How Does Pseudo-Labeling Affect the Generalization Error of the Semi-Supervised Gibbs Algorithm?

Haiyun He Cornell University
Gholamali Aminian The Alan Turing Institute
Yuheng Bu Univ. of Florida
Miguel Rodrigues Univ. College London
Vincent Y. F. Tan Nat. Univ. of Singapore

Abstract

We provide an exact characterization of the expected generalization error (gen-error) for semi-supervised learning (SSL) with pseudo-labeling via the Gibbs algorithm. The gen-error is expressed in terms of the symmetrized KL information between the output hypothesis, the pseudo-labeled dataset, and the labeled dataset. Distribution-free upper and lower bounds on the gen-error can also be obtained. Our findings offer new insights that the generalization performance of SSL with pseudo-labeling is affected not only by the information between the output hypothesis and input training data but also by the information *shared* between the *labeled* and *pseudo-labeled* data samples. This serves as a guideline to choose an appropriate pseudo-labeling method from a given family of methods. To deepen our understanding, we further explore two examples—mean estimation and logistic regression. In particular, we analyze how the ratio of the number of unlabeled to labeled data λ affects the gen-error under both scenarios. As λ increases, the gen-error for mean estimation decreases and then saturates at a value larger than when all the samples are labeled, and the gap can be quantified *exactly* with our analysis, and is dependent on the *cross-covariance* between the labeled and pseudo-labeled data samples. For logistic regression, the gen-error and the variance component of the excess risk also decrease as λ increases.

1 INTRODUCTION

There are several areas, like natural language processing, computer vision, and finance, where labeled data are rare but unlabeled data are abundant. In these situations, semi-supervised learning (SSL) techniques enable us to utilize both labeled and unlabeled data.

Self-training algorithms (Ouali et al., 2020) are a subcategory of SSL techniques. These algorithms use the supervised-learned model's confident predictions to predict the labels of unlabeled data. Entropy minimization and pseudo-labeling are two basic approaches used in self-training-based SSL. The entropy function may be viewed as a regularization term that penalizes uncertainty in the label prediction of unlabeled data in entropy minimization approaches (Grandvalet et al., 2005). The *manifold assumption* (Iscen et al., 2019)—where it is assumed that labeled and unlabeled features are drawn from a common data manifold—or the *cluster assumption* (Chapelle et al., 2003)—where it is assumed that similar data features have a similar label—are assumptions for adopting the entropy minimization algorithm. In contrast, in pseudo-labeling, which is the focus of this work, the model is trained using labeled data and then used to produce a pseudo-label for the unlabeled data (Lee et al., 2013). These pseudo-labels are then utilized to build another model, which is trained using both labeled and pseudo-labeled data in a supervised manner. Studying the generalization error (gen-error) of this procedure is critical to understanding and improving pseudo-labeling performance.

There have been various efforts to characterize the gen-error of SSL algorithms. In Rigollet (2007), an upper bound on the gen-error of binary classification under the cluster assumption is derived. Niu et al. (2013) provides an upper bound on gen-error based on the Rademacher complexity for binary classification with squared-loss mutual information regularization. Göpfert et al. (2019) employs the VC-dimension method to characterize the SSL gen-error. In Göpfert et al. (2019) and Zhu (2020), upper bounds for SSL gen-error using Bayes classifiers are provided. Zhu (2020) also provides an upper bound on the excess risk of SSL algorithm by assuming an exponentially concave function[1] based on the conditional mutual information. He et al. (2022) investigates the gen-error of iterative SSL techniques based

Correspondence to Haiyun He <haiyun.he@u.nus.edu>.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

[1] A function $f(x)$ is called β-exponentially concave function if $\exp(-\beta f(x))$ is concave

on pseudo-labels. An information-theoretic gen-error upper bound on self-training algorithms under the covariate-shift assumption is proposed by Aminian et al. (2022a). More discussions on related works are provided in Appendix A. These upper bounds on excess risk and gen-error do not entirely capture the impact of SSL, in particular pseudo-labeling and the relative number of labeled and unlabeled data, and thus constrain our ability to fully comprehend the performance of SSL.

In this paper, we are interested in characterizing the expected gen-error of pseudo-labeling-based SSL—using an appropriately-designed Gibbs algorithm—and studying how it depends on the output hypothesis, the labeled, and the pseudo-labeled data. Moreover, we intend to understand the effect of the ratio between the numbers of the unlabeled and labeled training data examples on the gen-error in different scenarios.

Our main contributions in this paper are as follows:

- We provide an exact characterization of the expected gen-error of Gibbs algorithm that models pseudo-labeling-based SSL. This characterization can be applied to obtain novel and informative upper and lower bounds.
- The characterization and bounds offer an insight that reducing the shared information between the labeled and pseudo-labeled samples can help to improve the generalization performance of pseudo-labeling-based SSL.
- We analyze the effect of the ratio of the number of unlabeled data to labeled data λ on the gen-error of Gibbs algorithm using a mean estimation example.
- Finally, we study the asymptotic behaviour of the Gibbs algorithm and analyze the effect of λ on the gen-error and the excess risk, applying our results to logistic regression.

2 SEMI-SUPERVISED LEARNING VIA THE GIBBS ALGORITHM

In this section, we formulate our problem using the Gibbs algorithm based on both the labeled and unlabeled training data with pseudo-labels. The Gibbs algorithm is a tractable and idealized model for learning algorithms with various approaches, e.g., stochastic optimization methods or relative entropy regularization (Raginsky et al., 2017).

2.1 Problem Formulation

Let $S_{\mathrm{l}} = \{\mathbf{Z}_i\}_{i=1}^n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be the labeled training dataset, where $\mathbf{X}_i \in \mathcal{X} = \mathbb{R}^d$ is the data feature, $Y_i \in \mathcal{Y} = [K]$ is the class label and each pair of $\mathbf{Z}_i = (\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$ is independently and identically distributed (i.i.d.) from $P_{\mathbf{Z}} = P_{\mathbf{X},Y} \in \mathcal{P}(\mathcal{Z})$. Conditioned on $\mathbf{X}_i$, each label Y_i is i.i.d. from $P_{Y|\mathbf{X}}$. Let $S_{\mathrm{u}} = \{\mathbf{X}_i\}_{i=n+1}^{n+m}$ be the unlabeled training dataset. For all $i \in [n+m]$, each $\mathbf{X}_i$ is i.i.d. from $P_{\mathbf{X}} \in \mathcal{P}(\mathcal{X})$. Based on the labeled dataset S_{l}, a pseudo-labeling method assigns a pseudo-label $\hat{Y}_i$ to each $\mathbf{X}_i \in S_{\mathrm{u}}$ and $\hat{Y}_i$ is drawn conditionally i.i.d. from $P_{\hat{Y}|\mathbf{X},S_{\mathrm{l}}}$. Let the pseudo-labeled data point be $\hat{\mathbf{Z}}_i = (\mathbf{X}_i, \hat{Y}_i)$ and the pseudo-labeled dataset be $\hat{S}_{\mathrm{u}} = \{\hat{\mathbf{Z}}_i\}_{i=n+1}^{n+m}$. For any labeled and pseudo-label datasets, we use $P_{S_{\mathrm{l}}}$, $P_{\hat{S}_{\mathrm{u}}}$ and $P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}$ to denote the joint distributions of the data samples in S_{l}, $\hat{S}_{\mathrm{u}}$, and $S_{\mathrm{l}} \cup \hat{S}_{\mathrm{u}}$, respectively. Note that $P_{S_{\mathrm{l}}} = P_{\mathbf{Z}}^{\otimes n}$.

Let $\mathbf{w} \in \mathcal{W}$ denote the output hypothesis. In semi-supervised learning, one considers the empirical risk based on both the labeled and unlabeled data. In this paper, by fixing a mixing weight $\eta \in \mathbb{R}_+$, for any loss function $l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$, the *total empirical risk* is the η-weighted sum of the empirical risks of the labeled and pseudo-labeled data (McLachlan, 2005; Chapelle et al., 2006)

$$L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) := L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}) + \eta L_{\mathrm{E}}(\mathbf{w}, \hat{S}_{\mathrm{u}}), \quad (1)$$

where $L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}) := \frac{1}{n}\sum_{i=1}^n l(\mathbf{w}, \mathbf{Z}_i)$ and $L_{\mathrm{E}}(\mathbf{w}, \hat{S}_{\mathrm{u}}) := \frac{1}{m}\sum_{i=n+1}^{n+m} l(\mathbf{w}, \hat{\mathbf{Z}}_i)$. Note that, by normalizing the weight η, minimizing the empirical risk $L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})$ is equivalent to minimizing

$$\bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \frac{1}{1+\eta} L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}) + \frac{\eta}{1+\eta} L_{\mathrm{E}}(\mathbf{w}, \hat{S}_{\mathrm{u}}). \quad (2)$$

The *population risk* under the true data distribution is

$$L_{\mathrm{P}}(\mathbf{w}, P_{S_{\mathrm{l}}}) := \mathbb{E}_{S_{\mathrm{l}} \sim P_{S_{\mathrm{l}}}}\big[L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}})\big]. \quad (3)$$

Under the i.i.d. assumption, such a definition reduces to $\mathbb{E}_{S_{\mathrm{l}} \sim P_{S_{\mathrm{l}}}}\big[L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}})\big] = \mathbb{E}_{\mathbf{Z} \sim P_{\mathbf{Z}}}\big[l(\mathbf{w}, \mathbf{Z})\big]$.

Any SSL algorithm can be characterized by a conditional distribution $P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}$, which is a stochastic map from the labeled dataset S_{l}, the pseudo-labeled dataset $\hat{S}_{\mathrm{u}}$ to the output hypothesis $\mathbf{W}$. For any training datasets $S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}$ and any SSL algorithm $P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}$, the *expected gen-error* is defined as the expected gap between the population risk and the empirical risk of the *labeled data* S_{l}, i.e.,

$$\begin{aligned}&\overline{\mathrm{gen}}(P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})\\ &\quad := \mathbb{E}_{\mathbf{W},S_{\mathrm{l}}}[L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}}) - L_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}})],\end{aligned} \quad (4)$$

which measures the extent to which the algorithm overfits to the labeled training data.

In particular, we consider the Gibbs algorithm (also known as the Gibbs posterior (Catoni, 2007)) to model a pseudo-labeling-based SSL algorithm. Given any $(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})$, *the* $(\alpha, \pi(\mathbf{w}), \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$*-Gibbs algorithm* (Gibbs, 1902; Jaynes, 1957) is

$$P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}(\mathbf{w}|S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \frac{\pi(\mathbf{w}) \exp\big(-\alpha \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})\big)}{\Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})},$$

where $\alpha \geq 0$ is the "inverse temperature", $\Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \int \pi(\mathbf{w}) \exp(-\alpha \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))\, \mathrm{d}\mathbf{w}$ is the partition function and $\pi(\mathbf{w})$ is the prior of $\mathbf{w}$. We provide more motivations for the Gibbs algorithm model in Appendix B.

Our goal—relying on the characterization of $P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}$—is to precisely quantify the gen-error in (4) as a function of various information-theoretic quantities.

If P is absolutely continuous with respect to Q and vice versa, let the *symmetrized KL-divergence* (also known as *Jeffrey's divergence* (Jeffreys, 1946)) be defined as $D_{\mathrm{SKL}}(P\|Q) := D(P\|Q) + D(Q\|P)$, where D is the Kullback–Leibler (KL) divergence (Cover, 1999).

For random variables X and Y with joint distribution $P_{X,Y}$, the mutual information is $I(X;Y) = D(P_{X,Y}\|P_X \otimes P_Y)$ and the *Lautum information* (Palomar and Verdú, 2008) is $L(X;Y) = D(P_X \otimes P_Y\|P_{X,Y})$. Similarly, the *symmetrized KL information* between X and Y (Aminian et al., 2015) is defined as $I_{\mathrm{SKL}}(X;Y) := D_{\mathrm{SKL}}(P_{X,Y}\|P_X \otimes P_Y) = I(X;Y) + L(X;Y)$. We define the *conditional symmetrized KL information* as $I_{\mathrm{SKL}}(X;Y|Z) := I(X;Y|Z) + L(X;Y|Z)$.

2.2 Main Results

One of our main results offers an exact closed-form information-theoretic expression for the gen-error of the $(\alpha, \pi(\mathbf{w}), \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs algorithm in terms of the symmetrized KL information defined above.

Theorem 1. *Under the $(\alpha, \pi(\mathbf{w}), \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs algorithm, the expected gen-error is*

$$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{1+\eta}{\alpha}\big(\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] + I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})\big), \quad (5)$$

where $\Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) := \mathbb{E}_{\mathbf{W}\sim\pi}[\exp(-\alpha \bar{L}_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))]$ and $\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\cdot] := \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\cdot] - \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}[\cdot]$.

The proof of Theorem 1 is provided in Appendix C, where we also show that by letting $\eta \to 0$, the result reduces to that for supervised learning (SL). In addition, we can even extend Theorem 1 to SSL based on other methods, e.g., entropy minimization (Amini and Gallinari, 2002; Grandvalet et al., 2005). This corollary is provided in Section D.

Theorem 1 can also be applied to derive novel bounds on the expected gen-error of the Gibbs algorithm as follows.

Proposition 1. *Assume that the loss function $l(\mathbf{w}, \mathbf{Z})$ is bounded in $[a, b] \subset \mathbb{R}_+$. Then, the expected gen-error of the $(\alpha, \pi(\mathbf{w}), \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs algorithm satisfies*

$$\left|\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) - \mathit{SKL}\right| \leq c(\eta, a, b)\sqrt{I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}, \quad (6)$$

where $c(\eta, a, b) := \frac{1}{\sqrt{2}}(1+\eta)(b-a)$ and $\mathit{SKL} := \frac{1+\eta}{\alpha}(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}))$.

The proof and more discussions are provided in Appendix E. In fact, we can further explore how the gen-error depends on the information that the output hypothesis contains about the labeled and unlabeled datasets by writing (see Appendix F)

$$\begin{aligned} &I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \\ &= I(\mathbf{W}; S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}) + D(P_{\mathbf{W}|\hat{S}_{\mathrm{u}}}\|P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}|P_{S_{\mathrm{l}}}P_{\hat{S}_{\mathrm{u}}}). \end{aligned} \quad (7)$$

From Theorem 1 and (7), we observe that for any (η, α), given $\hat{S}_{\mathrm{u}}$, as the dependency (or the information shared) between $\mathbf{W}$ and S_{l} decreases, the expected gen-error decreases, and the algorithm is less likely to overfit to the training data. This intuition dovetails with the results for supervised learning by Xu and Raginsky (2017), Russo and Zou (2019) and Aminian et al. (2021a). However, the difference here is that the quantities are *conditioned on* the pseudo-labeled data $\hat{S}_{\mathrm{u}}$, which reflects the impact of SSL.

Together with Proposition 1, we can observe that the gen-error is also dependent on the *information shared between the input labeled and pseudo-labeled data*. If the mutual information $I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})$ decreases, the expected gen-error is likely to decrease as well. This implies that pseudo-labels highly dependent on the labeled dataset may not be beneficial in terms of the generalization performance of an algorithm. In fact, in our subsequent example in Section 3, we exactly quantify the shared information using the cross-covariance between the labeled and pseudo-labeled data. This result sheds light on the future design of pseudo-labeling methods.

2.2.1 Special Cases of Theorem 1

It is instructive to elaborate how Theorem 1 specializes in some well-known learning settings such as transfer learning and SSL not reusing labeled data.

- **Case 1 ($\hat{S}_{\mathrm{u}}$ independent of S_{l})**: If the pseudo-labels $\{\hat{Y}_i\}_{i=n+1}^{n+m}$ are not generated based on the labeled dataset S_{l}, e.g. randomly assigned or generated from another domain (similar to transfer learning), the pseudo-labeled dataset $\hat{S}_{\mathrm{u}}$ is independent of S_{l}. According to the basic properties of mutual and Lautum information (Palomar and Verdú, 2008) and (7), we have

  $$I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) = I_{\mathrm{SKL}}(\mathbf{W}; S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}),$$

  and $\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] = 0$. That is,

  $$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{(1+\lambda) I_{\mathrm{SKL}}(\mathbf{W}; S_{\mathrm{l}}|\hat{S}_{\mathrm{u}})}{\alpha}, \quad (8)$$

  which corresponds to the result of transfer learning in Bu et al. (2022, Theorem 1).

- **Case 2** ($S_\mathrm{l} - \hat{S}_\mathrm{u} - \mathbf{W}$): When this Markov chain holds, meaning that the output hypothesis $\mathbf{W}$ is independent of S_l conditioned on $\hat{S}_\mathrm{u}$, we have

$$I(\mathbf{W}; S_\mathrm{l}|\hat{S}_\mathrm{u}) = 0,\ P_{\mathbf{W}|\hat{S}_\mathrm{u}} = P_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}$$
$$\text{and } D(P_{\mathbf{W}|\hat{S}_\mathrm{u}} \| P_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}} | P_{S_\mathrm{l}} P_{\hat{S}_\mathrm{u}}) = 0.$$

Thus, the expected gen-error $\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = \frac{1+\eta}{\alpha}\mathbb{E}_{\Delta_{S_\mathrm{l},\hat{S}_\mathrm{u}}}[\log \Lambda_{\alpha,\eta}(S_\mathrm{l}, \hat{S}_\mathrm{u})]$. However, this is a degenerate case. For example, it only occurs when $\eta \to \infty$, i.e., the labeled dataset S_l might be used for generating pseudo-labels for S_u but not used in the Gibbs algorithm to learn the output hypothesis $\mathbf{W}$. In this case, $\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = 0$.

