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

Haiyun He  
Gholamali Aminian  
Yuheng Bu  
Miguel Rodrigues  
Vincent Y. F. Tan  
Cornell University  
The Alan Turing Institute  
Univ. of Florida  
Univ. College London  
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 $\lambda$ affects the gen-error under both scenarios. As $\lambda$ 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 $\lambda$ 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 sub-category 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 based on the conditional mutual information. He et al. (2022) investigates the gen-error of iterative SSL techniques based

\begin{footnote}
A function $f(x)$ is called $\beta$-exponentially concave function if $\exp(-\beta f(x))$ is concave
\end{footnote}
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 behavior 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 $S_l = \{ S_l, \}^{n}_{i=1}$ be the labeled training dataset, where $X_i \in \mathcal{X} = \mathbb{R}^d$ is the data feature, $Y_i \in \mathcal{Y} = [K]$ is the class label and each pair of $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} = Z$ is independently and identically distributed (i.i.d.) from $P_X \in \mathcal{P}(\mathcal{X})$. Conditioned on $X_i$, each label $Y_i$ is i.i.d. from $P_{Y|X}$. Let $S_u = \{ X_i \}^{n+m}_{i=n+1}$ be the unlabeled training dataset. For all $i \in [n+m]$, each $X_i$ is i.i.d. from $P_X \in \mathcal{P}(\mathcal{X})$. Based on the labeled dataset $S_l$, a pseudo-labeling method assigns a pseudo-label $\hat{Y}_i$ to each $X_i \in S_u$ and $\hat{Y}_i$ is drawn conditionally i.i.d. from $P_{Y|X|S_l}$.

Let the pseudo-labeled data point be $Z_i = (X_i, \hat{Y}_i)$ and the pseudo-labeled dataset be $\hat{S}_u = \{ Z_i \}^{n+m}_{i=n+1}$. For any labeled and pseudo-label datasets, we use $P_{S_l}$, $P_{S_u}$ and $P_{S_l, \hat{S}_u}$ to denote the joint distributions of the data samples in $S_l, \hat{S}_u$, and $S_l \cup \hat{S}_u$, respectively. Note that $P_{S_l} = P^{\otimes n}_Z$.

Let $\omega \in \mathcal{W}$ 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 $\eta$-weighted sum of the empirical risks of the labeled and pseudo-labeled data [McLachlan 2005; Chapelle et al. 2006].

$$L_E(\omega, S_l, \hat{S}_u) := L_E(\omega, S_l) + \eta L_E(\omega, \hat{S}_u),$$

(1)

where $L_E(\omega, S_l) := \frac{1}{n} \sum_{i=1}^{n} l(\omega, Z_i)$ and $L_E(\omega, \hat{S}_u) := \frac{1}{m} \sum_{i=n+1}^{n+m} l(\omega, \hat{Z}_i)$. Note that, by normalizing the weight $\eta$, minimizing the empirical risk $L_E(\omega, S_l, \hat{S}_u)$ is equivalent to minimizing

$$L_E(\omega, S_l, \hat{S}_u) = \frac{1}{1 + \eta} L_E(\omega, S_l) + \frac{\eta}{1 + \eta} L_E(\omega, \hat{S}_u).$$

(2)

The population risk under the true data distribution is

$$L_P(\omega, S_l) := \mathbb{E}_{S_l \sim P_{S_l}} [L_E(\omega, S_l)].$$

(3)

Under the i.i.d. assumption, such a definition reduces to $\mathbb{E}_{S_l \sim P_{S_l}} [L_E(\omega, S_l)] = \mathbb{E}_{Z \sim P_Z} [l(\omega, Z)]$.

Any SSL algorithm can be characterized by a conditional distribution $P_{W|S_l, \hat{S}_u}$, which is a stochastic map from the labeled dataset $S_l$, the pseudo-labeled dataset $\hat{S}_u$ to the output hypothesis $W$. For any training datasets $S_l, \hat{S}_u$ and any SSL algorithm $P_{W|S_l, \hat{S}_u}$, 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.,

$$\mathbb{E}_{W,S_l}[L_P(W, S_l) - L_E(W, S_l)],$$

(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 [Caton 2007]) to model a pseudo-labeling-based SSL algorithm. Given any $(S_l, \hat{S}_u)$, the $(\alpha, \pi(w), \bar{L}_E(\omega, S_l, \hat{S}_u))$-Gibbs algorithm [Gibbs 1902; Jaynes 1957] is

$$P^\alpha_{W|S_l, \hat{S}_u}(w|S_l, \hat{S}_u) = \frac{\pi(w) \exp \left( - \alpha \bar{L}_E(w, S_l, \hat{S}_u) \right)}{\Lambda_{\alpha, \gamma}(S_l, \hat{S}_u)},$$

where $\Lambda_{\alpha, \gamma}(S_l, \hat{S}_u)$ is the partition function of $\alpha, \pi(w), \bar{L}_E(\omega, S_l, \hat{S}_u)$.
where $\alpha \geq 0$ is the “inverse temperature”, $A_{\alpha, \eta}(S_i, \hat{S}_u) = \int \pi(w) \exp(-\alpha L_E(w, S_i, \hat{S}_u)) \, dw$ is the partition function and $\pi(w)$ is the prior of $w$. We provide more motivations for the Gibbs algorithm model in Appendix [B].

Our goal—relying on the characterization of $P_{\alpha}^{\mathbf{w}|S_i, \hat{S}_u}$—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 $D_{SKL}(P||Q) := D(P||Q) + D(Q||P)$, 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 $Y$ [Ammin et al. 2015] is defined as $I_{SKL}(X;Y) := D_{SKL}(P_{X,Y}||P_X \otimes P_Y) = I(X;Y) + L(X;Y)$. We define the conditional symmetrized KL information as $I_{SKL}(X;Y|Z) := I(X;Y|Z) + L(X;Y|Z)$.

### 2.2 Main Results

One of our main results offers an exact closed-form information-theoretic expression for the gen-error of the $(\alpha, \pi(w), L_E(w, S_i, \hat{S}_u))$-Gibbs algorithm in terms of the symmetrized KL information defined above.

**Theorem 1.** Under the $(\alpha, \pi(w), L_E(w, S_i, \hat{S}_u))$-Gibbs algorithm, the expected gen-error is

\[
\begin{align*}
\text{gen}(P_{\alpha}^{\mathbf{w}|S_i, \hat{S}_u}, P_{S_i, \hat{S}_u}) &= 1 + \frac{\eta}{\alpha} \left( \mathbb{E}_{\Delta_{S_i, \hat{S}_u}} [\log \Lambda_{\alpha, \eta}(S_i, \hat{S}_u)] \right) \\
&+ I_{SKL}(\mathbf{w}, \hat{S}_u|S_i) - I_{SKL}(\hat{S}_u|S_i),
\end{align*}
\]

where $\Lambda_{\alpha, \eta}(S_i, \hat{S}_u) := \mathbb{E}_{\mathbf{w}|S_i} \left[ \exp(-\alpha L_E(\mathbf{w}, S_i, \hat{S}_u)) \right]$ and $\mathbb{E}_{\Delta_{S_i, \hat{S}_u}}[\cdot] := \mathbb{E}_{S_i, \hat{S}_u}[\cdot] - \mathbb{E}_{S_i}\mathbb{E}_{\hat{S}_u}[\cdot]$.

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 [Amin and Gallinar 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 $l(\mathbf{w}, Z)$ is bounded in $[a, b] \subset \mathbb{R}_+$. Then, the expected gen-error of the $(\alpha, \pi(w), L_E(w, S_i, \hat{S}_u))$-Gibbs algorithm satisfies

\[
\begin{align*}
\text{gen}(P_{\alpha}^{\mathbf{w}|S_i, \hat{S}_u}, P_{S_i, \hat{S}_u}) - SKL &\leq c(\eta, a, b) \sqrt{I(\hat{S}_u|S_i)},
\end{align*}
\]

where $c(\eta, a, b) := \frac{1}{\sqrt{2}}(1 + \eta)(b - a)$ and SKL := $1 + \frac{\eta}{\alpha} (I_{SKL}(\mathbf{w}, \hat{S}_u|S_i) - I_{SKL}(\hat{S}_u|S_i))$.

The proof and more discussions are provided in Appendix [E].

From Theorem 1 and (7), we observe that for any $(\eta, \alpha)$, given $\hat{S}_u$, as the dependency (or the information shared) between $\mathbf{w}$ and $S_i$ 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 [Amin et al. 2021a]. However, the difference here is that the quantities are conditioned on the pseudo-labeled data $\hat{S}_u$, which reflects the impact of SSL.

Together with Proposition 1, we observe that the gen-error is also dependent on the information shared between the input labeled and pseudo-labeled data. If the mutual information $I(\hat{S}_u; S_i)$ 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 [B] 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 ($\hat{S}_u$ independent of $S_i$):** If the pseudo-labels $\{Y_1^*\}_{i=1}^{n+m+1}$ are not generated based on the labeled dataset $S_i$, e.g., randomly assigned or generated from another domain (similar to transfer learning), the pseudo-labeled dataset $\hat{S}_u$ is independent of $S_i$. According to the basic properties of mutual and Lautum information [Palomar and Verdú 2008] and (7), we have

\[
I_{SKL}(\mathbf{w}, \hat{S}_u|S_i) - I_{SKL}(\hat{S}_u|S_i) = I_{SKL}(\mathbf{w}; S_i|\hat{S}_u),
\]

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

\[
\text{gen}(P_{\alpha}^{\mathbf{w}|S_i, \hat{S}_u}, P_{S_i, \hat{S}_u}) = \frac{(1 + \lambda)I_{SKL}(\mathbf{w}; S_i|\hat{S}_u)}{\alpha},
\]

which corresponds to the result of transfer learning in [Bu et al. 2022, Theorem 1].
• Case 2 \((S_1 - \hat{S}_u - W)\): When this Markov chain holds, meaning that the output hypothesis \(W\) is independent of \(S_1\) conditioned on \(\hat{S}_u\), we have

\[
I(W; S_1|\hat{S}_u) = 0, \quad P_{W|\hat{S}_u} = P_{W|S_1, \hat{S}_u}
\]

and

\[
D(P_{W|\hat{S}_u} \parallel P_{W|S_1, \hat{S}_u} P_{S_1}|P_{S_1}) = 0.
\]

Thus, the expected gen-error

\[
\text{gen}(P^\alpha_{W|S_1, \hat{S}_u}, P_{S_1, \hat{S}_u}) = \frac{1 + \alpha}{\alpha} \mathbb{E}_{\hat{S}_u} \{ \log \Lambda_{\alpha, \beta}(S_1|\hat{S}_u) \}.
\]

However, this is a degenerate case. For example, it only occurs when \(\eta \to \infty\), i.e., the labeled dataset \(S_1\) might be used for generating pseudo-labels for \(\hat{S}_u\) but not used in the Gibbs algorithm to learn the output hypothesis \(W\). In this case, \(\text{gen}(P^\alpha_{W|S_1, \hat{S}_u}, P_{S_1, \hat{S}_u}) = 0\).

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 \(S_1^{(n+m)} = \{Z_i\}_{i=1}^{n+m}\) contains \(n + m\) samples drawn i.i.d. from \(P_Z\). Then for any output hypothesis \(W_{\text{SL}}^{(n+m)}\) \(\in \mathcal{W}\), the population risk is given by

\[
\mathbb{E}_{Z \sim P_Z} \{l(W_{\text{SL}}^{(n+m)}, Z)\} \quad \text{(cf. (3))},
\]

and the empirical risk of the labeled data samples is given by

\[
L_E(W_{\text{SL}}^{(n+m)}, S_1^{(n+m)}) = \frac{1}{n + m} \sum_{i=1}^{n+m} l(W_{\text{SL}}^{(n+m)}, Z_i).
\]

From Aminian et al. (2021a, Theorem 1), under the \((\alpha, \pi(W_{\text{SL}}^{(n+m)}), L_E(W_{\text{SL}}^{(n+m)}, S_1^{(n+m)}))-\text{Gibbs}\) algorithm, the expected gen-error is

\[
\text{gen}_{\text{SSL}}(n+m) = \mathbb{E}[\mathbb{E}(W_{\text{SL}}^{(n+m)}, S_1^{(n+m)})] = \mathbb{E}[L_E(W_{\text{SL}}^{(n+m)}, S_1^{(n+m)})]
\]

In the SSL setup, in comparison with (2), under the \((\alpha, \pi(w), L_E(w, S_1, \hat{S}_u))-\text{Gibbs}\) algorithm, we let the mixing weight \(\eta\) 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 \(\text{gen}_{\text{all}}\) which is also evaluated over \(n + m\) data points as follows:

\[
\text{gen}_{\text{all}}(P^\alpha_{W|S_1, \hat{S}_u}, P_{S_1, \hat{S}_u}) := \mathbb{E}[L_E(W, P_{S_1}) - L_E(W_1, S_1, \hat{S}_u)].
\]

Consider the situation where the pseudo-labeling method is perfect such that \(P_{Y|x} = P_{Y|x}\) for all \(i \in [n+1:n+m]\).

