Information-Theoretic Generalization Bounds for Iterative Semi-Supervised Learning

Haiyun He, Hanshu Yan, and Vincent Y. F. Tan

Abstract

We consider iterative semi-supervised learning (SSL) algorithms that iteratively generate pseudo-labels for a large amount unlabelled data to progressively refine the model parameters. In particular, we seek to understand the behaviour of the generalization error of iterative SSL algorithms using information-theoretic principles. To obtain bounds that are amenable to numerical evaluation, we first work with a simple model—namely, the binary Gaussian mixture model. Our theoretical results suggest that when the class conditional variances are not too large, the upper bound on the generalization error decreases monotonically with the number of iterations, but quickly saturates. The theoretical results on the simple model are corroborated by extensive experiments on several benchmark datasets such as the MNIST and CIFAR datasets in which we notice that the generalization error improves after several pseudo-labelling iterations, but saturates afterwards.

1 Introduction

In real-life machine learning applications, it is relatively easy and cheap to obtain large amounts of unlabelled data, while the number of labelled data examples is usually small due to the high cost of annotating them with true labels. In light of this, semi-supervised learning (SSL) has come to the fore (Chapelle et al., 2006; Zhu, 2008; Van Engelen & Hoos, 2020). SSL makes use of the abundant unlabelled data to augment the performance of learning tasks with few labelled data examples. This has been shown to outperform supervised and unsupervised learning under certain conditions. For example, in a classification problem, the correlation between the additional unlabelled data and the labelled data may help to enhance the accuracy of classifiers. Among the plethora of SSL methods, pseudo-labelling (Lee et al., 2013) has been observed to be a simple and efficient way to improve the generalization performance empirically. In this paper, we consider the problem of pseudo-labelling a subset of the unlabelled data at each iteration based on the previous output parameter and then refining the model progressively, but we are interested in analysing this procedure theoretically. Our goal in this paper is to understand the impact of pseudo-labelling on the generalization error.

A learning algorithm can be viewed as a randomized map from the training dataset to the output model parameter. The output is highly data-dependent and may suffer from overfitting to the given dataset. In statistical learning theory, the generalization error is defined as the expected gap between the test and training losses, and is used to measure the extent to which the algorithms overfit to the training data. In SSL problems, the unlabelled data are expected to improve the generalization performance in a certain manner and thus, it is worthwhile to investigate the behaviour theoretically. In this paper, we leverage results in Bu et al. (2020), Wu et al. (2020) to derive an information-theoretic generalization error bound at each iteration for iterative SSL.

We state our main theoretical contribution informally as follows.

Theorem [Informal] For a $d$-variate binary Gaussian mixture model (bGMM) in which each component has variance $\sigma^2$, the generalization error across the different semi-supervised training iterations $|\text{gen}_t|$ can be bounded with high probability as follows:

$$|\text{gen}_t| \leq \text{const} \cdot \mathbb{E}\left[\sqrt{G_\sigma(F_\sigma^{(t-1)}(\alpha))}\right],$$

where $\alpha$ represents the correlation between the optimal and estimated parameter vectors, $F_\sigma^{(t)}$ is the iterated composition of the function $F_\sigma$ (sketched in Figure 3), and $G_\sigma$ (sketched in Figure 5) represents the KL-divergence between the pseudo-labelled and true data distributions.

As shown in Figure 1, the upper bound is monotonically decreasing in the iteration count $t$ and converges at around $t = 2$ with a sufficiently large amount of unlabelled data. In Section 4, we also
Figure 1: Upper bound on $|\text{gen}|$ as a function of $t$.

show that when the number of labelled data or the variance is large enough, using the unlabelled data does not help to significantly reduce the generalization error across iterations $t$. The behaviour of the empirical generalization error for the bGMM coincides with the upper bound. The results suggest that the proposed upper bound serves as a useful guide to understand how the generalization error changes across the semi-supervised training iterations and it can be used to establish conditions under which unlabelled data can help in terms of generalization. Experimental results on the MNIST and CIFAR datasets corroborate the phenomena for the bGMM that with few labelled data and abundant unlabelled data, the generalization error decreases quickly in the early pseudo-labelling iterations and saturates thereafter. For a more extensive literature review, please refer to Appendix A.

## 2 Problem Setup

Let the instance space be $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^{d+1}$, the model parameter space be $\Theta$ and the loss function be $l : \mathcal{Z} \times \Theta \to \mathbb{R}$, where $d \in \mathbb{N}$. We are given a labelled training dataset $S_l \equiv \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$ drawn from $\mathcal{Z}$, where each $Z_i = (X_i, Y_i)$ is independently and identically distributed (i.i.d.) from $P_Z = \mathcal{P}_X \times \mathcal{P}(Y)$ and $X_i$ is i.i.d. from $P_X \in \mathcal{P}(\mathcal{X})$. For any $i \in [n]$, $X_i$ is a vector of features and $Y_i$ is a label indicating the class to which $X_i$ belongs. However, in many real-life machine learning applications, we only have a limited number of labelled data while we have access to a large amount of unlabelled data, which are expensive to annotate. Then we can incorporate the unlabelled training data together with the labelled data to improve the performance of the model. This procedure is called semi-supervised learning (SSL). We are given an independent unlabelled training dataset $S_u \equiv \{X'_1, \ldots, X'_{m}\}, \tau \in \mathbb{N}$, where each $X'_i$ is i.i.d. generated from $P_X \in \mathcal{P}(\mathcal{X})$. Typically, $m \gg n$.