2.2.2 SSL vs. SL with $n+m$ labeled data

It is also instructive to elaborate on how SSL compares to SL with $n+m$ labeled data based on Theorem 1. In particular, assume that that the labeled training dataset $S_\mathrm{l}^{(n+m)} = \{\mathbf{Z}'_i\}_{i=1}^{n+m}$ contains $n+m$ samples drawn i.i.d. from $P_\mathbf{Z}$. Then for any output hypothesis $\mathbf{w}_\mathrm{SL}^{(n+m)} \in \mathcal{W}$, the population risk is given by $L_\mathrm{P}(\mathbf{w}_\mathrm{SL}^{(n+m)}, P_{S_\mathrm{l}^{(n+m)}}) = \mathbb{E}_{\mathbf{Z}'\sim P_\mathbf{Z}}[l(\mathbf{w}_\mathrm{SL}^{(n+m)}, \mathbf{Z}')]$ (cf. (3)), and the empirical risk of the labeled data samples is given by

$$L_\mathrm{E}(\mathbf{w}_\mathrm{SL}^{(n+m)}, S_\mathrm{l}^{(n+m)}) = \frac{1}{n+m}\sum_{i=1}^{n+m} l(\mathbf{w}_\mathrm{SL}^{(n+m)}, \mathbf{Z}'_i).$$

From Aminian et al. (2021a, Theorem 1), under the $(\alpha, \pi(\mathbf{w}_\mathrm{SL}^{(n+m)}), L_\mathrm{E}(\mathbf{w}_\mathrm{SL}^{(n+m)}, S_\mathrm{l}^{(n+m)}))$-Gibbs algorithm, the expected gen-error is

$$\begin{aligned}&\overline{\mathrm{gen}}_\mathrm{SL}^{(n+m)}\\ &:= \mathbb{E}[L_\mathrm{P}(\mathbf{W}_\mathrm{SL}^{(n+m)}, P_{S_\mathrm{l}^{(n+m)}}) - L_\mathrm{E}(\mathbf{W}_\mathrm{SL}^{(n+m)}, S_\mathrm{l}^{(n+m)})]\\ &= \frac{I_\mathrm{SKL}(\mathbf{W}_\mathrm{SL}^{(n+m)}; S_\mathrm{l}^{(n+m)})}{\alpha}. \end{aligned} \tag{9}$$

In the SSL setup, in comparison with (9), under the $(\alpha, \pi(\mathbf{w}), \bar{L}_\mathrm{E}(\mathbf{w}, S_\mathrm{l}, \hat{S}_\mathrm{u}))$-Gibbs algorithm, we let the mixing weight η of the empirical risk in (2) be the *ratio* of the number of unlabeled to labeled data, i.e., $\eta = \lambda := m/n$. We similarly define another expected gen-error $\overline{\mathrm{gen}}_\mathrm{all}$ which is also evaluated over $n+m$ data points as follows:

$$\begin{aligned}&\overline{\mathrm{gen}}_\mathrm{all}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}})\\ &:= \mathbb{E}[L_\mathrm{P}(\mathbf{W}, P_{S_\mathrm{l}}) - L_\mathrm{E}(\mathbf{W}, S_\mathrm{l}, \hat{S}_\mathrm{u})].\end{aligned}$$

Consider the situation where the pseudo-labeling method is perfect such that $P_{\hat{Y}_i|\mathbf{X}_i} = P_{Y|\mathbf{X}}$ for all $i \in [n+1 : n+m]$. Then any $\hat{\mathbf{Z}}_i \in \hat{S}_\mathrm{u}$ has the same distribution as that of any $\mathbf{Z} \in S_\mathrm{l}$. By applying the same technique in the proof of Theorem 1, we obtain (details provided in Appendix G)

$$\overline{\mathrm{gen}}_\mathrm{all}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = \frac{I_\mathrm{SKL}(\mathbf{W}; S_\mathrm{l}, \hat{S}_\mathrm{u})}{\alpha}. \tag{10}$$

However, only when $P_{\hat{Y}_i|\mathbf{X}_i,S_\mathrm{l}} = P_{\hat{Y}_i|\mathbf{X}_i} = P_{Y|\mathbf{X}}$ for any S_l, $I_\mathrm{SKL}(\mathbf{W}; S_\mathrm{l}, \hat{S}_\mathrm{u}) = I_\mathrm{SKL}(\mathbf{W}_\mathrm{SL}^{(n+m)}; S_\mathrm{l}^{(n+m)})$ and $\overline{\mathrm{gen}}_\mathrm{all} = \overline{\mathrm{gen}}_\mathrm{SL}^{(n+m)}$; otherwise, the gen-errors of SL in (9) and of SSL in (10) may not be equal. This is because the pseudo-label $\hat{Y}_i$'s are assumed to be generated based on the labeled dataset S_l and this dependence affects the gen-error. As we have discussed following Proposition 1, if the information shared between S_l and $\hat{S}_\mathrm{u}$ is large, the learning algorithm tends to overfit to the labeled training data.

If the $\hat{Y}_i$'s are independent of S_l, then $P_{\hat{Y}_i|\mathbf{X}_i,S_\mathrm{l}} = P_{\hat{Y}_i|\mathbf{X}_i}$ for any S_l. Consider the situation where the pseudo-labeling method is *close* to perfect, i.e., $P_{\hat{\mathbf{Z}}} = P_\mathbf{Z} + \epsilon\Delta$, where $\epsilon > 0$ is small and $\int_\mathcal{Z} \Delta(\mathbf{z})\,\mathrm{d}\mathbf{z} = 0$ such that $P_{\hat{\mathbf{Z}}} \geq 0$. We have

$$\overline{\mathrm{gen}}_\mathrm{all}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = \frac{I_\mathrm{SKL}(\mathbf{W}_\mathrm{SL}^{(n+m)}; S_\mathrm{l}^{(n+m)})}{\alpha} + \tilde{\epsilon},$$

where $\tilde{\epsilon}$ is proportional to ϵ and $\tilde{\epsilon} \to 0$ as $\epsilon \to 0$ (details are provided in Appendix G.2). Therefore, in this special case, we see that the gap between $\overline{\mathrm{gen}}_\mathrm{all}$ of SSL and $\overline{\mathrm{gen}}_\mathrm{SL}^{(n+m)}$ of SL can be directly quantified by the gap between the labeled and pseudo-labeled distributions.

2.3 Distribution-Free Upper and Lower Bounds

We can also build upon Theorem 1 and Xu and Raginsky (2017, Theorem 1) to provide distribution-free upper and lower bounds of the gen-error in (5). We next state an informal version of such bounds (the formal version and the proof are provided in Appendix H). Throughout this section, we let the mixing weight $\eta = \lambda = m/n$.

Proposition 2 (Informal Version)**.** *Assume $I(\hat{S}_\mathrm{u}; S_\mathrm{l}) = \gamma_{\alpha,\lambda} I(\mathbf{W}, \hat{S}_\mathrm{u}; S_\mathrm{l})$, for some $\gamma_{\alpha,\lambda} \in [0,1]$ that depends on λ and α. Assume that $l(\mathbf{w}, \mathbf{Z})$ is bounded in $[a,b] \subset \mathbb{R}_+$. Let $\gamma'_{\alpha,\lambda,a,b} = \frac{\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2(1-\gamma_{\alpha,\lambda})}$. Then the expected gen-error can be bounded as follows*

$$\begin{aligned}-\gamma'_{\alpha,\lambda,a,b}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{\gamma_{\alpha,\lambda}}}{1+\lambda}\right) &\leq \overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}})\\ &\leq \gamma'_{\alpha,\lambda,a,b}\left(\frac{1}{\sqrt{n}} + \frac{1}{(n+m)\sqrt{\gamma_{\alpha,\lambda}}}\right).\end{aligned} \tag{11}$$

Recall that for SL with n labeled data, the expected gen-error is of order $O(\frac{1}{n})$ (Aminian et al., 2021a, Theorem 2). This can be naturally extended to SL with $n+m$ labeled data, leading to the expected gen-error behaving as $O(\frac{1}{n+m})$. Compared to the SL case, we see that using the pseudo-labeled data may degrade the convergence rate of the expected gen-error and the extent to which it degrades depends on $\gamma_{\alpha,\lambda}$, which captures the impact of the amount of information shared among the output hypothesis, the input labeled and unlabeled data as well as the ratio λ. If $\gamma_{\alpha,\lambda}$ is small, it implies $I(\hat{S}_\mathrm{u}; S_\mathrm{l})$ is small and $I(\mathbf{W}; S_\mathrm{l}|\hat{S}_\mathrm{u})$ is large, which leads to a tight upper bound.

If we fix $0 < \lambda < \infty$ and assume $\hat{S}_{\mathrm{u}}$ is independent of S_{l}, then $\gamma_{\alpha,\lambda} = 0$ and

$$0 \le \overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) \le \frac{\alpha(b-a)^2}{2(n+m)},$$

which coincides with the result of transfer learning in Bu et al. (2022, Remark 3) and corresponds to the analysis of (8). Although it appears that using independent pseudo-labeled data $\hat{S}_{\mathrm{u}}$ does not degrade the convergence rate of the expected gen-error, it may affect the excess risk or the test accuracy, which taken in unison, represents the performance of a learning algorithm. The definition and further analysis of the excess risk is provided in Section 4.1.1.

When $\lambda \to 0$, the gen-error converges to that of SL with n labeled data and the distribution-free upper bound is of order $O(\frac{1}{n})$, given in Aminian et al. (2021a, Theorem 2).

3 AN APPLICATION TO MEAN ESTIMATION

To deepen our understanding of the expected gen-error in Theorem 1, we study a mean estimation example and analyze how $\lambda = m/n$ affects the gen-error.

3.1 Problem Setup

For any $(\mathbf{X}_i, Y_i) \in S_{\mathrm{l}}$ and $i \in [n]$, we assume that $Y_i \in \{-1, +1\}$, $\mathbb{E}[\mathbf{X}_i|Y_i = 1] = \boldsymbol{\mu}$, $\mathbb{E}[\mathbf{X}_i|Y_i = -1] = -\boldsymbol{\mu}$, $\|\boldsymbol{\mu}\|_2 = 1$ and $\mathrm{Cov}[Y_i\mathbf{X}_i] = \sigma^2\mathbf{I}_d$. Any $\mathbf{X} \in S_{\mathrm{u}}$ is drawn i.i.d. from the same distribution as $\mathbf{X}_i$. Consider the problem of learning $\boldsymbol{\mu}$ using S_{l} and $\hat{S}_{\mathrm{u}}$. We adopt the mean-squared loss $l(\mathbf{w}, \mathbf{z}) = \|\mathbf{w} - y\mathbf{x}\|_2^2$ and assume the prior $\pi(\mathbf{w})$ is uniform on $\mathcal{W}$. Based on $\mathbf{W}_0$ learned from the labeled dataset S_{l} (e.g., $\mathbf{W}_0 = \frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i$), we assign a pseudo-label to each $X_i \in S_{\mathrm{u}}$ as $\hat{Y}_i$ (e.g. $\hat{Y}_i = \mathrm{sgn}(\mathbf{W}_0^\top\mathbf{X}_i)$) and let $\boldsymbol{\mu}' = \mathbb{E}[\hat{Y}_i\mathbf{X}_i]$. Let us construct $S'_{\mathrm{l}} = \{Y_i\mathbf{X}_i\}_{i=1}^n$ and $\hat{S}'_{\mathrm{u}} = \{\hat{Y}_i\mathbf{X}_i\}_{i=n+1}^{n+m}$. The *empirical risk* is given by (cf. (2) by replacing $S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}$ with $S'_{\mathrm{l}}, \hat{S}'_{\mathrm{u}}$)

$$\bar{L}_{\mathrm{E}}(\mathbf{w}, S'_{\mathrm{l}}, \hat{S}'_{\mathrm{u}}) = \frac{1}{(1+\lambda)n}\sum_{i=1}^n \|Y_i\mathbf{X}_i - \mathbf{w}\|_2^2 + \frac{\lambda}{(1+\lambda)m}\sum_{j=n+1}^{n+m} \|\hat{Y}_j\mathbf{X}_j - \mathbf{w}\|_2^2.$$

The expected gen-error is equal to the right-hand side of (5) by replacing $S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}$ with $S'_{\mathrm{l}}, \hat{S}'_{\mathrm{u}}$.

The $(\alpha, \pi(\mathbf{W}), \bar{L}_{\mathrm{E}}(\mathbf{W}, S'_{\mathrm{l}}, \hat{S}'_{\mathrm{u}}))$-Gibbs algorithm is given by the following Gibbs posterior distribution

$$P^{\alpha}_{\mathbf{W}|S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}}}(\mathbf{W}|S'_{\mathrm{l}}, \hat{S}'_{\mathrm{u}}) = \mathcal{N}(\boldsymbol{\mu}_{n+m}, \sigma^2_{\mathrm{l,u}}\mathbf{I}_d),$$

where $\sigma^2_{\mathrm{l,u}} = \frac{1}{2\alpha}$ and

$$\boldsymbol{\mu}_{n,m} = \frac{1}{(1+\lambda)n}\sum_{i=1}^n Y_i\mathbf{X}_i + \frac{\lambda}{(1+\lambda)m}\sum_{j=n+1}^{n+m}\hat{Y}_j\mathbf{X}_j.$$

Details are provided in Appendix I. With this posterior Gaussian distribution $P^{\alpha}_{\mathbf{W}|S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}}}$, the expected gen-error in Theorem 1 can be exactly computed as follows

$$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big], \quad (12)$$

where $i \in [n]$ and $j \in [n+1 : n+m]$ run over the labeled and unlabeled datasets respectively. From the definition of the expected gen-error in (4), we obtain the same result, which corroborates the characterization of the expected gen-error in Theorem 1. See Appendix I for details.

Note that the second term in (12) is the trace of the cross-covariance between the labeled and pseudo-labeled data sample, i.e., for any $i \in [n]$ and $j \in [n+1 : n+m]$,

$$\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big] = \mathrm{tr}(\mathrm{Cov}[Y_i\mathbf{X}_i, \hat{Y}_j\mathbf{X}_j]).$$

This result shows that for any fixed (n, m), the expected gen-error decreases when the trace of the cross-covariance decreases, which corroborates our analyses in Section 2.2.

3.2 Effect of the Pseudo-labeling and the Ratio of Unlabeled to Labeled Samples λ

To futher study the effect of the pseudo-labeling method and the ratio $\lambda = m/n$, in this mean estimation example, we consider the class-conditional feature distribution of $\mathbf{X}_i|Y_i \sim \mathcal{N}(Y_i\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$ for any $i \in [n+m]$. Let $S'^{(n+m)}_{\mathrm{l}}$ be a dataset with $n+m$ independent copies of $Y\mathbf{X} \in S'_{\mathrm{l}}$. Similarly to (12), for the supervised $(\alpha, \pi(\mathbf{w}^{(n)}_{\mathrm{SL}}), L_{\mathrm{E}}(\mathbf{w}^{(n)}_{\mathrm{SL}}, S'_{\mathrm{l}}))$-Gibbs algorithm and the supervised $(\alpha, \pi(\mathbf{w}^{(n+m)}_{\mathrm{SL}}), L_{\mathrm{E}}(\mathbf{w}^{(n+m)}_{\mathrm{SL}}, S'^{(n+m)}_{\mathrm{l}}))$-Gibbs algorithm, the expected gen-errors are respectively given by (see details in Appendix I.1)

$$\overline{\mathrm{gen}}^{(n)}_{\mathrm{SL}} = \frac{2\sigma^2 d}{n} \quad \text{and} \quad \overline{\mathrm{gen}}^{(n+m)}_{\mathrm{SL}} = \frac{2\sigma^2 d}{n+m}.$$

Let $\overline{\mathrm{gen}}_{\mathrm{SSL}}$ denote $\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$ for brevity. By comparing $\overline{\mathrm{gen}}^{(n)}_{\mathrm{SL}}$ and $\overline{\mathrm{gen}}^{(n+m)}_{\mathrm{SL}}$, we observe that:

1. If $\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big] < 0$, then $\overline{\mathrm{gen}}_{\mathrm{SSL}} < \overline{\mathrm{gen}}^{(n+m)}_{\mathrm{SL}}$, which means that SSL with such pseudo-labeling method even has better generalization performance than SL with $n+m$ labeled data. This is the most desirable case in terms of the generalization error;
2. If $\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big] > \frac{\sigma^2 d}{n}$, then $\overline{\mathrm{gen}}_{\mathrm{SSL}} > \overline{\mathrm{gen}}^{(n)}_{\mathrm{SL}}$. This implies that if the cross-covariance between

the labeled and pseudo-labeled data is larger than a certain threshold, the pseudo-labeling method does not help to improve the generalization performance;

3. If $0 \leq \mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big] \leq \frac{\sigma^2 d}{n}$, then $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n+m)} \leq \overline{\mathrm{gen}}_{\mathrm{SSL}} \leq \overline{\mathrm{gen}}_{\mathrm{SL}}^{(n)}$. This implies that if the cross-covariance is sufficiently small, pseudo-labeling improves the generalization error.

The value of $\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big]$ also depends on the ratio λ. Next, to study the effect of λ, let the initial hypothesis learned from the labeled data S_l be $\mathbf{W}_0 = \frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$ and the pseudo-label for any $\mathbf{X}_i \in S_\mathrm{u}$ be $\hat{Y}_i = \mathrm{sgn}(\mathbf{W}_0^\top\mathbf{X}_i)$.

Inspired by the derivation techniques in He et al. (2022), we can rewrite the expected gen-error in (12) as

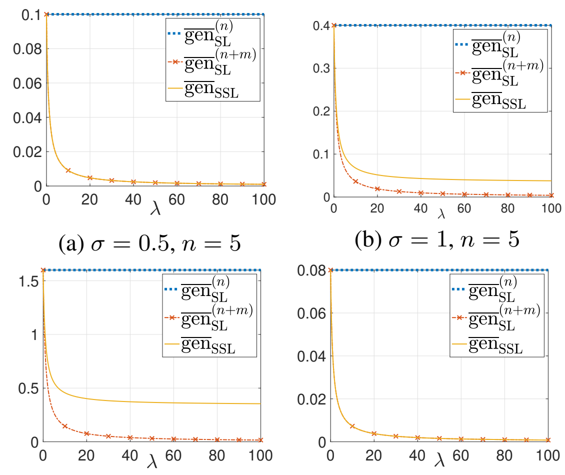
$$\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}E_n, \qquad (13)$$

where $E_n = \mathbb{E}[\tilde{g}J_\sigma(\gamma_n') + \|\boldsymbol{\mu}^\perp\|_2 K_\sigma(\gamma_n')]$, $\tilde{g} \sim \mathcal{N}(0,1)$, $\boldsymbol{\mu}^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top)$, J_σ and K_σ are functions with domain $[-1,1]$, and $\gamma_n' \in [-1,1]$ is a sequence of correlation coefficients. Details are presented in Appendix I.2, where we also prove that $E_n = O(d)$ and $E_n \geq 0$, which means that we always have $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n+m)} \leq \overline{\mathrm{gen}}_{\mathrm{SSL}}$ here. Furthermore, we observe that E_n does not depend on m and thus, the right-hand side of (13) converges to $2E_n$ when n is fixed and $\lambda \to \infty$.

In Figure 1, we numerically plot $\overline{\mathrm{gen}}_{\mathrm{SSL}}$ by varying λ and compare it with $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n)}$ and $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n+m)}$ for different values of noise level σ, and number of labeled data samples n. Under different choices of σ, we observe that $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n+m)} \leq \overline{\mathrm{gen}}_{\mathrm{SSL}} \leq \overline{\mathrm{gen}}_{\mathrm{SL}}^{(n)}$. Moreover, the gen-error of SSL monotonically decreases as λ increases, showing obvious improvement compared to $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n)}$. However, there exists a gap $\frac{2\lambda E_n}{1+\lambda}$ between $\overline{\mathrm{gen}}_{\mathrm{SSL}}$ and $\overline{\mathrm{gen}}_{\mathrm{SL}}^{(n+m)}$. For n large enough (e.g., $n = 100$) or noise level small enough (e.g., $\sigma = 0.5$), this gap is almost negligible, which means pseudo-labeling yields comparable generalization performance as the SL when *all* the samples are labeled.