Then any \(Z_i \in \hat{S}_u\) has the same distribution as that of any \(Z \in S_1\). By applying the same technique in the proof of Theorem 1 we obtain (details provided in Appendix C)

\[
\text{gen}_{\text{all}}(P^\alpha_{W|S_1, \hat{S}_u}, P_{S_1, \hat{S}_u}) = \frac{I_{\text{SKL}}(W; S_1, \hat{S}_u)}{\alpha}.
\]
If we fix $0 < \lambda < \infty$ and assume $\hat{S}_u$ is independent of $S_l$, then $\gamma_{\alpha, \lambda} = 0$ and

$$0 \leq \text{gen}(P_{W|S_l, S_u}^\alpha, P_{S_l, S_u}) \leq \frac{\alpha(b-a)^2}{2(n + m)},$$

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 $\hat{S}_u$ 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 $(X_i, Y_i) \in S_l$ and $i \in [n]$, we assume that $Y_i \in \{-1,+1\}$. $E[X_i|Y_i = 1] = \mu$, $E[X_i|Y_i = -1] = -\mu$, $\|\mu\|_2 = 1$ and $\text{Cov}[Y_iX_i] = \sigma^2 L$. Any $X \in S_u$ is drawn i.i.d. from the same distribution as $X_l$. Consider the problem of learning $\mu$ using $S_l$ and $S_u$. We adopt the mean-squared loss $l(w, z) = ||w - y||_2^2$ and assume the prior $\pi(w)$ is uniform on $W$. Based on $W_0$ learned from the labeled dataset $S_l$ (e.g., $W_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i X_i$), we assign a pseudo-label to each $X_i \in S_u$ as $\hat{Y}_i$ (e.g. $\hat{Y}_i = \text{sgn}(W_0^T X_i)$) and let $\hat{\mu} = E[\hat{Y}_i X_i]$. Let us construct $S_l' = \{Y_i X_i\}_{i=1}^n$ and $\hat{S}_u = \{\hat{Y}_i X_i\}_{i=n+1}^{n+m}$. The empirical risk is given by (cf. 2) by replacing $S_l, S_u$ with $S_l', \hat{S}_u'$

$$\tilde{L}_E(w, S_l', \hat{S}_u') = \frac{1}{(n+1)m} \sum_{i=1}^{n+m} \|Y_i X_i - w\|_2^2 + \frac{\lambda}{(1+\lambda)m} \sum_{j=n+1}^{n+m} \|\hat{Y}_j X_j - w\|_2^2.$$ 

The expected gen-error is equal to the right-hand side of (5) by replacing $S_l, \hat{S}_u$ with $S_l', \hat{S}_u'$. The $\alpha, \pi(W), L_E(W, S_l', \hat{S}_u')$-Gibbs algorithm is given by the following Gibbs posterior distribution

$$P_{W|S_l', \hat{S}_u'}^\alpha(W|S_l', \hat{S}_u') = \mathcal{N}(\mu_{n+m}, \sigma_{l,u}^2 I_d),$$

where $\sigma_{l,u}^2 = \frac{1}{2\lambda}$ and

$$\mu_{n+m} = \frac{1}{(1+\lambda)n} \sum_{i=1}^{n} Y_i X_i + \frac{\lambda}{(1+\lambda)m} \sum_{j=n+1}^{n+m} \hat{Y}_j X_j.$$ 

Details are provided in Appendix [I]. With this posterior Gaussian distribution $P_{W|S_l', \hat{S}_u'}$, the expected gen-error in Theorem [1] can be exactly computed as follows

$$\text{gen}(P_{W|S_l', \hat{S}_u'}^\alpha, P_{S_l, \hat{S}_u}) = \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} \mathbb{E}[(Y_i X_i - \mu) (\hat{Y}_j X_j - \mu')]_{i \in [n], j \in [n+1:n+m]},$$

(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 [3], 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 $j \in [n+1:n+m]$,

$$\mathbb{E}[(Y_i X_i - \mu)' (\hat{Y}_j X_j - \mu')] = tr(Cov[Y_i X_i, \hat{Y}_j X_j]).$$

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 3.2.

3.2 Effect of the Pseudo-labeling and the Ratio of Unlabeled to Labeled Samples $\lambda$

To further 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 $X_i | Y_i \sim \mathcal{N}(\mu, \sigma^2 I_d)$ for any $i \in [n+m]$. Let $S_l^{(n+m)}$ be a dataset with $n + m$ independent copies of $Y X_i \in S_l'$. Similarly to (12), for the supervised $(\alpha, \pi(w_{SL}^{(n)}), L_E(w_{SL}^{(n)}, S_l'))$-Gibbs algorithm and the supervised $(\alpha, \pi(w_{SL}^{(n+m)}), L_E(w_{SL}^{(n+m)}, S_l^{(n+m)}))$-Gibbs algorithm, the expected gen-errors are respectively given by (see details in Appendix [I])

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

Let $\text{gen}_{SSL}$ denote $\text{gen}(P_{W|S_l', \hat{S}_u'}^\alpha, P_{S_l, \hat{S}_u})$ for brevity. By comparing $\text{gen}_{SL}^{(n)}$ and $\text{gen}_{SL}^{(n+m)}$, we observe that:

1. If $\mathbb{E}[(Y_i X_i - \mu)' (\hat{Y}_j X_j - \mu')] < 0$, then $\text{gen}_{SSL} < \text{gen}_{SL}^{(n+m)}$, 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 $\mathbb{E}[(Y_i X_i - \mu)' (\hat{Y}_j X_j - \mu')] > \frac{\sigma^2 d}{n}$, then $\text{gen}_{SSL} > \text{gen}_{SL}^{(n)}$. This implies that if the cross-covariance between

Haiyun He, Gholamali Aminian, Yuheng Bu, Miguel Rodrigues, Vincent Y. F. Tan
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 $0 \leq \mathbb{E}[(Y_i X_i - \mu)^\top (\hat{Y}_j X_j - \mu')] \leq \frac{\sigma^2 d}{n}$, then $\overline{\text{gen}}_{\text{n+m}} \leq \overline{\text{gen}}_{\text{SSL}} \leq \overline{\text{gen}}_{\text{n}}$. This implies that if the cross-covariance is sufficiently small, pseudo-labeling improves the generalization error.

The value of $\mathbb{E}[(Y_i X_i - \mu)^\top (\hat{Y}_j X_j - \mu')]$ also depends on the ratio $\lambda$. Next, to study the effect of $\lambda$, let the initial hypothesis learned from the labeled data $S_0$ be $W_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i X_i \sim \mathcal{N}(\mu, \sigma^2 I_d)$ and the pseudo-label for any $X_i \in S_n$ be $\hat{Y}_i = \text{sgn}(W_0^\top X_i)$.

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

$$\overline{\text{gen}}(P_{\text{n+m}}, W_{S_l}, S_l, P_{S_l}, S_l) = \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} E_n,$$  \hspace{1cm} (13)

where $E_n = \mathbb{E}[\tilde{g} J_{\sigma}(\gamma'_{n}) + \|\mu^+\|^2 K_{\sigma}(\gamma'_n)]$, $\tilde{g} \sim \mathcal{N}(0, 1)$, $\mu^+ \sim \mathcal{N}(0, I_d - \mu \mu^\top)$, $J_{\sigma}$ and $K_{\sigma}$ are functions with domain $[-1, 1]$, and $\gamma'_n \in [-1, 1]$ is a sequence of correlation coefficients. Details are presented in Appendix A2 where we also prove that $E_n = O(d)$ and $E_n \geq 0$, which means that we always have $\overline{\text{gen}}(n+m) \leq \overline{\text{gen}}_{\text{SSL}}$ 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 $\overline{\text{gen}}_{\text{SSL}}$ by varying $\lambda$ and compare it with $\overline{\text{gen}}_{\text{n}}$ and $\overline{\text{gen}}_{\text{n+m}}$ for different values of noise level $\sigma$, and number of labeled data samples $n$. Under different choices of $\sigma$, we observe that $\overline{\text{gen}}_{\text{n+m}} \leq \overline{\text{gen}}_{\text{SSL}} \leq \overline{\text{gen}}_{\text{n}}$. Moreover, the gen-error of SSL monotonically decreases as $\lambda$ increases, showing obvious improvement compared to $\overline{\text{gen}}_{\text{n}}$. However, there exists a gap $\frac{2E_n}{\lambda+2}$ between $\overline{\text{gen}}_{\text{SSL}}$ and $\overline{\text{gen}}_{\text{n+m}}$. 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 $I(S_i; S_l)$) or the cross-covariance term $\text{tr}(\text{Cov}[Y_i X_i, \hat{Y}_j X_j])$ 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 $I(S_i; S_l)$. To make our ideas more concrete, for the mean estimation example that we have been discussing thus far, assume that the pseudo-labeling functions $\hat{Y} = \text{sgn}(W_0^\top X_i)1_{[\text{sgn}(W_0^\top X_i) \geq T]}$ (where $X_i \in S_l$) 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 $\text{tr}(\text{Cov}[Y_i X_i, \hat{Y}_j X_j])$ for $(X_i, Y_i) \in S_l$ and $X_j \in S_n$. From this figure, we observe that $T \geq 7$ approximately minimizes $\text{tr}(\text{Cov}[Y_i X_i, \hat{Y}_j X_j])$ 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.

### 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 \( S_1, \hat{S}_n \), assume that there exists a unique minimizer of the empirical risk \( \bar{L}_E(w, S_1, \hat{S}_n) \) (also known as the single-well case), i.e.,

\[
W^*(S_1, \hat{S}_n) = \arg\min_{w \in \mathcal{W}} \bar{L}_E(w, S_1, \hat{S}_n).
\]

Let the Hessian matrix of the empirical risk \( \bar{L}_E(w, S_1, \hat{S}_n) \) at \( w = W^*(S_1, \hat{S}_n) \) be defined as

\[
H^*(S_1, \hat{S}_n) = \nabla_w^2 \bar{L}_E(w, S_1, \hat{S}_n)|_{w=W^*(S_1, \hat{S}_n)}.
\]

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

**Theorem 2.** Under the \((\alpha, \pi(w), \bar{L}_E(w, S_1, \hat{S}_n))\)-Gibbs algorithm, as \( \alpha \to \infty \), if the Hessian matrix \( H^*(S_1, \hat{S}_n) \) is not singular and \( W^*(S_1, \hat{S}_n) \) is unique, the expected gen-error converges to

\[
\text{gen}(P_{\infty|S_1, \hat{S}_n}, P_{S_1, \hat{S}_n}) = \mathbb{E}_{S_1, \hat{S}_n} \left[ W^*(S_1, \hat{S}_n)^T H^*(S_1, \hat{S}_n) W^*(S_1, \hat{S}_n) \right] 
+ \mathbb{E}_{S_1} \mathbb{E}_{\hat{S}_n} \left[ (W^*(S_1, \hat{S}_n) - 2E_{S_1|\hat{S}_n} [W^*(S_1, \hat{S}_n)])^T \right] 
H^*(S_1, \hat{S}_n) W^*(S_1, \hat{S}_n) 
- \mathbb{E}_{\Delta_{S_1, \hat{S}_n}} \left[ \bar{L}_E(W^*(S_1, \hat{S}_n), S_1, \hat{S}_n) \right] 
- \mathbb{E}_{\Delta_{(W, S_1, \hat{S}_n)}} \left[ \nabla_w^T H^*(S_1, \hat{S}_n) W \right],
\]

where the expectations \( \mathbb{E}_{\Delta_{S_1, \hat{S}_n}}[\cdot] := \mathbb{E}_{S_1, \hat{S}_n}[\cdot] - \mathbb{E}_{S_1} \mathbb{E}_{\hat{S}_n}[\cdot] \) and \( \mathbb{E}_{\Delta_{(W, S_1, \hat{S}_n)}}[\cdot] := \mathbb{E}_{W, S_1, \hat{S}_n}[\cdot] - \mathbb{E}_{W, S_1} \mathbb{E}_{\hat{S}_n}[\cdot] \).

