum_{i=1}^{m} \psi^{\star -1}(I(W;Z_i^t|D_s)) \\ &\leq \psi^{\star -1}\Big(\frac{1}{m}\sum_{i=1}^{m} I(W;Z_i^t|D_s)\Big) \\ &\leq \psi^{\star -1}\Big(\frac{I(W,D_t|D_s)}{m}\Big), \end{aligned} \tag{78}$$

where the inequality follows due to the concavity of $\psi^{\star -1}$ function and the Independence between Z_i^t. □

Theorem 3. ***(restated)*** *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W,Z)$ is σ-sub-Gaussian under the distribution $P_Z^t \otimes P_W$. Then, the following upper bound holds*

$$|\overline{gen}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| \leq \sqrt{\frac{2\sigma^2}{m} I(W;D_t|D_s)}.$$

Proof. For σ-subgaussian assumption, we have $\psi^{\star -1}(y) = \sqrt{2\sigma^2 y}$ in Theorem 8 and this completes the proof. □

Remark 6. *Similar upper bound on the expected transfer generalization error in Theorem 3 holds by considering a different assumption that the loss function $\ell(w,Z)$ is σ-sub-Gaussian under the distribution P_Z^t for all $w \in \mathcal{W}$.*

C.3 Other Tail Distributions

Using Theorem 8, we can also provide upper bounds on the expected transfer generalization error for any general learning algorithms under sub-Exponential and sub-Gamma assumptions.

Corollary 1 (Sub-Exponential)**.** *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W,Z)$ (σ_e^2, b)-sub-Exponential under distribution $P_Z^t \otimes P_W$. Then the following upper bound holds*

$$|\overline{gen}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| \leq \begin{cases} \sqrt{2\sigma_e^2 \frac{I(W;D_t|D_s)}{m}}, & \text{if } \frac{I(W;D_t|D_s)}{m} \leq \frac{\sigma_e^2}{2b}; \\ b\frac{I(W;D_t|D_s)}{m} + \frac{\sigma_e^2}{2b}, & \text{otherwise.} \end{cases} . \tag{79}$$

Corollary 2 (Sub-Gamma)**.** *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W,Z)$ is $\Gamma(\sigma_s^2, c_s)$-sub-Gamma under distribution $P_Z^t \otimes P_W$. Then, the following upper bound holds*

$$|\overline{gen}(P_{W|D_s,D_t}, P_{D_s}, P_{D_t})| \leq \sqrt{2\sigma_s^2 \frac{I(W;D_t|D_s)}{m}} + c_s \frac{I(W;D_t|D_s)}{m}. \tag{80}$$

D Distribution-free Upper Bound on Generalization Error

Theorem 4. ***(restated)*** *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W,Z)$ is σ_α-sub-Gaussian under the distribution $P_Z^t \otimes P_{W_\alpha}$. If we further assume $C_\alpha \leq \frac{L(W_\alpha;D_t|D_s)}{I(W_\alpha;D_t|D_s)}$ for some $C_\alpha \geq 0$, then for the α-weighted Gibbs algorithm and $0 < \alpha < 1$,*

$$\overline{gen}_\alpha(P_{D_s}, P_{D_t}) \leq \frac{2\sigma_\alpha^2 \gamma\alpha}{(1+C_\alpha)m}.$$

Proof. By equation (26) in Theorem 3, we have

$$\begin{aligned} \overline{\text{gen}}_\alpha(P_{D_s}, P_{D_t}) &= \frac{I_{\text{SKL}}(W_\alpha; D_t|D_s)}{\gamma\alpha} \\ &\leq \sqrt{\frac{2\sigma^2 I(W_\alpha; D_t|D_s)}{m}}. \end{aligned} \tag{81}$$

As we have $I(W_\alpha; D_t|D_s)(1+C_\alpha) \leq I_{\text{SKL}}(W_\alpha; D_t|D_s)$ in the assumption, the following upper bound holds:

$$\frac{I(W_\alpha; D_t|D_s)(1+C_\alpha)}{\gamma\alpha} \leq \sqrt{\frac{2\sigma_\alpha^2 I(W_\alpha; D_t|D_s)}{m}}, \tag{82}$$

which implies that

$$I(W_\alpha; D_t|D_s) \leq \frac{2\sigma_\alpha^2 \gamma^2\alpha^2}{(1+C_\alpha)^2 m}. \tag{83}$$

Combining (83) with (81) completes the proof. □

Theorem 5. ***(restated)*** *Suppose that the target training samples $D_t = \{Z_j^t\}_{j=1}^m$ are i.i.d generated from the distribution P_Z^t, and the non-negative loss function $\ell(W_c, w_\phi, Z)$ is σ_β-sub-Gaussian under distribution $P_Z^t \otimes P_{W_c^t|W_\phi=w_\phi}$ for all $w_\phi \in \mathcal{W}_\phi$. If we further assume $C_\beta \leq \frac{L(W_c^t;D_t|W_\phi)}{I(W_c^t;D_t|W_\phi)}$ for some $C_\beta \geq 0$, then for the two-stage Gibbs algorithm,*

$$P^\gamma_{W_c^t|D_t,W_\phi}(w_c^t|d_t, w_\phi) = \frac{\pi(w_c^t) e^{-\gamma L_E^{S_2}(w_\phi, w_c^t, d_t)}}{V_\beta(w_\phi, d_t, \gamma)},$$

we have

$$\overline{gen}_\beta(P_{D_s}, P_{D_t}) \leq \frac{2\sigma_\beta^2 \gamma}{(1+C_\beta)m}.$$

Proof. Using Theorem 3 by considering $W = (W_C^t, W_\phi)$,

$$|\overline{\text{gen}}_\beta(P_{D_s}, P_{D_t})| \leq \sqrt{\frac{2\sigma^2}{m} I(W_c^t, W_\phi; D_t|D_s)}.$$

Now, based on chain rule for mutual information we have

$$\begin{aligned} I(W_c^t, W_\phi; D_t|D_s) &= I(W_\phi; D_t|D_s) + I(W_c^t; D_t|D_s, W_\phi) \\ &= I(W_c^t; D_t|W_\phi), \end{aligned}$$

where $I(W_\phi; D_t|D_s) = 0$ due to the fact that W_ϕ is independent from D_t given D_s, and $I(W_c^t; D_t|W_\phi, D_s) = I(W_c^t; D_t|W_\phi)$ since $D_s \perp (W_c^t, D_t)|W_\phi$.

Using Theorem 2, it can be shown that

$$\overline{\text{gen}}_\beta(P_{D_s}, P_{D_t}) = \frac{I_{\text{SKL}}(D_t; W_c^t|W_\phi)}{\gamma} \leq \sqrt{\frac{2\sigma_\beta^2}{m} I(W_c^t; D_t|W_\phi)}. \tag{84}$$

As we have $I(W_c^t; D_t|W_\phi)(1 + C_\beta) \leq I_{\text{SKL}}(W_c^t; D_t|W_\phi)$, the following bound holds:

$$\frac{I(W_c^t; D_t|W_\phi)(1 + C_\beta)}{\gamma} \leq \sqrt{\frac{2\sigma_\beta I(W_c^t; D_t|W_\phi)}{m}}, \tag{85}$$

which implies that

$$I(W_c^t; D_t|W_\phi) \leq \frac{2\sigma_\beta^2 \gamma^2}{(1 + C_\beta)^2 m}. \tag{86}$$

Combining (86) with (84) completes the proof. □

We could provide distribution-free upper bounds under sub-Exponential and sub-Gamma assumption using similar approach as in Theorem 4 and Theorem 5 for α-weighted Gibbs algorithm and two-stage Gibbs algorithm, respectively.