3.3 Choosing a Pseudo-labeling Method from a Given Family of Methods

We have shown that the generalization error decreases as the information shared between the input labeled and pseudo-labeled data (represented by the mutual information $I(\hat{S}_\mathrm{u}; S_\mathrm{l})$ or the cross-covariance term $\mathrm{tr}(\mathrm{Cov}[Y_i\mathbf{X}_i, \hat{Y}_j\mathbf{X}_j])$ in the mean estimation example) decreases. This result serves as a guideline for choosing an appropriate pseudo-labeling method. For example, if we are given a set of pseudo-labeling functions $\{f_i\}_{i=1}^C$ with $f_i : \mathcal{X} \to \mathcal{Y}$, we can choose the best one f_{i^*} that yields the minimum mutual information $I(\hat{S}_\mathrm{u}; S_\mathrm{l})$. To make our ideas more concrete, for the mean estimation example that



(a) $\sigma = 0.5$, $n = 5$ (b) $\sigma = 1$, $n = 5$

(c) $\sigma = 2$, $n = 5$ (d) $\sigma = 2$, $n = 100$

Figure 1: Gen-error of mean estimation vs. $\lambda = m/n$.

we have been discussing thus far, assume that the pseudo-labeling functions $\hat{Y} = \mathrm{sgn}(\mathbf{W}_0^\top\mathbf{X})\mathbf{1}_{\{|\mathbf{W}_0^\top\mathbf{X}| \geq T\}}$ (where $\mathbf{X} \in S_\mathrm{u}$) are indexed by various "confidence" thresholds $T \in \mathbb{R}_+$ (T plays the role of the index i in $\{f_i\}_{i=1}^C$). For instance, let us consider the case where $n = 5$, $\sigma = 1$ (cf. Figure 1b). As shown in Figure 2, one can choose the threshold T that minimizes $\mathrm{tr}(\mathrm{Cov}[Y_i\mathbf{X}_i, \hat{Y}_j\mathbf{X}_j])$ for $(\mathbf{X}_i, Y_i) \in S_\mathrm{l}$ and $\mathbf{X}_j \in S_\mathrm{u}$. From this figure, we observe that $T \geq 7$ approximately minimizes $\mathrm{tr}(\mathrm{Cov}[Y_i\mathbf{X}_i, \hat{Y}_j\mathbf{X}_j])$ and hence, the generalization error. We have thus exhibited a concrete way to choose one pseudo-labeling method from a given family of methods via our characterization of generalization error.

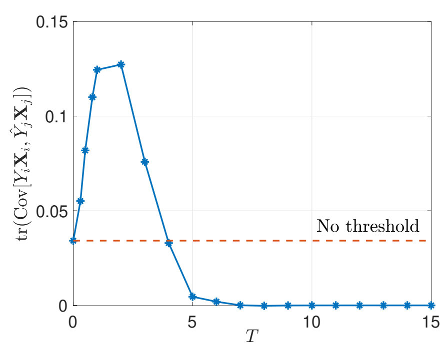


Figure 2: The cross-covariance term between labelled and pseudo-labelled data for different confidence threshold T.

4 PARTICULARIZING THE GIBBS ALGORITHM TO EMPIRICAL RISK MINIMIZATION

In this section, we consider the asymptotic behavior of the expected gen-error and excess risk for the Gibbs algorithm under SSL as the "inverse temperature" $\alpha \to \infty$. It is known that the Gibbs algorithm converges to empirical risk minimization (ERM) as $\alpha \to \infty$; see Appendix B.

Given any $S_\mathrm{l}, \hat{S}_\mathrm{u}$, assume that there exists a unique minimizer of the empirical risk $\bar{L}_\mathrm{E}(\mathbf{W}, S_\mathrm{l}, \hat{S}_\mathrm{u})$ (also known as the single-well case), i.e.,

$$\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u}) = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \bar{L}_\mathrm{E}(\mathbf{w}, S_\mathrm{l}, \hat{S}_\mathrm{u}).$$

Let the Hessian matrix of the empirical risk $\bar{L}_\mathrm{E}(\mathbf{w}, S_\mathrm{l}, \hat{S}_\mathrm{u})$ at $\mathbf{w} = \mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})$ be defined as

$$H^*(S_\mathrm{l}, \hat{S}_\mathrm{u}) = \nabla^2_\mathbf{w} \bar{L}_\mathrm{E}(\mathbf{w}, S_\mathrm{l}, \hat{S}_\mathrm{u})|_{\mathbf{w}=\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})}.$$

Under the single-well case, we obtain a characterization of the asymptotic expected gen-error in the following theorem.

Theorem 2. *Under the $(\alpha, \pi(\mathbf{w}), \bar{L}_\mathrm{E}(\mathbf{w}, S_\mathrm{l}, \hat{S}_\mathrm{u}))$-Gibbs algorithm, as $\alpha \to \infty$, if the Hessian matrix $H^*(S_\mathrm{l}, \hat{S}_\mathrm{u})$ is not singular and $\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})$ is unique, the expected gen-error converges to*

$$\begin{aligned}
&\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) \\
&= \mathbb{E}_{\hat{S}_\mathrm{u},S_\mathrm{l}}\Big[\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})^\top H^*(S_\mathrm{l}, \hat{S}_\mathrm{u})\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})\Big] \\
&\quad + \mathbb{E}_{\hat{S}_\mathrm{u}}\mathbb{E}_{S_\mathrm{l}}\Big[\Big(\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u}) - 2\mathbb{E}_{S_\mathrm{l}|\hat{S}_\mathrm{u}}[\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})]\Big)^\top \\
&\quad\quad \cdot H^*(S_\mathrm{l}, \hat{S}_\mathrm{u})\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u})\Big] \\
&\quad - \mathbb{E}_{\Delta_{S_\mathrm{l},\hat{S}_\mathrm{u}}}\Big[\bar{L}_\mathrm{E}(\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u}), S_\mathrm{l}, \hat{S}_\mathrm{u})\Big] \\
&\quad - \mathbb{E}_{\Delta_{(\mathbf{W},\hat{S}_\mathrm{u}),S_\mathrm{l}}}\Big[\mathbf{W}^\top H^*(S_\mathrm{l}, \hat{S}_\mathrm{u})\mathbf{W}\Big]\Big),
\end{aligned} \tag{14}$$

where the expectations $\mathbb{E}_{\Delta_{S_\mathrm{l},\hat{S}_\mathrm{u}}}[\cdot] := \mathbb{E}_{S_\mathrm{l},\hat{S}_\mathrm{u}}[\cdot] - \mathbb{E}_{S_\mathrm{l}}\mathbb{E}_{\hat{S}_\mathrm{u}}[\cdot]$ and $\mathbb{E}_{\Delta_{(\mathbf{W},\hat{S}_\mathrm{u}),S_\mathrm{l}}}[\cdot] := \mathbb{E}_{\mathbf{W},\hat{S}_\mathrm{u},S_\mathrm{l}}[\cdot] - \mathbb{E}_{\mathbf{W},\hat{S}_\mathrm{u}}\mathbb{E}_{S_\mathrm{l}}[\cdot]$.

The proof of Theorem 2 is provided in Appendix J. Theorem 2 shows that the gen-error of the Gibbs algorithm under SSL when $\alpha \to \infty$ depends strongly on the second derivatives of the empirical risk (a.k.a. loss landscape) and the relationship between S_l and $\hat{S}_\mathrm{u}$. The loss landscape is an important tool to understand the dynamics of the learning process in deep learning. In the extreme case where S_l and $\hat{S}_\mathrm{u}$ are independent, $\mathbb{E}_{\Delta_{S_\mathrm{l},\hat{S}_\mathrm{u}}}\big[\bar{L}_\mathrm{E}(\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u}), S_\mathrm{l}, \hat{S}_\mathrm{u})\big] = 0$ and a simplified expression for the asymptotic gen-error is provided in Appendix K.

4.1 Semi-Supervised Maximum Likelihood Estimation

In particular, by letting the loss function be the *negative log-likelihood*, this algorithm becomes semi-supervised maximum likelihood estimation (SS-MLE). In this part, we consider the SS-MLE in the asymptotic regime where $n, m \to \infty$. Throughout this section, we let the mixing weight in (2) $\eta = \lambda = m/n$.

We aim to fit the training data with a parametric family $p(\cdot|\mathbf{w})$, where $\mathbf{w} \in \mathcal{W}$. Consider the negative log-loss function $l(\mathbf{w}, \mathbf{z}) = -\log p(\mathbf{z}|\mathbf{w})$. For the single-well case, the unique minimizer is

$$\begin{aligned}
\mathbf{W}^*(S_\mathrm{l}, \hat{S}_\mathrm{u}) = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \bigg(&-\frac{1}{1+\lambda}\cdot\frac{1}{n}\sum_{i=1}^n \log p(\mathbf{Z}_i|\mathbf{w}) \\
&- \frac{\lambda}{1+\lambda}\cdot\frac{1}{m}\sum_{i=n+1}^{n+m}\log p(\hat{\mathbf{Z}}_i|\mathbf{w})\bigg).
\end{aligned} \tag{15}$$

Given any labeled dataset S_l, we use $P_{\hat{\mathbf{Z}}|S_\mathrm{l}} = \mathbb{E}_{\mathbf{W}_0|S_\mathrm{l}}[P_{\hat{\mathbf{Z}}|\mathbf{W}_0}]$ to denote the conditional distribution of the pseudo-labeled data (i.e., $\hat{\mathbf{Z}}_i \overset{\text{i.i.d.}}{\sim} P_{\hat{\mathbf{Z}}|S_\mathrm{l}}$), where $\mathbf{W}_0$ is the initial hypothesis learned only from S_l and used to generate pseudo-labels for S_u. Assume that $\mathbf{W}_0$ is learned from S_l using MLE, i.e.,

$$\mathbf{W}_0 = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} -\frac{1}{n}\sum_{i=1}^n \log p(\mathbf{Z}_i|\mathbf{w}).$$

We have $\mathbf{W}_0 \xrightarrow{\text{p}} \mathbf{w}^*_{P_\mathbf{Z}} = \arg\min_{\mathbf{w}\in\mathcal{W}} D(P_\mathbf{Z}\|p(\cdot|\mathbf{w}))$ as $n \to \infty$, where $\mathbf{w}^*_{P_\mathbf{Z}}$ depends only on the true distribution of the labeled data $P_\mathbf{Z}$. Then we have $P_{\hat{\mathbf{Z}}|\mathbf{W}_0} \xrightarrow{\text{p}} P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_\mathbf{Z}}}$, $P_{\hat{\mathbf{Z}}|S_\mathrm{l}} \xrightarrow{\text{p}} P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_\mathbf{Z}}}$ and $P_{\hat{\mathbf{Z}}} = \mathbb{E}_{\mathbf{W}_0}[P_{\hat{\mathbf{Z}}|\mathbf{W}_0}] \xrightarrow{\text{p}} P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_\mathbf{Z}}}$, which means $\{\hat{\mathbf{Z}}_i\}_{i=n+1}^{n+m}$ become independent of one another and of S_l. We analogously define the minimizer

$$\begin{aligned}
\mathbf{w}^*_\lambda = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \bigg(&\frac{n}{n+m} D\big(P_\mathbf{Z}\|p(\cdot|\mathbf{w})\big) \\
&+ \frac{m}{n+m} D\big(P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_\mathbf{Z}}}\|p(\cdot|\mathbf{w})\big)\bigg)
\end{aligned} \tag{16}$$

and the *landscapes* of the loss function

$$\begin{aligned}
J_\mathrm{l}(\mathbf{w}) &:= \mathbb{E}_{\mathbf{Z}\sim P_\mathbf{Z}}[-\nabla^2_\mathbf{w}\log p(\mathbf{Z}|\mathbf{w})], \\
J_\mathrm{u}(\mathbf{w}) &:= \mathbb{E}_{\hat{\mathbf{Z}}\sim P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_\mathbf{Z}}}}[-\nabla^2_\mathbf{w}\log p(\hat{\mathbf{Z}}|\mathbf{w})], \\
\mathcal{I}_\mathrm{l}(\mathbf{w}) &:= \mathbb{E}_{\mathbf{Z}\sim P_\mathbf{Z}}[\nabla_\mathbf{w}\log p(\mathbf{Z}|\mathbf{w})\nabla_\mathbf{w}\log p(\mathbf{Z}|\mathbf{w})^\top].
\end{aligned}$$

Let $J(\mathbf{w}) = \frac{n}{n+m}J_\mathrm{l}(\mathbf{w}) + \frac{m}{n+m}J_\mathrm{u}(\mathbf{w})$. Given any ratio $\lambda > 0$, as $n, m \to \infty$, by the law of large numbers, the Hessian matrix $H^*(S_\mathrm{l}, \hat{S}_\mathrm{u})$ converges as follows

$$H^*(S_\mathrm{l}, \hat{S}_\mathrm{u}) \xrightarrow{\text{p}} J(\mathbf{w}^*_\lambda),$$

which is independent of $(S_\mathrm{l}, \hat{S}_\mathrm{u})$. Leveraging these asymptotic approximations, from Theorem 2, we obtain the following characterization of the gen-error of SS-MLE.

Corollary 1. *In the asymptotic regime where $n, m \to \infty$, the expected gen-error of SS-MLE is given by*

$$\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}}) = \frac{\mathrm{tr}(J(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_\mathrm{l}(\mathbf{w}^*_\lambda))}{n+m}.$$

The proof of Corollary 1 is provided in Appendix K. For the extreme cases where $\lambda \to 0$, we have $\mathbf{w}^*_\lambda \to$

$\mathbf{w}^*_{P_\mathbf{Z}}$, $J(\mathbf{w}^*_\lambda) \to J_l(\mathbf{w}^*_{P_\mathbf{Z}})$ and $\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u}) \to \frac{1}{n}\operatorname{tr}(J_l(\mathbf{w}^*_{P_\mathbf{Z}})^{-1}\mathcal{I}_l(\mathbf{w}^*_{P_\mathbf{Z}})) = O(\frac{d}{n})$, which means the gen-error degenerates to that of the SL case with n labeled data.

For the other case where $\lambda \to \infty$, from Appendix K, we have $\mathbf{W}_{\mathrm{ML}}(S_l,\hat{S}_u) \to \mathbb{E}_{S_l}[\mathbf{W}_{\mathrm{ML}}(S_l,\hat{S}_u)]$ and $\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u}) \to 0$.

For $\lambda \in (0,\infty)$, we have $\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u}) = O(\frac{d}{n+m})$, which is order-wise the same as the gen-error of SL with $n+m$ labeled data. The intuition is that for large n, the pseudo-labeled samples only depend on the labeled data *distribution* instead of the labeled samples. However, the performance of an algorithm depends not only on the gen-error and but also on the excess risk. Even when the gen-error is small, the bias of the excess risk may be high.

4.1.1 Excess Risk as $\alpha, n, m \to \infty$

In this section, we discuss the excess risk of SS-MLE when $n, m \to \infty$. The *excess risk* is defined as the gap between the expectation and the minimum of the population risk, i.e.,

$$\begin{aligned}\mathcal{E}_r(P_\mathbf{W}) &:= \mathbb{E}_\mathbf{W}[L_P(\mathbf{W}, P_{S_l})] - L_P(\mathbf{w}^*_l, P_{S_l}) \qquad (17)\\ &= \overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u})\\ &\quad + \mathbb{E}_{\mathbf{W},S_l}[L_E(\mathbf{W}, S_l) - L_E(\mathbf{w}^*_l, S_l)]. \qquad (18)\end{aligned}$$

where $\mathbf{w}^*_l = \arg\min_{\mathbf{w}\in\mathcal{W}} L_P(\mathbf{W}, P_{S_l})$ is the optimal hypothesis. The second term in (18) is known as the *estimation error*. We observe that the excess risk depends on both the gen-error and estimation error. When the gen-error is controlled to be sufficiently small, but if the estimation error is large, the excess risk can still be large.

Corollary 2. *In the asymptotic regime where $n, m \to \infty$, the excess risk of SS-MLE is given by*

$$\begin{aligned}\mathcal{E}_r(P_\mathbf{W}) &= \frac{1}{2}\operatorname{tr}((\mathbf{w}^*_\lambda - \mathbf{w}^*_l)(\mathbf{w}^*_\lambda - \mathbf{w}^*_l)^\top J_l(\mathbf{w}^*_l))\\ &+ \frac{\operatorname{tr}(J_l(\mathbf{w}^*_l)J(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_l(\mathbf{w}^*_\lambda)J(\mathbf{w}^*_\lambda)^{-1})}{2(1+\lambda)(n+m)}. \qquad (19)\end{aligned}$$

The proof of Corollary 2 is provided in Appendix L. In (19), the first term represents the bias caused by learning with the mixture of labeled and pseudo-labeled data. When $\lambda \to 0$, the bias converges to 0. As λ increases, the bias increases. The second term represents the variance component of the excess risk, which is of order $O(\frac{d}{m+n})$ for $0 < \lambda < \infty$, the same as that for the gen-error in Corollary 1.