The proof of Theorem 2 is provided in Appendix D. 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 \( S_1 \) and \( \hat{S}_n \). The loss landscape is an important tool to understand the dynamics of the learning process in deep learning. In the extreme case where \( S_1 \) and \( \hat{S}_n \) are independent, \( \mathbb{E}_{\Delta_{S_1, \hat{S}_n}} \left[ \bar{L}_E(W^*(S_1, \hat{S}_n), S_1, \hat{S}_n) \right] = 0 \) 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 \( p(\cdot|w) \), where \( w \in \mathcal{W} \). Consider the negative log-loss function \( l(w, z) = -\log p(z|w) \). For the single-well case, the unique minimizer is

\[
W^*(S_1, \hat{S}_n) = \arg\min_{w \in \mathcal{W}} \left( -\frac{1}{1+\lambda} \cdot \frac{1}{n} \sum_{i=1}^{n} \log p(Z_i|w) \right) - \frac{\lambda}{1+\lambda} \cdot \frac{1}{m} \sum_{i=n+1}^{n+m} \log p(Z_i|w) \right).
\]

Given any labeled dataset \( S_1 \), we use \( P_{Z|S_1} = \mathbb{E}_{W_0|S_1} [P_{Z|W_0}] \) to denote the conditional distribution of the pseudo-labeled data (i.e., \( Z_i \overset{i.i.d.}{\sim} P_{Z|S_1} \)), where \( W_0 \) is the initial hypothesis learned only from \( S_1 \) and used to generate pseudo-labels for \( S_0 \). Assume that \( W_0 \) is learned from \( S_1 \) using MLE, i.e.,

\[
W_0 = \arg\min_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \log p(Z_i|w).
\]

We have \( W_0 \overset{P}{\to} w_{\hat{S}_n}^* = \arg\min_{w \in \mathcal{W}} D(P_Z||p(\cdot|w)) \) as \( n \to \infty \), where \( w_{\hat{S}_n}^* \) depends only on the true distribution of the labeled data \( P_Z \). Then we have \( P_{Z|w_0} \overset{P}{\to} P_{Z|w_{\hat{S}_n}^*} \) and \( P_Z = \mathbb{E}_{W_0}[P_{Z|W_0}] \overset{P}{\to} P_{Z|w_{\hat{S}_n}^*} \), which means \( \{Z_i\}_{i=n+1}^{n+m} \) become independent of one another and of \( S_1 \). We analogously define the minimizer

\[
w_{\lambda}^* = \arg\min_{w \in \mathcal{W}} \left( \frac{n}{n+m} D(P_Z||p(\cdot|w)) \right) + \frac{m}{n+m} D(P_{Z|w_{\hat{S}_n}^*}||p(\cdot|w)) \right)
\]

and the landscapes of the loss function

\[
J_1(w) := \mathbb{E}_{Z \sim P_{Z|w_{\hat{S}_n}^*}} [-\nabla_w^2 \log p(Z|w)]
\]

\[
J_u(w) := \mathbb{E}_{Z \sim P_{Z|w_{\hat{S}_n}^*}} [-\nabla_w \log p(Z|w)]
\]

\[
I_4(w) := \mathbb{E}_{Z \sim P_{Z|w_{\hat{S}_n}^*}} [\nabla_w \log p(Z|w) \nabla_w \log p(Z|w)^T].
\]

Let \( J(w) = \frac{n}{n+m} J_1(w) + \frac{m}{n+m} J_u(w) \). Given any ratio \( \lambda > 0 \), as \( n, m \to \infty \), by the law of large numbers, the Hessian matrix \( H^*(S_1, \hat{S}_n) \) converges as follows

\[
H^*(S_1, \hat{S}_n) \overset{P}{\to} J(w_{\lambda}^*),
\]

which is independent of \( (S_1, \hat{S}_n) \). 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

\[
\text{gen}(P_{\infty|S_1, \hat{S}_n}, P_{S_1, \hat{S}_n}) = \frac{\text{tr}(J(w_{\lambda}^*)) - \text{I}_4(w_{\lambda}^*)}{n+m}.
\]

The proof of Corollary 1 is provided in Appendix K. For the extreme cases where \( \lambda \to 0 \), we have \( w_{\lambda}^* \to
We have 

\[ w^*_L : J(w^*_L) \rightarrow J_i(w^*_L) \quad \text{and} \quad \text{gen}(P^{\infty}_{W|S_i,S_u}, P_{S_i,S_u}) \rightarrow \frac{1}{n} \text{tr}(J_i(w^*_L) - 1)J_i(w^*_L) = O\left(\frac{d}{mn}\right), \]

which means the gen-error degenerates to that of the SL case with \( n \) labeled data.

For the other case where \( \lambda \rightarrow \infty \), from Appendix \[ \text{we have } W_{\text{ML}}(S_i, S_u) \rightarrow \mathbb{E}_{S_i}[W_{\text{ML}}(S_i, S_u)] \text{ and } \text{gen}(P^{\infty}_{W|S_i,S_u}, P_{S_i,S_u}) \rightarrow 0. \]

For \( \lambda \in (0, \infty) \), we have 

\[ \text{gen}(P^{\infty}_{W|S_i,S_u}, P_{S_i,S_u}) = O\left(\frac{d}{n+m}\right), \]

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 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 \( \alpha, n, m \rightarrow \infty \)

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

\[ \mathcal{E}_r(P_{W}) := \mathbb{E}_{W}[L_P(W, P_S)] - L_P(w^*_P, P_S). \]

\[ = \text{gen}(P_{W|S_i,S_u}, P_{S_i,S_u}) + \mathbb{E}_{W,S_i}[L_E(W, S_i) - L_E(w^*_P, S_i)]. \]

where \( w^*_P = \arg\min_{w \in W} L_P(W, P_S) \) is the optimal hypothesis. The second term in \[ (13) \] 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 \( n, m \rightarrow \infty \), the excess risk of SS-MLE is given by

\[ \mathcal{E}_r(P_{W}) = \frac{1}{2} \text{tr}((w^*_L - w^*_P)(w^*_L - w^*_P)^\top J_i(w^*_L)) \]

\[ + \frac{\text{tr}(J_i(w^*_L) - 1)J_i(w^*_L)(J_i(w^*_L)^{-1} - 1)}{2(1 + \lambda)(n+m)}. \]

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

### 4.2 An Application to Logistic Regression

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

\[ l(w, z) = -\log p(y|x, w) + \frac{\nu}{2}||w||^2_2 \]

\[ = \log(1 + e^{-yw^\top x}) + \frac{\nu}{2}||w||^2_2. \]

We assume that there exists a unique minimizer of the empirical risk in this example. We also assume that the initial hypothesis \( W_0 \) is learned from the labeled dataset \( S_i \):

\[ W_0 = \arg\min_{w \in W} \left( \frac{1}{n} \sum_{i=1}^n \log(1 + e^{-y_i w^\top x_i}) + \frac{\nu}{2}||w||^2_2 \right). \]

Consider the case when \( n, m \rightarrow \infty \) and \( \lambda > 0 \). Then

\[ W_0 \xrightarrow{P_S} w^*_0 = \arg\min_{w \in W} \left( D(P_Z || P_X) \frac{P_X}{1 + e^{-yw^\top x}} + \frac{\nu}{2}||w||^2_2 \right). \]

Let the pseudo label for any \( X_i \in S_u \) be defined as \( \hat{Y}_i = \text{sgn}(X_i^\top w_0) \). The conditional distribution of the pseudo-labeled data sample given \( W_0 \) converges as follows

\[ P_{Z|w_0}(z|W_0) \xrightarrow{P_S} \]

\[ P_{Z|w_0}(z|w_0^*) = P_X(x)I\{\hat{y} = \text{sgn}(X^\top w_0^*)\}. \]

Let us rewrite the minimizer \( w^*_0 \) in \[ (16) \] as

\[ w^*_0 = \arg\min_{w \in W} \left( \frac{n}{n+m} D(P_Z || p(\cdot|w)) \right. \]

\[ \left. + \frac{m}{n+m} D(P_Z || p(\cdot|w)) + \frac{\nu}{2}||w||^2_2 \right). \]

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

\[ n \cdot \text{gen}(P^{\infty}_{W|S_i,S_u}, P_{S_i,S_u}) \rightarrow \frac{\text{tr}(J_i(w^*_0) + \nu I_d - 1)J_i(w^*_0))}{1 + \lambda}. \]

Details of the derivation are provided in Appendix \[ \text{We focus on the right-hand side, which depends on the ratio } \lambda \text{ instead of the individual } m, n. \text{ As mentioned in Section 4.1, when } \lambda \rightarrow 0, n \cdot \text{gen}(P^{\infty}_{W|S_i,S_u}, P_{S_i,S_u}) \rightarrow d. \text{ On the other hand, the excess risk } \mathcal{E}_r(P_W) \text{ of this example is given by Corollary 2 where } J_i(w^*_0) \text{ is replaced by } J_i(w^*_0) + \nu I_d. \text{ Intuitively, as the regularization parameter } \nu \text{ increases, the gen-error decreases.} \]

For different values of \( \lambda \), we can numerically calculate the hypothesis \( w^*_0 \), the expected gen-error and the excess risk. Consider the example of a dataset which contains two classes of Gaussian samples. Let \( P_X = \mathcal{N}(Y \mu I_d, I_d) \) and \( P_Y = \text{Unif}((-1, +1)) \), where \( \mu \in \mathbb{R} \) and \( I_d \) is the all-ones vector in \( \mathbb{R}^d \). For notational simplicity, let
Empirical gen-error

We also implement experiments (which the inverse temperature to labeled data. This sheds light in our quest to design good labeled data and the ratio of the number of unlabeled data the information shared between the labeled and pseudo-labeled data, e.g., pseudo-labeling methods, which should penalize the dependence between the labeled and pseudo-labeled data, e.g., $I(S_l; S_u)$, 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.

## 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., $I(S_l; S_u)$, 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


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


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 initial 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, $\alpha$-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 $(\alpha, \pi(W), L_{\xi}(W, S_{l}, S_{u}))$-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:

\[ W_{k+1} = W_k - \beta \nabla L_E(W_k, S_l, \hat{S}_u) + \sqrt{\frac{2 \beta}{\gamma}} \zeta_k, \quad k = 0, 1, \ldots \]

where \( \zeta_k \) 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 \( P_{W_k|S_l, \hat{S}_u} \) induced by SGLD algorithm is close to the \((\gamma, \pi(W_0), \bar{L}_E(W_k, S_l, \hat{S}_u))-\text{Gibbs distribution in } 2\text{-Wasserstein distance for sufficiently large iterations}, k.