In the following, we consider the iterative self-training with pseudo-labelling in SSL setup, as shown in Figure 2. Let $t \in [0 : \tau]$ denote the iteration counter. In the initial round ($t = 0$), the labelled data $S_l$ are first used to learn an initial model parameter $\theta_0 \in \Theta$. Next, we split the unlabelled dataset $S_u$ into $\tau$ disjoint equal-size sub-datasets $\{S_{u,k}\}_{k=1}^\tau$, where $S_{u,k} = \{X'_{(k-1)m+1}, \ldots, X'_{km}\}$. In each subsequent round $t \in [1 : \tau]$, based on $\theta_{t-1}$ trained from the previous round, we use a predictor $f_{\theta_{t-1}} : \mathcal{X} \to \mathcal{Y}$ to assign a pseudo-label $Y'_i$ to the unlabelled sample $X'_i$ for all $i \in [(t-1)m + 1 : tm] := \{(t-1)m, (t-1)m + 1, \ldots, tm\}$. Let $S_{u,t} = \{(X'_i, Y'_i)\}_{i={(t-1)m+1}}^{tm}$ denote the $t$th pseudo-labelled dataset. After pseudo-labelling, both the labelled data $S_l$ and the pseudo-labelled data $S_{u,t}$ are used to learn a new model parameter $\theta_t$. The procedure is then repeated iteratively until the maximum number of iterations $\tau$ is reached.

![Figure 2: Paradigm of iterative self-training with pseudo-labelling in SSL.](image)

Under the setup of iterative SSL, during each iteration $t$, our goal is to find a model parameter $\theta_t \in \Theta$ that minimizes the population risk with respect to the underlying data distribution

$$L_{P_Z}(\theta_t) := \mathbb{E}_{Z \sim P_Z} [l(\theta_t, Z)].$$

(2)
Since $P_Z$ is unknown, $L_{P_Z}(\theta_t)$ cannot be computed directly. Hence, we instead minimize the empirical risk. The procedure is termed empirical risk minimization (ERM). For any model parameter $\theta_t \in \Theta$, the empirical risk of the labelled data is defined as

$$L_{S_t}(\theta_t) := \frac{1}{n} \sum_{i=1}^{n} l(\theta_t, Z_i),$$  \hspace{1cm} (3)

and for $t \geq 1$, the empirical risk of pseudo-labelled data $\hat{S}_{u,t}$ as

$$L_{\hat{S}_{u,t}}(\theta_t) := \frac{1}{m} \sum_{i=1}^{m} l(\theta_t, (X'_i, Y'_i)).$$  \hspace{1cm} (4)

We set $L_{\hat{S}_{u,t}}(\theta_t) = 0$ for $t = 0$. For a fixed weight $w \in [0, 1]$, the total empirical risk can be defined as the following linear combination of $L_{S_t}(\theta_t)$ and $L_{\hat{S}_{u,t}}(\theta_t)$:

$$L_{S_t,\hat{S}_{u,t}}(\theta_t) := wL_{S_t}(\theta_t) + (1-w)L_{\hat{S}_{u,t}}(\theta_t).$$  \hspace{1cm} (5)

An SSL algorithm can be characterized by a randomized map from the labelled and unlabelled training data $S_t$, $\hat{S}_u$ to a model parameter $\theta$ according to a conditional distribution $P_{\theta_t|S_t,\hat{S}_u}$. Then at each iteration $t$, we can use the sequence of conditional distributions $\{P_{\theta_t|S_t,\hat{S}_u}\}_{k=0}^{t}$ with $P_{\theta_t|S_t} = P_{\theta_t|S_t}$ to represent an iterative SSL algorithm. The generalization error at the $t$-th iteration is defined as the expected gap between the population risk of $\theta_t$ and the empirical risk on the training data:

$$\text{gen}_t(P_Z, P_X, \{P_{\theta_k|S_t,\hat{S}_u}\}_{k=0}^{t-1}, \{f_{\theta_k}\}_{k=0}^{t-1}) := \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_t,\hat{S}_{u,t}}(\theta_t)]$$  \hspace{1cm} (6)

$$= w \left( \mathbb{E}_{\theta_t} \mathbb{E}_{Z}[l(\theta_t, Z) | \theta_t] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)] \right)$$

$$+ (1-w) \left( \mathbb{E}_{\theta_t} \mathbb{E}_{Z}[l(\theta_t, Z) | \theta_t] - \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta_t, X'_i,Y'_i}[l(\theta_t, (X'_i, Y'_i))] \right).$$  \hspace{1cm} (7)

When $t = 0$ and $w = 1$, the definition of the generalization error above reduces to that of vanilla supervised learning. The generalization error $\text{gen}_t$ is used to measure the extent to which the iterative learning algorithm overfits the training data at the $t$-th iteration. Instead of focusing on the total generalization error induced during the entire process, we are more interested in the following questions. How does $\text{gen}_t$ evolve as the iteration count $t$ increases? Do the unlabelled data examples in $\hat{S}_u$ help to improve the generalization error?

### 3 Preliminaries

Inspired by the information-theoretic generalization results in [Bu et al., 2020, Theorem 1] and [Wu et al., 2020, Theorem 1], we derive an upper bound on the generalization error $\text{gen}_t$ for any $t \in [0 : \tau]$ in terms of the mutual information between input data samples (either labelled or pseudo-labelled) and the output model parameter $\theta_t$, as well as the KL-divergence between the underlying data distributions and the joint distribution of feature vectors and pseudo-labels.