4.2 An Application to Logistic Regression

To re-emphasize, we let the mixing weight in (2) $\eta = \lambda = m/n$. We now apply SS-MLE to logistic regression to study the effect of λ on the gen-error and excess risk. For any hypothesis $\mathbf{w}$, let $p(y|\mathbf{x},\mathbf{w})$ be the conditional likelihood of a label upon seeing a feature sample under $\mathbf{w}$. Assume that the label $Y \in \{-1,+1\}$. For any $\mathbf{z} \in \mathcal{X}\times\{-1,+1\}$, the underlying distribution is $P_\mathbf{Z}(\mathbf{z}) = P_{Y|\mathbf{X}}(y|\mathbf{x})P_\mathbf{X}(\mathbf{x})$ and the logistic regression model uses $p(y|\mathbf{x},\mathbf{w})$ to approximate $P_{Y|\mathbf{X}}(y|\mathbf{x})$. Let $P_{\mathbf{Z}|\mathbf{w}}(\mathbf{z}|\mathbf{w}) = p(y|\mathbf{x},\mathbf{w})P_\mathbf{X}(\mathbf{x})$. To stabilize the solution $\mathbf{w}$, we consider a regularized version of the logistic regression model, where for a fixed $\nu > 0$, the objective function can be expressed as

$$\begin{aligned}l(\mathbf{w},\mathbf{z}) &= -\log p(y|\mathbf{x},\mathbf{w}) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\\ &= \log(1+e^{-y\mathbf{w}^\top\mathbf{x}}) + \frac{\nu}{2}\|\mathbf{w}\|_2^2.\end{aligned}$$

We assume that there exists a unique minimizer of the empirical risk in this example. We also assume that the initial hypothesis $\mathbf{W}_0$ is learned from the labeled dataset S_l:

$$\mathbf{W}_0 = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}}\left(\frac{1}{n}\sum_{i=1}^n \log(1+e^{-Y_i\mathbf{w}^\top\mathbf{X}_i}) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\right).$$

Consider the case when $n, m \to \infty$ and $\lambda > 0$. Then

$$\mathbf{W}_0 \xrightarrow{\mathrm{p}} \mathbf{w}^*_0 = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}}\left(D\left(P_\mathbf{Z}\middle\|\frac{P_\mathbf{X}}{1+e^{-Y\mathbf{w}^\top\mathbf{X}}}\right) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\right).$$

Let the pseudo label for any $\mathbf{X}_i \in S_u$ be defined as $\hat{Y}_i = \operatorname{sgn}(\mathbf{X}_i^\top\mathbf{W}_0)$. The conditional distribution of the pseudo-labeled data sample given $\mathbf{W}_0$ converges as follows

$$P_{\hat{\mathbf{Z}}|\mathbf{W}_0}(\hat{\mathbf{z}}|\mathbf{W}_0) \xrightarrow{\mathrm{p}} P_{\hat{\mathbf{Z}}|\mathbf{w}^*_0}(\hat{\mathbf{z}}|\mathbf{w}^*_0) = P_\mathbf{X}(\mathbf{x})\mathbb{1}\{\hat{y} = \operatorname{sgn}(\mathbf{x}^\top\mathbf{w}^*_0)\}.$$

Let us rewrite the minimizer $\mathbf{w}^*_\lambda$ in (16) as

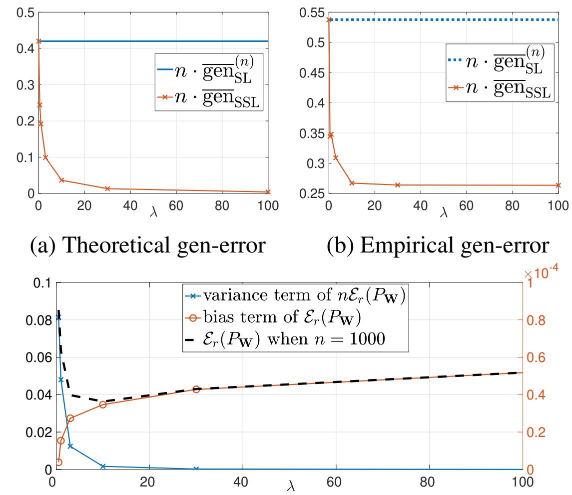
$$\begin{aligned}\mathbf{w}^*_\lambda = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}}\Bigg(&\frac{n}{n+m}D\big(P_\mathbf{Z}\|p(\cdot|\mathbf{w})\big)\\ &+ \frac{m}{n+m}D\big(P_{\hat{\mathbf{Z}}|\mathbf{w}^*_0}\|p(\cdot|\mathbf{w})\big) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\Bigg).\end{aligned}$$

Recall the expected gen-error in Corollary 1, which can also be rewritten as

$$n\cdot\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u}) = \frac{\operatorname{tr}((J(\mathbf{w}^*_\lambda)+\nu\mathbf{I}_d)^{-1}\mathcal{I}_l(\mathbf{w}^*_\lambda))}{1+\lambda}.$$

Details of the derivation are provided in Appendix M. We focus on the right-hand side, which depends on the ratio λ instead of the individual m, n. As mentioned in Section 4.1, when $\lambda \to 0$, $n\cdot\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_l,\hat{S}_u}, P_{S_l,\hat{S}_u}) \to d$. On the other hand, the excess risk $\mathcal{E}_r(P_\mathbf{W})$ of this example is given by Corollary 2 where $J(\mathbf{w}^*_\lambda)$ is replaced by $J(\mathbf{w}^*_\lambda)+\nu\mathbf{I}_d$. Intuitively, as the regularization parameter ν increases, the gen-error decreases.

For different values of λ, we can numerically calculate the hypothesis $\mathbf{w}^*_\lambda$, the expected gen-error and the excess risk. Consider the example of a dataset which contains two classes of Gaussian samples. Let $P_{\mathbf{X}|Y} = \mathcal{N}(Y\mu\mathbf{1}_d, \mathbf{I}_d)$ and $P_Y = \mathrm{Unif}(\{-1,+1\})$, where $\mu \in \mathbb{R}_+$ and $\mathbf{1}_d$ is the all-ones vector in $\mathbb{R}^d$. For notational simplicity, let



(a) Theoretical gen-error

(b) Empirical gen-error

(c) Theoretical excess risk

Figure 3: Theoretical and empirical gen-error and excess risk vs. $\lambda = m/n$ when $\mu = 2$ and $\nu = 0.001$.

$\overline{\mathrm{gen}}_{\mathrm{SSL}}$ denote $\overline{\mathrm{gen}}(P^{\infty}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$. In Figure 3, we plot $n \cdot \overline{\mathrm{gen}}_{\mathrm{SSL}}$ and $\mathcal{E}_{\mathrm{r}}(P_{\mathbf{W}})$ versus λ when $\mu = 2$ and $d = 2$. We also implement experiments ($n = 1,000$) for 40 times on synthetic data and plot the average empirical gen-error for $\lambda \in \{0, 0.5, 1, 3, 10, 30, 100\}$. In both theoretical and empirical plots, we observe that as the ratio λ increases, the gen-error of SSL decreases and is smaller than that of SL with only n labeled data. We also observe similar behaviours of the empirical gen-error of logistic regression on MNIST dataset (see Appendix M). On the other hand, the variance component of the excess risk behaves similarly as the gen-error while the bias component increases. By taking $n = 1,000$, we can see the excess risk first decreases and then increases as λ increases. In this setting, our results suggest that when λ exceeds some threshold, it may not be beneficial to the performance of the learning algorithm by utilizing any more unlabeled data.

5 CONCLUDING REMARKS

To develop a comprehensive understanding of SSL, we present an exact characterization of the expected gen-error of pseudo-labeling-based SSL via the Gibbs algorithm. Our results reveal that the expected gen-error is influenced by *the information shared between the labeled and pseudo-labeled data* and the ratio of the number of unlabeled data to labeled data. This sheds light in our quest to design good pseudo-labeling methods, which should penalize the dependence between the labeled and pseudo-labeled data, e.g., $I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})$, so as to improve the generalization performance. To understand the ERM counterpart of pseudo-labeling-based SSL, we also investigate the asymptotic regime, in which the inverse temperature $\alpha \to \infty$. Finally, we present two examples—mean estimation and logistic regression—as applications of our theoretical findings.

Acknowledgements

Haiyun He and Vincent Tan are supported by a Singapore National Research Foundation (NRF) Fellowship (Grant Number: A-0005077-01-00). Gholamali Aminian is supported by the UKRI Prosperity Partnership Scheme (FAIR) under the EPSRC Grant EP/V056883/1. M. R. D. Rodrigues and Gholamali Aminian are also supported by the Alan Turing Institute.

References

Massih-Reza Amini and Patrick Gallinari. Semi-supervised logistic regression. In *ECAI*, volume 2, page 11, 2002.

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. *Advances in Neural Information Processing Systems*, 34, 2021a.

Gholamali Aminian, Laura Toni, and Miguel RD Rodrigues. Information-theoretic bounds on the moments of the generalization error of learning algorithms. In *IEEE International Symposium on Information Theory (ISIT)*, 2021b.

Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pages 7433–7449. PMLR, 2022a.

Gholamali Aminian, Saeed Masiha, Laura Toni, and Miguel RD Rodrigues. Learning algorithm generalization error bounds via auxiliary distributions. *arXiv preprint arXiv:2210.00483*, 2022b.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

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.

KB Athreya and Chii-Ruey Hwang. Gibbs measures asymptotics. *Sankhya A*, 72(1):191–207, 2010.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

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.

Yuheng Bu, Gholamali Aminian, Laura Toni, Gregory W Wornell, and Miguel Rodrigues. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 8673–8699. PMLR, 2022.

Vittorio Castelli and Thomas M Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

Olivier Catoni. PAC-Bayesian supervised classification: the thermodynamics of statistical learning. *arXiv preprint arXiv:0712.0248*, 2007.

Olivier Chapelle, Jason Weston, and Bernhard Scholkopf. Cluster kernels for semi-supervised learning. *Advances in neural information processing systems*, pages 601–608, 2003.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. The MIT Press, 2006.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in $\mathbb{R}^n$. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Robert Dupre, Jiri Fajtl, Vasileios Argyriou, and Paolo Remagnino. Improving dataset volumes and model accuracy with semi-supervised iterative self-learning. *IEEE Transactions on Image Processing*, 29:4337–4348, 2019.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, f-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 2021.

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

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518. PMLR, 2019.

Yves Grandvalet, Yoshua Bengio, et al. Semi-supervised learning by entropy minimization. *CAP*, 367:281–296, 2005.

Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

Haiyun He, Hanshu Yan, and Vincent YF Tan. Information-theoretic characterization of the generalization error for iterative semi-supervised learning. *Journal of Machine Learning Research*, 23:1–52, 2022.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

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.

Ming Ji, Tianbao Yang, Binbin Lin, Rong Jin, and Jiawei Han. A simple algorithm for semi-supervised learning with improved generalization error bound. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 835–842, 2012.

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.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, 2013.

Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Multi-class learning using unlabeled samples: Theory and algorithm. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2880–2886, 2019.

Peter A Markowich and Cédric Villani. On the trend to equilibrium for the Fokker-Planck equation: an interplay between physics and functional analysis. *Mat. Contemp*, 19:1–29, 2000.

Geoffrey J McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2005.

Todd K Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

Gang Niu, Wittawat Jitkrittum, Bo Dai, Hirotaka Hachiya, and Masashi Sugiyama. Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013.

Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

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

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.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7), 2007.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Daniel Russo and James Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1): 302–323, 2019.

Matthias Seeger. Learning with labeled and unlabeled data. 2000. URL `http://infoscience.epfl.ch/record/161327`.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in Neural Information Processing Systems*, 21:1513–1520, 2008.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pages 3437–3452. PMLR, 2020.

Martin Szummer and Tommi Jaakkola. Information regularization with partially labeled data. *Advances in Neural Information processing systems*, 15, 2002.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 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.

Jingge Zhu. Semi-supervised learning: the case when unlabeled data is equally useful. In *Conference on Uncertainty in Artificial Intelligence*, pages 709–718. PMLR, 2020.

Xiaojin Zhu and Andrew B Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2008.

A RELATED WORKS

Semi-supervised learning: The SSL approaches can be partitioned into six main classes: generative models, low-density separation methods, graph-based methods, self-training and co-training (Zhu, 2008). Among all these, SSL first appeared as self-training in which the model is first trained on the labeled data and annotates the unlabeled data to improve the intial model (Chapelle et al., 2006). SSL has gradually gained more attention after the well-known expectation-minimization (EM) algorithm Moon (1996) was proposed in 1996. One key problem of interest in SSL is whether the unlabeled data can help to improve the performance of learning algorithms. Many existing works have studied this problem either theoretically (e.g., providing bounds) or empirically (e.g., proposing new algorithms). One classical work by Castelli and Cover (1996) set out to study SSL under a traditional setup with unknown mixture of known distributions and characterized the error probability by the fisher information of the labeled and unlabeled data. Szummer and Jaakkola (2002) proposed an algorithm that utilizes the unlabeled data to learn the marginal data distribution to augment learning the class conditional distribution. Amini and Gallinari (2002) empirically showed that semi-supervised logistic regression based on EM algorithm has higher accuracy than the naive Bayes classifier. Singh et al. (2008) studied the benefit of unlabeled data on the excess risk based on the number of unlabeled data and the margin between classes. Ji et al. (2012) developed a simple algorithm based on top eigenfunctions of integral operator derived from both labeled and unlabeled examples that can improve the regression error bound. Li et al. (2019) showed how the unlabeled data can improve the Rademacher-complexity-based generalization error bound of a multi-class classification problem. In deep learning, Berthelot et al. (2019) introduced an effective algorithm that generates low-entropy labels for unlabeled data and then mixes them up with the labeled data to train the model. Sohn et al. (2020) showed that augmenting the confidently pseudo-labeled images can help to improve the accuracy of their model. For a more comprehensive overview of SSL, one can refer to the report by Seeger (2000) and the book by Chapelle et al. (2006).

Pseudo-labeling: Pseudo-labeling is one of the approaches in self-training (Zhu and Goldberg, 2009). Due to the reliance of pseudo-labeling on the quality of the pseudo labels, pseudo-labeling approach might perform poorly. Seeger (2000) stated that there exists a tradeoff between robustness of the learning algorithm and the information gain from the pseudo-labels. Rizve et al. (2020) offered an uncertainty-aware pseudo-labeling strategy to circumvent this difficulty. In Wei et al. (2020), a theoretical framework for combining input-consistency regularization with self-training algorithms in deep neural networks is provided. Dupre et al. (2019) empirically showed that iterative pseudo-labeling with a confidence threshold can improve the test accuracy in early stage. Arazo et al. (2020) showed that the method of generating soft labels for unlabeled data plus mixup augmentation can outperform consistency regularization methods.

Despite the plenty of works on SSL and pseudo-labeling, our work provides a new viewpoint of understanding the effect of pseudo-labeling method on the generalization error using information-theoretic quantities.

Information-theoretic upper bounds: Russo and Zou (2019); Xu and Raginsky (2017) proposed to use the mutual information between the input training set and the output hypothesis to upper bound the expected generalization error. This paves a new way of understanding and improving the generalization performance of a learning algorithm from an information-theoretic viewpoint. Tighter upper bounds by considering the individual sample mutual information is proposed by Bu et al. (2020). Asadi et al. (2018) proposed using chaining mutual information, and Hafez-Kolahi et al. (2020); Haghifam et al. (2020); Steinke and Zakynthinou (2020) advocated the conditioning and processing techniques. Information-theoretic generalization error bounds using other information quantities are also studied, e.g., f-divergences, α-Rényi divergence and generalized Jensen-Shannon divergence Esposito et al. (2021); Aminian et al. (2022b, 2021b).

B MOTIVATIONS FOR GIBBS ALGORITHM

There are different motivations for the Gibbs algorithm as discussed in Raginsky et al. (2017); Kuzborskij et al. (2019); Asadi and Abbe (2020); Aminian et al. (2021a). Following are an overview of the most prominent motivations for the Gibbs algorithm:

Empirical Risk Minimization: The $(\alpha, \pi(\mathbf{W}), \bar{L}_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs algorithm can be viewed as a randomized version of empirical risk minimization (ERM). As the inverse temperature $\alpha \to \infty$, then hypothesis generated by the Gibbs algorithm converges to the hypothesis corresponding to standard ERM.

SGLD Algorithm: As discussed in Chiang et al. (1987) and Markowich and Villani (2000), the Stochastic Gradient Langevin Dynamics (SGLD) algorithm can be viewed as the discrete version of the continuous-time Langevin diffusion, and it is defined as follows:

$$\mathbf{W}_{k+1} = \mathbf{W}_k - \beta \nabla \bar{L}_{\mathrm{E}}(\mathbf{W}_k, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) + \sqrt{\frac{2\beta}{\gamma}}\, \zeta_k, \quad k = 0, 1, \cdots, \tag{20}$$

where ζ_k is a standard Gaussian random vector and $\beta > 0$ is the step size. In Raginsky et al. (2017), it is proved that under some conditions on the loss function, the conditional distribution $P_{\mathbf{W}_k|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}$ induced by SGLD algorithm is close to the $(\gamma, \pi(\mathbf{W}_0), \bar{L}_{\mathrm{E}}(\mathbf{W}_k, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs distribution in 2-Wasserstein distance for sufficiently large iterations, k.

C PROOF OF THEOREM 1

Under the $(\alpha, \pi(\mathbf{W}), \bar{L}_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))$-Gibbs algorithm, we have

$$\begin{aligned}
&I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \\
&= \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}} + \log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] - \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}} + \log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] && (21)\\
&= \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}[-\alpha \bar{L}_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) - \log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] - \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}[-\alpha \bar{L}_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) - \log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}[\log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] - \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}[\log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] && (22)\\
&= \frac{\alpha}{1+\eta}(\mathbb{E}_{\mathbf{W}}[L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})] - \mathbb{E}_{\mathbf{W},S_{\mathrm{l}}}[L_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}})]) + \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}[\log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] - \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}[\log P_{\hat{S}_{\mathrm{u}}|S_{\mathrm{l}}}] \\
&\quad - \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] + \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] && (23)\\
&= \frac{\alpha}{1+\eta}\overline{\mathrm{gen}}(P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) + I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - \mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})], && (24)
\end{aligned}$$

where $\Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \int \pi(\mathbf{w}) \exp(-\alpha \bar{L}_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))\mathrm{d}\mathbf{w}$ and $\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\cdot] = \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\cdot] - \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}[\cdot]$.

Thus, the expected gen-error is given by

$$\overline{\mathrm{gen}}(P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{(1+\eta)(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) + \mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})])}{\alpha}. \tag{25}$$

Remark 1. *(Apply Theorem 1 to SL) When $\eta \to 0$ and only the labeled dataset S_{l} is used for learning the output hypothesis $\mathbf{W}$, the setup recovers back to SL and we have $\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] = 0$. From Theorem 1 and* (7)*, the expected gen-error becomes*

$$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{I_{\mathrm{SKL}}(\mathbf{W}; S_{\mathrm{l}})}{\alpha}, \tag{26}$$

reducing to SL with n labeled data examples as in Aminian et al. (2021a).

D EXACT CHARACTERIZATION OF EXPECTED GEN-ERROR UNDER ENTROPY MINIMIZATION

Let us recall SSL by entropy minimization in Amini and Gallinari (2002) and Grandvalet et al. (2005). In these works, by letting the loss function be the negative log-likelihood $P_{Y|X;\mathbf{W}}$ under hypothesis $\mathbf{W}$, the authors considered minimizing the regularized empirical risk as follows:

$$L^{\mathrm{EM}}_{\mathrm{E}}(\mathbf{W}; S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \frac{1}{n+m}\Bigg(\underbrace{-\sum_{i=1}^{n} \log P_{Y|X;\mathbf{W}}(Y_i|X_i;\mathbf{W})}_{\text{Empirical risk of } S_{\mathrm{l}}} \underbrace{- \sum_{i=n+1}^{n+m}\sum_{k=1}^{K} P_{Y|X;\mathbf{W}}(k|X_i;\mathbf{W}) \log P_{Y|X;\mathbf{W}}(k|X_i;\mathbf{W})}_{\text{Regularizer: empirical entropy of } S_{\mathrm{u}}}\Bigg).$$

Under the $(\alpha, \pi(\mathbf{W}), L^{\mathrm{EM}}_{\mathrm{E}}(\mathbf{W}; S_{\mathrm{l}}, S_{\mathrm{u}}))$-Gibbs algorithm, the posterior distribution of $\mathbf{W}$ can be denoted as $P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},S_{\mathrm{u}}}$. The output hypothesis $\mathbf{W}$ only depends on the labeled data S_{l} and unlabeled data S_{u} instead of pseudo-labeled data $\hat{S}_{\mathrm{u}}$. Since S_{l} and S_{u} are independent, according to (8), by replacing λ with $\frac{m}{n}$, we can characterize the expected gen-error (cf. (4)) inspired by Theorem 1 in the following corollary.