sub-Exponential: For α-weighted Gibbs algorithm, we assume that the loss function is $(\sigma_{\alpha,e}^2, b_\alpha)$-sub-Exponential under distribution $P_Z^t \otimes P_{W_\alpha}$. And for two-stage Gibbs algorithm, we assume that the loss function is $(\sigma_{\beta,e}^2, b_\beta)$-sub-Exponential under distribution $P_Z^t \otimes P_{W_c^t|W_\phi = w_\phi}$ for all $w_\phi \in \mathcal{W}_\phi$. We provide the results in Table 2. Denote $B_\alpha \triangleq \lceil \frac{\gamma\alpha b_\alpha}{1+C_\alpha} \rceil$, $B_\beta \triangleq \lceil \frac{\gamma b_\beta}{1+C_\beta} \rceil$, $I_\alpha \triangleq \frac{2b_\alpha I(W_\alpha; D_t|D_s)}{\sigma_{\alpha,e}^2}$ and $I_\beta \triangleq \frac{2b_\beta I(W_c^t; D_t|W_\phi)}{\sigma_{\beta,e}^2}$ in Table 2.

sub-Gamma: For α-weighted Gibbs algorithm, we assume that the loss function is $\Gamma(\sigma_{\alpha,s}^2, c_{\alpha,s})$-sub-Gamma under distribution $P_Z^t \otimes P_{W_\alpha}$ and $m > \frac{\gamma\alpha c_{\alpha,s}}{(1+C_\alpha)}$. For two-stage Gibbs algorithm, we assume that the loss function is $\Gamma(\sigma_{\beta,s}^2, c_{\beta,s})$-sub-Gamma under distribution $P_Z^t \otimes P_{W_c^t|W_\phi = w_\phi}$ for all $w_\phi \in \mathcal{W}_\phi$ and $m > \frac{\gamma c_{\beta,s}}{(1+C_\beta)}$. We provide the results in Table 2.

Table 2: Distribution-free Upper Bounds under different Tail Distributions.

	sub-Exponential	**sub-Gamma**
α-weighted Gibbs Algorithm	$\begin{cases} \frac{2\sigma_{\alpha,e}^2 \gamma\alpha}{m(1+C_\alpha)}, & \text{if } m \geq I_\alpha; \\ \frac{\sigma_{\alpha,e}^2}{2b_\alpha}\left(\frac{\gamma\alpha b_\alpha}{(m(1+C_\alpha)-\gamma\alpha b_\alpha)} + 1\right), & \text{if } B_\alpha < m < I_\alpha \end{cases}$	$\frac{2\sigma_{\alpha,s}^2 \gamma\alpha}{(1+C_\alpha)m-\gamma\alpha c_{\alpha,s}}\left(1 + \frac{\gamma\alpha c_{\alpha,s}}{(1+C_\alpha)m-\gamma\alpha c_{\alpha,s}}\right)$
Two-stage Gibbs Algorithm	$\begin{cases} \frac{2\sigma_{\beta,e}^2 \gamma}{m(1+C_\beta)}, & \text{if } m \geq I_\beta; \\ \frac{\sigma_{\beta,e}^2}{2b_\beta}\left(\frac{\gamma b_\beta}{(m(1+C_E)-\gamma b_\beta)} + 1\right), & \text{if } B_\beta < m < I_\beta \end{cases}$	$\frac{2\sigma_{\beta,s}^2 \gamma}{(1+C_\beta)m-\gamma c_{\beta,s}}\left(1 + \frac{\gamma c_{\beta,s}}{(1+C_\beta)m-\gamma c_{\beta,s}}\right)$

E Exact Characterization of Generalization Error Based on Symmetrized KL divergence

We first present the following Lemma to prove the results related to symmetrized KL divergence.

Lemma 3. *Denote the $(\gamma, \pi(w), L_E(w, d_t))$-Gibbs algorithm as $P^{\gamma}_{W|D_t}$ and the $(\gamma, \pi(w), L_P(w, P_{D_t}))$-Gibbs algorithm as $P_W^{\gamma, L_{P_{D_t}}}$. Then, the following equality holds for these two Gibbs distributions with the same inverse temperature and prior distribution*

$$\mathbb{E}_{\Delta(P^{\gamma}_{W|D_t=d_t}, P_W^{\gamma, L_{P_{D_t}}})}[L_P(W, P_{D_t}) - L_E(W, d_t)] = \frac{D_{\mathrm{SKL}}(P^{\gamma}_{W|D_t=d_t} \| P_W^{\gamma, L_{P_{D_t}}})}{\gamma}, \tag{87}$$

where $\mathbb{E}_{\Delta(P^{\gamma}_{W|D_t=d_t}, P_W^{\gamma, L_{P_{D_t}}})}[f(W)] = \mathbb{E}_{P^{\gamma}_{W|D_t=d_t}}[f(W)] - \mathbb{E}_{P_W^{\gamma, L_{P_{D_t}}}}[f(W)]$.

Proof.

$$\begin{aligned} D_{\mathrm{SKL}}(P^{\gamma}_{W|D_t=d_t} \| P_W^{\gamma, L_{P_{D_t}}}) &= \int_{\mathcal{W}} (P^{\gamma}_{W|D_t=d_t} - P_W^{\gamma, L_{P_{D_t}}}) \log \left(\frac{P^{\gamma}_{W|D_t=d_t}}{P_W^{\gamma, L_{P_{D_t}}}} \right) dw \\ &= \int_{\mathcal{W}} (P^{\gamma}_{W|D_t=d_t} - P_W^{\gamma, L_{P_{D_t}}}) \log(e^{-\gamma(L_E(w,d_t) - L_P(w, P_{D_t}))}) dw \\ &= \gamma \mathbb{E}_{\Delta(P^{\gamma}_{W|D_t=d_t}, P_W^{\gamma, L_{P_{D_t}}})}[L_P(W, P_{D_t}) - L_E(W, d_t)]. \end{aligned} \tag{88}$$

□

Using Lemma 3, we provide different characterizations of α-weighted Gibbs algorithm and two-stage Gibbs algorithm using symmetrized KL divergence.

E.1 α-weighted Gibbs Algorithm

Theorem 6. ***(restated)*** *The expected transfer generalization error of the α-weighted Gibbs algorithm in (10) is given by:*

$$\overline{gen}_{\alpha}(P_{D_s}, P_{D_t}) = \frac{D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s,D_t} \| P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s} | P_{D_s} P_{D_t})}{\gamma \alpha}, \tag{89}$$

where $P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s}$ is the $(\gamma, \pi(w_\alpha), L_\alpha(w_\alpha, d_s, P_{D_t}))$-Gibbs algorithm with $L_\alpha(w, d_s, P_{D_t}) \triangleq \alpha L_P(w_\alpha, P_{D_t}) + (1-\alpha) L_E(w_\alpha, d_s)$.

Proof. Applying Lemma 3 to the α-weighted Gibbs algorithm and $(\gamma, \pi(w_\alpha), L_\alpha(w, d_s, P_{D_t}))$-Gibbs algorithm gives

$$\begin{aligned} &\frac{D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s=d_s,D_t=d_t} \| P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s=d_s})}{\gamma} \\ &= \mathbb{E}_{\Delta\left(P^{\gamma}_{W_\alpha|D_s=d_s,D_t=d_t}, P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s=d_s}\right)}[L_\alpha(W_\alpha, d_s, P_{D_t}) - L_E(W_\alpha, d_s, d_t)] \\ &= \alpha \mathbb{E}_{\Delta\left(P^{\gamma}_{W_\alpha|D_s=d_s,D_t=d_t}, P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s=d_s}\right)}[L_P(W_\alpha, P_{D_t}) - L_E(W_\alpha, d_t)]. \end{aligned} \tag{90}$$

Notice the fact that

$$\mathbb{E}_{P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s=d_s}}[L_P(W_\alpha, P_{D_t})] = \mathbb{E}_{P_{D_t}}\left[\mathbb{E}_{P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s=d_s}}[L_E(W_\alpha, D_t)]\right],$$

and taking expectation over D_s and D_t, we have

$$\begin{aligned} D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s,D_t} \| P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|D_s} | P_{D_s} P_{D_t}) &= \mathbb{E}_{P_{D_s} P_{D_t}}[D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|d_s,d_t} \| P^{\gamma, L_\alpha(w_\alpha, d_s, P_{D_t})}_{W_\alpha|d_s})], \\ &= \gamma \alpha \overline{\mathrm{gen}}_{\alpha}(P_{D_s}, P_{D_t}). \end{aligned}$$

□

In the following, we provide an explanation for the existence of two different characterizations of the expected transfer generalization error, i.e., Theorem 6 and Theorem 1.