We denote an $R$-sub-Gaussian random variable $L \in \mathbb{R}$ [Vershynin, 2018] as $L \sim \text{subG}(R)$. Furthermore, let us recall the following non-standard information quantities.

**Definition 1.** For arbitrary random variables $X, Y$ and $U$, define the disintegrated mutual information [Negrea et al., 2017] [Haghifam et al., 2020] between $X$ and $Y$ given $U$ as $I_U(X;Y) := D(P_{X,Y|U}P_{X|U} \otimes P_{Y|U})$, and the disintegrated KL-divergence between $P_X$ and $P_Y$ given $U$ as $D_U(P_X|P_Y) := D(P_{X|U}P_{Y|U})$. These are $(\sigma(U))$-measurable random variables. It follows immediately that the conditional mutual information $I(X;Y|U) = E_U[I_U(X;Y)]$ and the conditional KL-divergence $D(P_{X|U}P_{Y|U}) = E_U[D_U(P_X|P_Y)]$.

Let $\theta(t) = (\theta_0, \ldots, \theta_t)$ for any $t \in [0 : \tau]$. In iterative SSL, we can upper bound the generalization error as shown in Theorem [1] to follow by applying the law of total expectation.

**Theorem 1** (Generalization error upper bound for iterative SSL). Suppose $l(\theta, Z) \sim \text{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0 : \tau]$,

$$|\text{gen}_t(P_Z, P_X, \{P_{\theta_k|S_t,\hat{S}_u}\}_{k=0}^{t-1}, \{f_{\theta_k}\}_{k=0}^{t-1})| \leq \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta(\tau-1)} \left[ \sqrt{2R^2 I_{\theta(\tau-1)}(\theta_t; Z_i)} \right]$$

$$+ \frac{1-w}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta(\tau-1)} \left[ \sqrt{2R^2 \left( I_{\theta(\tau-1)}(\theta_t; X'_i, Y'_i) + D_{\theta(\tau-1)}(P_{X'_i,Y'_i}|P_Z) \right)} \right].$$  \hspace{1cm} (8)
we find it convenient to show that \( X \) information-theoretic quantities (just as in Russo & Zou (2016); Xu & Raginsky (2017); Bu et al. (2020)), information between the individual input data sample \( Z \) and the output model parameter \( \theta_t \) measures the extent to which the algorithm is sensitive to the input data, and the KL-divergence between the underlying \( P_Z \) and pseudo-labelled distribution \( P_{X_i,Y_i} \) measures how well the algorithm generalizes to the true data distribution. As \( n \to \infty \) and \( m \to \infty \), we show that the disintegrated mutual information \( I_{\theta(t-1)}(\theta_t; X_i', Y_i') \) tends to 0 (in probability), which means that there are sufficient training data such that the algorithm can generalize well. On the other hand, the impact on the generalization error of pseudo-labelling is reflected in the KL-divergence \( D_{\theta(t-1)}(P_{X_i,Y_i} \| P_Z) \) and this term does not necessarily vanish as \( n, m \to \infty \). We quantify this precisely in Remark 1 in Section 4.

In iterative learning algorithms, it is usually difficult to directly calculate the mutual information and KL-divergence between the input and the final output \( \{ \text{Paninski} 2003 \text{, Nguyen et al. 2010, McAllester \& Stratos 2020} \} \). However, by applying the law of total expectation and conditioning the information-theoretic quantities on the output model parameters \( \theta^{(t-1)} = \{ \theta_1, \ldots, \theta_{t-1} \} \) from previous iterations, we are able to calculate the upper bound iteratively. In the next section, we apply the iterated generalization error bound to a classification problem under a specific generative model—the bGMM. This simple model allows us to derive a tractable upper bound on the generalization error as a function of iteration number \( t \) that we can compute numerically.

4 Main Results

We now particularize the iterative semi-supervised classification setup to the bGMM. We calculate the term in (3) to understand the effect of multiple self-training rounds on the generalization error.

Fix a unit vector \( \mu \in \mathbb{R}^d \) and a scalar \( \sigma \in \mathbb{R}_+ = (0, \infty) \). Under the bGMM with mean \( \mu \) and standard deviation \( \sigma \) (bGMM(\( \mu, \sigma \))), we assume that the distribution of any labelled data example \( (X, Y) \) is specified as follows. Let \( \mathcal{Y} = \{-1, +1\} \), \( Y \sim P_Y \), where \( P_Y(-1) = P_Y(1) = \frac{1}{2} \), and \( X \mid Y \sim N(Y \mu, \sigma^2 I_d) \), where \( I_d \) is the identity matrix of size \( d \times d \). In anticipation of leveraging Theorem 1 together with the sub-Gaussianity of the loss function for the bGMM to derive generalization bounds in terms of information-theoretic quantities (just as in Russo & Zou (2016); Xu & Raginsky (2017); Bu et al. (2020)), we find it convenient to show that \( X \) and \( l(\theta, (X,Y)) \) are bounded w.h.p. By defining the \( \ell_\infty \) ball \( B^\ell_\infty := \{ x \in \mathbb{R}^d : \| x - y \mu \|_\infty \leq r \} \), we see that

\[
\Pr(X \in B^\ell_\infty) = \left(1 - 2\Phi\left(-\frac{r}{\sigma}\right)\right)^d = 1 - \delta_{r,d},
\]

where \( \Phi(\cdot) \) is the Gaussian cumulative distribution function. By choosing \( r \) appropriately, the failure probability \( \delta_{r,d} \) can be made arbitrarily small.