\section*{C PROOF OF THEOREM 1}

Under the \((\alpha, \pi(W), \bar{L}_E(W, S_l, \hat{S}_u))-\text{Gibbs algorithm, we have}

\[
I_{\text{SKL}}(W, \hat{S}_u; S_l) = \mathbb{E}_{W, S_l, S_u} \left[ \log P_{W,S_l,S_u}^\alpha + \log P_{S_u|S_l} - \mathbb{E}_{W, S_u} \left[ \log P_{W,S_l,S_u}^\alpha + \log P_{S_u|S_l} \right] \right] = \mathbb{E}_{W, S_u} \left[ -\alpha \bar{L}_E(W, S_l, \hat{S}_u) - \log \Lambda_{\alpha, \eta}(S_l, \hat{S}_u) \right] + \mathbb{E}_{S_u} \left[ \log P_{S_u|S_l} - \mathbb{E}_{S_u} \left[ \log P_{S_u|S_l} \right] \right]
\]

\[
= \frac{1}{1 + \eta} (\mathbb{E}_{W} \left[ L_P(W, S_l) \right] - \mathbb{E}_{W, S_l} \left[ \bar{L}_E(W, S_l) \right] + \mathbb{E}_{S_u} \left[ \log P_{S_u|S_l} \right] - \mathbb{E}_{S_u} \left[ \log P_{S_u|S_l} \right])
\]

\[
= \frac{\alpha}{1 + \eta} \mathbb{E}_{\alpha, \eta} (P_{W|S_l}, P_{S_u|S_l}) + I_{\text{SKL}}(\hat{S}_u; S_l) - \mathbb{E}_{S_u, S_l} \left[ \log \Lambda_{\alpha, \eta}(S_l, \hat{S}_u) \right],
\]

where \( \Lambda_{\alpha, \eta}(S_l, \hat{S}_u) = \int \pi(w) \exp(-\alpha \bar{L}_E(W, S_l, \hat{S}_u)) dw \) and \( \mathbb{E}_{S_u, S_l} \left[ \cdot \right] = \mathbb{E}_{S_u} \left[ \mathbb{E}_{S_l} \left[ \cdot \right] \right] - \mathbb{E}_{S_l} \left[ \mathbb{E}_{S_u} \left[ \cdot \right] \right].

Thus, the expected gen-error is given by

\[
\mathbb{E}(P_{W|S_l}, P_{S_u|S_l}) = \frac{(1 + \eta)(I_{\text{SKL}}(W, \hat{S}_u; S_l) - I_{\text{SKL}}(\hat{S}_u; S_l) + \mathbb{E}_{S_u, S_l} \left[ \log \Lambda_{\alpha, \eta}(S_l, \hat{S}_u) \right])}{\alpha},
\]

\section*{D EXACT CHARACTERIZATION OF EXPECTED GEN-ERROR UNDER ENTROPY MINIMIZATION}

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

\[
L_{\text{EM}}^{\alpha}(W; S_l, \hat{S}_u) = \frac{1}{n + m} \left( -\sum_{i=1}^n \log P_{Y|X,W}(Y_i|X; W) - \sum_{i=n+1}^{n+m} \sum_{k=1}^K P_{Y|X,W}(k|X; W) \log P_{Y|X,W}(k|X; W) \right).
\]

Under the \((\alpha, \pi(W), L_{\text{EM}}^{\alpha}(W; S_l, S_u))-\text{Gibbs algorithm, the posterior distribution of } W \text{ can be denoted as } P_{W|S_l, S_u}^\alpha. \text{ The output hypothesis } W \text{ only depends on the labeled data } S_l \text{ and unlabeled data } S_u \text{ instead of pseudo-labeled data } \hat{S}_u. \text{ Since } S_l \text{ and } S_u \text{ are independent, according to (8), by replacing } \lambda \text{ with } \frac{\alpha}{\eta}, \text{ 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 \((\alpha, \pi(W), L_{\EM}(W; S_1, S_u))\)-Gibbs algorithm), the expected gen-error is

\[
\text{gen}(\hat{P}_W|S_1, S_u, P_{S_1}, S_u) = \frac{(n + m)I_{\SKL}(W; S_1|S_u)}{n \alpha}.
\]

E  PROOF OF PROPOSITION 1

We provide the proof of Proposition 1 and a novel lower bound on \(\text{gen}(P^\alpha_W|S_1, \hat{S}_u, P_{S_1}, S_u)\).

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

Assume that the loss function \(l(W, Z)\) is bounded in \([a, b] \subset \mathbb{R}_+\) for any \(W \in W\) and \(Z \in Z\). Then we have \(\log \Lambda_{\alpha, \eta}(S_1, \hat{S}_u) \in [-\alpha b, -\alpha a]\). According to Hoeffding’s lemma, the loss function \(l(W, Z)\) is \(\frac{b-a}{2}\)-sub-Gaussian and \(\log \Lambda_{\alpha, \eta}(S_1, \hat{S}_u)\) is \(\frac{\alpha(b-a)}{2}\)-sub-Gaussian. From the Donsker–Varadhan representation, we have

\[
|\mathbb{E}_{S_1, S_u}[\log \Lambda_{\alpha, \eta}(S_1, \hat{S}_u)]| \leq \sqrt{\frac{\alpha^2(b-a)^2}{2}} I(\hat{S}_u; S_1),
\]

From Theorem 1, we can directly obtain

\[
\frac{|\text{gen}(P^\alpha_W|S_1, \hat{S}_u, P_{S_1}, S_u) - (1 + \eta)(I_{\SKL}(W, \hat{S}_u; S_1) - I_{\SKL}(\hat{S}_u; S_1))|}{\alpha} \leq \frac{(1 + \eta)(b-a)}{\sqrt{2}} \sqrt{I(\hat{S}_u; S_1)}. \tag{28}
\]

We can also provide another lower bound on \(\text{gen}(P^\alpha_W|S_1, \hat{S}_u, P_{S_1}, S_u)\). 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 \(L(X; Y; Z) = L(X; Y) + D(P_{Z|X,Y}||P_Z|X,Y)P_X P_Y \geq L(X; Y)\). Thus, we have

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

Then the gen-error is lower bounded by

\[
\text{gen}(P^\alpha_W|S_1, \hat{S}_u, P_{S_1}, S_u) \geq \frac{1 + \eta}{\alpha} \mathbb{E}_{S_1, S_u}[\log \Lambda_{\alpha, \eta}(S_1, \hat{S}_u)]. \tag{30}
\]

The sign of the lower bound depends on the loss function, the prior distribution \(\pi(W)\) and the data distributions. As \(\eta \to 0\), or \(\eta \to \infty\), or \(P_{S_1, S_u} = P_{S_1} \otimes P_{S_u}\), the lower bound vanishes to 0.

F  PROOF OF Eq. (7)

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

\[
I_{\SKL}(W, \hat{S}_u; S_1) - I_{\SKL}(\hat{S}_u; S_1)
= I(W, \hat{S}_u; S_1) - I(\hat{S}_u; S_1) + L(W, \hat{S}_u; S_1) - L(\hat{S}_u; S_1)
= I(W; S_1|\hat{S}_u) + L(W, \hat{S}_u; S_1) - L(\hat{S}_u; S_1)
= I(W; S_1|\hat{S}_u) + D(P_W|S_u||P_W|S_1, S_u)\|P_{S_1} P_{S_u}), \tag{31}
\]

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

$$I_{\text{SKL}}(W; S_t, \hat{S}_u) = \mathbb{E}_{W, S_t, \hat{S}_u} [\log P_{W|S_t, \hat{S}_u}^\alpha - \mathbb{E}_{W, S_t, \hat{S}_u} [\log P_{W|S_t, \hat{S}_u}^\alpha] \quad (34)$$

$$= \frac{\alpha}{n + m} \left( \mathbb{E}_{W, S_t, \hat{S}_u} [L_{E}(W, S_t)] - \mathbb{E}_{W, S_t, \hat{S}_u} [L_{E}(W, S_t)] \right) + \frac{\alpha}{n + m} \left( \mathbb{E}_{W, S_t, \hat{S}_u} [L_{E}(W, \hat{S}_u)] - \mathbb{E}_{W, S_t, \hat{S}_u} [L_{E}(W, \hat{S}_u)] \right) \quad (35)$$

$$= \frac{\alpha}{n + m} \sum_{i=1}^{n} \left( \mathbb{E}_{W, Z} [l(W, \tilde{Z})] - \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] \right)$$

$$+ \frac{\alpha}{n + m} \sum_{i=n+1}^{n+m} \left( \mathbb{E}_{W, Z} [l(W, \tilde{Z}')] - \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] \right) \quad (36)$$

$$= \alpha \mathbb{E}_{W, Z} [l(W, \tilde{Z})] - \frac{\alpha}{n + m} \left( \sum_{i=1}^{n} \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] + \sum_{i=n+1}^{n+m} \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] \right) \quad (37)$$

$$= \alpha \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] + \frac{\alpha}{n + m} \mathbb{E}_{W, \tilde{Z}_t} [l(W, \tilde{Z}_t)] \quad (38)$$

where (37) follows since $P_{\tilde{Z}_t} = P_{\tilde{Z}}$. Thus, [10] is proved.

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

Without loss of generality, let $S_1 = Z_1 = (X_1, Y_1)$ and $\hat{S}_u = Z_2 = (X_2, \hat{Y}_2)$. With the assumption that $P_{\tilde{Z}_t} = P_{\tilde{Z}} + \epsilon \Delta$, the joint distributions $P_{W, \tilde{Z}_t}$ and $P_{W, \tilde{Z}_t}$ are given by

$$P_{W, \tilde{Z}_t} (\cdot, \cdot) = P_{\tilde{Z}_t} (\cdot) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) = (P_{\tilde{Z}_t} (\cdot) + \epsilon \Delta (\cdot)) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) \quad (39)$$

$$P_{W, \tilde{Z}_t} (\cdot, \cdot) = P_{\tilde{Z}_t} (\cdot) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) = P_{\tilde{Z}_t} (\cdot) \sum_z (P_{\tilde{Z}_t} (z) + \epsilon \Delta (z)) P_{W|\tilde{Z}_t, z} (\cdot|z, z) \quad (40)$$

$$= P_{\tilde{Z}_t} (\cdot) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) + \epsilon P_{\tilde{Z}_t} (\cdot) \sum_z \Delta (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) \quad (41)$$

$$= P_{\tilde{Z}_t} (\cdot, \cdot) - \epsilon \Delta (\cdot) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) + \epsilon P_{\tilde{Z}_t} (\cdot) \sum_z \Delta (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) \quad (42)$$

$$= P_{\tilde{Z}_t} (\cdot, \cdot) - \epsilon \Delta (\cdot) + \epsilon \Delta (\cdot), \quad (43)$$

where [12] follows since $P_{W|\tilde{Z}_t, z} (\cdot|z, z) = P_{W|\tilde{Z}_t, z} (\cdot|z, z)$, $\Delta (\cdot, \cdot) = -\Delta (\cdot) \sum_z P_{\tilde{Z}_t} (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z) + P_{\tilde{Z}_t} (\cdot) \sum_z \Delta (z) P_{W|\tilde{Z}_t, z} (\cdot|z, z)$ and $\sum_z \Delta (w, z) = 0$.

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

$$\text{gen}(P_{W|S_1, S_2} (z), P_{S_1} (z)) = \frac{I_{\text{SKL}}(W_{\text{SL}}; S_1^{(2)}(z))}{\alpha} = \mathbb{E}_{W, Z} [l(W, Z)] - \mathbb{E}_{W, Z} [l(W, Z_1)]. \quad (44)$$

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

$$\text{gen}_{\text{all}}(P_{W|S_1, \hat{S}_u}, P_{S_1}, P_{\hat{S}_u})$$

$$= \mathbb{E}_{W, Z} [l(W, Z)] - \frac{1}{2} \left( \mathbb{E}_{W, Z} [l(W, Z_1)] + \mathbb{E}_{W, \hat{Z}_2} [l(W, \hat{Z}_2)] \right) \quad (45)$$

$$= \mathbb{E}_{W, Z} [l(W, Z)] - \mathbb{E}_{W, Z} [l(W, Z_1)] - \frac{\epsilon}{2} \mathbb{E}_{(W, \hat{Z}_2)} [l(W, \hat{Z}_2)] \quad (46)$$

$$= \frac{I_{\text{SKL}}(W_{\text{SL}}; S_1^{(2)})}{\alpha} - \frac{\epsilon}{2} \mathbb{E}_{(W, \hat{Z}_2)} [l(W, \hat{Z}_2)] \quad (47)$$
Without loss of generality, we can take \( \tilde{\epsilon} \) proportional to \( \epsilon \) and \( \tilde{\epsilon} \to 0 \) as \( \epsilon \to 0 \). The result can be directly extended to \( S_l \) with \( n \) data samples and \( \hat{S}_u \) with \( m \) data samples.