For an arbitrary conditional distribution on hypothesis space $Q_{W_\alpha|D_s}$, we can write

$$I(W_\alpha; D_t|D_s) = D(P_{W_\alpha,D_t|D_s}\|Q_{W_\alpha|D_s}\otimes P_{D_t}|P_{D_s}) - D(P_{W_\alpha|D_s}\|Q_{W_\alpha|D_s}|P_{D_s}), \tag{91}$$

$$L(W_\alpha; D_t|D_s) = \mathbb{E}_{P_{D_s}}\left[\mathbb{E}_{P_{D_t}\otimes P_{W_\alpha|D_s}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})]\right] + D(P_{W_\alpha|D_s}\|Q_{W_\alpha|D_s}|P_{D_s}). \tag{92}$$

Thus, the symmetrized KL information can be written as

$$\begin{aligned} I_{\mathrm{SKL}}(W_\alpha; D_t|D_s) &= I(W_\alpha; D_t|D_s) + L(W_\alpha; D_t|D_s) \\ &= D(P_{W_\alpha,D_t|D_s}\|Q_{W_\alpha|D_s}\otimes P_{D_t}|P_{D_s}) + \mathbb{E}_{P_{D_s}}\left[\mathbb{E}_{P_{D_t}\otimes P_{W_\alpha|D_s}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})]\right], \end{aligned} \tag{93}$$

which holds for all $Q_{W_\alpha|D_s}$. We compare this expression with the following representation:

$$D(P_{W_\alpha,D_t|D_s}\|Q_{W_\alpha|D_s}\otimes P_{D_t}|P_{D_s}) + D(Q_{W_\alpha|D_s}\otimes P_{D_t}\|P_{W_\alpha,D_t|D_s}|P_{D_s}). \tag{94}$$

The difference between these two expressions is as follows:

$$\begin{aligned} &I_{\mathrm{SKL}}(W_\alpha; D_t|D_s) - \left(D(P_{W_\alpha,D_t|D_s}\|Q_{W_\alpha|D_s}\otimes P_{D_t}|P_{D_s}) + D(Q_{W_\alpha|D_s}\otimes P_{D_t}\|P_{W_\alpha,D_t|D_s}|P_{D_s})\right) \\ &= \mathbb{E}_{P_{D_s}}\left[\mathbb{E}_{P_{D_t}\otimes P_{W_\alpha|D_s}}[\log(Q_{W_\alpha|D_s}/P_{W|D_t,D_s})]\right] - D(Q_{W_\alpha|D_s}\otimes P_{D_t}\|P_{W_\alpha,D_t|D_s}|P_{D_s}) \\ &= \mathbb{E}_{P_{D_s}}\left[\mathbb{E}_{P_{D_t}\otimes P_{W_\alpha|D_s}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})] - \mathbb{E}_{P_{D_t}\otimes Q_{W_\alpha|D_s}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})]\right] \\ &= \mathbb{E}_{P_{D_s}}\left[\mathbb{E}_{\Delta(P_{W_\alpha|D_s},Q_{W_\alpha|D_s})}[\mathbb{E}_{P_{D_t}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})]]\right]. \end{aligned} \tag{95}$$

Thus, if $Q_{W_\alpha|D_s}$ satisfies the following condition

$$\mathbb{E}_{\Delta(P_{W_\alpha|D_s}-Q_{W_\alpha|D_s})}[\mathbb{E}_{P_{D_t}}[\log(Q_{W_\alpha|D_s}/P_{W_\alpha|D_t,D_s})]] = 0, \tag{96}$$

then we have

$$I_{\mathrm{SKL}}(W_\alpha; D_t) = D(P_{W_\alpha,D_t|D_s}\|Q_{W_\alpha|D_s}\otimes P_{D_t}|P_{D_s}) + D(Q_{W_\alpha|D_s}\otimes P_{D_t}\|P_{W_\alpha,D_t|D_s}|P_{D_s}). \tag{97}$$

Now, if we set $(\gamma, \pi(w), L_E(w, d_s, d_t))$-Gibbs algorithm as $P_{W_\alpha|D_t,D_s}$, then it can be verified that using $(\gamma, \pi(w), L_\alpha(w_\alpha, d_s, P_{D_t}))$-Gibbs algorithm as $Q_{W_\alpha|D_s}$ would satisfy the condition in (96). Thus, we can represent the expected transfer generalization error using both symmetrized KL information and divergence.

E.2 Two-stage Gibbs Algorithm

Theorem 7. ***(restated)*** *The expected transfer generalization error of the two-stage Gibbs algorithm in (12) is given by:*

$$\overline{gen}_\beta(P_{D_s}, P_{D_t}) = \frac{D_{\mathrm{SKL}}(P^\gamma_{W_c^t|D_t,W_\phi}\|P^{\gamma,L_P(w_\phi,w_c^t,P_{D_t})}_{W_c^t|W_\phi}|P_{D_t}P_{W_\phi})}{\gamma},$$

where $P^{\gamma,L_P(w_\phi,w_c^t,P_{D_t})}_{W_c^t|W_\phi}$ *is the* $(\gamma, \pi(w_c^t), L_P(w_\phi, w_c^t, P_{D_t}))$*-Gibbs algorithm.*

Proof. Applying Lemma 3 to the two-stage Gibbs algorithm and $(\gamma, \pi(w_c^t), L_P(w_\phi, w_c^t, P_{D_t}))$-Gibbs algorithm, we have

$$\begin{aligned} &\frac{D_{\mathrm{SKL}}(P^\gamma_{W_c^t|D_t=d_t,W_\phi=w_\phi}\|P^{\gamma,L_P(w_\phi,W_c^t,P_{D_t})}_{W_c^t|W_\phi=w_\phi})}{\gamma} \\ &= \mathbb{E}_{\Delta\left(P^\gamma_{W_c^t|D_t=d_t,W_\phi=w_\phi},P^{\gamma,L_P(w_\phi,w_c^t,P_{D_t})}_{W_c^t|W_\phi=w_\phi}\right)}\left[L_P(W_c^t, w_\phi, P_{D_t}) - L_E(W_c^t, w_\alpha, d_t)\right]. \end{aligned} \tag{98}$$

Notice the fact that

$$\mathbb{E}_{P^{\gamma,L_P(w_\phi,w_c^t,P_{D_t})}_{W_c^t|W_\phi=w_\phi}}[L_P(W_c^t, w_\phi, P_{D_t})] = \mathbb{E}_{P_{D_t}}\left[\mathbb{E}_{P^{\gamma,L_P(w_\phi,w_c^t,P_{D_t})}_{W_c^t|W_\phi=w_\phi}}[L_E(W_c^t, w_\phi, d_t)]\right],$$

and taking expectation over W_ϕ and D_t completes the proof. □

F Asymptotic Behavior of Generalization Error for Gibbs Algorithm

We first provide a summary of different variables used in the Appendix F in Table 3.