The random vector \( X \) is distributed according to the mixture distribution \( p_\mu = \frac{1}{2} N(\mu, \sigma^2 I_d) + \frac{1}{2} N(-\mu, \sigma^2 I_d) \). In the unlabelled dataset \( S_u \), each \( X'_i \) for \( i \in [1 : \tau m] \) is drawn i.i.d. from \( p_\mu \).

For any \( \theta \in \Theta \), under the bGMM(\( \theta, \sigma \)), the joint distribution of any pair of \( (X, Y) \in \mathcal{Z} \) is given by \( N(Y \theta, \sigma^2 I_d) \otimes P_Y \). Let the loss function be the negative log-likelihood, which can be expressed as

\[
l(\theta, (x, y)) = -\log (P_Y(y)p_\theta(x|y)) = -\log \frac{1}{2\sqrt{\pi} \sigma^d} + \frac{1}{2\sigma^2}(x - y \theta)^\top(x - y \theta).
\]

The minimizer of \( \min_{\theta \in \Theta} E_{(X,Y) \sim N(Y \mu, \sigma^2 I_d) \otimes P_Y} [l(\theta, (X,Y))] \) is equal to \( \mu \). To show that \( \theta \) is bounded with high probability, define the set \( \Theta_{\mu,c} := \{ \theta \in \Theta : \| \theta - \mu \|_\infty \leq c \} \) for some \( c > 0 \). For any \( \theta \in \Theta_{\mu,c} \), we have

\[
\min_{(x, y) \in \mathcal{Z}} l(\theta, (x, y)) = -\log \frac{1}{2\sqrt{\pi} \sigma^d} =: c_1, \quad \text{and}
\]

\[
\max_{x \in B^\ell_\infty, y \in \mathcal{Y}} l(\theta, (x, y)) \leq -\log \frac{1}{2\sqrt{\pi} \sigma^d} + \frac{d(c + r)^2}{2\sigma^2} =: c_2.
\]

For any \( (X, Y) \) from the bGMM(\( \mu, \sigma \)) and any \( \theta \in \Theta_{\mu,c} \), the probability that \( l(\theta, (X,Y)) \) belongs to the interval \( [c_1, c_2] \) is bounded by

\[
\Pr \{ l(\theta, (X,Y)) \in [c_1, c_2] \} \geq 1 - \delta_{r,d}.
\]
Thus, according to Hoeffding’s lemma, with probability at least $1 - \delta_{rd}$, \( l(\theta, (X, Y)) \sim \text{subG}((c_2 - c_1)/2) \) under \((X, Y) \sim N(Y\mu, \sigma^2 I_d) \otimes P_Y \) for all \( \theta \in \Theta_{\mu,c} \), i.e., for all \( \lambda \in \mathbb{R} \),

\[
\mathbb{E}_{X,Y} \left[ \exp \left( \lambda(l(\theta, (X, Y)) - \mathbb{E}_{X,Y}[l(\theta, (X, Y))]) \right) \right] \leq \exp \left( \frac{\lambda^2(c_2 - c_1)^2}{8} \right). \tag{14}
\]

Under this setup, the iterative SSL procedure is shown in Figure 2 but the labelled dataset \( S_t \) is only used to train in the initial round \( t = 0 \); we discuss the use of \( S_t \) in all iterations in Corollary 3. The algorithm operates in the following steps.

- **Step 1: Initial round \( t = 0 \) with \( S_1 \):** By minimizing the empirical risk of labelled dataset \( S_1 \)

\[
L_{S_1}(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta, (X_i, Y_i)) \leq \frac{1}{2\sigma^2 n} \sum_{i=1}^{n} (X_i - Y_i\theta)^T (X_i - Y_i\theta), \tag{15}
\]

where \( \leq \) means that both sides differ by a constant independent of \( \theta \), we obtain the minimizer

\[
\theta_0 = \arg \min_{\theta \in \Theta} L_{S_1}(\theta) = \frac{1}{n} \sum_{i=1}^{n} Y_iX_i. \tag{16}
\]

- **Step 2: Pseudo-label data in \( S_t \):** At each iteration \( t \in [1 : \tau] \), for any \( i \in [(t-1)m + 1 : tm] \), we use \( \theta_{t-1} \) to assign a pseudo-label for \( X'_i \), that is, \( Y'_i = f_{\theta_{t-1}}(X'_i) = \text{sgn}(\theta_{t-1}'X'_i) \).

- **Step 3: Refine the model:** We then use the pseudo-labelled dataset \( \hat{S}_{n,t} \) to train the new model. By minimizing the empirical risk of \( \hat{S}_{n,t} \)

\[
L_{\hat{S}_{n,t}}(\theta) = \frac{1}{m} \sum_{i=(t-1)m+1}^{tm} l(\theta, (X'_i, Y'_i)) \leq \frac{1}{2\sigma^2 m} \sum_{i=(t-1)m+1}^{tm} (X'_i - Y'_i\theta)^T (X'_i - Y'_i\theta), \tag{17}
\]

we obtain the new model parameter

\[
\theta_t = \frac{1}{m} \sum_{i=(t-1)m+1}^{tm} Y'_iX'_i = \frac{1}{m} \sum_{i=(t-1)m+1}^{tm} \text{sgn}(\theta_{t-1}'X'_i)X'_i. \tag{18}
\]

If \( t < \tau \), go back to Step 2.