## Proof of Proposition 3

The formal version of Proposition 2 is stated as follows.

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

\[
\Pr(l(W, Z) \geq \lambda) \leq e^{-\frac{2\gamma_{\alpha,\lambda}}{2}(b-a)^2}.
\]

Since \( L(W, \hat{S}_u; S_l) - L(\hat{S}_u; S_l) = D(P_{lW|\hat{S}_u||W|\hat{S}_u,S_l}|P_{lW|S_l,S_u},P_{S_l,S_u}) \geq 0 \), we always have

\[
I(W, \hat{S}_u; S_l) - I(\hat{S}_u; S_l) \leq I(W, \hat{S}_u; S_l) - I(\hat{S}_u; S_l) + L(W, \hat{S}_u; S_l) - L(\hat{S}_u; S_l) = I_{SKL}(W, \hat{S}_u; S_l) - I_{SKL}(\hat{S}_u; S_l).
\]

Without loss of generality, we can take \( C_\alpha = 0 \). Recall \( \lambda = m/n \). Then we have (i.e., the informal version in Proposition 3)

\[
\Pr(l(W, Z) \geq \lambda) \leq e^{-\frac{2\gamma_{\alpha,\lambda}}{2}(b-a)^2}.
\]

**Proof.** If \( l(W, Z) \) is \( \sigma \)-sub-Gaussian under \( P_Z \) for every \( W \in W \), from Xu and Raginsky [2017] Theorem 1), we have

\[
\Pr(l(W, Z) \geq \lambda) \leq e^{-\frac{2\gamma_{\alpha,\lambda}}{2}(b-a)^2}.
\]

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

\[
(1 + \lambda)(I_{SKL}(W, \hat{S}_u; S_l) - I_{SKL}(\hat{S}_u; S_l)) \leq \frac{b-a}{2}(I(W, \hat{S}_u; S_l) - I(\hat{S}_u; S_l)) \leq \frac{(b-a)^2}{2n}.
\]

Furthermore, by Donsker–Varadhan representation, \( -\E_{\Delta_{S_l,\hat{S}_u}}[\log \Lambda_{\alpha,\lambda}(S_l, \hat{S}_u)] \) can be upper bounded as follows

\[
-\E_{\Delta_{S_l,\hat{S}_u}}[\log \Lambda_{\alpha,\lambda}(S_l, \hat{S}_u)] = E_{\hat{S}_u}[\log \Lambda_{\alpha,\lambda}(S_l, \hat{S}_u)] - E_{S_l}[\log \Lambda_{\alpha,\lambda}(S_l, \hat{S}_u)] \leq \frac{\alpha(b-a)^2}{2}.
\]

Then (55) can be further upper bounded as

\[
\frac{1}{\alpha}(I_{SKL}(W, \hat{S}_u; S_l) - I_{SKL}(\hat{S}_u; S_l)) \leq \frac{b-a}{2}(I(W, \hat{S}_u; S_l) - I(\hat{S}_u; S_l)) + \frac{(b-a)^2}{2n}.
\]
Suppose \( I(W, \hat{S}_u; S_l) - I(I(W, \hat{S}_u; S_l)) + C_\alpha \leq I_{\text{SKL}}(W, \hat{S}_u; S_l) - I_{\text{SKL}}(\hat{S}_u; S_l) \) for some constant \( C_\alpha \geq 0 \) and \( I(\hat{S}_u; S_l) = \gamma_{\alpha, \lambda} I(W, \hat{S}_u; S_l) \), where \( \gamma_{\alpha, \lambda} \) depends on \( \lambda \) and \( 0 \leq \gamma_{\alpha, \lambda} \leq 1 \). Then we have

\[
(1 + \lambda)(1 + C_\alpha) I(W, \hat{S}_u; S_l) - I(\hat{S}_u; S_l) \leq \sqrt{\frac{(b-a)^2}{2n} I(W, \hat{S}_u; S_l)} + \sqrt{\frac{(1 + \lambda)(b-a)^2}{2} I(\hat{S}_u; S_l)}
\]

(58)

\[
(1 + \lambda)(1 + C_\alpha)(1 - \gamma_{\alpha, \lambda}) I(W, \hat{S}_u; S_l) \leq \sqrt{\frac{(b-a)^2}{2n} I(W, \hat{S}_u; S_l)} + \sqrt{\frac{(1 + \lambda)(b-a)^2}{2} \gamma_{\alpha, \lambda} I(\hat{S}_u; S_l)}
\]

(59)

\[
\sqrt{I(W, \hat{S}_u; S_l)} \leq \frac{\alpha}{(1 + \lambda)(1 + C_\alpha)(1 - \gamma_{\alpha, \lambda})} \left( \sqrt{\frac{(b-a)^2}{2n}} + \frac{1 + \lambda}{\sqrt{2}} \sqrt{\gamma_{\alpha, \lambda}} \right).
\]

(60)

Thus, we have

\[
\overline{\alpha}(P_\alpha^{\beta} | W, S_\alpha, S_u \rightarrow P_{S_\alpha, S_u} \rightarrow P_{S_\alpha, S_u}) \leq \frac{\alpha(b-a)^2}{2\sqrt{n}(1 + C_\alpha)(1 - \gamma_{\alpha, \lambda})} + \frac{\alpha(b-a)^2}{2n(1 + \lambda)(1 + C_\alpha)(1 - \gamma_{\alpha, \lambda})}.
\]

(61)

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

\[
\overline{\alpha}(P_\alpha^{\beta} | W, S_\alpha, S_u \rightarrow P_{S_\alpha, S_u} \rightarrow P_{S_\alpha, S_u}) \geq -\frac{(1 + \lambda)(b-a)}{\sqrt{2}} \sqrt{I(\hat{S}_u; S_l)}
\]

(62)

\[
= \frac{(1 + \lambda)(b-a)}{\sqrt{2}} \sqrt{\gamma_{\alpha, \lambda} I(W, \hat{S}_u; S_l)}
\]

(63)

\[
\geq -\frac{\alpha(b-a)^2}{2(1 + C_\alpha)(1 - \gamma_{\alpha, \lambda})} \left( \sqrt{\frac{1}{n} + (1 + \lambda) \sqrt{\gamma_{\alpha, \lambda}}} \right).
\]

(64)

\[
\square
\]

# I PROOFS OF MEAN ESTIMATION EXAMPLE

The \((\alpha, \pi(W), \hat{L}_E(W, S_\alpha', \hat{S}_u'))\)-Gibbs algorithm is given by the following Gibbs posterior distribution

\[
P_\alpha^{\beta}(W | Y_i X_i i=1, \hat{(Y_i X_i)^T i=n+1}) = \frac{\pi(W)}{\Lambda_{\alpha, \lambda}(S_\alpha', \hat{S}_u')} \exp \left[ -\frac{\alpha}{(1 + \lambda)n} \sum_{i=1}^{n} \left( W^T W - 2W^T Y_i X_i + X_i^T X_i \right) \right.
\]

\[
- \left. \frac{\alpha \lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} \left( W^T W - 2W^T \hat{Y}_i X_i + \hat{X}_i^T X_i \right) \right]
\]

(65)

\[
= \frac{1}{\sqrt{2\pi \sigma_{1,u}^2}} \exp \left( -\frac{1}{2\sigma_{1,u}^2} \| W - \mu_{n,m} \|^2_2 \right),
\]

(66)

where \( \sigma_{1,u}^2 = \frac{1}{\sigma^2} \).

\[
\mu_{n,m} = \frac{1}{(1 + \lambda)n} \sum_{i=1}^{n} Y_i X_i + \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} \hat{Y}_i X_i,
\]

(67)

and

\[
\frac{\pi(W)}{\Lambda_{\alpha, \lambda}(S_\alpha', \hat{S}_u')} = \frac{1}{\sqrt{2\pi \sigma_{1,u}^2}} \exp \left( -\frac{\alpha}{(1 + \lambda)n} \sum_{i=1}^{n} \mu_{n,m}^T X_i - \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} \hat{X}_i^T X_i \right)
\]

(68)

\[
= \frac{1}{\sqrt{2\pi \sigma_{1,u}^2}} \exp \left( -\frac{1}{(1 + \lambda)^2 n^2} \sum_{i,j \in [n]^2} X_i^T X_j + \frac{\lambda^2}{(1 + \lambda)^2 m^2} \sum_{i,j \in [n+1:n+m]^2} X_i^T X_j \right),
\]

(69)
where \((a)\) follows since it can be seen that \(P_{W|S_i',\hat{S}_u}^\alpha\) is a Gaussian distribution. Thus, the output hypothesis \(W\) can be written as

\[
W = \frac{1}{(1 + \lambda)n} \sum_{i=1}^{n} (Y_i X_i) + \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} \hat{Y}_i X_i + N
\]

(69)

\[
= \frac{1}{(1 + \lambda)\lambda} \sum_{i=1}^{n} (Y_i X_i - \mu) + \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} (\hat{Y}_i X_i - \mu') + \frac{\mu + \lambda \mu'}{1 + \lambda} + N
\]

(70)

where \(N \sim N(0, \sigma^2 u_1 I_d)\) is independent of \((S_i, \hat{S}_u)\), \(T(S'_i) = [Y_1 X_1 - \mu; \ldots; Y_1 X_1 - \mu] \in \mathbb{R}^{nd \times 1}\), \(A = \frac{1}{(1 + \lambda)n} I_d, \ldots, \frac{1}{(1 + \lambda)n} I_d \in \mathbb{R}^{nd \times nd}\), \(N_G|\hat{S}_u \sim N(\mu_{NG}, \sigma^2 u_1 I_d)\) and \(\mu_{NG} = \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} (\hat{Y}_i X_i - \mu') + \frac{\mu + \lambda \mu'}{1 + \lambda}\).

First, let us calculate the expected gen-error according to [5] in Theorem [1]. Let \(T = T(S'_i)\) for simplicity. We have

\[
I_{SKL}(W, \hat{S}_u; S_i) - I_{SKL}(\hat{S}_u; S_i)
\]

\[
= I(W; S_i|\hat{S}_u) + D(P_W||P_{\hat{W}}) - I_{SKL}(\hat{S}_u; S_i)
\]

\[
= \frac{1}{2\sigma^2_{1u}} \left( E_{S_i,\hat{S}_u} E_{W|S_i,\hat{S}_u} [-(W - \mu_{nm})^T (W - \mu_{nm})] + E_{S_i} E_{\hat{S}_u} E_{W|\hat{S}_u} [(W - \mu_{nm}, m)^T (W - \mu_{nm}, m)] \right)
\]

\[
= \frac{1}{2\sigma^2_{1u}} \left( E_{S_i,\hat{S}_u} [AT]^T (AT) + 2(AT)^T \mu_{NG} + \mu_{NG}^T \mu_{NG} \right)
\]

\[
+ E_{S_i} E_{\hat{S}_u} [AT]^T (AT + \mu_{NG} - 2\mu_{NG} (AT + \mu_{NG} + (AT + \mu_{NG})^T (AT + \mu_{NG}))]
\]

\[
\equiv \frac{1}{\sigma^2_{1u}} E_{S_i,\hat{S}_u} [AT]^T (AT)
\]

\[
= \frac{1}{\sigma^2_{1u}} tr(A^T \Sigma_{AT} T T^T)
\]

\[
= \frac{2\lambda \alpha^2 d}{(1 + \lambda)^2 n},
\]