Corollary 3. *Under the semi-supervised Gibbs algorithm with entropy minimization (i.e., the* $(\alpha, \pi(\mathbf{W}), L_{\mathrm{E}}^{EM}(\mathbf{W}; S_{\mathrm{l}}, S_{\mathrm{u}}))$*-Gibbs algorithm), the expected gen-error is*

$$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},S_{\mathrm{u}}}, P_{S_{\mathrm{l}},S_{\mathrm{u}}}) = \frac{(n+m) I_{\mathrm{SKL}}(\mathbf{W}; S_{\mathrm{l}}|S_{\mathrm{u}})}{n\alpha}.$$

E PROOF OF PROPOSITION 1

We provide the proof of Proposition 1 and a novel lower bound on $\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$.

Proof. A *σ-sub-Gaussian random variable* L is such that its cumulant generating function $\Lambda_L(s) := \mathbb{E}[\exp(s(L-\mathbb{E}[L]))] \leq \exp(s^2\sigma^2/2)$ for all $s \in \mathbb{R}$ (Vershynin, 2018).

Assume that the loss function $l(\mathbf{W}, \mathbf{Z})$ is bounded in $[a, b] \subset \mathbb{R}_+$ for any $\mathbf{W} \in \mathcal{W}$ and $\mathbf{Z} \in \mathcal{Z}$. Then we have $\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \in [-\alpha b, -\alpha a]$. According to Hoeffding's lemma, the loss function $l(\mathbf{W}, \mathbf{Z})$ is $\frac{b-a}{2}$-sub-Gaussian and $\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})$ is $\frac{\alpha(b-a)}{2}$-sub-Gaussian. From the Donsker–Varadhan representation, we have

$$|\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})]| \leq \sqrt{\frac{\alpha^2(b-a)^2}{2} I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}, \tag{27}$$

From Theorem 1, we can directly obtain

$$\left| \overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) - \frac{(1+\eta)(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}))}{\alpha} \right| \leq \frac{(1+\eta)(b-a)}{\sqrt{2}} \sqrt{I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}. \tag{28}$$

□

We can also provide another lower bound on $\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$. For any random variable Z, from the chain rule of the mutual information, we have $I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \geq I(X; Y)$. We can also expand the lautum information $L(X; Y, Z)$ as $L(X; Y, Z) = L(X; Y) + D(P_{Z|Y} \| P_{Z|X,Y} | P_X P_Y) \geq L(X; Y)$. Thus, we have

$$I_{\mathrm{SKL}}(X; Y, Z) \geq I_{\mathrm{SKL}}(X; Y). \tag{29}$$

Then the gen-error is lower bounded by

$$\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) \geq \frac{1+\eta}{\alpha} \mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\eta}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})]. \tag{30}$$

The sign of the lower bound depends on the loss function, the prior distribution $\pi(\mathbf{W})$ and the data distributions. As $\eta \to 0$, or $\eta \to \infty$, or $P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}} = P_{S_{\mathrm{l}}} \otimes P_{\hat{S}_{\mathrm{u}}}$, the lower bound vanishes to 0.

F PROOF OF Eq. (7)

From the definition of symmetrized KL information in Section 2.1, we have

$$\begin{aligned}
&I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \\
&= I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) + L(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - L(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \qquad (31) \\
&= I(\mathbf{W}; S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}) + L(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - L(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \qquad (32) \\
&= I(\mathbf{W}; S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}) + D(P_{\mathbf{W}|\hat{S}_{\mathrm{u}}} \| P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}} | P_{S_{\mathrm{l}}} P_{\hat{S}_{\mathrm{u}}}), \qquad (33)
\end{aligned}$$

where (32) follows from the chain rule of mutual information and (33) follows from the expansion of lautum information (Palomar and Verdú, 2008, Eq. (52)).

G PROOFS FOR COMPARISON TO SL WITH $n+m$ LABELED DATA

G.1 Proof of Eq. (10)

Let $\bar{\mathbf{Z}}$, $\bar{\mathbf{Z}}'$ be independent copies of $\mathbf{Z}_i$ and $\hat{\mathbf{Z}}_i$. Then $\bar{\mathbf{Z}} \sim P_{\mathbf{Z}}$ and $\bar{\mathbf{Z}}' \sim P_{\hat{\mathbf{Z}}}$. Recall that with a perfect pseudo-labeling method, $P_{\mathbf{Z}} = P_{\hat{\mathbf{Z}}}$. For $\lambda = \frac{m}{n}$, we have

$$I_{\mathrm{SKL}}(\mathbf{W}; S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \mathbb{E}_{\mathbf{W},S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}] - \mathbb{E}_{\mathbf{W}}\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}] \tag{34}$$

$$= \frac{\alpha n}{n+m}\big(\mathbb{E}_{\mathbf{W}}\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[L_{\mathrm{E}}(\mathbf{W},S_{\mathrm{l}})] - \mathbb{E}_{\mathbf{W},S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[L_{\mathrm{E}}(\mathbf{W},S_{\mathrm{l}})]\big)$$
$$+ \frac{\alpha m}{n+m}\big(\mathbb{E}_{\mathbf{W}}\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[L_{\mathrm{E}}(\mathbf{W},\hat{S}_{\mathrm{u}})] - \mathbb{E}_{\mathbf{W},S_{\mathrm{l}}^{n},\hat{S}_{\mathrm{u}}}[L_{\mathrm{E}}(\mathbf{W},\hat{S}_{\mathrm{u}})]\big) \tag{35}$$

$$= \frac{\alpha}{n+m}\sum_{i=1}^{n}\big(\mathbb{E}_{\mathbf{W}}\mathbb{E}_{\bar{\mathbf{Z}}}[l(\mathbf{W},\bar{\mathbf{Z}})] - \mathbb{E}_{\mathbf{W},\mathbf{Z}_i}[l(\mathbf{W},\mathbf{Z}_i)]\big)$$
$$+ \frac{\alpha}{n+m}\sum_{i=n+1}^{n+m}\big(\mathbb{E}_{\mathbf{W}}\mathbb{E}_{\bar{\mathbf{Z}}'}[l(\mathbf{W},\bar{\mathbf{Z}}')] - \mathbb{E}_{\mathbf{W},\hat{\mathbf{Z}}_i}[l(\mathbf{W},\hat{\mathbf{Z}}_i)]\big) \tag{36}$$

$$= \alpha\mathbb{E}_{\mathbf{W}}\mathbb{E}_{\bar{\mathbf{Z}}}[l(\mathbf{W},\bar{\mathbf{Z}})] - \frac{\alpha}{n+m}\left(\sum_{i=1}^{n}\mathbb{E}_{\mathbf{W},\mathbf{Z}_i}[l(\mathbf{W},\mathbf{Z}_i)] + \sum_{i=n+1}^{n+m}\mathbb{E}_{\mathbf{W},\hat{\mathbf{Z}}_i}[l(\mathbf{W},\hat{\mathbf{Z}}_i)]\right) \tag{37}$$

$$= \alpha\overline{\mathrm{gen}}_{\mathrm{all}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}). \tag{38}$$

where (37) follows since $P_{\mathbf{Z}} = P_{\hat{\mathbf{Z}}}$. Thus, (10) is proved.

G.2 Proof for $\overline{\mathrm{gen}}_{\mathrm{all}}$ when the pseudo-labeling is close to perfect

Without loss of generality, let $S_{\mathrm{l}} = \mathbf{Z}_1 = (\mathbf{X}_1, Y_1)$ and $\hat{S}_{\mathrm{u}} = \hat{\mathbf{Z}}_2 = (\mathbf{X}_2, \hat{Y}_2)$. With the assumption that $P_{\hat{\mathbf{Z}}} = P_{\mathbf{Z}} + \epsilon\Delta$, the joint distributions $P_{\mathbf{W},\hat{\mathbf{Z}}_2}$ and $P_{\mathbf{W},\mathbf{Z}_1}$ are given by

$$P_{\mathbf{W},\hat{\mathbf{Z}}_2}(\cdot,*) = P_{\hat{\mathbf{Z}}}(*)\sum_{\mathbf{z}}P_{\mathbf{Z}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|*,\mathbf{z}) = (P_{\mathbf{Z}}(*) + \epsilon\Delta(*))\sum_{\mathbf{z}}P_{\mathbf{Z}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|*,\mathbf{z}) \tag{39}$$

$$P_{\mathbf{W},\mathbf{Z}_1}(\cdot,*) = P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}P_{\hat{\mathbf{Z}}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) = P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}(P_{\mathbf{Z}}(\mathbf{z}) + \epsilon\Delta(\mathbf{z}))P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) \tag{40}$$

$$= P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}P_{\mathbf{Z}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) + \epsilon P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}\Delta(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) \tag{41}$$

$$= P_{\mathbf{W},\hat{\mathbf{Z}}_2}(\cdot,*) - \epsilon\Delta(*)\sum_{\mathbf{z}}P_{\mathbf{Z}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|*,\mathbf{z}) + \epsilon P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}\Delta(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) \tag{42}$$

$$= P_{\mathbf{W},\hat{\mathbf{Z}}_2}(\cdot,*) + \epsilon\Delta'(\cdot,*), \tag{43}$$

where (42) follows since $P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*) = P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|*,\mathbf{z})$, $\Delta'(\cdot,*) = -\Delta(*)\sum_{\mathbf{z}}P_{\mathbf{Z}}(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|*,\mathbf{z}) + P_{\mathbf{Z}}(*)\sum_{\mathbf{z}}\Delta(\mathbf{z})P_{\mathbf{W}|\hat{\mathbf{Z}}_2,\mathbf{Z}_1}(\cdot|\mathbf{z},*)$ and $\sum_{\mathbf{w},\mathbf{z}}\Delta'(\mathbf{w},\mathbf{z}) = 0$.

First recall that in the SL setting with 2 labeled training data samples, the expected gen-error is given by

$$\overline{\mathrm{gen}}(P_{\mathbf{W}^{(2)}_{\mathrm{SL}}|S^{(2)}_{\mathrm{l}}}, P_{S^{(2)}_{\mathrm{l}}}) = \frac{I_{\mathrm{SKL}}(\mathbf{W}^{(2)}_{\mathrm{SL}}; S^{(2)}_{\mathrm{l}})}{\alpha} = \mathbb{E}_{\mathbf{W}}\mathbb{E}_{\mathbf{Z}}[l(\mathbf{W},\mathbf{Z})] - \mathbb{E}_{\mathbf{W},\mathbf{Z}_1}[l(\mathbf{W},\mathbf{Z}_1)]. \tag{44}$$

Next, in the SSL setting with this assumption, the expected gen-error is given by

$$\overline{\mathrm{gen}}_{\mathrm{all}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$$
$$= \mathbb{E}_{\mathbf{W}}\mathbb{E}_{\mathbf{Z}}[l(\mathbf{W},\mathbf{Z})] - \frac{1}{2}\big(\mathbb{E}_{\mathbf{W},\mathbf{Z}_1}[l(\mathbf{W},\mathbf{Z}_1)] + \mathbb{E}_{\mathbf{W},\hat{\mathbf{Z}}_2}[l(\mathbf{W},\hat{\mathbf{Z}}_2)]\big) \tag{45}$$

$$= \mathbb{E}_{\mathbf{W}}\mathbb{E}_{\mathbf{Z}}[l(\mathbf{W},\mathbf{Z})] - \mathbb{E}_{\mathbf{W},\mathbf{Z}_1}[l(\mathbf{W},\mathbf{Z}_1)] - \frac{\epsilon}{2}\mathbb{E}_{(\mathbf{W},\hat{\mathbf{Z}}_2)\sim\Delta'}[l(\mathbf{W},\hat{\mathbf{Z}}_2)] \tag{46}$$

$$= \frac{I_{\mathrm{SKL}}(\mathbf{W}^{(2)}_{\mathrm{SL}}; S^{(2)}_{\mathrm{l}})}{\alpha} - \frac{\epsilon}{2}\mathbb{E}_{(\mathbf{W},\hat{\mathbf{Z}}_2)\sim\Delta'}[l(\mathbf{W},\hat{\mathbf{Z}}_2)] \tag{47}$$

$$= \overline{\mathrm{gen}}(P_{\mathbf{W}_{\mathrm{SL}}^{(2)}|S_{\mathrm{l}}^{(2)}}, P_{S_{\mathrm{l}}^{(2)}}) + \tilde{\epsilon}, \tag{48}$$

where $\tilde{\epsilon}$ is proportional to ϵ and $\tilde{\epsilon} \to 0$ as $\epsilon \to 0$. The result can be directly extended to S_{l} with n data samples and $\hat{S}_{\mathrm{u}}$ with m data samples.

H PROOF OF PROPOSITION 2

The formal version of Proposition 2 is stated as follows.

Proposition 3 (Formal Version)**.** *For any $0 < \lambda < \infty$, suppose $(I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}))(1 + C_\alpha) \le I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})$ for some constant $C_\alpha \ge 0$ and $I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) = \gamma_{\alpha,\lambda} I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}})$, where $\gamma_{\alpha,\lambda}$ depends on λ and $0 \le \gamma_{\alpha,\lambda} \le 1$. If $l(\mathbf{W}, \mathbf{Z})$ is bounded within $[a, b] \subset \mathbb{R}_+$ for any $\mathbf{W} \in \mathcal{W}$ and $\mathbf{Z} \in \mathcal{Z}$, the expected gen-error can be bounded as follows*

$$\begin{aligned} \frac{-\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}\left(\frac{1}{\sqrt{n}} + (1+\lambda)\sqrt{\gamma_{\alpha,\lambda}}\right) &\le \overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) \\ &\le \frac{\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2\sqrt{n}(1+C_\alpha)(1-\gamma_{\alpha,\lambda})} + \frac{\alpha(b-a)^2}{2n(1+\lambda)(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}. \end{aligned} \tag{49}$$

Since $L(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - L(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) = D(P_{\mathbf{W}|\hat{S}_{\mathrm{u}}} \| P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}} | P_{S_{\mathrm{l}}} P_{\hat{S}_{\mathrm{u}}}) \ge 0$, we always have

$$I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \le I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) + L(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - L(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) = I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}). \tag{50}$$

Without loss of generality, we can take $C_\alpha = 0$. Recall $\lambda = m/n$. Then we have (i.e., the informal version in Proposition 2)

$$\frac{-\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2(1-\gamma_{\alpha,\lambda})}\left(\frac{1}{\sqrt{n}} + \frac{n\sqrt{\gamma_{\alpha,\lambda}}}{n+m}\right) \le \overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) \le \frac{\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2\sqrt{n}(1-\gamma_{\alpha,\lambda})} + \frac{\alpha(b-a)^2}{2(n+m)(1-\gamma_{\alpha,\lambda})}. \tag{51}$$

Proof. If $l(\mathbf{W}, \mathbf{Z})$ is σ-sub-Gaussian under $P_{\mathbf{Z}}$ for every $\mathbf{w} \in \mathcal{W}$, from (Xu and Raginsky, 2017, Theorem 1), we have

$$\begin{aligned} |\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})| &= |\mathbb{E}_{\mathbf{W},S_{\mathrm{l}}}[L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}}) - L_{\mathrm{E}}(\mathbf{W}, S_{\mathrm{l}})]| \\ &\le \sqrt{\frac{2\sigma^2 I(\mathbf{W}; S_{\mathrm{l}})}{n}} \qquad (52) \\ &\le \sqrt{\frac{2\sigma^2 I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{n}}. \qquad (53) \end{aligned}$$

With the assumption that the loss function $l(\mathbf{W}, \mathbf{Z})$ is bounded within $[a, b] \subset \mathbb{R}_+$ for any $\mathbf{W} \in \mathcal{W}$ and $\mathbf{Z} \in \mathcal{Z}$, we have $\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \in [-\alpha b, -\alpha a]$. According to Hoeffding's lemma, the loss function $l(\mathbf{W}, \mathbf{Z})$ is $\frac{b-a}{2}$-sub-Gaussian and $\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})$ is $\frac{\alpha(b-a)}{2}$-sub-Gaussian. Thus, from Theorem 1, we have

$$\frac{(1+\lambda)(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) + \mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})])}{\alpha} \le \sqrt{\frac{(b-a)^2 I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{2n}} \tag{54}$$

$$\Longrightarrow \frac{(1+\lambda)(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}))}{\alpha} \le \sqrt{\frac{(b-a)^2 I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{2n}} - \frac{(1+\lambda)\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})]}{\alpha}. \tag{55}$$

Furthermore, by Donsker–Varadhan representation, $-\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})]$ can be upper bounded as follows

$$-\mathbb{E}_{\Delta_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] = \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] - \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\log \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})] \le \sqrt{\frac{\alpha^2(b-a)^2 I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{2}}. \tag{56}$$

Then (55) can be further upper bounded as

$$\frac{(1+\lambda)(I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}))}{\alpha} \le \sqrt{\frac{(b-a)^2 I(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{2n}} + \sqrt{\frac{(1+\lambda)^2(b-a)^2 I(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}})}{2}}. \tag{57}$$

Suppose $(I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})-I(\hat{S}_{\mathrm{u}};S_{\mathrm{l}}))(1+C_\alpha)\le I_{\mathrm{SKL}}(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})-I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}};S_{\mathrm{l}})$ for some constant $C_\alpha\ge 0$ and $I(\hat{S}_{\mathrm{u}};S_{\mathrm{l}})=\gamma_{\alpha,\lambda}I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})$, where $\gamma_{\alpha,\lambda}$ depends on λ and $0\le\gamma_{\alpha,\lambda}\le 1$. Then we have

$$\frac{(1+\lambda)(1+C_\alpha)(I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})-I(\hat{S}_{\mathrm{u}};S_{\mathrm{l}}))}{\alpha}\le\sqrt{\frac{(b-a)^2I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}{2n}}+\sqrt{\frac{(1+\lambda)^2(b-a)^2I(\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}{2}} \tag{58}$$

$$\frac{(1+\lambda)(1+C_\alpha)(1-\gamma_{\alpha,\lambda})I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}{\alpha}\le\sqrt{\frac{(b-a)^2I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}{2n}}+\sqrt{\frac{(1+\lambda)^2(b-a)^2\gamma_{\alpha,\lambda}I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}{2}} \tag{59}$$

$$\sqrt{I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})}\le\frac{\alpha}{(1+\lambda)(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}\bigg(\sqrt{\frac{(b-a)^2}{2n}}+\frac{(1+\lambda)(b-a)\sqrt{\gamma_{\alpha,\lambda}}}{\sqrt{2}}\bigg). \tag{60}$$

Thus, we have

$$\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}},P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})\le\frac{\alpha(b-a)^2\sqrt{\gamma_{\alpha,\lambda}}}{2\sqrt{n}(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}+\frac{\alpha(b-a)^2}{2n(1+\lambda)(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}. \tag{61}$$