To state our result succinctly, we first define some non-standard notations and functions. From \[16\], we know that \( \theta_0 \sim N(\mu, \frac{1}{\sigma^2} I_d) \) and inspired by [Oymak & Gurung 2021], we can decompose \( \theta_0 \) as

\[
\theta_0 = (1 + \frac{\sigma}{\sqrt{\alpha}}\xi_0)\mu + \frac{\sigma}{\sqrt{\alpha}}\mu^+, \tag{19}
\]

where \( \xi_0 \sim N(0,1) \), \( \mu^+ \sim N(0,1 - \mu^-) \), and \( \mu^- \) is perpendicular to \( \mu \) and independent of \( \xi_0 \) (the details of this decomposition are provided in Appendix C).

Given two vectors \((a, b)\), define their correlation as \( \rho(a, b) := \frac{(a, b)}{|a| |b|} \) in \([-1, 1]\). The correlation between the estimated parameter \( \theta_0 \) and true parameter \( \mu \) is given by

\[
\alpha(\xi_0, \mu^+) := \rho(\theta_0, \mu) = \frac{1 + \frac{\sigma}{\sqrt{\alpha}}\xi_0}{\sqrt{(1 + \frac{\sigma}{\sqrt{\alpha}}\xi_0)^2 + \frac{\sigma^2}{\alpha}||\mu^-||^2_2}}. \tag{20}
\]

Let \( \beta(\xi_0, \mu^+) = \sqrt{1 - \alpha(\xi_0, \mu^+)^2} \). We abbreviate \( \alpha(\xi_0, \mu^+) \) and \( \beta(\xi_0, \mu^+) \) to \( \alpha \) and \( \beta \) respectively in the following. We can decompose the normalized vector \( \theta_0/||\theta_0||_2 \) as follows

\[
\tilde{\theta}_0 := \frac{\theta_0}{||\theta_0||_2} = \alpha \mu + \beta \nu, \tag{21}
\]

where \( \nu = \mu^+/||\mu^+||_2 \). Let \( \theta_0^\perp := (2\beta^2 - 2\alpha\beta)\nu/\sigma \), which is a vector perpendicular to \( \theta_0 \).

Define the KL-divergence between the pseudo-labelled data distribution and the true data distribution after the first iteration \( G_\sigma : [-1,1] \times \mathbb{R} \times \mathbb{R}^d \to [0, \infty) \) as

\[
G_\sigma(\alpha, \xi_0, \mu^+) := \Phi \left( \frac{\alpha}{\sigma} \right) p_{\tilde{g} + \frac{\sigma}{\sqrt{\alpha}}\xi_0 \leq \frac{\sigma}{\sqrt{\alpha}} \otimes p_{\mu^+} + \xi_0^\perp} + \Phi \left( \frac{\alpha}{\sigma} \right) p_{\tilde{g}^\perp \leq \frac{\sigma}{\sigma} \otimes p_{\mu^+}} | p_{\tilde{g} \otimes p_{\mu^+}} \right), \tag{22}
\]

where \( \tilde{g} \sim N(0,1), \tilde{g}^\perp \sim N(0, I_d - \theta_0\theta_0^\perp) \), \( \tilde{g}^\perp \) is independent of \( \tilde{g} \) and perpendicular to \( \theta_0 \). Note that \( p_{\tilde{g} + \frac{\sigma}{\sqrt{\alpha}}\xi_0 \leq \frac{\sigma}{\sqrt{\alpha}} \} \) is the Gaussian probability density function with mean \( \frac{\sigma}{\sqrt{\alpha}} \xi_0 \) and variance \( 1 \).
to the interval $(-\infty, -\frac{2}{\sigma})$, and similarly for $p_{y_{\bar{y}}} \leq \frac{2}{\sigma}$. In general, when $G_\sigma(\alpha, \xi, \mu^\perp)$ is small, so is the generalization error.

Let $Q(\cdot) := 1 - \Phi(\cdot)$. Define the correlation evolution function $F_\sigma : [-1, 1] \to [-1, 1]$ that quantifies the increase to the correlation (between the current model parameter and the optimal one) and improvement to the generalization error as the iteration counter increases from $t$ to $t + 1$:

$$F_\sigma(x) := \left(1 + \frac{2\sigma^2(x - x^2)}{\sigma} \frac{\exp(-\frac{x^2}{2\sigma^2})}{(1 - 2Q(\frac{x}{\sigma}) + \frac{2\sigma^2}{\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}))^2}\right)^{-\frac{1}{2}}. \quad (22)$$

![Figure 3: $F_\sigma^{(t)}(x)$ versus $x$ for different $t$ when $\sigma = 0.5$.](image)

![Figure 4: $F_\sigma(x)$ versus $x$ for $\sigma = 0.3$ and 0.5.](image)

![Figure 5: $G_\sigma(\alpha)$ versus $\alpha$ for different $\sigma$.](image)

The $t^{th}$ iterate of the function $F_\sigma$ is defined as $F_\sigma^{(t)} := F_\sigma \circ F_\sigma^{(t-1)}$ with $F_\sigma^{(0)}(x) = x$. As shown in Figure 3 for any fixed $\sigma$, we can see that $F_\sigma^{(2)}(x) \geq F_\sigma(x) \geq x$ for $x \geq 0$ and $F_\sigma^{(2)}(x) < F_\sigma(x) < x$ for $x < 0$. It can also be easily deduced that for any $t \in [0 : \tau]$, $F_\sigma^{(t+1)}(x) \geq F_\sigma^{(t)}(x)$ for any $x \geq 0$ and $F_\sigma^{(t+1)}(x) < F_\sigma^{(t)}(x)$ for any $x < 0$. This important observation implies that if the correlation $\alpha$, defined in \[19\], is positive, $F_\sigma^{(t)}(\alpha)$ increases with $t$; and vice versa. Moreover, as shown in Figure 4 by varying $\sigma$, we can see that smaller $\sigma$ results in a larger $|F_\sigma(x)|$.