Table 3: Notations and Definitions

Notation	Definition
$\|W\|_H^2$	$W^\top HW$
$\hat{W}_\alpha(D_s, D_t)$	$\arg\min_{w\in\mathcal{W}} L_E(w, D_s, D_t)$
$\hat{W}_\alpha(D_s)$	$\arg\min_{w\in\mathcal{W}} L_\alpha(w, D_s, P_{D_t})$
$\boldsymbol{w}_\alpha^*$	$\arg\min_{\boldsymbol{w}\in\mathcal{W}} nD(P_Z^s\|f(\cdot\|\boldsymbol{w})) + mD(P_Z^t\|f(\cdot\|\boldsymbol{w}))$
$\hat{W}_c^t(D_t, W_\phi)$	$\arg\min_{w_c\in\mathcal{W}_c} L_E^{S2}(W_\phi, w_c, D_t)$
$\hat{W}_c^t(W_\phi)$	$\arg\min_{w_c\in\mathcal{W}_c} L_P(W_\phi, w_c, P_{D_t})$
$[\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{s*}]$	$\arg\min_{[\boldsymbol{w}_\phi,\boldsymbol{w}_c]\in\mathcal{W}} D(P_{Z^s}\|f(\cdot\|\boldsymbol{w}_\phi, \boldsymbol{w}_c))$
$\boldsymbol{w}_c^{st*}$	$\arg\min_{\boldsymbol{w}_c\in\mathcal{W}_c} D(P_{Z^t}\|f(\cdot\|\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c))$

F.1 Generalization Error

Proposition 1. ***(restated)*** *If the Hessian matrices $H^*(D_s, D_t) = H^*(D_s) = H^*$ are independent of D_s and D_t, then the generalization error of the α-weighted-ERM algorithm is*

$$\overline{gen}_\alpha(P_{D_t}, P_{D_s}) = \frac{\mathbb{E}_{P_{D_s,D_t}}[\|\hat{W}_\alpha(D_s, D_t) - \hat{W}_\alpha(D_s)\|_{H^*}^2]}{\alpha}.$$

Proof. It is shown in Hwang (1980) that if the following Hessian matrices

$$H^*(D_s, D_t) \triangleq \nabla_w^2 L_E(w, D_s, D_t)\big|_{w=\hat{W}_\alpha(D_s,D_t)}, \tag{99}$$
$$H^*(D_s) \triangleq \nabla_w^2 L_\alpha(w, D_s, P_{D_t})\big|_{w=\hat{W}_\alpha(D_s)} \tag{100}$$

are not singular, then, as $\gamma \to \infty$

$$P_{W_\alpha|D_s,D_t}^\gamma \to \mathcal{N}(\hat{W}_\alpha(D_s, D_t), \frac{1}{\gamma} H^*(D_s, D_t)^{-1}),$$
$$\text{and} \quad P_{W_\alpha|D_s}^{\gamma,L_\alpha} \to \mathcal{N}(\hat{W}_\alpha(D_s), \frac{1}{\gamma} H^*(D_s)^{-1}), \tag{101}$$

and we use $P_{W_\alpha|D_s}^{\gamma,L_\alpha}$ to denote $P_{W_\alpha|D_s}^{\gamma,L_\alpha(w_\alpha,d_s,P_{D_t})}$.

Thus, the conditional symmetrized KL divergence in Theorem 6 can be evaluated directly using Gaussian ap-

proximations under the assumption that $H^*(D_s, D_t) = H^*(D_s) = H^*$,

$$
\begin{aligned}
&D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s,D_t} \| P^{\gamma,L_\alpha}_{W_\alpha|D_s} | P_{D_s} P_{D_t}) \\
&= \mathbb{E}_{P_{D_t,D_s}}\Big[\mathbb{E}_{P^{\gamma}_{W_\alpha|D_s,D_t}}\big[\log \frac{P^{\gamma}_{W_\alpha|D_s,D_t}}{P^{\gamma,L_\alpha}_{W_\alpha|D_s}}\big] - \mathbb{E}_{P^{\gamma,L_\alpha}_{W_\alpha|D_s}}\big[\log \frac{P^{\gamma}_{W_\alpha|D_s,D_t}}{P^{\gamma,L_\alpha}_{W_\alpha|D_s}}\big]\Big] \\
&= \mathbb{E}_{P_{D_t,D_s}}\Big[\mathbb{E}_{\Delta(P^{\gamma}_{W_\alpha|D_s,D_t}, P^{\gamma,L_\alpha}_{W_\alpha|D_s})}\big[-\frac{\gamma}{2}(W_\alpha - \hat{W}_\alpha(D_s,D_t))^\top H^*(W_\alpha - \hat{W}_\alpha(D_s,D_t)) \\
&\qquad + \frac{\gamma}{2}(W_\alpha - \hat{W}_\alpha(D_s))^\top H^*(W_\alpha - \hat{W}_\alpha(D_s))\big]\Big] \\
&= \gamma \mathbb{E}_{P_{D_t,D_s}}\Big[\mathbb{E}_{\Delta(P^{\gamma}_{W_\alpha|D_s,D_t}, P^{\gamma,L_\alpha}_{W_\alpha|D_s})}\big[W_\alpha^\top H^* \hat{W}_\alpha(D_s,D_t) - W_\alpha^\top H^* \hat{W}_\alpha(D_s)\big]\Big] \\
&= \gamma \mathbb{E}_{P_{D_t,D_s}}\big[\hat{W}_\alpha(D_s,D_t)^\top H^* \hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s,D_t)^\top H^* \hat{W}_\alpha(D_s) \\
&\qquad - \hat{W}_\alpha(D_s)^\top H^* \hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s)^\top H^* \hat{W}_\alpha(D_s)\big] \\
&= \gamma \mathbb{E}_{P_{D_t,D_s}}\big[(\hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s))^\top H^* (\hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s))\big]. \qquad (102)
\end{aligned}
$$

Thus,

$$
\overline{\mathrm{gen}}_\alpha(P_{D_s}, P_{D_t}) = \frac{D_{\mathrm{SKL}}(P^{\gamma}_{W_\alpha|D_s,D_t} \| P^{\gamma,L_\alpha}_{W_\alpha|D_s} | P_{D_s} P_{D_t})}{\gamma\alpha} = \frac{\mathbb{E}_{P_{D_s,D_t}}[\|\hat{W}_\alpha(D_s,D_t) - \hat{W}_\alpha(D_s)\|^2_{H^*}]}{\alpha}. \qquad \square
$$

Proposition 2. ***(restated)*** *If Hessian matrices $H^*_c(D_t, W_\phi) = H^*_c(W_\phi) = H^*_c$ are independent of D_s, D_t, then the generalization error of the two-stage-ERM algorithm is*

$$
\overline{gen}_\beta(P_{D_t}, P_{D_s}) = \mathbb{E}_{D_s,D_t,W_\phi}[\|\hat{W}^t_c(D_t,W_\phi) - \hat{W}^t_c(W_\phi)\|^2_{H^*_c}].
$$

Proof. It is shown in Hwang (1980) that if the following Hessian matrices

$$
H^*_c(D_t,W_\phi) \triangleq \nabla^2_{w_c} L^{S2}_E(W_\phi, w_c, D_t)\big|_{w_c=\hat{W}^t_c(D_t,W_\phi)} \qquad (103)
$$

$$
H^*_c(W_\phi) \triangleq \nabla^2_{w_c} L_P(W_\phi, w_c, P_{D_t})\big|_{w_c=\hat{W}^t_c(W_\phi)} \qquad (104)
$$

are not singular, then, as $\gamma \to \infty$

$$
\begin{aligned}
P^{\gamma}_{W^t_c|D_t,W_\phi} &\to \mathcal{N}(\hat{W}^t_c(D_t,W_\phi), \frac{1}{\gamma} H^*_c(D_t,W_\phi)^{-1}), \\
P^{\gamma,L_P}_{W^t_c|W_\phi} &\to \mathcal{N}(\hat{W}^t_c(W_\phi), \frac{1}{\gamma} H^*(D_s)^{-1}), \qquad (105)
\end{aligned}
$$

where we use $P^{\gamma,L_P}_{W^t_c|W_\phi}$ to denote $P^{\gamma,L_P(w_\phi,w^t_c,P_{D_t})}_{W^t_c|W_\phi}$. Thus, the conditional symmetrized KL divergence in Theorem 7 can be evaluated directly using Gaussian approximations under the assumption that $H^*_c(D_t,W_\phi) = H^*_c(W_\phi) = H^*_c$.