On the other hand, in (6), since SKL ≥ 0, we have

$$\begin{aligned}\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}},P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})&\ge-\frac{(1+\lambda)(b-a)}{\sqrt{2}}\sqrt{I(\hat{S}_{\mathrm{u}};S_{\mathrm{l}})} &(62)\\&=-\frac{(1+\lambda)(b-a)}{\sqrt{2}}\sqrt{\gamma_{\alpha,\lambda}I(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})} &(63)\\&\ge\frac{-\alpha(b-a)^2\gamma_{\alpha,\lambda}}{2(1+C_\alpha)(1-\gamma_{\alpha,\lambda})}\bigg(\sqrt{\frac{1}{n}}+(1+\lambda)\sqrt{\gamma_{\alpha,\lambda}}\bigg). &(64)\end{aligned}$$

□

I PROOFS OF MEAN ESTIMATION EXAMPLE

The $(\alpha,\pi(\mathbf{W}),\bar{L}_{\mathrm{E}}(\mathbf{W},S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}}))$-Gibbs algorithm is given by the following Gibbs posterior distribution

$$\begin{aligned}&P^\alpha_{\mathbf{W}|S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}}}(\mathbf{W}|(Y_i\mathbf{X}_i)_{i=1}^n,(\hat{Y}_i\mathbf{X}_i)_{i=n+1}^{n+m})\\&=\frac{\pi(\mathbf{W})}{\Lambda_{\alpha,\lambda}(S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}})}\exp\bigg[-\frac{\alpha}{(1+\lambda)n}\sum_{i=1}^n\big(\mathbf{W}^\top\mathbf{W}-2\mathbf{W}^\top Y_i\mathbf{X}_i+\mathbf{X}_i^\top\mathbf{X}_i\big)\\&\qquad-\frac{\alpha\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\big(\mathbf{W}^\top\mathbf{W}-2\mathbf{W}^\top\hat{Y}_i\mathbf{X}_i+\mathbf{X}_i^\top\mathbf{X}_i\big)\bigg] &(65)\\&=\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{l,u}}}\exp\bigg(-\frac{1}{2\sigma^2_{\mathrm{l,u}}}\big\|\mathbf{W}-\boldsymbol{\mu}_{n,m}\big\|_2^2\bigg), &(66)\end{aligned}$$

where $\sigma^2_{\mathrm{l,u}}=\frac{1}{2\alpha}$,

$$\boldsymbol{\mu}_{n,m}=\frac{1}{(1+\lambda)n}\sum_{i=1}^nY_i\mathbf{X}_i+\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\hat{Y}_i\mathbf{X}_i, \tag{67}$$

and

$$\begin{aligned}\frac{\pi(\mathbf{W})}{\Lambda_{\alpha,\lambda}(S'_{\mathrm{l}},\hat{S}'_{\mathrm{u}})}&=\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{l,u}}}\exp\bigg(-\alpha\bigg(\boldsymbol{\mu}_{n,m}^\top\boldsymbol{\mu}_{n,m}-\frac{1}{(1+\lambda)n}\sum_{i=1}^n\mathbf{X}_i^\top\mathbf{X}_i-\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\mathbf{X}_i^\top\mathbf{X}_i\bigg)\bigg)\\&=\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{l,u}}}\exp\bigg(-\alpha\bigg(\frac{1}{(1+\lambda)^2n^2}\sum_{i,j\in[n]^2}\mathbf{X}_i^\top\mathbf{X}_i+\frac{\lambda^2}{(1+\lambda)^2m^2}\sum_{i,j\in[n+1:n+m]^2}\mathbf{X}_i^\top\mathbf{X}_i\end{aligned}$$

$$+\frac{2\lambda}{(1+\lambda)^2nm}\sum_{i=1}^{n}\sum_{j=n+1}^{n+m}(Y_i\mathbf{X}_i)^\top(\hat{Y}_j\mathbf{X}_j)-\frac{1}{(1+\lambda)n}\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{X}_i-\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\mathbf{X}_i^\top\mathbf{X}_i\Big)\Big). \tag{68}$$

It can be seen that $P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}$ is a Gaussian distribution. Thus, the output hypothesis $\mathbf{W}$ can be written as

$$\begin{aligned}\mathbf{W}&=\frac{1}{(1+\lambda)n}\sum_{i=1}^{n}Y_i\mathbf{X}_i+\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\hat{Y}_i\mathbf{X}_i+N &(69)\\&=\frac{1}{(1+\lambda)n}\sum_{i=1}^{n}(Y_i\mathbf{X}_i-\boldsymbol{\mu})+\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}(\hat{Y}_i\mathbf{X}_i-\boldsymbol{\mu}')+\frac{\boldsymbol{\mu}+\lambda\boldsymbol{\mu}'}{1+\lambda}+N &(70)\\&=\mathbf{A}\mathbf{T}(S_{\mathrm{l}}')+N_G &(71)\end{aligned}$$

where $N\sim\mathcal{N}(0,\sigma_{\mathrm{l,u}}^2\mathbf{I}_d)$ is independent of $(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})$, $\mathbf{T}(S_{\mathrm{l}}')=[Y_1\mathbf{X}_1-\boldsymbol{\mu};\dots;Y_1\mathbf{X}_1-\boldsymbol{\mu}]\in\mathbb{R}^{nd\times 1}$, $\mathbf{A}=[\frac{1}{(1+\lambda)n}\mathbf{I}_d,\dots,\frac{1}{(1+\lambda)n}\mathbf{I}_d]\in\mathbb{R}^{d\times nd}$, $N_G|\hat{S}_{\mathrm{u}}\sim\mathcal{N}(\boldsymbol{\mu}_{N_G},\sigma_{\mathrm{l,u}}^2\mathbf{I}_d)$ and $\boldsymbol{\mu}_{N_G}=\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}(\hat{Y}_i\mathbf{X}_i-\boldsymbol{\mu}')+\frac{\boldsymbol{\mu}+\lambda\boldsymbol{\mu}'}{1+\lambda}$.

First, let us calculate the expected gen-error according to (5) in Theorem 1. Let $\mathbf{T}=\mathbf{T}(S_{\mathrm{l}}')$ for simplicity. We have

$$\begin{aligned}&I_{\mathrm{SKL}}(\mathbf{W},\hat{S}_{\mathrm{u}};S_{\mathrm{l}})-I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}};S_{\mathrm{l}})\\&=I(\mathbf{W};S_{\mathrm{l}}|\hat{S}_{\mathrm{u}})+D(P_{\mathbf{W}|\hat{S}_{\mathrm{u}}}\|P_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}|P_{S_{\mathrm{l}}}P_{\hat{S}_{\mathrm{u}}}) &(72)\\&=\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}]-\mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|\hat{S}_{\mathrm{u}}}[\log P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}] &(73)\\&=\frac{1}{2\sigma_{\mathrm{l,u}}^2}\Big(\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[-(\mathbf{W}-\boldsymbol{\mu}_{n,m})^\top(\mathbf{W}-\boldsymbol{\mu}_{n,m})]+\mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|\hat{S}_{\mathrm{u}}}[(\mathbf{W}-\boldsymbol{\mu}_{n,m})^\top(\mathbf{W}-\boldsymbol{\mu}_{n,m})]\Big) &(74)\\&=\frac{1}{2\sigma_{\mathrm{l,u}}^2}\Big(\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[\boldsymbol{\mu}_{n,m}^\top\boldsymbol{\mu}_{n,m}]+\mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{\mathbf{W}|\hat{S}_{\mathrm{u}}}[-2\mathbf{W}^\top\boldsymbol{\mu}_{n,m}+\boldsymbol{\mu}_{n,m}^\top\boldsymbol{\mu}_{n,m}]\Big) &(75)\\&=\frac{1}{2\sigma_{\mathrm{l,u}}^2}\Big(\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[(\mathbf{A}\mathbf{T})^\top(\mathbf{A}\mathbf{T})+2(\mathbf{A}\mathbf{T})^\top\boldsymbol{\mu}_{N_G}+\boldsymbol{\mu}_{N_G}^\top\boldsymbol{\mu}_{N_G}] &(76)\\&\quad+\mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}[-2\mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{A}\mathbf{T}]^\top(\mathbf{A}\mathbf{T}+\boldsymbol{\mu}_{N_G})-2\boldsymbol{\mu}_{N_G}^\top(\mathbf{A}\mathbf{T}+\boldsymbol{\mu}_{N_G})+(\mathbf{A}\mathbf{T}+\boldsymbol{\mu}_{N_G})^\top(\mathbf{A}\mathbf{T}+\boldsymbol{\mu}_{N_G})]\Big) &(77)\\&\overset{(a)}{=}\frac{1}{\sigma_{\mathrm{l,u}}^2}\mathbb{E}[(\mathbf{A}\mathbf{T})^\top(\mathbf{A}\mathbf{T})] &(78)\\&=\frac{1}{\sigma_{\mathrm{l,u}}^2}\operatorname{tr}(\mathbf{A}^\top\mathbf{A}\mathbb{E}[\mathbf{T}\mathbf{T}^\top]) &(79)\\&=\frac{2\alpha\sigma^2 d}{(1+\lambda)^2 n}, &(80)\end{aligned}$$

where (a) follows since $\mathbb{E}_{S_{\mathrm{l}}'}[\mathbf{A}\mathbf{T}(S_{\mathrm{l}}')]=0$, and

$$\begin{aligned}&\mathbb{E}_{\Delta_{S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}}[\log\Lambda_{\alpha,\lambda}(S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}')]\\&=\alpha\mathbb{E}_{\Delta_{S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}}\Big[\frac{1}{(1+\lambda)^2n^2}\sum_{i,j\in[n]^2}\mathbf{X}_i^\top\mathbf{X}_i+\frac{\lambda^2}{(1+\lambda)^2m^2}\sum_{i,j\in[n+1:n+m]^2}\mathbf{X}_i^\top\mathbf{X}_i\\&\quad+\frac{2\lambda}{(1+\lambda)^2nm}\sum_{i=1}^{n}\sum_{j=n+1}^{n+m}(Y_i\mathbf{X}_i)^\top(\hat{Y}_j\mathbf{X}_j)-\frac{1}{(1+\lambda)n}\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{X}_i-\frac{\lambda}{(1+\lambda)m}\sum_{i=n+1}^{n+m}\mathbf{X}_i^\top\mathbf{X}_i\Big] &(81)\\&=\alpha\mathbb{E}_{\Delta_{S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}}\Big[\frac{2\lambda}{(1+\lambda)^2nm}\sum_{i=1}^{n}\sum_{j=n+1}^{n+m}(Y_i\mathbf{X}_i)^\top(\hat{Y}_j\mathbf{X}_j)\Big] &(82)\\&=\frac{2\alpha\lambda}{(1+\lambda)^2nm}\sum_{i=1}^{n}\sum_{j=n+1}^{n+m}\big(\mathbb{E}_{S_{\mathrm{l}}',\hat{S}_{\mathrm{u}}'}[(Y_i\mathbf{X}_i-\boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')]+\boldsymbol{\mu}^\top\boldsymbol{\mu}'-\boldsymbol{\mu}^\top\boldsymbol{\mu}\big) &(83)\end{aligned}$$

$$\overset{\text{(b)}}{=} \frac{2\alpha\lambda}{(1+\lambda)^2}\mathbb{E}\big[(Y_i\mathbf{X}_i-\boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')\big], \tag{84}$$

where (b) follows since $\mathbb{E}[(Y_i\mathbf{X}_i-\boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')]$ is symmetric for any $i\in[n]$ and $j\in[n+1:n+m]$. By letting $\lambda=\frac{m}{n}$ and combining (80) and (84), the expected gen-error of this example is given by

$$\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}},P_{S_\mathrm{l},\hat{S}_\mathrm{u}})=\frac{2\sigma^2 d}{n+m}+\frac{2m}{n+m}\mathbb{E}\big[(Y_i\mathbf{X}_i-\boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')\big]. \tag{85}$$

From the definition of gen-error in (4), we can also derive the same result as follows. Let $\bar{Y}\bar{\mathbf{X}}$ be an independent copy of $Y_i\mathbf{X}_i$ for any $i\in[n]$.

$$\begin{aligned}
&\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}},P_{S_\mathrm{l},\hat{S}_\mathrm{u}})\\
&=\mathbb{E}_{\mathbf{W}}\mathbb{E}_{\bar{Y}\bar{\mathbf{X}}}[\|\bar{Y}\bar{\mathbf{X}}-\mathbf{W}\|_2^2]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{\mathbf{W},Y_i\mathbf{X}_i}[\|Y_i\mathbf{X}_i-\mathbf{W}\|_2^2] &(86)\\
&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[2\mathbf{W}^\top(Y_i\mathbf{X}_i-\bar{Y}\bar{\mathbf{X}})-\mathbf{X}_i^\top\mathbf{X}_i+\bar{\mathbf{X}}^\top\bar{\mathbf{X}}] &(87)\\
&\overset{\text{(c)}}{=}\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[2(\boldsymbol{\mu}_{n,m}+N)^\top(Y_i\mathbf{X}_i-\bar{Y}\bar{\mathbf{X}})\big] &(88)\\
&=\frac{1}{n}\sum_{i=1}^n\mathbb{E}\big[2\boldsymbol{\mu}_{n,m}^\top(Y_i\mathbf{X}_i-\bar{Y}\bar{\mathbf{X}})\big] &(89)\\
&=\frac{1}{n}\sum_{i=1}^n 2\mathbb{E}\Bigg[\bigg(\frac{1}{n+m}\sum_{j=1}^n(Y_j\mathbf{X}_j-\boldsymbol{\mu})+\frac{1}{n+m}\sum_{j=n+1}^{n+m}(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')+\frac{\boldsymbol{\mu}+\lambda\boldsymbol{\mu}'}{1+\lambda}\bigg)^\top\big((Y_i\mathbf{X}_i-\boldsymbol{\mu})-(\bar{Y}\bar{\mathbf{X}}-\boldsymbol{\mu})\big)\Bigg] &(90)\\
&=\frac{1}{n}\sum_{i=1}^n\bigg(\frac{2}{n+m}\mathbb{E}[(Y_i\mathbf{X}_i-\boldsymbol{\mu})^\top(Y_i\mathbf{X}_i-\boldsymbol{\mu})]+\frac{2}{n+m}\sum_{j=n+1}^{n+m}\mathbb{E}[(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')^\top(Y_i\mathbf{X}_i-\boldsymbol{\mu})]\bigg) &(91)\\
&=\frac{2\sigma^2 d}{n+m}+\frac{1}{n}\sum_{i=1}^n\frac{2}{n+m}\sum_{j=n+1}^{n+m}\mathbb{E}[(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')^\top(Y_i\mathbf{X}_i-\boldsymbol{\mu})] &(92)\\
&\overset{\text{(d)}}{=}\frac{2\sigma^2 d}{n+m}+\frac{2m}{n+m}\mathbb{E}[(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')^\top(Y_i\mathbf{X}_i-\boldsymbol{\mu})], &(93)
\end{aligned}$$

where (c) follows from (69), (d) follows since $\mathbb{E}[(\hat{Y}_j\mathbf{X}_j-\boldsymbol{\mu}')^\top(Y_i\mathbf{X}_i-\boldsymbol{\mu})]$ is symmetric for any $i\in[n]$ and $j\in[n+1:n+m]$.

I.1 Mean Estimation under Supervised Learning

Under the supervised $(\alpha,\pi(\mathbf{W}_\mathrm{SL}^{(n)}),L_\mathrm{E}(\mathbf{W}_\mathrm{SL}^{(n)},S_\mathrm{l}'))$-Gibbs algorithm, the posterior distribution $P_{\mathbf{W}_\mathrm{SL}^{(n)}|S_\mathrm{l}'}$ is given by

$$P_{\mathbf{W}_\mathrm{SL}^{(n)}|S_\mathrm{l}'}=\mathcal{N}\bigg(\frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i,\sigma_{\mathrm{l,u}}^2\mathbf{I}_d\bigg). \tag{94}$$

According to (Aminian et al., 2021a, Theorem 1), with the similar techniques of obtaining (80), the expected gen-error is given by

$$\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}_\mathrm{SL}^{(n)}|S_\mathrm{l}'},P_{S_\mathrm{l}'})=\frac{I_\mathrm{SKL}(\mathbf{W}_\mathrm{SL}^{(n)};S_\mathrm{l}')}{\alpha}=2\,\mathrm{tr}(\mathbf{A}_1\mathbb{E}[\mathbf{T}(S_\mathrm{l}')\mathbf{T}(S_\mathrm{l}')^\top]\mathbf{A}_1^\top)=\frac{2\sigma^2 d}{n}. \tag{95}$$

where $\mathbf{T}(S_\mathrm{l}')=[Y_1\mathbf{X}_1-\boldsymbol{\mu};\ldots;Y_1\mathbf{X}_1-\boldsymbol{\mu}]\in\mathbb{R}^{nd\times 1}$ and $\mathbf{A}_1=[\frac{1}{n}\mathbf{I}_d,\ldots,\frac{1}{n}\mathbf{I}_d]\in\mathbb{R}^{d\times nd}$.