By applying the result in Theorem 1, the following theorem provides an upper bound for the generalization error at each iteration $t$ for $m$ large enough.

**Theorem 2.** Fix any $\sigma \in \mathbb{R}_+$, $d \in \mathbb{N}$, $\epsilon \in \mathbb{R}$, and $\delta \in (0, 1)$. With probability at least $1 - \delta$, the absolute generalization error at $t = 0$ can be upper bounded as follows

$$|\text{gen}_t(P_\mathbf{Z}, P_{\mathbf{X}, \mathbf{P}_0|S_t,S_{t-1}})| \leq \sqrt{(c_2 - c_1)^2 d} \log \frac{n}{n - 1}. \quad (23)$$

For each $t \in [1 : \tau]$, for $m$ large enough, with probability at least $1 - \delta$,

$$|\text{gen}_t(P_\mathbf{Z}, P_{\mathbf{X}, \mathbf{P}_0|S_t,S_{t-1}})_{k=0} \{P_\mathbf{0}_k|S_t,S_{t-1}\}^1_{k=0} \{f_{\mathbf{0}}\}_{k=0}^{1-1}| \leq \sqrt{(c_2 - c_1)^2 \frac{2}{d}} \mathbb{E}_{\xi_0,\mu^\perp} \left[\sqrt{G_\sigma(F_\sigma^{(t)}(\alpha(\xi_0,\mu^\perp),\xi_0,\mu^\perp))} + \epsilon\right]. \quad (24)$$

The proof of Theorem 2 is provided in Appendix 3. Several remarks are in order.

First, to gain more insight, we numerically plot $G_\sigma(\alpha, \xi_0, \mu^\perp)$ when $d = 2$ and $\mu = (1, 0)$ in Figure 5. Under these settings, $G_\sigma(\alpha, \xi_0, \mu^\perp)$ depends only on $\alpha$ and hence, we can rewrite it as $G_\sigma(\alpha)$. As shown in Figure 5 for all $\sigma_1 > \sigma_2$, there exists an $\alpha_0 \in [-1, 1]$ such that for all $\alpha > \alpha_0 = \alpha_0(\sigma_1, \sigma_2)$, $G_{\sigma_1}(\alpha) > G_{\sigma_2}(\alpha)$. From \[19\], we can see that $\alpha$ is close to 1 of high probability, which means that $\sigma \to G_\sigma(\alpha)$ is monotonically increasing in $\sigma$ with high probability. As a result, $\mathbb{E}_{\alpha}[\sqrt{G_\sigma(\alpha)}]$ increases as $\sigma$ increases. This is consistent with the intuition that when the training data has larger variance, it is more difficult to generalize well. Moreover, for $\alpha > 0$, $G_\sigma(\alpha)$ decreases as $\alpha$ increases. Since $F_\sigma^{(t)}(\alpha)$ is increasing in $t$ for $\alpha > 0$, then $G_\sigma(F_\sigma^{(t)}(\alpha))$ is decreasing in $t$, which implies that the upper bound in \[24\] is also decreasing in $t$.

**Remark 1.** As $n \to \infty$, $\theta_0 \to \mu$ and $\alpha = \rho(\theta_0, \mu) \to 1$ almost surely, which means that the estimator converges to the optimal classifier for this bGMM. However, since there is no margin between two groups of data samples, the error probability $\text{Pr}(\hat{Y}^t_i \neq Y^t_i) \to Q(1/\sigma) > 0$ (which is the Bayes error rate) and the disintegrated KL-divergence $D_{\xi_0,\mu^\perp}(P_{\mathbf{X},\mathbf{Y}|\mathbf{P}_0}) || P_{\mathbf{X,Y}})$ between the estimated and underlying distributions cannot converge to 0. We discuss the other extreme case in which $\alpha = -1$ in Remark 2 in Appendix 4 of the supplementary material.
Second, by letting \( \epsilon \to 0 \), we compare the upper bounds for \( |\text{gen}_{\mu}| \) and \( |\text{gen}_1| \), as shown in Figures 6(a) and 6(b). For any fixed \( \sigma \), when \( n \) is sufficiently small, the upper bound for \( |\text{gen}_{\mu}| \) is greater than that for \( |\text{gen}_1| \). As \( n \) increases, the upper bound for \( |\text{gen}_1| \) surpasses that of \( |\text{gen}_{\mu}| \), as shown in Figure 6(b). This is consistent with the intuition that when the labelled data is limited, using the unlabelled data can help improve the generalization performance. However, as the number of labelled data increases, using the unlabelled data may degrade the generalization performance, if the distributions corresponding to classes +1 and −1 have a large overlap. This is because the labelled data is already effective in learning the unknown parameter \( \theta_t \) well and additional pseudo-labelled data does not help to further boost the generalization performance. Furthermore, by comparing Figures 6(a) and 6(b), we can see that for smaller \( \sigma \), the improvement from \( |\text{gen}_{\mu}| \) to \( |\text{gen}_1| \) is more pronounced. The intuition is that when \( \sigma \) decreases, the data samples have smaller variance and thus the pseudo-labelling is more accurate. In this case, unlabelled data can improve the generalization performance. Let us examine the effect of \( n \), the number of labelled training samples. By expanding \( \alpha \), defined in (19), using a Taylor series, we have

\[
\alpha = 1 - \frac{\sigma^2}{2n} \|\mu^\perp\|_2^2 + o\left(\frac{1}{n}\right). \tag{25}
\]