$$
\begin{aligned}
&D_{\mathrm{SKL}}(P^{\gamma}_{W^t_c|D_t,W_\phi} \| P^{\gamma,L_P}_{W^t_c|W_\phi} | P_{D_t} P_{W_\phi}) \\
&= \mathbb{E}_{P_{D_t,W_\phi}}\Big[\mathbb{E}_{P^{\gamma}_{W^t_c|D_t,W_\phi}}\big[\log \frac{P^{\gamma}_{W^t_c|D_t,W_\phi}}{P^{\gamma,L_P}_{W^t_c|W_\phi}}\big] - \mathbb{E}_{P^{\gamma,L_P}_{W^t_c|W_\phi}}\big[\log \frac{P^{\gamma}_{W^t_c|D_t,W_\phi}}{P^{\gamma,L_P}_{W^t_c|W_\phi}}\big]\Big] \\
&= \mathbb{E}_{P_{D_t,W_\phi}}\Big[\mathbb{E}_{\Delta(P^{\gamma}_{W^t_c|D_t,W_\phi}, P^{\gamma,L_P}_{W^t_c|W_\phi})}\big[-\frac{\gamma}{2}(W^t_c - \hat{W}^t_c(D_t,W_\phi))^\top H^*_c (W^t_c - \hat{W}^t_c(D_t,W_\phi)) \\
&\qquad + \frac{\gamma}{2}(W^t_c - \hat{W}^t_c(W_\phi))^\top H^*_c (W^t_c - \hat{W}^t_c(W_\phi))\big]\Big] \\
&= \gamma \mathbb{E}_{P_{D_t,W_\phi}}\Big[\mathbb{E}_{\Delta(P^{\gamma}_{W^t_c|D_t,W_\phi}, P^{\gamma,L_P}_{W^t_c|W_\phi})}\big[(W^t_c)^\top H^*_c \hat{W}^t_c(D_t,W_\phi) - (W^t_c)^\top H^*_c \hat{W}^t_c(W_\phi)\big]\Big] \\
&= \gamma \mathbb{E}_{P_{D_t,W_\phi}}\big[\hat{W}^t_c(D_t,W_\phi)^\top H^*_c \hat{W}^t_c(D_t,W_\phi) - \hat{W}^t_c(D_t,W_\phi)^\top H^*_c \hat{W}^t_c(W_\phi) \\
&\qquad - \hat{W}^t_c(W_\phi)^\top H^*_c \hat{W}^t_c(D_t,W_\phi) - \hat{W}^t_c(W_\phi)^\top H^*_c \hat{W}^t_c(W_\phi)\big] \\
&= \gamma \mathbb{E}_{P_{D_t,W_\phi}}\big[(\hat{W}^t_c(D_t,W_\phi) - \hat{W}^t_c(W_\phi))^\top H^*_c (\hat{W}^t_c(D_t,W_\phi) - \hat{W}^t_c(W_\phi))\big]. \qquad (106)
\end{aligned}
$$

Thus,

$$\overline{\text{gen}}_\beta(P_{D_t},P_{D_s})=\frac{D_{\text{SKL}}(P^{\gamma}_{W^t_c|D_t,W_\phi}\|P^{\gamma,L_P}_{W^t_c|W_\phi}|P_{D_t}P_{W_\phi})}{\gamma}=\mathbb{E}_{D_t,W_\phi}[\|\hat{W}^t_c(D_t,W_\phi)-\hat{W}^t_c(W_\phi)\|^2_{H^*_c}]. \qquad \square$$

F.2 Regularity Conditions for MLE

In this section, we present the regularity conditions required by the asymptotic normality (Van der Vaart, 2000) of maximum likelihood estimates.

Assumption 1. ***Regularity Conditions for MLE:***

1. *$f(z|\boldsymbol{w})\neq f(z|\boldsymbol{w}')$ for $\boldsymbol{w}\neq\boldsymbol{w}'$.*
2. *$\mathcal{W}$ is an open subset of $\mathbb{R}^d$.*
3. *The function $\log f(z|\boldsymbol{w})$ is three times continuously differentiable with respect to $\boldsymbol{w}$.*
4. *There exist functions $F_1(z):\mathcal{Z}\to\mathbb{R}$, $F_2(z):\mathcal{Z}\to\mathbb{R}$ and $M(z):\mathcal{Z}\to\mathbb{R}$, such that*

$$\mathbb{E}_{Z\sim f(z|\boldsymbol{w})}[M(Z)]<\infty,$$

and the following inequalities hold for any $\boldsymbol{w}\in\mathcal{W}$,

$$\left|\frac{\partial\log f(z|\boldsymbol{w})}{\partial w_i}\right|<F_1(z),\qquad \left|\frac{\partial^2\log f(z|\boldsymbol{w})}{\partial w_i\partial w_j}\right|<F_1(z),$$
$$\left|\frac{\partial^3\log f(z|\boldsymbol{w})}{\partial w_i\partial w_j\partial w_k}\right|<M(z),\qquad i,j,k=1,2,\cdots,d.$$

5. *The following inequality holds for an arbitrary $\boldsymbol{w}\in\mathcal{W}$,*

$$0<\mathbb{E}_{Z\sim f(z|\boldsymbol{w})}\left[\frac{\partial\log f(z|\boldsymbol{w})}{\partial w_i}\frac{\partial\log f(z|\boldsymbol{w})}{\partial w_j}\right]<\infty,\quad i,j=1,2,\cdots,d.$$

F.3 Generalization error in MLE

α-weighted ERM: We use the following notations to denote the expectation of the Hessian matrices and the Fisher information matrices,

$$J_s(\boldsymbol{w}_\alpha)\triangleq\mathbb{E}_{P_Z^s}\big[-\nabla^2_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)\big],\quad J_t(\boldsymbol{w}_\alpha)\triangleq\mathbb{E}_{P_Z^t}\big[-\nabla^2_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)\big],$$
$$\mathcal{I}_s(\boldsymbol{w}_\alpha)\triangleq\mathbb{E}_{P_Z^s}\big[\nabla_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)\nabla_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)^\top\big],\quad \mathcal{I}_t(\boldsymbol{w}_\alpha)\triangleq\mathbb{E}_{P_Z^t}\big[\nabla_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)\nabla_{\boldsymbol{w}_\alpha}\log f(Z|\boldsymbol{w}_\alpha)^\top\big],$$
$$\bar{J}(\boldsymbol{w}_\alpha)=\frac{n}{m+n}J_s(\boldsymbol{w}_\alpha)+\frac{m}{m+n}J_t(\boldsymbol{w}_\alpha),\quad \bar{\mathcal{I}}(\boldsymbol{w}_\alpha)=\frac{n}{m+n}\mathcal{I}_s(\boldsymbol{w}_\alpha)+\frac{m}{m+n}\mathcal{I}_t(\boldsymbol{w}_\alpha).$$

Lemma 4. *Under Assumption 1, for any fixed source samples d_s, if we let $m\to\infty$, then the α-weighted ERM satisfies*

$$\sqrt{m}\big(\hat{W}_\alpha(d_s,D_t)-\hat{W}_\alpha(d_s)\big)\to\mathcal{N}\big(0,\alpha^2\widetilde{J}(\hat{W}_\alpha(d_s))^{-1}\mathcal{I}_t(\hat{W}_\alpha(d_s))\widetilde{J}(\hat{W}_\alpha(d_s))^{-1}\big), \tag{107}$$

where $\widetilde{J}(\hat{W}_\alpha(d_s))\triangleq\alpha J_t(\hat{W}_\alpha(d_s))+(1-\alpha)\nabla^2_w L_E(w,d_s)\big|_{w=\hat{W}_\alpha(d_s)}$, and $\mathcal{I}_t(\hat{W}_\alpha(d_s))$ is the covariance matrix of $\nabla_w\log f(Z^t|\hat{W}_\alpha(d_s))$.