Similarly, under the supervised $(\alpha,\pi(\mathbf{W}_\mathrm{SL}^{(n+m)}),L_\mathrm{E}(\mathbf{W}_\mathrm{SL}^{(n+m)},S_\mathrm{l}'^{(n+m)}))$-Gibbs algorithm, where $S_\mathrm{l}'^{(n+m)}$ contains $n+m$ i.i.d. $Y_i\mathbf{X}_i$ samples and $L_\mathrm{E}(\mathbf{W}_\mathrm{SL}^{(n+m)},S_\mathrm{l}'^{(n+m)})=\frac{1}{n+m}\sum_{i=1}^{n+m}l(\mathbf{W}_\mathrm{SL}^{(n+m)},\mathbf{Z}_i')$, the expected gen-error (cf. (9)) is given by

$$\overline{\text{gen}}(P^{\alpha}_{\mathbf{W}_{\text{SL}}^{(n+m)}|S_{\text{l}}'^{(n+m)}}, P_{S_{\text{l}}'^{(n+m)}}) = \frac{I_{\text{SKL}}(\mathbf{W}_{\text{SL}}^{(n+m)}; S_{\text{l}}'^{(n+m)})}{\alpha} = \frac{2\sigma^2 d}{n+m}. \tag{96}$$

I.2 Proofs for Eq. (13)

Let us rewrite the gen-error in (12) as follows: for any $i \in [n]$ and $j \in [n+1:n+m]$.

$$\overline{\text{gen}}(P^{\alpha}_{\mathbf{W}|S_{\text{l}},\hat{S}_{\text{u}}}, P_{S_{\text{l}},\hat{S}_{\text{u}}}) = \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\hat{Y}_j\mathbf{X}_j - \boldsymbol{\mu}')\big]$$

$$= \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\mathbb{E}\big[(Y_i\mathbf{X}_i - \boldsymbol{\mu})^\top(\text{sgn}(\mathbf{W}_0^\top\mathbf{X}_j)\mathbf{X}_j - \boldsymbol{\mu}')\big] \tag{97}$$

$$= \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\big(\mathbb{E}\big[(\text{sgn}(\mathbf{W}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top Y_i\mathbf{X}_i)\big] - \mathbb{E}\big[(\text{sgn}(\mathbf{W}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top)\big]\boldsymbol{\mu}\big). \tag{98}$$

Let $\mathbf{W}_0 = \frac{1}{n}\sum_{i=1}^n Y_i\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$. Using the proof idea from (He et al., 2022), we can decompose it as

$$\mathbf{W}_0 = \boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\xi} = \left(1 + \frac{\sigma}{\sqrt{n}}\xi_0\right)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^\perp, \tag{99}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\xi_0 \sim \mathcal{N}(0,1)$, $\boldsymbol{\mu}^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top)$ and $\boldsymbol{\mu}^\perp$ is perpendicular to $\boldsymbol{\mu}$ and independent of ξ_0. The normalized $\mathbf{W}_0$ can be written as

$$\overline{\mathbf{W}}_0 = \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|_2} = \gamma_n\boldsymbol{\mu} + \bar{\gamma}_n\boldsymbol{\upsilon} \tag{100}$$

where $\boldsymbol{\upsilon} = \boldsymbol{\mu}^\perp/\|\boldsymbol{\mu}^\perp\|_2$, $\gamma_n^2 + \bar{\gamma}_n^2 = 1$ and

$$\gamma_n = \gamma_n(\xi_0, \boldsymbol{\mu}^\perp) := \frac{1 + \frac{\sigma}{\sqrt{n}}\xi_0}{\sqrt{(1 + \frac{\sigma}{\sqrt{n}}\xi_0)^2 + \frac{\sigma^2}{n}\|\boldsymbol{\mu}^\perp\|_2^2}}. \tag{101}$$

For any $i \in [n+m]$, since $Y_i\mathbf{X}_i \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$, we have

$$Y_i\mathbf{X}_i = \boldsymbol{\mu} + \sigma\mathbf{g}_i = \boldsymbol{\mu} + \tilde{g}_i\boldsymbol{\mu} + \boldsymbol{\mu}_i^\perp, \tag{102}$$

where $\mathbf{g}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\tilde{g}_i \sim \mathcal{N}(0,1)$, $\boldsymbol{\mu}_i^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top)$ and $\tilde{g}_i$ is independent of $\boldsymbol{\mu}_i^\perp$. Given any $Y_i\mathbf{X}_i$ for $i \in [1:n]$, we have

$$\mathbf{W}_0|Y_i\mathbf{X}_i = \frac{1}{n}Y_i\mathbf{X}_i + \frac{n-1}{n}\boldsymbol{\mu} + \frac{\sqrt{n-1}}{n}\sigma\boldsymbol{\xi}' \tag{103}$$

$$= \frac{1}{n}(\boldsymbol{\mu} + \sigma\mathbf{g}_i) + \frac{n-1}{n}\boldsymbol{\mu} + \frac{\sqrt{n-1}}{n}\sigma(\xi_0'\boldsymbol{\mu} + \boldsymbol{\mu}'^\perp) \tag{104}$$

$$= \left(1 + \frac{\sqrt{n-1}}{n}\sigma\xi_0' + \frac{\sigma}{n}\tilde{g}_i\right)\boldsymbol{\mu} + \left(\frac{\sqrt{n-1}}{n}\sigma\|\boldsymbol{\mu}'^\perp\|_2 + \frac{\sigma}{n}\|\boldsymbol{\mu}_i^\perp\|_2\right)\boldsymbol{\upsilon}, \tag{105}$$

where $\boldsymbol{\xi}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, $\xi_0' \sim \mathcal{N}(0,1)$, $\boldsymbol{\mu}'^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top)$ and $\boldsymbol{\mu}'^\perp$ is perpendicular to $\boldsymbol{\mu}$ and independent of ξ_0'. The normalized version is given by

$$\overline{\mathbf{W}}_0|Y_i\mathbf{X}_i = \frac{\mathbf{W}_0}{\|\mathbf{W}_0\|_2}\big|Y_i\mathbf{X}_i = \gamma_n'\boldsymbol{\mu} + \bar{\gamma}_n'\boldsymbol{\upsilon} \tag{106}$$

where $\gamma_n'^2 + \bar{\gamma}_n'^2 = 1$ and

$$\gamma_n' = \gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp) := \frac{1 + \frac{\sqrt{n-1}}{n}\sigma\xi_0' + \frac{\sigma}{n}\tilde{g}_i}{\sqrt{(1 + \frac{\sqrt{n-1}}{n}\sigma\xi_0' + \frac{\sigma}{n}\tilde{g}_i)^2 + (\frac{\sqrt{n-1}}{n}\sigma\|\boldsymbol{\mu}'^\perp\|_2 + \frac{\sigma}{n}\|\boldsymbol{\mu}_i^\perp\|_2)^2}}. \tag{107}$$

Define the correlation evolution function $F_\sigma : [-1,1] \to [-1,1]$:

$$F_\sigma(x) := J_\sigma(x)/\sqrt{J_\sigma(x)^2 + K_\sigma(x)^2}, \tag{108}$$

where $J_\sigma(x) := 1 - 2Q(\frac{x}{\sigma}) + \frac{2\sigma x}{\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2})$ and $K_\sigma := \frac{2\sigma\sqrt{1-x^2}}{\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2})$.

For any $j \in [n+1 : n+m]$, we decompose the Gaussian random vector $\mathbf{g}_j \sim \mathcal{N}(0, \mathbf{I}_d)$ in another way

$$\mathbf{g}_j = \tilde{g}_j \overline{\mathbf{W}}_0 + \tilde{\mathbf{g}}_j^\perp, \tag{109}$$

where $\tilde{g}_j \sim \mathcal{N}(0,1)$, $\tilde{\mathbf{g}}_j^\perp \sim \mathcal{N}(0, \mathbf{I}_d - \bar{\mathbf{W}}_0\bar{\mathbf{W}}_0^\top)$, $\tilde{g}_j$ and $\tilde{\mathbf{g}}_j^\perp$ are mutually independent and $\tilde{\mathbf{g}}_j^\perp \perp \bar{\mathbf{W}}_0$. Then we decompose $\mathbf{X}_j$ and $\bar{\mathbf{W}}_0^\top \mathbf{X}_j$ as

$$\mathbf{X}_j = Y_j\boldsymbol{\mu} + \sigma\tilde{g}_j\overline{\mathbf{W}}_0 + \sigma\tilde{\mathbf{g}}_j^\perp, \text{ and} \tag{110}$$

$$\overline{\mathbf{W}}_0^\top \mathbf{X}_j = Y_j\gamma_n + \sigma\tilde{g}_j. \tag{111}$$

Then we have

$$\begin{aligned}&\mathbb{E}[\operatorname{sgn}(\bar{\mathbf{W}}_0^\top\mathbf{X}_j)\mathbf{X}_j \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j = -1]\\ &= -\mathbb{E}[\operatorname{sgn}(-\gamma_n + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} + \sigma\mathbb{E}[\operatorname{sgn}(-\gamma_n + \sigma\tilde{g}_j)\tilde{g}_j|\xi_0, \boldsymbol{\mu}^\perp]\overline{\mathbf{W}}_0,\end{aligned} \tag{112}$$

where (112) follows since $\tilde{\mathbf{g}}^\perp$ is independent of $\tilde{g}_j$ and $\mathbb{E}[\tilde{\mathbf{g}}^\perp] = 0$. Since $\tilde{g}_j \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}\big[\operatorname{sgn}\big(\overline{\mathbf{W}}_0^\top\mathbf{X}_j'\big)\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = -1\big] = \bigg(1 - 2\mathrm{Q}\bigg(\frac{\gamma_n}{\sigma}\bigg)\bigg)\boldsymbol{\mu} + \frac{2\sigma}{\sqrt{2\pi}}\exp\bigg(-\frac{\gamma_n^2}{2\sigma^2}\bigg)\overline{\mathbf{W}}_0, \tag{113}$$

and similarly,

$$\mathbb{E}\big[\operatorname{sgn}\big(\overline{\mathbf{W}}_0^\top\mathbf{X}_j'\big)\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = 1\big] = \bigg(2\mathrm{Q}\bigg(-\frac{\gamma_n}{\sigma}\bigg) - 1\bigg)\boldsymbol{\mu} + \frac{2\sigma}{\sqrt{2\pi}}\exp\bigg(-\frac{\gamma_n^2}{2\sigma^2}\bigg)\overline{\mathbf{W}}_0. \tag{114}$$

Recall the definitions of J_σ and K_σ. Then we have

$$\mathbb{E}[\operatorname{sgn}(\mathbf{W}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top]\boldsymbol{\mu} = \mathbb{E}[\operatorname{sgn}(\overline{\mathbf{W}}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top]\boldsymbol{\mu} = \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}[J_\sigma(\gamma_n(\xi_0, \boldsymbol{\mu}^\perp))] \tag{115}$$

and similarly

$$\begin{aligned}&\mathbb{E}[\operatorname{sgn}(\mathbf{W}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top Y_i\mathbf{X}_i] = \mathbb{E}[\operatorname{sgn}(\overline{\mathbf{W}}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top Y_i\mathbf{X}_i]\\ &= \mathbb{E}\big[\mathbb{E}[\operatorname{sgn}(\overline{\mathbf{W}}_0^\top\mathbf{X}_j)\mathbf{X}_j^\top|Y_i\mathbf{X}_i]Y_i\mathbf{X}_i\big] &(116)\\ &= \mathbb{E}_{\tilde{g}_i,\boldsymbol{\mu}_i^\perp}\big[((1+\tilde{g}_i)\boldsymbol{\mu} + \boldsymbol{\mu}_i^\perp)^\top\mathbb{E}_{\xi_0',\boldsymbol{\mu}'^\perp}[J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\boldsymbol{\mu} + K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\boldsymbol{v}|\tilde{g}_i, \boldsymbol{\mu}_i^\perp]\big] &(117)\\ &= \mathbb{E}_{\tilde{g}_i,\boldsymbol{\mu}_i^\perp}\Big[(1+\tilde{g}_i)\mathbb{E}_{\xi_0',\boldsymbol{\mu}'^\perp}\big[J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))|\tilde{g}_i, \boldsymbol{\mu}_i^\perp\big] + \|\boldsymbol{\mu}_i^\perp\|_2\mathbb{E}_{\xi_0',\boldsymbol{\mu}'^\perp}\big[K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))|\tilde{g}_i, \boldsymbol{\mu}_i^\perp\big]\Big] &(118)\\ &= \mathbb{E}\Big[(1+\tilde{g}_i)J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp)) + \|\boldsymbol{\mu}_i^\perp\|_2K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\Big]. &(119)\end{aligned}$$

By plugging (115) and (119) back to (98), we have

$$\begin{aligned}&\overline{\mathrm{gen}}(P^\alpha_{\mathbf{W}|S_\mathrm{l},\hat{S}_\mathrm{u}}, P_{S_\mathrm{l},\hat{S}_\mathrm{u}})\\ &= \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\bigg(\mathbb{E}\Big[(1+\tilde{g}_i)J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp)) + \|\boldsymbol{\mu}_i^\perp\|_2K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\Big] - \mathbb{E}\big[J_\sigma(\gamma_n(\xi_0, \boldsymbol{\mu}^\perp))\big]\bigg) &(120)\\ &= \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m}\bigg(\underbrace{\mathbb{E}\Big[\tilde{g}_iJ_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp)) + \|\boldsymbol{\mu}_i^\perp\|_2K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\Big]}_{E_n}\bigg), &(121)\end{aligned}$$

where (121) follows since $\mathbb{E}[J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))] = \mathbb{E}[J_\sigma(\gamma_n(\xi_0, \boldsymbol{\mu}^\perp))]$. Since $J_\sigma(x) \in [J_\sigma(-1), J_\sigma(1)]$ and $K_\sigma(x) \in [0, \frac{2\sigma}{\sqrt{2\pi}}]$ for any $x \in [-1,1]$, we have that $E_n = O(d)$.

Remark 2 (Sign of E_n)**.** *First, since $\|\mu_i^\perp\|_2 \geq 0$ and $K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp) \geq 0$, we have*

$$\mathbb{E}\big[\|\boldsymbol{\mu}_i^\perp\|_2 K_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))\big] \geq 0. \tag{122}$$

Next, for some constant $g \geq 0$, by fixing $\xi_0', \boldsymbol{\mu}'^\perp, \boldsymbol{\mu}_i^\perp$, we have

$$\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, g, \boldsymbol{\mu}_i^\perp) \geq \gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, -g, \boldsymbol{\mu}_i^\perp). \tag{123}$$

Since $J_\sigma(x)$ is an odd increasing function on $[-1, 1]$, we have

$$J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, g, \boldsymbol{\mu}_i^\perp)) \geq J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, -g, \boldsymbol{\mu}_i^\perp)) \tag{124}$$

and

$$\mathbb{E}_{\xi_0', \boldsymbol{\mu}'^\perp, \boldsymbol{\mu}_i^\perp}[J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, g, \boldsymbol{\mu}_i^\perp))] \geq \mathbb{E}_{\xi_0', \boldsymbol{\mu}'^\perp, \boldsymbol{\mu}_i^\perp}[J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, -g, \boldsymbol{\mu}_i^\perp))]. \tag{125}$$

Thus, by recalling that $\tilde{g}_i \sim \mathcal{N}(0,1)$, we can deduce that

$$\mathbb{E}[\tilde{g}_i J_\sigma(\gamma_n'(\xi_0', \boldsymbol{\mu}'^\perp, \tilde{g}_i, \boldsymbol{\mu}_i^\perp))] \geq 0. \tag{126}$$

In conclusion, $E_n \geq 0$.

J PROOF OF THEOREM 2

In this case, there exists a unique minimizer of the empirical risk, i.e.,

$$\mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) = \arg\min_{\mathbf{w} \in \mathcal{W}} \bigg(\frac{1}{1+\lambda} L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}) + \frac{\lambda}{1+\lambda} L_{\mathrm{E}}(\mathbf{w}, \hat{S}_{\mathrm{u}})\bigg). \tag{127}$$

According to (Athreya and Hwang, 2010), if the following Hessian matrix

$$H^*(S_{\mathrm{l}}, \hat{S}_u) = \nabla_{\mathbf{w}}^2 \bigg(\frac{1}{1+\lambda} L_{\mathrm{E}}(\mathbf{w}, S_{\mathrm{l}}) + \frac{\lambda}{1+\lambda} L_{\mathrm{E}}(\mathbf{w}, \hat{S}_{\mathrm{u}})\bigg)\bigg|_{\mathbf{w} = \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})} \tag{128}$$

is not singular, then as $\alpha \to \infty$

$$P_{\mathbf{W}|S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}}^\alpha \xrightarrow{\mathrm{d}} \mathcal{N}\bigg(\mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}), \frac{1}{\alpha} H^*(S_{\mathrm{l}}, \hat{S}_u)^{-1}\bigg) \tag{129}$$

and

$$\sqrt{\det\bigg(\frac{\alpha H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})}{2}\bigg)} e^{\alpha \bar{L}_{\mathrm{E}}(\mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}), S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})} \Lambda_{\alpha,\lambda}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \to \sqrt{\pi^d}. \tag{130}$$

Then we have

$$\mathbb{E}_{\mathbf{W}|S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}}[\mathbf{W}] = \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \text{ and } \mathbb{E}_{\mathbf{W}|\hat{S}_{\mathrm{u}}}[\mathbf{W}] = \mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})]. \tag{131}$$

By applying Theorem 1, we use the Gaussian approximation to simplify the symmetrized KL information as follows

$$\begin{aligned}
&I_{\mathrm{SKL}}(\mathbf{W}, \hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) - I_{\mathrm{SKL}}(\hat{S}_{\mathrm{u}}; S_{\mathrm{l}}) \\
&= \mathbb{E}_{\mathbf{W}, \hat{S}_{\mathrm{u}}, S_{\mathrm{l}}}[\log P_{\mathbf{W}|S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}}^\alpha] - \mathbb{E}_{\mathbf{W}, \hat{S}_{\mathrm{u}}} \mathbb{E}_{S_{\mathrm{l}}}[\log P_{\mathbf{W}|S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}}^\alpha] && (132) \\
&= \mathbb{E}_{\mathbf{W}, \hat{S}_{\mathrm{u}}, S_{\mathrm{l}}}\Big[-\frac{\alpha}{2}(\mathbf{W} - \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))^\top H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})(\mathbf{W} - \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))\Big] \\
&\quad - \mathbb{E}_{\mathbf{W}, \hat{S}_{\mathrm{u}}} \mathbb{E}_{S_{\mathrm{l}}}\Big[-\frac{\alpha}{2}(\mathbf{W} - \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))^\top H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}})(\mathbf{W} - \mathbf{W}^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))\Big] \\
&\quad + \mathbb{E}_{S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}}\bigg[\log \frac{\sqrt{\det(\alpha H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))}}{\sqrt{2\pi}}\bigg] - \mathbb{E}_{S_{\mathrm{l}}} \mathbb{E}_{\hat{S}_{\mathrm{u}}}\bigg[\log \frac{\sqrt{\det(\alpha H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}))}}{\sqrt{2\pi}}\bigg] && (133)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] \\
&\quad + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\operatorname{tr}\Big(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^\top + \mathbf{W}\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top - \mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top)\Big)\Big] \\
&\quad - \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\operatorname{tr}\Big(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^\top + \mathbf{W}\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top - \mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top)\Big)\Big] \\
&\quad + \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big] - \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big] \qquad (134) \\
&= \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\operatorname{tr}\Big(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top)\Big)\Big] \\
&\quad - \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\operatorname{tr}\Big(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\big(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]^\top + \mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top \\
&\quad - \mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top)\Big)\Big] \\
&\quad + \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big] - \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big] \qquad (135) \\
&= \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[\frac{\alpha}{2}\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\alpha\Big(\frac{1}{2}\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\Big)^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big] \\
&\quad + \mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big] - \mathbb{E}_{S_{\mathrm{l}}}\mathbb{E}_{\hat{S}_{\mathrm{u}}}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}))}\big]. \qquad (136)
\end{aligned}
$$

From (130), we have

$$
\begin{aligned}
0 &= \mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\Big[\log\sqrt{\det\Big(\frac{\alpha H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}{2}\Big)} + \alpha\bar{L}_{\mathrm{E}}(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})),S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) + \log\Lambda_{\alpha,\lambda}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big] \\
&= \mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})} + \alpha\bar{L}_{\mathrm{E}}(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}),S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) + \log\Lambda_{\alpha,\lambda}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\big], \qquad (137)
\end{aligned}
$$

which means

$$
\mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\big[\log\Lambda_{\alpha,\lambda}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\big] = -\mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\Big[\log\sqrt{\det(H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\Big] - \mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\big[\alpha\bar{L}_{\mathrm{E}}(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}),S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\big]. \qquad (138)
$$