It can be seen that as \( n \) increases, \( \alpha \) converges to 1 in probability. Suppose the dimension \( d = 2 \) and \( \mu = (1, 0) \). Then \( \mu^\perp = [0, \mu_2^\perp] \) where \( \mu_2^\perp \sim \mathcal{N}(0, 1) \). The upper bound for the absolute generalization error at \( t = 1 \) can be rewritten as

\[
|\text{gen}_1| \lesssim \sqrt{\frac{(c_2 - c_1)^2}{2}} \int_{-\sqrt{\frac{2\pi}{\pi\sigma}}}^{\sqrt{\frac{2\pi}{\pi\sigma}}} e^{-\frac{y^2}{2\sigma^2}} \sqrt{G_x(1 - y^2)} \, dy, \tag{26}
\]

which is a decreasing function of \( n \), as shown in Figures 6(a) and 6(b).

Third, given any pair of \((\xi_0, \mu^\perp)\), if \( \alpha(\xi_0, \mu^\perp) > 0 \), \( F_\theta^{(t)}(\alpha(\xi_0, \mu^\perp)) > F_\theta^{(t-1)}(\alpha(\xi_0, \mu^\perp)) \) for all \( t \in [1 : \tau] \), as shown in Figure 3. This means that if the quality of the labelled data \( S_l \) is reasonably good, by using \( \theta_t \) which is learned from \( S_l \), the generated pseudo-labels for the unlabelled data are largely correct. Then the subsequent parameters \( \theta_t \), \( t \geq 1 \) learned from the large number of pseudo-labelled data examples can improve the generalization error. Therefore, the upper bound for \( |\text{gen}_{\mu}| \) decreases as \( t \) increases. In Figure 6(c), we plot the theoretical upper bound in (24) by ignoring \( \epsilon \). Unfortunately it is computationally difficult to numerically calculate the bound in (24) for high dimensions \( d \) (due to the need for high-dimensional numerical integration), but we can still gain insight from the result for \( d = 2 \). It is shown that the upper bound for \( |\text{gen}_{\mu}| \) decreases as \( t \) increases and finally converges to a non-zero constant. The gap between the upper bounds for \( |\text{gen}_1| \) and for \( |\text{gen}_{\mu}| \) decreases as \( t \) increases and shrinks to almost 0 for \( t \geq 2 \). The intuition is that as \( m \to \infty \), there are sufficient data at each iteration and the algorithm can converge at very early stage. In the empirical simulation, we let \( d = 50 \), \( \mu = (1, 0, \ldots, 0) \) and iteratively run the self-training procedure for 20 iterations and 2000 rounds. We find that the behaviour of the empirical generalization error (the red ‘-x’ line) is similar to the theoretical upper bound (the blue ‘-o’ line), which almost converges to its final value at \( t = 2 \). This result shows that the theoretical upper bound in (24) serves as a useful rule-of-thumb for how the generalization error changes over iterations. In Figure 6(d), we plot the theoretical bound and result from the empirical simulation based on the toy example for \( d = 2 \) but larger \( n \) and \( \sigma \). This figure shows that when we increase \( n \) and \( \sigma \), using unlabelled data may not be able to improve the generalization performance. The intuition is that for \( n \) large enough, merely using the labelled data can yield sufficiently low generalization error and for subsequent iterations with the pseudo-labelled data, the reduction in the test loss is negligible but the training loss will decrease more significantly (thus causing the generalization error to increase). When \( \sigma \) is larger, the data samples have larger variance and the classes have a larger overlap, and thus, the initial
which means that for large enough $n$. We find that when $w = \frac{n}{n + m}$ in (5). We can extend Theorem 2 to Corollary 3 as follows. Similarly to $F_\sigma$, let us define the enhanced correlation evolution function $\tilde{F}_{\sigma, \xi_0, \mu^+}(x) : [-1, 1] \rightarrow [-1, 1]$ as follows:

$$
\tilde{F}_{\sigma, \xi_0, \mu^+}(x) = \left(1 + \frac{w \sigma x^+ \|x\|_2^2 + (1 - w)(\frac{2\sigma \sqrt{\mathcal{E}}}{\sqrt{2\pi}} \exp(-\frac{x^2}{2\sigma^2}))^2}{(w(1 + \frac{\sigma}{\sqrt{n}} \xi_0) + (1 - w)(1 - 2Q(\frac{\xi_0}{\sigma}) + \frac{2\sigma}{\sqrt{2\pi}} \exp(-\frac{\xi_0^2}{2\sigma^2})))^2}\right)^{-\frac{1}{2}}. \quad (27)
$$