Proof. By using a Taylor expansion of the first derivative of the weighted log-likelihood $L_E(\hat{W}_\alpha(d_s,D_t),d_s,D_t)$ around $\hat{W}_\alpha(d_s)$, we obtain

$$\begin{aligned}0&=\nabla_w L_E(w,d_s,D_t)\big|_{w=\hat{W}_\alpha(d_s,D_t)}\\&\approx\nabla_w L_E(w,d_s,D_t)\big|_{w=\hat{W}_\alpha(d_s)}+\nabla^2_w L_E(w,d_s,D_t)\big|_{w=\hat{W}_\alpha(d_s)}(\hat{W}_\alpha(d_s,D_t)-\hat{W}_\alpha(d_s)).\end{aligned} \tag{108}$$

From the Taylor series expansion formula, the following approximation can be obtained

$$-\nabla_w^2 L_E(w, d_s, D_t)\big|_{w=\hat{W}_\alpha(d_s)}(\hat{W}_\alpha(d_s, D_t) - \hat{W}_\alpha(d_s)) \approx \nabla_w L_E(w, d_s, D_t)\big|_{w=\hat{W}_\alpha(d_s)}. \tag{109}$$

By the law of large numbers, when $m \to \infty$, it can be shown that

$$-\nabla_w^2 L_E(\hat{W}_\alpha(d_s), D_t) = \frac{1}{m}\sum_{i=1}^{m} \nabla_w^2 \log f(Z_i^t|\hat{W}_\alpha(d_s)) \to -J_t(\hat{W}_\alpha(d_s)). \tag{110}$$

Thus, the LHS of (109) can be written as

$$\nabla_w^2 L_E(w, d_s, D_t)\big|_{w=\hat{W}_\alpha(d_s)} = \nabla_w^2\big[\alpha L_E(w, D_t) + (1-\alpha)L_E(w, d_s)\big]\big|_{w=\hat{W}_\alpha(d_s)} \to \widetilde{J}(\hat{W}_\alpha(d_s)), \tag{111}$$

where $\widetilde{J}(\hat{W}_\alpha(d_s)) = \alpha J_t(\hat{W}_\alpha(d_s)) + (1-\alpha)\nabla_w^2 L_E(w, d_s)\big|_{w=\hat{W}_\alpha(d_s)}$.

As for the RHS of (109), note that

$$\sqrt{m}\nabla_w L_E(w, D_t)\big|_{w=\hat{W}_\alpha(d_s)} = -\frac{1}{\sqrt{m}}\sum_{i=1}^{m} \nabla_w \log f(Z_i^t|\hat{W}_\alpha(d_s)), \tag{112}$$

by multivariate central limit theorem

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\Big(-\nabla_w \log f(Z_i^t|\hat{W}_\alpha(d_s)) + \mathbb{E}_{Z^t}[\nabla_w \log f(Z^t|\hat{W}_\alpha(d_s))]\Big) \to \mathcal{N}(0, \mathcal{I}_t(\hat{W}_\alpha(d_s))), \tag{113}$$

where $\mathcal{I}_t(\hat{W}_\alpha(d_s))$ is the covariance matrix of $\nabla_w \log f(Z^t|\hat{W}_\alpha(d_s))$.

Due to the definition of $\hat{W}_\alpha(d_s)$, we have $\nabla_w L_E(w, d_s, P_{D_t})\big|_{w=\hat{W}_\alpha(d_s)} = 0$, i.e.,

$$(1-\alpha)\nabla_w L_E(\hat{W}_\alpha(d_s), d_s) = \alpha\mathbb{E}_{Z^t}[\nabla_w \log f(Z^t|\hat{W}_\alpha(d_s))]. \tag{114}$$

Thus, the RHS of (109) will converge to

$$\sqrt{m}\nabla_w L_E(w, D_s, D_t)\big|_{w=\hat{W}_\alpha(D_s)} \to \mathcal{N}\big(0, \alpha^2\mathcal{I}_t(\hat{W}_\alpha(d_s))\big). \tag{115}$$

Combining with (110), when $m \to \infty$, we obtain

$$\sqrt{m}\big(\hat{W}_\alpha(d_s, D_t) - \hat{W}_\alpha(d_s)\big) \to \mathcal{N}\big(0, \alpha^2\widetilde{J}(\hat{W}_\alpha(d_s))^{-1}\mathcal{I}_t(\hat{W}_\alpha(d_s))\widetilde{J}(\hat{W}_\alpha(d_s))^{-1}\big). \tag{116}$$

□

In the main body of the paper, we further let $n \to \infty$, then $\hat{W}_\alpha(d_s) \to \boldsymbol{w}_\alpha^*$, and $\widetilde{J}(\hat{W}_\alpha(d_s)) \to \bar{J}(\boldsymbol{w}_\alpha^*)$, $\mathcal{I}_t(\hat{W}_\alpha(d_s)) \to \mathcal{I}_t(\boldsymbol{w}_\alpha^*)$. For $\alpha = \frac{m}{m+n}$, using Lemma 4, we can show that

$$\hat{W}_\alpha(D_s, D_t) - \hat{W}_\alpha(D_s) \to \mathcal{N}\big(0, \frac{m}{(m+n)^2}\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\mathcal{I}_t(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\big). \tag{117}$$

In addition, the Hessian matrix $H^*(D_s, D_t) \to \bar{J}(\boldsymbol{w}_\alpha^*)$ as $m, n \to \infty$, which is independent of the samples D_s, D_t. Proposition 1 gives

$$\overline{\text{gen}}_\alpha(P_{D_t}, P_{D_s}) = \frac{\text{tr}(\mathcal{I}_t(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1})}{n+m} = \mathcal{O}(\frac{d}{m+n}).$$

Two-stage ERM:

We use the following notations to denote the expectation of the Hessian matrix and the Fisher information matrix with respect to $\boldsymbol{w}_c$,

$$J_c^t(\boldsymbol{w}_\phi, \boldsymbol{w}_c) \triangleq \mathbb{E}_{P_Z^t}\big[-\nabla_{\boldsymbol{w}_c}^2 \log f(Z|[\boldsymbol{w}_\phi, \boldsymbol{w}_c])\big],$$
$$\mathcal{I}_c^t(\boldsymbol{w}_\phi, \boldsymbol{w}_c) \triangleq \mathbb{E}_{P_Z^t}[\nabla_{\boldsymbol{w}_c} \log f(Z|[\boldsymbol{w}_\phi, \boldsymbol{w}_c])]\nabla_{\boldsymbol{w}_c}^\top \log f(Z|[\boldsymbol{w}_\phi, \boldsymbol{w}_c])].$$

Lemma 5. *Under Assumption 1, for any fixed $\hat{\boldsymbol{w}}_\phi$, if we let $m \to \infty$, then the two-stage ERM satisfies*

$$\sqrt{m}\big((\hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi) - \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))\big) \to \mathcal{N}\big(0, J_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))^{-1}\mathcal{I}_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))J_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))^{-1}\big). \tag{118}$$

Proof. For any fixed $\hat{\boldsymbol{w}}_\phi$, using a Taylor expansion of the gradient with respect to $\boldsymbol{w}_c$ of the log-likelihood $L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi), D_t)$ around $\hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)$, we obtain

$$\begin{aligned}
0 &= \nabla_{\boldsymbol{w}_c} L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi), D_t) \\
&\approx \nabla_{\boldsymbol{w}_c} L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi), D_t) + \nabla_{\boldsymbol{w}_c}^2 L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi), D_t)(\hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi) - \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)).
\end{aligned}$$

From the Taylor series expansion formula, the following approximation can be obtained

$$-\nabla_{\boldsymbol{w}_c}^2 L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi), D_t)(\hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi) - \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)) \approx \nabla_{\boldsymbol{w}_c} L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi), D_t). \tag{119}$$

By the law of large numbers, when $m \to \infty$, it can be shown that

$$-\nabla_{\boldsymbol{w}_c}^2 L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi), D_t) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{w}_c}^2 \log f(Z_i^t|[\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)]) \to -J_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)). \tag{120}$$

As for the RHS of (119), note that $\mathbb{E}_{P_Z^t}[\nabla_{\boldsymbol{w}_c} \log f(Z|[\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)])] = 0$ due to the definition of $\hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)$, by multivariate central limit theorem, we have

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{n}\Big(-\nabla_{\boldsymbol{w}_c} \log f(Z_i^t|[\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)])\Big) \to \mathcal{N}(0, \mathcal{I}_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))), \tag{121}$$

where $\mathcal{I}_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)) = \mathbb{E}_{P_Z^t}[\nabla_{\boldsymbol{w}_c} \log f(Z|[\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)])\nabla_{\boldsymbol{w}_c}^\top \log f(Z|[\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)])]$.