Therefore, by applying Theorem 1, the expected gen-error can be rewritten as

$$
\begin{aligned}
\overline{\mathrm{gen}}(P^{\alpha}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) &= \frac{1+\lambda}{2}\Big(\mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big] \\
&\quad + \mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\Big(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - 2\mathbb{E}_{S_{\mathrm{l}}|\hat{S}_{\mathrm{u}}}[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\Big)^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big] \\
&\quad - \mathbb{E}_{\Delta(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})}\big[\bar{L}_{\mathrm{E}}(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}),S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\big]\Big). \qquad (139)
\end{aligned}
$$

K PROOF OF COROLLARY 1

When S_{l} and $\hat{S}_{\mathrm{u}}$ are independent, we can simplify the asymptotic gen-error as

$$
\overline{\mathrm{gen}}(P^{\infty}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})
$$

$$= \frac{1+\lambda}{2}\Bigg(\mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\mathbf{W}^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}\Big]$$
$$+ 2\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\Big(\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\Big)^\top H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big]\Bigg). \tag{140}$$

Let the MLE optimizer be denoted as $\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) = \mathbf{W}^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})$. Recall the definition

$$\mathbf{w}_\lambda^* = \arg\min_{\mathbf{w}\in\mathcal{W}}\left(\frac{n}{n+m}D\big(P_{\mathbf{Z}}\|p(\cdot|\mathbf{w})\big) + \frac{m}{n+m}D\big(P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_{\mathbf{Z}}}}\|p(\cdot|\mathbf{w})\big)\right) \tag{141}$$

and then we have

$$\mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}}\mathbb{E}_{\hat{\mathbf{Z}}\sim P_{\hat{\mathbf{Z}}|\mathbf{w}^*_{P_{\mathbf{Z}}}}}\left[\nabla_{\mathbf{w}}\Big(-\frac{n}{n+m}\log p(\mathbf{Z}|\mathbf{w}) - \frac{m}{n+m}\log p(\hat{\mathbf{Z}}|\mathbf{w})\Big)\Big|_{\mathbf{w}=\mathbf{w}^*_\lambda}\right] = 0. \tag{142}$$

According to Theorem 2 and (140), the asymptotic gen-error of MLE is given by

$$\overline{\mathrm{gen}}(P^\infty_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})$$
$$= \frac{n+m}{2n}\Bigg(\mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}},S_{\mathrm{l}}}\Big[-\mathbf{W}^\top J(\mathbf{w}^*_\lambda)\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W},\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\mathbf{W}^\top J(\mathbf{w}^*_\lambda)\mathbf{W}\Big]$$
$$+ 2\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)^\top J(\mathbf{w}^*_\lambda)\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})\Big]\Bigg) \tag{143}$$
$$= \frac{n+m}{2n}\Bigg(\mathbb{E}_{\mathbf{W}}\Big[-\mathbf{W}^\top J(\mathbf{w}^*_\lambda)\mathbf{W}\Big] + \mathbb{E}_{\mathbf{W}}\Big[\mathbf{W}^\top J(\mathbf{w}^*_\lambda)\mathbf{W}\Big]$$
$$+ 2\,\mathrm{tr}\Big(\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)^\top J(\mathbf{w}^*_\lambda)\Big]\Big)\Bigg) \tag{144}$$
$$= \frac{n+m}{n}\,\mathrm{tr}\Big(\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)^\top\Big]J(\mathbf{w}^*_\lambda)\Big). \tag{145}$$

Fix any pseudo-labeled data set $\hat{s}_{\mathrm{u}}$ and let

$$\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}}) = \arg\min_{\mathbf{w}\in\mathcal{W}}\left(\frac{n}{n+m}D\big(P_{\mathbf{Z}}\|p(\cdot|\mathbf{w})\big) - \frac{1}{n+m}\sum_{i=n+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\right). \tag{146}$$

Then we have given any ratio $m/n > 0$, as $m\to\infty$,

$$\hat{\mathbf{W}}(\hat{S}_{\mathrm{u}}) \xrightarrow{\mathrm{p}} \mathbf{w}^*_\lambda, \tag{147}$$

and

$$\mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}}\left[\nabla_{\mathbf{w}}\Big(-\frac{n}{n+m}\log p(\mathbf{Z}|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\Big)\Big|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})}\right] = 0. \tag{148}$$

As $n\to\infty$, by central limit theorem, $\mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{s}_{\mathrm{u}})] = \hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})$ (cf. (15) and (146)).

By applying Taylor expansion to $\nabla_{\mathbf{w}}\bar{L}_{\mathrm{E}}(\mathbf{w},S_{\mathrm{l}},\hat{s}_{\mathrm{u}})|_{\mathbf{w}=\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{s}_{\mathrm{u}})}$ around $\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{s}_{\mathrm{u}}) = \hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})$, we have

$$0 = \nabla_{\mathbf{w}}\Big(-\frac{1}{n+m}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\Big)\Big|_{\mathbf{w}=\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{s}_{\mathrm{u}})}$$
$$\approx \nabla_{\mathbf{w}}\Big(-\frac{1}{n+m}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\Big)\Big|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})}$$
$$+ \nabla^2_{\mathbf{w}}\Big(-\frac{1}{n+m}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\Big)\Big|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})}(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{s}_{\mathrm{u}}) - \hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})). \tag{149}$$

By multivariate central limit theorem, as $n \to \infty$, the first term in (149) converges as follows

$$\nabla_{\mathbf{w}}\bigg(-\frac{1}{n+m}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\bigg)\bigg|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})} \xrightarrow{\mathrm{d}} \mathcal{N}\bigg(0, \frac{n}{(n+m)^2}\mathcal{I}_{\mathrm{l}}(\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}}))\bigg). \tag{150}$$

By the law of large numbers, as $n \to \infty$, the second term in (149) converges as follows

$$\begin{aligned}&\nabla^2_{\mathbf{w}}\bigg(-\frac{1}{n+m}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\bigg)\bigg|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})}\\ &\xrightarrow{\mathrm{p}} \frac{n}{n+m}J_{\mathrm{l}}(\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})) - \nabla^2_{\mathbf{w}}\bigg(\frac{1}{n+m}\sum_{i=m+1}^{n+m}\log p(\hat{\mathbf{z}}_i|\mathbf{w})\bigg)\bigg|_{\mathbf{w}=\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})} := \tilde{J}(\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})). \end{aligned}\tag{151}$$

Given any ratio $m/n > 0$, as $n, m \to \infty$, according to (147),

$$\mathcal{I}_{\mathrm{l}}(\hat{\mathbf{W}}(\hat{S}_{\mathrm{u}})) \xrightarrow{\mathrm{p}} \mathcal{I}_{\mathrm{l}}(\mathbf{w}^*_\lambda) \text{ and } \tilde{J}(\hat{\mathbf{W}}(\hat{S}_{\mathrm{u}})) \xrightarrow{\mathrm{p}} J(\mathbf{w}^*_\lambda). \tag{152}$$

Thus, we have

$$\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \xrightarrow{\mathrm{d}} \mathcal{N}\bigg(\mathbf{w}^*_\lambda, \frac{nJ(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}^*_\lambda)J(\mathbf{w}^*_\lambda)^{-1}}{(n+m)^2}\bigg). \tag{153}$$

Finally, the expected gen-error in (145) can be rewritten as

$$\overline{\mathrm{gen}}(P^{\infty}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{n+m}{n}\,\mathrm{tr}\left(\frac{n}{(n+m)^2}J(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}^*_\lambda)\right) \tag{154}$$

$$= \frac{\mathrm{tr}(J(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}^*_\lambda))}{n+m}. \tag{155}$$

L PROOF OF COROLLARY 2

By applying Taylor expansion of $L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})$ around $\mathbf{W} = \mathbf{w}^*_{\mathrm{l}}$, we have the following approximation

$$L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})$$

$$\approx L_{\mathrm{P}}(\mathbf{w}^*_{\mathrm{l}}, P_{S_{\mathrm{l}}}) + (\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})^\top \nabla_{\mathbf{W}} L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})|_{\mathbf{W}=\mathbf{w}^*_{\mathrm{l}}} + \frac{1}{2}(\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})^\top \nabla^2_{\mathbf{W}} L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})|_{\mathbf{W}=\mathbf{w}^*_{\mathrm{l}}}(\mathbf{W}-\mathbf{w}^*_{\mathrm{l}}) \tag{156}$$

$$= L_{\mathrm{P}}(\mathbf{w}^*_{\mathrm{l}}, P_{S_{\mathrm{l}}}) + \frac{1}{2}\,\mathrm{tr}\big((\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})(\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})^\top J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})\big). \tag{157}$$

Thus, the excess risk can be approximated as follows:

$$\mathcal{E}_{\mathrm{r}}(P_{\mathbf{W}}) = \mathbb{E}_{\mathbf{W}}[L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})] - L_{\mathrm{P}}(\mathbf{w}^*_{\mathrm{l}}, P_{S_{\mathrm{l}}})$$

$$\approx \frac{1}{2}\,\mathrm{tr}\big(\mathbb{E}_{\mathbf{W}}[(\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})(\mathbf{W}-\mathbf{w}^*_{\mathrm{l}})^\top]J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})\big) \tag{158}$$

$$= \frac{1}{2}\,\mathrm{tr}\big(\mathbb{E}_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}[(\hat{\mathbf{W}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})-\mathbf{w}^*_{\mathrm{l}})(\hat{\mathbf{W}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})-\mathbf{w}^*_{\mathrm{l}})^\top]J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})\big) + \frac{\mathrm{tr}(J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})\mathbb{E}_{S_{\mathrm{l}},S_{\mathrm{u}}}[\mathrm{Cov}(\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}})])}{2} \tag{159}$$

$$= \frac{1}{2}\,\mathrm{tr}((\mathbf{w}^*_\lambda-\mathbf{w}^*_{\mathrm{l}})(\mathbf{w}^*_\lambda-\mathbf{w}^*_{\mathrm{l}})^\top J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})) + \frac{\mathrm{tr}(J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})\,\mathrm{Cov}(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})))}{2} \tag{160}$$

$$= \frac{1}{2}\,\mathrm{tr}((\mathbf{w}^*_\lambda-\mathbf{w}^*_{\mathrm{l}})(\mathbf{w}^*_\lambda-\mathbf{w}^*_{\mathrm{l}})^\top J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})) + \frac{\mathrm{tr}(J_{\mathrm{l}}(\mathbf{w}^*_{\mathrm{l}})J(\mathbf{w}^*_\lambda)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}^*_\lambda)J(\mathbf{w}^*_\lambda)^{-1})}{2(1+\lambda)(n+m)} \tag{161}$$

where (160) follows since when $\alpha \to \infty$, $\mathrm{Cov}(\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) = \frac{1}{\alpha}H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})^{-1} \to 0$ and from (153).

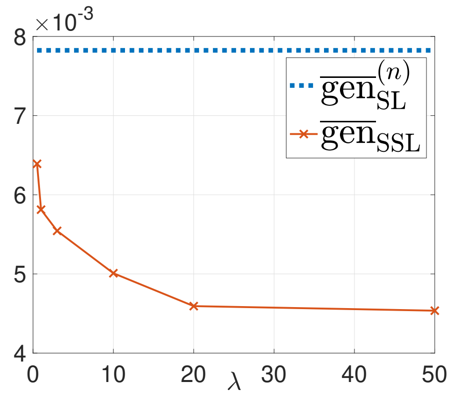


Figure 4: Empirical gen-error of logistic regression on MNIST dataset.

M PROOF OF LOGISTIC REGRESSION EXAMPLE

The first and second derivatives of the loss function are as follows

$$\nabla_{\mathbf{w}} l(\mathbf{w},\mathbf{z}) = \nabla_{\mathbf{w}} \log(1+\exp(-y\mathbf{w}^\top \mathbf{x})) + \nu \mathbf{w} = \frac{-y\mathbf{x}e^{-y\mathbf{w}^\top\mathbf{x}}}{1+e^{-y\mathbf{w}^\top\mathbf{x}}} + \nu\mathbf{w}, \text{ and} \tag{162}$$

$$\nabla^2_{\mathbf{w}} l(\mathbf{w},\mathbf{z}) = \nabla^2_{\mathbf{w}} \log(1+\exp(-y\mathbf{w}^\top \mathbf{x})) + \nu \mathbf{I}_d = \frac{\mathbf{x}\mathbf{x}^\top e^{-y\mathbf{w}^\top\mathbf{x}}}{(1+e^{-y\mathbf{w}^\top\mathbf{x}})^2} + \nu\mathbf{I}_d. \tag{163}$$

The expected Hessian matrices J_{l}, J_{u} and expected product of the first derivative $\mathcal{I}_{\mathrm{l}}$ are given as follows:

$$J_{\mathrm{l}}(\mathbf{w}) = \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}}\left[\frac{\mathbf{X}\mathbf{X}^\top e^{-Y\mathbf{w}^\top\mathbf{X}}}{(1+e^{-Y\mathbf{w}^\top\mathbf{X}})^2}\right] = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\frac{\mathbf{X}\mathbf{X}^\top}{e^{-\mathbf{w}^\top\mathbf{X}}+e^{\mathbf{w}^\top\mathbf{X}}+2}\right], \tag{164}$$

$$J_{\mathrm{u}}(\mathbf{w}) = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\frac{\mathbf{X}\mathbf{X}^\top e^{-\operatorname{sgn}(\mathbf{X}^\top\mathbf{w}_0^*)\mathbf{w}^\top\mathbf{X}}}{(1+e^{-\operatorname{sgn}(\mathbf{X}^\top\mathbf{w}_0^*)\mathbf{w}^\top\mathbf{X}})^2}\right] = \mathbb{E}_{\mathbf{X}\sim P_{\mathbf{X}}}\left[\frac{\mathbf{X}\mathbf{X}^\top}{e^{-\mathbf{w}^\top\mathbf{X}}+e^{\mathbf{w}^\top\mathbf{X}}+2}\right], \tag{165}$$

$$\mathcal{I}_{\mathrm{l}}(\mathbf{w}) = \mathbb{E}_{\mathbf{Z}\sim P_{\mathbf{Z}}}\left[\frac{\mathbf{X}\mathbf{X}^\top e^{-2Y\mathbf{w}^\top\mathbf{X}}}{(1+e^{-Y\mathbf{w}^\top\mathbf{X}})^2}\right]. \tag{166}$$

We can see that $J(\mathbf{w}) = J_{\mathrm{l}}(\mathbf{w}) = J_{\mathrm{u}}(\mathbf{w})$.

Recall the proof of Corollary 1 in Appendix K. In the logistic regression with l_2 regularization, the unique minimizer of the empirical risk in (15) is rewritten as

$$\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \left(-\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{\lambda}{1+\lambda}\frac{1}{m}\sum_{i=n+1}^{n+m}\log p(\hat{\mathbf{Z}}_i|\mathbf{w}) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\right) \tag{167}$$

and the Hessian matrix of the empirical risk at $\mathbf{w} = \hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})$ is rewritten as

$$H^*(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) = \nabla^2_{\mathbf{w}}\left(-\frac{1}{1+\lambda}\frac{1}{n}\sum_{i=1}^{n}\log p(\mathbf{Z}_i|\mathbf{w}) - \frac{\lambda}{1+\lambda}\frac{1}{m}\sum_{i=n+1}^{n+m}\log p(\hat{\mathbf{Z}}_i|\mathbf{w})\right)\Bigg|_{\mathbf{w}=\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})} + \nu\mathbf{I}_d. \tag{168}$$

Recall the definition of $\mathbf{w}^*_\lambda$ with regularization

$$\mathbf{w}^*_\lambda = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}}\left(\frac{n}{n+m}D\big(P_{\mathbf{Z}}\|p(\cdot|\mathbf{w})\big) + \frac{m}{n+m}D\big(P_{\hat{\mathbf{Z}}|\mathbf{w}_0^*}\|p(\cdot|\mathbf{w})\big) + \frac{\nu}{2}\|\mathbf{w}\|_2^2\right). \tag{169}$$

Given any ratio $\lambda > 0$, as $n, m \to \infty$, the Hessian matrix converges as follows

$$H^*(S_{\mathrm{l}}, \hat{S}_{\mathrm{u}}) \xrightarrow{\mathrm{p}} J(\mathbf{w}_\lambda^*) + \nu \mathbf{I}_d. \tag{170}$$

Then the asymptotic expected gen-error in (145) is rewritten as

$$\begin{aligned}&\overline{\mathrm{gen}}(P^{\infty}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}})\\ &= \frac{n+m}{n}\,\mathrm{tr}\left(\mathbb{E}_{\hat{S}_{\mathrm{u}}}\mathbb{E}_{S_{\mathrm{l}}}\Big[\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)\big(\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) - \mathbb{E}_{S_{\mathrm{l}}}[\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}})]\big)^{\top}\Big](J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)\right).\end{aligned} \tag{171}$$

By redefining $\hat{\mathbf{W}}(\hat{s}_{\mathrm{u}})$ in (146) with the l_2 regularization term $\frac{\nu}{2}\|\mathbf{w}\|_2^2$, we can similarly obtain

$$\hat{\mathbf{W}}_{\mathrm{ML}}(S_{\mathrm{l}},\hat{S}_{\mathrm{u}}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{w}_\lambda^*, \frac{n(J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}_\lambda^*)(J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)^{-1}}{(n+m)^2}\right). \tag{172}$$

Then the expected gen-error in (145) can be rewritten as

$$\overline{\mathrm{gen}}(P^{\infty}_{\mathbf{W}|S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}, P_{S_{\mathrm{l}},\hat{S}_{\mathrm{u}}}) = \frac{\mathrm{tr}((J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}_\lambda^*))}{n+m}. \tag{173}$$

Similarly, the excess risk in (19) can be rewritten as

$$\mathcal{E}_{\mathrm{r}}(P_{\mathbf{W}}) = \frac{1}{2}\,\mathrm{tr}((\mathbf{w}_\lambda^* - \mathbf{w}_{\mathrm{l}}^*)(\mathbf{w}_\lambda^* - \mathbf{w}_{\mathrm{l}}^*)^{\top} J_{\mathrm{l}}(\mathbf{w}_{\mathrm{l}}^*)) + \frac{\mathrm{tr}(J_{\mathrm{l}}(\mathbf{w}_{\mathrm{l}}^*)(J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)^{-1}\mathcal{I}_{\mathrm{l}}(\mathbf{w}_\lambda^*)(J(\mathbf{w}_\lambda^*) + \nu\mathbf{I}_d)^{-1})}{2(1+\lambda)(n+m)}, \tag{174}$$

where $\mathbf{w}_{\mathrm{l}}^* = \arg\min_{\mathbf{w}\in\mathcal{W}} L_{\mathrm{P}}(\mathbf{W}, P_{S_{\mathrm{l}}})$.

In addition to the experiments on the synthetic datasets, we implement an logistic regression experiment on "0–1" digit pair in MNIST dataset by setting $n = 200$, $\lambda \in \{0.5, 1, 3, 10, 20, 50\}$ and $\nu = 5$. In Figure 4, we observe that the gen-error decreases as λ increases.