(80)

where \((a)\) follows since \(E_{S'_i}[AT(S'_i)] = 0\), and

\[
E_{S'_i,\hat{S}_u} [\log \Lambda_{\alpha,\lambda}(S'_i,\hat{S}_u)]
\]

\[
= \alpha E_{S'_i,\hat{S}_u} \left[ \frac{1}{(1 + \lambda)^2 n^2} \sum_{i,j \in [n]} X_i^T X_j + \frac{\lambda^2}{(1 + \lambda)^2 m^2} \sum_{i,j \in [n+1:n+m]} X_i^T X_j \right]
\]

\[
+ \frac{2\lambda}{(1 + \lambda)^2 n m} \sum_{i,j \in [n]} \sum_{n+1}^{n+m} (Y_i X_i)^T (\hat{Y}_j X_j) - \frac{1}{(1 + \lambda)n} \sum_{i=1}^{n} X_i^T X_i - \frac{\lambda}{(1 + \lambda)m} \sum_{i=n+1}^{n+m} X_i^T X_i
\]

\[
= \alpha E_{S'_i,\hat{S}_u} \left[ \frac{2\lambda}{(1 + \lambda)^2 n m} \sum_{i,j \in [n]} \sum_{n+1}^{n+m} (Y_i X_i)^T (\hat{Y}_j X_j) \right]
\]

\[
= \frac{2\alpha \lambda}{(1 + \lambda)^2 n m} \sum_{i,j \in [n]} \sum_{n+1}^{n+m} (E_{S'_i,\hat{S}_u} [(Y_i X_i - \mu)^T (\hat{Y}_j X_j - \mu^T)] + \mu^T \mu - \mu^T \mu)\]

(83)
\[
\frac{2\alpha}{(1+\lambda)^2}\mathbb{E}[(Y_i^\top X_i - \mu_1^\top (\hat{Y}_j^\top X_j - \mu_1')],
\]

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

\[
\text{gen}(P^\alpha_{W[SL,S_l]}, P_{S_l}) = \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m} \mathbb{E}[(Y_i^\top X_i - \mu)^\top (\hat{Y}_j^\top X_j - \mu')].
\]

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

\[
\begin{align*}
\text{gen}(P^\alpha_{W[SL,S_l]}, P_{S_l}) &= \mathbb{E}_{W,Y} \mathbb{E}_{\hat{Y} \hat{X}} [\|\hat{Y} \hat{X} - W\|^2] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W,Y,X_i} [\|Y_i^\top X_i - W\|^2] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[2W^\top (Y_i^\top X_i - \hat{Y} \hat{X}) - X_i^\top X_i + \hat{X}^\top \hat{X}] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[2(\mu_{n,m} + N)^\top (Y_i^\top X_i - \hat{Y} \hat{X})] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[2\mu_{n,m}^\top (Y_i^\top X_i - \hat{Y} \hat{X})] \\
&= \frac{1}{n} \sum_{i=1}^n 2\mathbb{E}\left[\left(\frac{1}{n+m} \sum_{j=1}^n (Y_j^\top X_j - \mu) + \frac{1}{n+m} \sum_{j=n+1}^{n+m} (\hat{Y}_j^\top X_j - \mu') + \frac{\mu + \lambda \mu'}{1+\lambda}\right)^\top (Y_i^\top X_i - \mu) - (\hat{Y} \hat{X} - \mu)\right] \\
&= \frac{1}{n} \sum_{i=1}^n \left(\frac{2}{n+m} \mathbb{E}[(Y_i^\top X_i - \mu)^\top (Y_i^\top X_i - \mu)] + \frac{2}{n+m} \sum_{j=n+1}^{n+m} \mathbb{E}[(\hat{Y}_j^\top X_j - \mu')^\top (Y_i^\top X_i - \mu)]\right) \\
&= \frac{2\sigma^2 d}{n+m} + \frac{1}{n} \sum_{i=1}^n \frac{2}{n+m} \sum_{j=n+1}^{n+m} \mathbb{E}[(\hat{Y}_j^\top X_j - \mu')^\top (Y_i^\top X_i - \mu)] \\
&= \frac{2\sigma^2 d}{n+m} + \frac{2m}{n+m} \mathbb{E}[(\hat{Y}_j^\top X_j - \mu')^\top (Y_i^\top X_i - \mu)],
\end{align*}
\]

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

### I.1 Mean Estimation under Supervised Learning

Under the supervised \((\alpha, \pi(W_{SL}^{(n)}), L_E(W_{SL}^{(n)}, S_1))\)-Gibbs algorithm, the posterior distribution \(P_{W_{SL}^{(n)}|S_1}\) is given by

\[
P_{W_{SL}^{(n)}|S_1} = \mathcal{N}\left(\frac{1}{n} \sum_{i=1}^n Y_i^\top X_i, \sigma^2_{1,n} I_d\right).
\]

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

\[
\text{gen}(P^\alpha_{W_{SL}^{(n)}|S_1}, P_{S_1}) = \frac{I_{SKL}(W_{SL}^{(n)}, S_1)}{\alpha} = 2 \text{tr}(A_1 \mathbb{E}[(T(S_1^\top T(S_1^\top)^\top A_1)] = \frac{2\sigma^2 d}{n}.
\]

where \(T(S_1) = [Y_i^\top X_i - \mu; \ldots; Y_i^\top X_i - \mu] \in \mathbb{R}^{n \times d\times 1}\) and \(A_1 = [\frac{1}{n} I_d, \ldots, \frac{1}{n} I_d] \in \mathbb{R}^{d \times nd}\).

Similarly, under the supervised \((\alpha, \pi(W_{SL}^{(n+m)}), L_E(W_{SL}^{(n+m)}, S_1^{(n+m)}))\)-Gibbs algorithm, where \(S_1^{(n+m)}\) contains \(n + m\) i.i.d. \(Y_i^\top X_i\) samples and \(L_E(W_{SL}^{(n+m)}, S_1^{(n+m)}) = \frac{1}{n+m} \sum_{i=1}^{n+m} l(W_{SL}^{(n+m)}; Z_i)\), the expected gen-error (cf. (9)) is given
Let us rewrite the gen-error in (12) as follows: for any $i \in [n]$ and $j \in [n + 1 : n + m]$.

\[
\gen(P^n_{W|S_i, S_j}, P^\gamma_{S_i, S_j}) = \frac{\gamma}{n + m} \left( \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} \mathbb{E}[\langle Y_i X_i - \mu \rangle^T (\hat{Y}_j X_j - \mu')] \right) 
\]

\[
= \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} \mathbb{E}[\langle Y_i X_i - \mu \rangle^T (\text{sgn}(W_0^T X_j) X_j - \mu')] 
\]

\[
= \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} (\mathbb{E}[\langle \text{sgn}(W_0^T X_j) X_j^T Y_i X_i \rangle] - \mathbb{E}[\langle \text{sgn}(W_0^T X_j) X_j^T \rangle] \mu) 
\]

(96)

1.2 Proofs for Eq. (13)

Let $W_0 = \frac{1}{n} \sum_{i=1}^{n} Y_i X_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{n} I_d)$. Using the proof idea from [He et al., 2022], we can decompose it as

\[
W_0 = \mu + \frac{\sigma}{\sqrt{n}} \xi = \left(1 + \frac{\sigma}{\sqrt{n}} \xi_0\right) \mu + \frac{\sigma}{\sqrt{n}} \mu^\perp, 
\]

(99)

where $\xi \sim \mathcal{N}(0, I_d)$, $\xi_0 \sim \mathcal{N}(0, 1)$, $\mu^\perp \sim \mathcal{N}(0, I_d - \mu \mu^T)$ and $\mu^\perp$ is perpendicular to $\mu$ and independent of $\xi_0$. The normalized $W_0$ can be written as

\[
\bar{W}_0 = \frac{W_0}{\|W_0\|_2} = \gamma_n \mu + \bar{\gamma}_n \nu
\]

(100)

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

\[
\gamma_n = \gamma_n(\xi_0, \mu^\perp) := \frac{1 + \frac{\sigma}{\sqrt{n}} \xi_0}{\sqrt{(1 + \frac{\sigma}{\sqrt{n}} \xi_0)^2 + \frac{\sigma^2}{n} \|\mu^\perp\|_2^2}}. 
\]

(101)

For any $i \in [n + m]$, since $Y_i X_i \sim \mathcal{N}(\mu, \sigma^2 I_d)$, we have

\[
Y_i X_i = \mu + \sigma g_i = \mu + \tilde{g}_i \mu + \mu^\perp_i, 
\]

(102)

where $g_i \sim \mathcal{N}(0, I_d)$, $\tilde{g}_i \sim \mathcal{N}(0, 1)$, $\mu^\perp_i \sim \mathcal{N}(0, I_d - \mu \mu^T)$ and $\tilde{g}_i$ is independent of $\mu^\perp_i$. Given any $Y_i X_i$ for $i \in [1 : n]$, we have

\[
W_0 | Y_i X_i = \frac{1}{n} Y_i X_i + \frac{n-1}{n} \mu + \sqrt{n-1} \frac{\sigma \xi'}{n} 
\]

\[
= \frac{1}{n} (\mu + \sigma g_i) + \frac{n-1}{n} \mu + \sqrt{n-1} \frac{\sigma (\xi'_0 \mu + \mu^\perp)}{n} 
\]

\[
= \left(1 + \frac{\sqrt{n-1}}{n} \sigma \xi'_0 + \frac{\sigma}{n} \tilde{g}_i\right) \mu + \left(\frac{\sqrt{n-1}}{n} \sigma \|\mu^\perp\|_2 + \frac{\sigma}{n} \|\mu^\perp\|_2\right) \nu, 
\]

(103)

where $\xi'_0 \sim \mathcal{N}(0, I_d)$, $\xi'_0 \sim \mathcal{N}(0, 1)$, $\mu^\perp \sim \mathcal{N}(0, I_d - \mu \mu^T)$ and $\mu^\perp$ is perpendicular to $\mu$ and independent of $\xi'_0$. The normalized version is given by

\[
\bar{W}_0 | Y_i X_i = \frac{W_0}{\|W_0\|_2} | Y_i X_i = \gamma'_n \mu + \bar{\gamma}'_n \nu 
\]

(104)

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

\[
\gamma'_n = \gamma'_n(\xi'_0, \mu^\perp, \tilde{g}_i, \mu^\perp_i) := \frac{1 + \sqrt{n-1} \sigma \xi'_0 + \bar{\xi}_0 \tilde{g}_i}{\sqrt{(1 + \frac{\sqrt{n-1}}{n} \sigma \xi'_0 + \frac{\sigma}{n} \tilde{g}_i)^2 + (\frac{\sqrt{n-1}}{n} \sigma \|\mu^\perp\|_2 + \frac{\sigma}{n} \|\mu^\perp\|_2)^2}}. 
\]

(105)
Define the correlation evolution function $F_\sigma : [-1, 1] \to [-1, 1]$:

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

(108)

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

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

$$g_j = \tilde{g}_j \mathbf{W}_0 + \bar{g}_j^\perp,$$

(109)

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

$$X_j = Y_j \mu + \sigma \tilde{g}_j \mathbf{W}_0 + \sigma \bar{g}_j^\perp,$$

(110)

and

$$\mathbf{W}_0^T X_j = Y_j \gamma_n + \sigma \tilde{g}_j.$$  

(111)

Then we have

$$E[\text{sgn}(\mathbf{W}_0^T X_j) X_j \mid \xi_0, \mu^\perp, Y_j = -1] = -E[\text{sgn}(\gamma_n + \sigma \tilde{g}_j) \xi_0, \mu^\perp] \mu + \sigma E[\text{sgn}(\gamma_n + \sigma \tilde{g}_j) \tilde{g}_j | \xi_0, \mu^\perp] \mathbf{W}_0,$$

(112)

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

$$E[\text{sgn}(\mathbf{W}_0^T X_j) X_j' \mid \xi_0, \mu^\perp, Y_j' = 1] = \left(1 - 2Q\left(\frac{\gamma_n}{\sigma}\right)\right) \mu + \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma_n^2}{2\sigma^2}\right) \mathbf{W}_0,$$

(113)

and similarly,

$$E[\text{sgn}(\mathbf{W}_0^T X_j') X_j' \mid \xi_0, \mu^\perp, Y_j' = 1] = \left(2Q\left(\frac{\gamma_n}{\sigma}\right) - 1\right) \mu + \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma_n^2}{2\sigma^2}\right) \mathbf{W}_0.$$

(114)

Recall the definitions of $J_\sigma$ and $K_\sigma$. Then we have

$$E[\text{sgn}(\mathbf{W}_0^T X_j) X_j') \mu = E[\text{sgn}(\mathbf{W}_0^T X_j) X_j') \mu = E[\gamma_n (\xi_0, \mu^\perp)]$$