Thus, the RHS of (119) will converge to

$$\sqrt{m}\nabla_{\boldsymbol{w}_c} L_E^{S2}(\hat{\boldsymbol{w}}_\phi, \boldsymbol{w}_c, D_t)\big|_{\boldsymbol{w}_c = \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi)} \to \mathcal{N}\big(0, \mathcal{I}_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))\big). \tag{122}$$

When $m \to \infty$, we obtain

$$\sqrt{m}\big((\hat{W}_c^t(D_t, \hat{\boldsymbol{w}}_\phi) - \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))\big) \to \mathcal{N}\big(0, J_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))^{-1}\mathcal{I}_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))J_c^t(\hat{\boldsymbol{w}}_\phi, \hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi))^{-1}\big). \tag{123}$$

□

In the main body of the paper, we further let $n \to \infty$, then $\hat{\boldsymbol{w}}_\phi \to \boldsymbol{w}_\phi^{s*}$, and $\hat{\boldsymbol{w}}_c^t(\hat{\boldsymbol{w}}_\phi) \to \boldsymbol{w}_c^{st*}$. Using Lemma 5, we can show that

$$\hat{W}_c^t(D_t, \hat{W}_\phi) - \hat{W}_c^t(\hat{W}_\phi) \to \mathcal{N}\big(0, \frac{J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}}{m}\big).$$

As the Hessian matrix $H_c^*(D_t, W_\phi) = H_c^*(W_\phi) \to J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})$ as $m, n \to \infty$. By Proposition 2, we have

$$\overline{\text{gen}}_\beta(P_{D_t}, P_{D_s}) = \frac{\text{tr}\big(\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}\big)}{m} = \mathcal{O}(\frac{d_c}{m}). \tag{124}$$

F.4 Excess risk

α-weighted ERM: In the following lemma, we characterize the variance of the α-weighted ERM algorithm.

Lemma 6. *Under Assumption 1, if we let $m, n \to \infty$, then the α-weighted ERM satisfies*

$$\sqrt{m+n}\big(\hat{W}_\alpha(D_s, D_t) - \boldsymbol{w}_\alpha^*\big) \to \mathcal{N}\big(0, \bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\bar{\mathcal{I}}_t(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\big). \tag{125}$$

Proof. By using a Taylor expansion of the first derivative of the weighted log-likelihood $L_E(\hat{W}_\alpha(D_s,D_t),D_s,D_t)$ around $\boldsymbol{w}_\alpha^*$, we obtain

$$0=\nabla_w L_E(w,D_s,D_t)\big|_{w=\hat{W}_\alpha(D_s,D_t)}\approx\nabla_w L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}+\nabla_w^2 L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}(\hat{W}_\alpha(D_s,D_t)-\boldsymbol{w}_\alpha^*).$$

From the Taylor series expansion formula, the following approximation can be obtained

$$-\nabla_w^2 L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}(\hat{W}_\alpha(D_s,D_t)-\boldsymbol{w}_\alpha^*)\approx\nabla_w L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}. \tag{126}$$

By the law of large numbers, when $m,n\to\infty$, it can be shown that

$$-\nabla_w^2 L_E(\boldsymbol{w}_\alpha^*,D_t)=\frac{1}{m}\sum_{i=1}^m\nabla_w^2\log f(Z_i^t|\boldsymbol{w}_\alpha^*)\to -J_t(\boldsymbol{w}_\alpha^*), \tag{127}$$

$$-\nabla_w^2 L_E(\boldsymbol{w}_\alpha^*,D_s)=\frac{1}{n}\sum_{i=1}^n\nabla_w^2\log f(Z_i^s|\boldsymbol{w}_\alpha^*)\to -J_s(\boldsymbol{w}_\alpha^*). \tag{128}$$

Thus, the LHS of (126) converges to

$$\nabla_w^2 L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}\to\bar{J}(\boldsymbol{w}_\alpha^*), \tag{129}$$

where $\bar{J}(\boldsymbol{w}_\alpha^*)\triangleq\alpha J_t(\boldsymbol{w}_\alpha^*)+(1-\alpha)J_t(\boldsymbol{w}_\alpha^*)$.

As for the RHS of (126), by multivariate central limit theorem

$$\frac{1}{\sqrt{m}}\sum_{i=1}^n\Big(-\nabla_w\log f(Z_i^t|\boldsymbol{w}_\alpha^*)+\mathbb{E}_{Z^t}[\nabla_w\log f(Z^t|\boldsymbol{w}_\alpha^*)]\Big)\to\mathcal{N}(0,\mathcal{I}_t(\boldsymbol{w}_\alpha^*)), \tag{130}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\Big(-\nabla_w\log f(Z_i^s|\boldsymbol{w}_\alpha^*)+\mathbb{E}_{Z^s}[\nabla_w\log f(Z^s|\boldsymbol{w}_\alpha^*)]\Big)\to\mathcal{N}(0,\mathcal{I}_s(\boldsymbol{w}_\alpha^*)), \tag{131}$$

where $\mathcal{I}_t(\boldsymbol{w}_\alpha^*)$ and $\mathcal{I}_s(\boldsymbol{w}_\alpha^*)$ are the covariance matrix of $\nabla_w\log f(Z^t|\boldsymbol{w}_\alpha^*)$ and $\nabla_w\log f(Z^s|\boldsymbol{w}_\alpha^*)$, respectively.

Due to the definition of $\boldsymbol{w}_\alpha^*$, we have

$$(1-\alpha)\mathbb{E}_{Z^s}[\nabla_w\log f(Z^s|\boldsymbol{w}_\alpha^*)]+\alpha\mathbb{E}_{Z^t}[\nabla_w\log f(Z^t|\boldsymbol{w}_\alpha^*)]=0. \tag{132}$$

Thus, the RHS of (126) will converge to

$$\nabla_w L_E(w,D_s,D_t)\big|_{w=\boldsymbol{w}_\alpha^*}\to\mathcal{N}\big(0,\frac{\alpha^2}{m}\mathcal{I}_t(\boldsymbol{w}_\alpha^*)+\frac{(1-\alpha)^2}{n}\mathcal{I}_s(\boldsymbol{w}_\alpha^*)\big). \tag{133}$$

When $m,n\to\infty$, we obtain

$$\big(\hat{W}_\alpha(D_s,D_t)-\boldsymbol{w}_\alpha^*\big)\to\mathcal{N}\big(0,\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}(\frac{\alpha^2}{m}\mathcal{I}_t(\boldsymbol{w}_\alpha^*)+\frac{(1-\alpha)^2}{n}\mathcal{I}_s(\boldsymbol{w}_\alpha^*))\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\big). \tag{134}$$

For $\alpha=\frac{m}{m+n}$, if we denote $\bar{\mathcal{I}}(\boldsymbol{w}_\alpha)=\frac{n}{m+n}\mathcal{I}_s(\boldsymbol{w}_\alpha)+\frac{m}{m+n}\mathcal{I}_t(\boldsymbol{w}_\alpha)$, we have

$$\big(\hat{W}_\alpha(D_s,D_t)-\boldsymbol{w}_\alpha^*\big)\to\mathcal{N}\big(0,\frac{1}{m+n}\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\bar{\mathcal{I}}(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\big). \tag{135}$$

□

Thus, the variance term in the excess risk can be computed as:

$$\mathrm{tr}(J_t(\boldsymbol{w}_t^*)\mathrm{Cov}(\hat{W}_\alpha(D_s,D_t)))=\frac{\mathrm{tr}(J_t(\boldsymbol{w}_t^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1}\bar{\mathcal{I}}(\boldsymbol{w}_\alpha^*)\bar{J}(\boldsymbol{w}_\alpha^*)^{-1})}{m+n}=\mathcal{O}(\frac{d}{m+n}). \tag{136}$$

Two-stage ERM: We use the following notations to denote the expectation of the Hessian matrix and the Fisher information matrix with respect to $\boldsymbol{w}_\phi$,

$$\begin{aligned}J_{c,\phi}^t(\boldsymbol{w}_\phi,\boldsymbol{w}_c)&\triangleq\mathbb{E}_{P_Z^t}\big[-\nabla^2_{\boldsymbol{w}_c,\boldsymbol{w}_\phi}\log f(Z|[\boldsymbol{w}_\phi,\boldsymbol{w}_c])\big],\\ J_\phi^s(\boldsymbol{w}_\phi)&\triangleq\mathbb{E}_{P_Z^s}\big[-\nabla^2_{\boldsymbol{w}_\phi}\log f(Z|[\boldsymbol{w}_\phi,\boldsymbol{w}_c])\big],\\ \mathcal{I}_\phi^s(\boldsymbol{w}_\phi,\boldsymbol{w}_c)&\triangleq\mathbb{E}_{P_Z^s}[\nabla_{\boldsymbol{w}_\phi}\log f(Z|[\boldsymbol{w}_\phi,\boldsymbol{w}_c])]\nabla^\top_{\boldsymbol{w}_\phi}\log f(Z|[\boldsymbol{w}_\phi,\boldsymbol{w}_c])].\end{aligned}$$

In the following lemma, we characterize the variance of the two-stage ERM algorithm.

Lemma 7. *Under Assumption 1, if we let $m, n \to \infty$, then the two-stage ERM satisfies*

$$\begin{aligned}&(\hat{W}_c^t(\hat{W}_\phi, D_t) - \boldsymbol{w}_c^{st*}) \to \mathcal{N}\Big(0, J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1} \\ &\big(\frac{1}{m}\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) + \frac{1}{n} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} \mathcal{I}_\phi^s(\boldsymbol{w}_\phi^{s*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})\big) J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}\Big).\end{aligned} \tag{137}$$

Proof. By using a Taylor expansion of the gradient with respect to $\boldsymbol{w}_c$ of the log-likelihood $L_E^{S2}(\hat{W}_\phi(D_s), \hat{W}_c^t(\hat{W}_\phi, D_t), D_t)$ around $[\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}]$, we obtain

$$\begin{aligned}0 &= \nabla_{\boldsymbol{w}_c} L_E^{S2}(\hat{W}_\phi(D_s), \hat{W}_c^t(\hat{W}_\phi, D_t), D_t) \\ &\approx \nabla_{\boldsymbol{w}_c} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t) + \nabla^2_{\boldsymbol{w}_c, \boldsymbol{w}_\phi} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t)(\hat{W}_\phi(D_s) - \boldsymbol{w}_\phi^{s*}) \\ &\quad + \nabla^2_{\boldsymbol{w}_c} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t)(\hat{W}_c^t(\hat{W}_\phi, D_t) - \boldsymbol{w}_c^{st*}).\end{aligned}$$

From the Taylor series expansion formula, the following approximation can be obtained

$$\begin{aligned}&- \nabla^2_{\boldsymbol{w}_c} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t)(\hat{W}_c^t(\hat{W}_\phi, D_t) - \boldsymbol{w}_c^{st*}) \\ &\approx \nabla_{\boldsymbol{w}_c} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t) + \nabla^2_{\boldsymbol{w}_c, \boldsymbol{w}_\phi} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t)(\hat{W}_\phi(D_s) - \boldsymbol{w}_\phi^{s*}).\end{aligned} \tag{138}$$

By the law of large numbers, when $m \to \infty$, it can be shown that

$$- \nabla^2_{\boldsymbol{w}_c} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t) = \frac{1}{m}\sum_{i=1}^m \nabla^2_{\boldsymbol{w}_c} \log f(Z_i^t | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}]) \to -J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}), \tag{139}$$

$$- \nabla^2_{\boldsymbol{w}_c, \boldsymbol{w}_\phi} L_E^{S2}(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}, D_t) = \frac{1}{m}\sum_{i=1}^m \nabla^2_{\boldsymbol{w}_c, \boldsymbol{w}_\phi} \log f(Z_i^t | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}]) \to -J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}). \tag{140}$$

As for the first term in the RHS of (138), note that $\mathbb{E}_{P_Z^t}[\nabla_{\boldsymbol{w}_c} \log f(Z | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}])] = 0$, by multivariate central limit theorem, we have

$$\frac{1}{\sqrt{m}}\sum_{i=1}^n \Big(- \nabla_{\boldsymbol{w}_c} \log f(Z_i^t | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}])\Big) \to \mathcal{N}(0, \mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})). \tag{141}$$

When $n \to \infty$, due to the asymptotic normality of maximum likelihood estimate, we have

$$\sqrt{n}(\hat{W}_\phi(D_s) - \boldsymbol{w}_\phi^{s*}) \to \mathcal{N}(0, J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} \mathcal{I}_\phi^s(\boldsymbol{w}_\phi^{s*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1}), \tag{142}$$

where $\mathcal{I}_\phi^s(\boldsymbol{w}_\phi^{s*}) = \mathbb{E}_{P_Z^s}[\nabla_{\boldsymbol{w}_\phi} \log f(Z | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{s*}]) \nabla_{\boldsymbol{w}_\phi}^\top \log f(Z | [\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{s*}])]$.

Thus, the RHS of (138) converges to

$$\mathcal{N}\Big(0, \ \frac{1}{m}\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) + \frac{1}{n} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} \mathcal{I}_\phi^s(\boldsymbol{w}_\phi^{s*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})\Big) \tag{143}$$

when $m, n \to \infty$.

Thus, we obtain

$$\begin{aligned}&(\hat{W}_c^t(\hat{W}_\phi, D_t) - \boldsymbol{w}_c^{st*}) \to \mathcal{N}\Big(0, J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1} \\ &\big(\frac{1}{m}\mathcal{I}_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) + \frac{1}{n} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} \mathcal{I}_\phi^s(\boldsymbol{w}_\phi^{s*}) J_\phi^s(\boldsymbol{w}_\phi^{s*})^{-1} J_{c,\phi}^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})\big) J_c^t(\boldsymbol{w}_\phi^{s*}, \boldsymbol{w}_c^{st*})^{-1}\Big).\end{aligned} \tag{144}$$

□

Note that $\mathrm{Cov}(\hat{W}_\phi(D_s))$ can be characterized by the asymptotic normality of maximum likelihood estimate. Thus, the variance term in the excess risk can be computed as:

$$\mathrm{tr}\big(J_t(\boldsymbol{w}_\phi^{t*}, \boldsymbol{w}_c^{t*}) \mathrm{Cov}(\hat{W}_\phi(D_s), \hat{W}_c^t(D_t, \hat{W}_\phi))\big) = \mathcal{O}(\frac{d_c}{m} + \frac{d}{n}). \tag{145}$$