(115)

and similarly

$$E[\text{sgn}(\mathbf{W}_0^T X_j) X_j') Y_i X_i] = E[\text{sgn}(\mathbf{W}_0^T X_j) X_j') Y_i X_i]$$

(116)

$$= E[\text{sgn}(\mathbf{W}_0^T X_j) X_j') Y_i X_i]$$

$$= E[\gamma_n (\xi_0, \mu^\perp), \xi_0, \mu^\perp)] \mathbf{W}_0^T \mathbf{W}_0 E[\gamma_n (\xi_0, \mu^\perp), \xi_0, \mu^\perp)]$$

(117)

$$= E[\gamma_n (\xi_0, \mu^\perp), \xi_0, \mu^\perp)] \mathbf{W}_0^T \mathbf{W}_0$$

(118)

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

$$\begin{align*}
\mathbb{E}[\min(P^\omega_{\mathbf{W}} | s_i, s_u, P^\omega_{s_i, s_u}) &= \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} \left(E[(1 + \tilde{g}_i) J_\sigma(\gamma_n (\xi_0, \mu^\perp, \tilde{g}_i, \tilde{g}_i)] + \mu^\perp_1 ||2 K_\sigma(\gamma_n (\xi_0, \mu^\perp, \tilde{g}_i, \tilde{g}_i)]\right) \\
&= \frac{2\sigma^2 d}{n + m} + \frac{2m}{n + m} \left(E[(1 + \tilde{g}_i) J_\sigma(\gamma_n (\xi_0, \mu^\perp, \tilde{g}_i, \tilde{g}_i)] + \mu^\perp_1 ||2 K_\sigma(\gamma_n (\xi_0, \mu^\perp, \tilde{g}_i, \tilde{g}_i)]\right),
\end{align*}$$

(121)

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

Haiyun He, Gholamali Aminian, Yuheng Bu, Miguel Rodrigues, Vincent Y. F. Tan
Thus, by recalling that \( \tilde{g}_i \sim \mathcal{N}(0, 1) \), we have
\[
E[\tilde{g}_i J_\sigma(\gamma_n(\xi_n, \mu_{\perp}^l, g, \mu^l_i))] \geq 0.
\] (126)

In conclusion, \( E_n \geq 0 \).

**J PROOF OF THEOREM**

In this case, there exists a unique minimizer of the empirical risk, i.e.,
\[
W^*(S_1, \hat{S}_u) = \arg\min_{w \in \mathcal{W}} \left( \frac{1}{1 + \lambda} L_E(w, S_1) + \frac{\lambda}{1 + \lambda} L_E(w, \hat{S}_u) \right).
\] (127)

According to [Athreya and Hwang, 2010], if the following Hessian matrix
\[
H^*(S_1, \hat{S}_u) = \nabla^2_w \left( \frac{1}{1 + \lambda} L_E(w, S_1) + \frac{\lambda}{1 + \lambda} L_E(w, \hat{S}_u) \right)
\] (128)
is not singular, then as \( \alpha \to \infty \)
\[
P^\alpha_{W|S_1, \hat{S}_u} \overset{d}{\to} \mathcal{N} \left( W^*(S_1, \hat{S}_u), \frac{1}{\alpha} H^*(S_1, \hat{S}_u)^{-1} \right)
\] (129)

and
\[
\sqrt{\det \left( \frac{\alpha H^*(S_1, \hat{S}_u)}{2} \right)} e^{\alpha L_E(W^*(S_1, \hat{S}_u), S_1, \hat{S}_u)} \Lambda_{\alpha}(S_1, \hat{S}_u) \to \sqrt{\pi^d}.
\] (130)

Then we have
\[
E_{W|S_1, \hat{S}_u}[W] = W^*(S_1, \hat{S}_u) \text{ and } E_{W|S_1}[W] = E_{S_1|\hat{S}_u}[W^*(S_1, \hat{S}_u)],
\] (131)

By applying Theorem\[1\] we use the Gaussian approximation to simplify the symmetrized KL information as follows
\[
I_{SKL}(W, \hat{S}_u; S_1) - I_{SKL}(\hat{S}_u; S_1)
= E_{W, \hat{S}_u, S_1} \left[ \log P^\alpha_{W|S_1, \hat{S}_u} - E_{W, \hat{S}_u} E_{S_1} \left[ \log P^\alpha_{W|S_1, \hat{S}_u} \right] \right]
= E_{W, \hat{S}_u, S_1} \left[ -\frac{\alpha}{2} (W - W^*(S_1, \hat{S}_u))^\top H^*(S_1, \hat{S}_u)(W - W^*(S_1, \hat{S}_u)) \right]
- E_{W, \hat{S}_u} E_{S_1} \left[ -\frac{\alpha}{2} (W - W^*(S_1, \hat{S}_u))^\top H^*(S_1, \hat{S}_u)(W - W^*(S_1, \hat{S}_u)) \right]
+ E_{S_1, \hat{S}_u} \left[ \log \frac{\sqrt{\det(\alpha H^*(S_1, \hat{S}_u))}}{\sqrt{2\pi}} \right] - E_{S_1} E_{\hat{S}_u} \left[ \log \frac{\sqrt{\det(\alpha H^*(S_1, \hat{S}_u))}}{\sqrt{2\pi}} \right]
\] (133)
\[
E_{\mathbf{W}, \hat{S}_u, S_u} \left[ - \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right] + E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right]
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \text{tr} \left( H^*(S_l, \hat{S}_u)(\mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}^T + \mathbf{W} \mathbf{W}^*(S_l, \hat{S}_u)^\top - \mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}(S_l, \hat{S}_u)^\top) \right) \right]
\]

\[
- E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \text{tr} \left( H^*(S_l, \hat{S}_u)(\mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}^T + \mathbf{W} \mathbf{W}^*(S_l, \hat{S}_u)^\top - \mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}(S_l, \hat{S}_u)^\top) \right) \right]
\]

\[
+ E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right] - E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right]
\]

\[
= E_{\mathbf{W}, \hat{S}_u, S_u} \left[ - \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right] + E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right]
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \text{tr} \left( H^*(S_l, \hat{S}_u)(\mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}^T + \mathbf{W} \mathbf{W}^*(S_l, \hat{S}_u)^\top - \mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}(S_l, \hat{S}_u)^\top) \right) \right]
\]

\[
- E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \text{tr} \left( H^*(S_l, \hat{S}_u)(\mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}^T + \mathbf{W} \mathbf{W}^*(S_l, \hat{S}_u)^\top - \mathbf{W}^*(S_l, \hat{S}_u) \mathbf{W}(S_l, \hat{S}_u)^\top) \right) \right]
\]

\[
+ E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right] - E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right]
\]

\[
= E_{\mathbf{W}, \hat{S}_u, S_u} \left[ - \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right] + E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right]
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \frac{\alpha}{2} \mathbf{W}^*(S_l, \hat{S}_u)^\top H^*(S_l, \hat{S}_u) \mathbf{W}^*(S_l, \hat{S}_u) \right]
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_u} \left[ \alpha \left( \frac{1}{2} \mathbf{W}^*(S_l, \hat{S}_u) - E_{S_l, \hat{S}_u}[\mathbf{W}^*(S_l, \hat{S}_u)] \right)^\top \right] H^*(S_l, \hat{S}_u) \mathbf{W}^*(S_l, \hat{S}_u)
\]

\[
+ E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right] - E_{S_l, \hat{S}_u} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right].
\]

From [130], we have

\[
0 = E_{\Delta(S_l, \hat{S}_u)} \left[ \log \sqrt{\frac{\alpha H^*(S_l, \hat{S}_u)}{2}} + \alpha \bar{L}_E(\mathbf{W}^*(S_l, \hat{S}_u)), S_l, \hat{S}_u \right] + \log \Lambda_{\alpha, \lambda}(S_l, \hat{S}_u)
\]

\[
= E_{\Delta(S_l, \hat{S}_u)} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} + \alpha \bar{L}_E(\mathbf{W}^*(S_l, \hat{S}_u)), S_l, \hat{S}_u \right] + \log \Lambda_{\alpha, \lambda}(S_l, \hat{S}_u),
\]

which means

\[
E_{\Delta(S_l, \hat{S}_u)} \left[ \log \Lambda_{\alpha, \lambda}(S_l, \hat{S}_u) \right] = -E_{\Delta(S_l, \hat{S}_u)} \left[ \log \sqrt{\det(H^*(S_l, \hat{S}_u))} \right] - E_{\Delta(S_l, \hat{S}_u)} \left[ \alpha \bar{L}_E(\mathbf{W}^*(S_l, \hat{S}_u)), S_l, \hat{S}_u \right].
\]

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

\[
\tilde{\text{gen}} \left( P_{\mathbf{W}|S_l, \hat{S}_u}, P_{S_l, \hat{S}_u} \right) = \frac{1 + \lambda}{2} \left( E_{\mathbf{W}, \hat{S}_u, S_l} \left[ - \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right] + E_{\mathbf{W}, \hat{S}_u, S_l} \left[ \mathbf{W}^T H^*(S_l, \hat{S}_u) \mathbf{W} \right] \right)
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_l} \left[ \mathbf{W}^*(S_l, \hat{S}_u)^\top H^*(S_l, \hat{S}_u) \mathbf{W}^*(S_l, \hat{S}_u) \right]
\]

\[
+ E_{\mathbf{W}, \hat{S}_u, S_l} \left[ \left( \mathbf{W}^*(S_l, \hat{S}_u) - 2E_{S_l, \hat{S}_u}[\mathbf{W}^*(S_l, \hat{S}_u)] \right)^\top H^*(S_l, \hat{S}_u) \mathbf{W}^*(S_l, \hat{S}_u) \right]
\]

\[
- E_{\Delta(S_l, \hat{S}_u)} \left[ \bar{L}_E(\mathbf{W}^*(S_l, \hat{S}_u)), S_l, \hat{S}_u \right] \right). \]

K PROOF OF COROLLARY[1]

When \( S_l \) and \( \hat{S}_u \) are independent, we can simplify the asymptotic gen-error as

\[
\tilde{\text{gen}} \left( P_{\mathbf{W}|S_l, \hat{S}_u}, P_{S_l, \hat{S}_u} \right) \]
According to Theorem 2 and (140), the asymptotic gen-error of MLE is given by

\[
\sum_{n=1}^{\infty} \frac{1}{n} \left( \frac{m}{n} D(P_Z || p(\cdot | w)) + \frac{m}{n + m} D(P_{Z|w} || p(\cdot | w)) \right)
\]

and then we have

\[
\mathbb{E}_{Z \sim P_Z} \mathbb{E}_{Z \sim P_{Z|w}} \left[ \nabla_w \left(- \frac{n}{n + m} \log p(Z|w) - \frac{m}{n + m} \log p(Z|w) \right) \right]_{w = w^*_\lambda} = 0.
\]

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

\[
\text{gen}(P_{\infty|S, \hat{S}}) = \frac{n + m}{2n} \left( \mathbb{E}_{w, \hat{S}, S, \hat{S}} \left[ - W^T J(w^*_\lambda) W \right] + \mathbb{E}_{w, \hat{S}, S, \hat{S}} \left[ W^T J(w^*_\lambda) W \right] \right)
\]

\[
+ 2 \mathbb{E}_{\hat{S}, S, \hat{S}} \left[ \left( W_{ML}(S, \hat{S}) - \mathbb{E}_S [W_{ML}(S, \hat{S})] \right)^T J(w^*_\lambda) \mathbb{E}_{P_{S|\hat{S}}} [W_{ML}(S, \hat{S})] \right)
\]

\[
= \frac{n + m}{2n} \left( \mathbb{E}_w \left[ - W^T J(w^*_\lambda) W \right] + \mathbb{E}_w \left[ W^T J(w^*_\lambda) W \right] \right)
\]

\[
+ 2 \text{tr} \left( \mathbb{E}_{\hat{S}, S, \hat{S}} \left[ \left( W_{ML}(S, \hat{S}) - \mathbb{E}_S [W_{ML}(S, \hat{S})] \right) (W_{ML}(S, \hat{S}) - \mathbb{E}_S [W_{ML}(S, \hat{S})])^T J(w^*_\lambda) \right] \right)
\]

\[
= \frac{n + m}{n} \text{tr} \left( \mathbb{E}_{\hat{S}, S, \hat{S}} \left[ \left( W_{ML}(S, \hat{S}) - \mathbb{E}_S [W_{ML}(S, \hat{S})] \right) (W_{ML}(S, \hat{S}) - \mathbb{E}_S [W_{ML}(S, \hat{S})])^T J(w^*_\lambda) \right] \right).
\]

Fix any pseudo-labeled data set \( \hat{S} \) and let

\[
\hat{W}(\hat{S}) = \arg \min_{w \in W} \left( \frac{n}{n + m} D(P_Z || p(\cdot | w)) - \frac{1}{n + m} \sum_{i=n+1}^{n+m} \log p(\hat{Z}_i|w) \right).
\]

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

\[
\hat{W}(\hat{S}) \rightarrow w^*_\lambda,
\]

and

\[
\mathbb{E}_{Z \sim P_Z} \left[ \nabla_w \left(- \frac{n}{n + m} \log p(Z|w) - \frac{1}{n + m} \sum_{i=n+1}^{n+m} \log p(\hat{Z}_i|w) \right) \right]_{w = \hat{W}(\hat{S})} = 0.
\]

As \( n \to \infty \), by central limit theorem, \( \mathbb{E}_{S} [\hat{W}_{ML}(S, \hat{S})] = \hat{W}(\hat{S}) \) (cf. (15) and (146)).

By applying Taylor expansion to \( \nabla_w \hat{L}_E(w, S, \hat{S}) |_{w = \hat{W}_{ML}(S, \hat{S})} \) around \( \hat{W}_{ML}(S, \hat{S}) = \hat{W}(\hat{S}) \), we have

\[
0 = \nabla_w \left(- \frac{1}{n + m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n + m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) |_{w = \hat{W}_{ML}(S, \hat{S})}
\]

\[
\approx \nabla_w \left(- \frac{1}{n + m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n + m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) |_{w = \hat{W}(\hat{S})}
\]

\[
+ \nabla^2_w \left(- \frac{1}{n + m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n + m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) |_{w = \hat{W}(\hat{S})} (\hat{W}_{ML}(S, \hat{S}) - \hat{W}(\hat{S})).
\]
By multivariate central limit theorem, as \( n \to \infty \), the first term in (149) converges as follows

\[
\nabla_w \left( - \frac{1}{n+m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n+m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) \bigg|_{w=\hat{W}(\hat{s}_u)} \xrightarrow{d} \mathcal{N} \left( 0, \frac{n}{(n+m)^2} \mathcal{I}_l(\hat{W}(\hat{s}_u)) \right).
\]

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

\[
\nabla_w^2 \left( - \frac{1}{n+m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n+m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) \bigg|_{w=\hat{W}(\hat{s}_u)} \xrightarrow{p} \frac{n}{n+m} J_l(\hat{W}(\hat{s}_u)) - \nabla_w^2 \left( - \frac{1}{n+m} \sum_{i=1}^{n} \log p(Z_i|w) - \frac{1}{n+m} \sum_{i=m+1}^{n+m} \log p(\hat{Z}_i|w) \right) \bigg|_{w=\hat{W}(\hat{s}_u)} =: \hat{J}(\hat{W}(\hat{s}_u)).
\]

Thus, we have

\[
\hat{W}_{ML}(S_1, \hat{S}_u) \xrightarrow{d} \mathcal{N} \left( w_\lambda^*, \frac{n J(w_\lambda^*)^{-1} \mathcal{I}_l(w_\lambda^*) J(w_\lambda^*)}{(n+m)^2} \right).
\]

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

\[
\mathbb{E} (P_{\infty}^W_{w|S_1,\hat{S}_u}, P_{S_1,\hat{S}_u}) = \frac{n+m}{n} \text{tr} \left( \frac{n}{(n+m)^2} J(w_\lambda^*)^{-1} \mathcal{I}_l(w_\lambda^*) \right)
\]

\[
= \frac{\text{tr}(J(w_\lambda^*)^{-1} \mathcal{I}_l(w_\lambda^*))}{n+m}.
\]

### L PROOF OF COROLLARY 2

By applying Taylor expansion of \( L_P(W, P_{S_1}) \) around \( W = w_1^* \), we have the following approximation

\[
L_P(W, P_{S_1}) \approx L_P(w_1^*, P_{S_1}) + (W - w_1^*)^\top \nabla_W L_P(W, P_{S_1}) |_{W=w_1^*} + \frac{1}{2} (W - w_1^*)^\top \nabla_W^2 L_P(W, P_{S_1}) |_{W=w_1^*} (W - w_1^*)
\]

\[
= L_P(w_1^*, P_{S_1}) + \frac{1}{2} \text{tr} \left( (W - w_1^*) (W - w_1^*)^\top J_l(w_1^*) \right).
\]

Thus, the excess risk can be approximated as follows:

\[
\mathcal{E}_l(P_W) = \mathbb{E}_W [L_P(W, P_{S_1})] - L_P(w_1^*, P_{S_1})
\]

\[
\approx \frac{1}{2} \text{tr} \left( \mathbb{E}_W [(W - w_1^*) (W - w_1^*)^\top] J_l(w_1^*) \right)
\]

\[
= \frac{1}{2} \text{tr} \left( \mathbb{E}_{S_1, \hat{S}_u} [(W(S_1, \hat{S}_u) - w_1^*) (W(S_1, \hat{S}_u) - w_1^*)^\top] J_l(w_1^*) \right) + \frac{\text{tr}(J_l(w_1^*) \text{Cov}(W|S_1, \hat{S}_u))}{2}
\]

\[
= \frac{1}{2} \text{tr} \left( (w_\lambda^* - w_1^*) (w_\lambda^* - w_1^*)^\top J_l(w_1^*) \right) + \frac{\text{tr}(J_l(w_1^*) \text{Cov}(W_{ML}(S_1, \hat{S}_u)))}{2}
\]

\[
= \frac{1}{2} \text{tr} \left( (w_\lambda^* - w_1^*) (w_\lambda^* - w_1^*)^\top J_l(w_1^*) \right) + \frac{\text{tr}(J_l(w_1^*) J(w_\lambda^*)^{-1} \mathcal{I}_l(w_\lambda^*) J(w_\lambda^*)^{-1})}{2(1+\lambda)(n+m)}
\]

where (160) follows since when \( \alpha \to \infty \), \( \text{Cov}(W|S_1, \hat{S}_u) = \frac{1}{\alpha} J^*(S_1, \hat{S}_u) \to 0 \) and from (153).
M PROOF OF LOGISTIC REGRESSION EXAMPLE

The first and second derivatives of the loss function are as follows
\[
\nabla_w l(w, z) = \nabla_w \log(1 + \exp(-yw^T x)) + \nu w = \frac{-yxe^{-yw^T x}}{1 + e^{-yw^T x}} + \nu w, \quad \text{and} \quad (162)
\]
\[
\nabla_w^2 l(w, z) = \nabla_w^2 \log(1 + \exp(-yw^T x)) + \nu I_d = \frac{xx^Te^{-yw^T x}}{(1 + e^{-yw^T x})^2} + \nu I_d. \quad (163)
\]

The expected Hessian matrices \(J_l, J_u\) and expected product of the first derivative \(\mathcal{I}_l\) are given as follows:
\[
J_l(w) = \mathbb{E}_{Z \sim P_Z} \left[ \frac{XX^Te^{-yw^T x}}{(1 + e^{-yw^T x})^2} \right] = \mathbb{E}_{X \sim P_X} \left[ \frac{XX^T}{e^{-w^TX} + e^{w^TX} + 2} \right], \quad (164)
\]
\[
J_u(w) = \mathbb{E}_{X \sim P_X} \left[ \frac{XX^Te^{-sgn(X^Tw_0^T)w^T x}}{(1 + e^{-sgn(X^Tw_0^T)w^TX})^2} \right] = \mathbb{E}_{X \sim P_X} \left[ \frac{XX^T}{e^{-w^TX} + e^{w^TX} + 2} \right], \quad (165)
\]
\[
\mathcal{I}_l(w) = \mathbb{E}_{Z \sim P_Z} \left[ \frac{XX^Te^{-2yw^T x}}{(1 + e^{-yw^T x})^2} \right]. \quad (166)
\]

We can see that \(J(w) = J_l(w) = J_u(w)\).

Recall the proof of Corollary 1 in Appendix [13]. In the logistic regression with \(l_2\) regularization, the unique minimizer of the empirical risk in [13] is rewritten as
\[
\hat{W}_{ML}(S_l, \hat{S}_u) = \arg \min_{w \in \mathcal{W}} \left( -\frac{1}{1 + \lambda} \frac{1}{n} \sum_{i=1}^{n} \log p(Z_i | w) - \frac{\lambda}{1 + \lambda m} \frac{1}{m} \sum_{i=n+1}^{n+m} \log p(Z_i | w) + \frac{\nu}{2} \| w \|^2 \right) \quad (167)
\]
and the Hessian matrix of the empirical risk at \(w = \hat{W}_{ML}(S_l, \hat{S}_u)\) is rewritten as
\[
H^*(S_l, \hat{S}_u) = \nabla_w^2 \left( -\frac{1}{1 + \lambda} \frac{1}{n} \sum_{i=1}^{n} \log p(Z_i | w) - \frac{\lambda}{1 + \lambda m} \frac{1}{m} \sum_{i=n+1}^{n+m} \log p(Z_i | w) \right) \bigg|_{w=\hat{W}_{ML}(S_l, \hat{S}_u)} + \nu I_d. \quad (168)
\]

Recall the definition of \(w^*_\lambda\) with regularization
\[
w^*_\lambda = \arg \min_{w \in \mathcal{W}} \left( \frac{n}{n + m} D(P_Z \| p(\cdot | w)) + \frac{m}{n + m} D(P_{\hat{Z}|w_0} \| p(\cdot | w)) + \frac{\nu}{2} \| w \|^2 \right), \quad (169)
\]
Given any ratio $\lambda > 0$, as $n, m \to \infty$, the Hessian matrix converges as follows

$$H^*(S_l, \hat{S}_u) \overset{p}{\to} J(w^*_\lambda) + \nu I_d.$$  

(170)

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

$$\text{gen}(P^\infty_{W|S_l, \hat{S}_u}, P_{S_l}, \hat{S}_u) = \frac{n + m}{n} \text{tr} \left( \frac{1}{n} E_{S_u} E_{S_l} \left[ (\hat{W}_{\text{ML}}(S_l, \hat{S}_u) - E_{S_l} [\hat{W}_{\text{ML}}(S_l, \hat{S}_u)]) (\hat{W}_{\text{ML}}(S_l, \hat{S}_u) - E_{S_l} [\hat{W}_{\text{ML}}(S_l, \hat{S}_u)])^\top \right] (J(w^*_\lambda) + \nu I_d) \right).$$

(171)

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

$$\hat{W}_{\text{ML}}(S_l, \hat{S}_u) \overset{d}{\to} \mathcal{N} \left( w^*_\lambda, \frac{n(J(w^*_\lambda) + \nu I_d)^{-1} I_l(w^*_\lambda) (J(w^*_\lambda) + \nu I_d)^{-1}}{(n + m)^2} \right).$$

(172)

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

$$\text{gen}(P^\infty_{W|S_l, \hat{S}_u}, P_{S_l}, \hat{S}_u) = \frac{\text{tr} ((J(w^*_\lambda) + \nu I_d)^{-1} I_l(w^*_\lambda))}{n + m}.$$  

(173)

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

$$\mathcal{E}(P_W) = \frac{1}{2} \text{tr} ((w^*_\lambda - w^*_\ell)^\top (w^*_\lambda - w^*_\ell)^\top J_l(w^*_\ell)) + \frac{\text{tr} (J_l(w^*_\ell) (J(w^*_\lambda) + \nu I_d)^{-1} I_l(w^*_\lambda) (J(w^*_\lambda) + \nu I_d)^{-1})}{2(1 + \lambda)(n + m)},$$

(174)

where $w^*_\ell = \arg\min_{w \in \mathcal{W}} L_P(W, P_{S_l})$.

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 $\lambda$ increases.