**Corollary 3.** Fix any $\sigma \in \mathbb{R}_+, \ d \in \mathbb{N}, \ e \in \mathbb{R}_+$ and $\delta \in (0, 1)$. For $m$ large enough, with probability at least $1 - \delta$, the absolute generalization error at any $t \in [1 : \tau]$ can be upper bounded as follows:

$$
\left|\text{gen}_t(P_Z, P_X, \{P_{\theta_k}|S_l, S_u\}_{k=1}^t, \{f_{\theta_k}|S_l, S_u\}_{k=1}^t)\right| \leq w \sqrt{\frac{(c_2 - c_1)^2 d}{4}} \log \frac{n}{n - 1} + (1 - w) \mathbb{E}_{\xi_0, \mu^+} \left[\sqrt{G_{\sigma}(\tilde{F}_{\sigma, \xi_0, \mu^+}^{(t-1)}(\alpha(\xi_0, \mu^+)), \xi_0, \mu^+)} + e\right]. \quad (28)
$$

The details are provided in Appendix D and the proof of Corollary 3 is provided in Appendix E. It can be seen from Figure 1 that the new upper bound for $|\text{gen}_t|$ remains as a decreasing function of $t$. We find that when $n = 10$, $m = 1000$, the upper bound is almost the same as that one in Figure 6(c) which means that for large enough $m \equiv \frac{n}{n}$, reusing the labelled data does not necessarily help to improve the generalization performance. Moreover, when $m = 100$, the upper bound is higher than that for $m = 1000$, which coincides with the intuition that increasing the number of unlabelled data helps to reduce the generalization error.

## 5 Experimental Results

In Sections 3 and 4, we theoretically analyse the upper bound of generalization error across the iterations for iterative self-training and especially for the case of bGMM classification. In this section, we conduct experiments on real datasets to demonstrate that our theoretical results on the bGMM example can also reflect the training dynamics on complicated tasks.

We train deep neural networks via a iterative self-learning strategy (under the same setting as that for Corollary 3) to perform binary and multi-class classification tasks. In the first iteration, we only use the labelled data to optimize the deep neural network (DNN) and train the model for a relatively large number of epochs so that the training loss will converge to a small value and the model is initialized well. In the following iterations, we first sample a subset of unlabelled data from the whole set and generate pseudo-labels for them via the model trained in the previous iteration. Then, we update the model for a small number of epochs with both the labelled and pseudo-labelled data.

**Experimental settings:** For binary classification, we collect pairs of classes of images, i.e., “automobile” and “truck”, “horse” and “ship”, from the CIFAR10 (Krizhevsky 2009) dataset. In this dataset, each class has 5000 images for training and 1000 images for testing. We use the whole set of images in the selected pair of categories and divide them into two sets, i.e., the labelled training set with 500 images and the unlabelled training set with 9500 images. We train a convolutional neural network, ResNet-10 (He et al. 2016), to minimize the cross-entropy loss via the self-learning strategy to perform the binary classification. The model is trained for 100 epochs in the first iteration and 20 epochs in the following iterations; we use the Adam (Kingma & Ba 2015) optimizer with a learning rate of 0.001. In each iteration after the initial one, we sample 2500 unlabelled images assign them pseudo-labels. The complete training procedure lasts for 100 self-training iterations.

We further validate our theoretical contributions on a multi-class classification problem in which we train a ResNet-6 model with the cross-entropy loss to perform 10-class handwritten digits classification on the MNIST (LeCun et al. 1998) dataset. We sample 51000 images from the training set, which contains 6000 images for each of the ten classes. We divide them into two sets, i.e., a labelled training set with 1000 images and an unlabelled set with 50000 images. The optimizer and training iterations follow those in the aforementioned binary classification tasks.

**Experimental observations:** We perform each experiment 3 times and report the average test and training (cross entropy) losses, the generalization error, and test and training accuracies in Figures 7–9. The generalization error appears to have relatively large reduction in the early training iterations and then...
fluctuates around a constant value afterwards. For example, in Figure 7, the generalization error converges to around 0.25 after 30 iterations; in Figure 8, it converges to around 0.4 after 10 iterations; in Figure 9, it converges to around 0.1 after 12 iterations. These results corroborate the theoretical and empirical analyses in the bGMM case, which again verifies the validity of the proposed generalization error bound in Theorem 2 and Corollary 3 on benchmark datasets. It also reveals that the generalization performance of iterative self-training on real datasets from relatively distinguishable classes can be quickly improved with the help of unlabelled data. We also show that the test accuracy increases with the iterations and has significant improvement compared to the initial iteration when only labelled data are used. In Figure 7, the highest accuracy has about a 4% increase from the initial point; in Figure 8, there is about a 10% increase; and in Figure 9, there is about a 3% increase. Thus, these numerical results suggest that via iterative self-training with pseudo-labelling, not only can we improve the generalization error as the iteration count increases, but we can also enhance the test accuracy. In addition, apart from the “horse-ship” and “automobile-truck” pairs (that are relatively easy to distinguish based on the high classification accuracy and low loss as shown in Figures 7 and 8), we also perform another experiment (detailed in Appendix F) on a harder-to-distinguish pair, “cat” and “dog” (see Table 1), whose results show that the generalization error does not decrease with the iterations even though the classification accuracy increases. This again corroborates the results in Figure 6(d) for the bGMM with large variance.

Figure 7: Binary classification on the “horse” and “ship” classes from the CIFAR10 dataset. Figure 8: Binary classification on the “automobile” and “truck” classes from the CIFAR10 dataset. Figure 9: 10-class classification on the MNIST handwritten digits dataset.

References

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 1567–1578, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 11192–11203, 2019.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (eds.). *Semi-Supervised Learning*. The MIT Press, 2006. ISBN 9780262033589. URL http://dblp.uni-trier.de/db/books/collections/CSZ2006.html.
Nitesh V Chawla and Grigoris Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 23:331–366, 2005.

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

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Generalization error bounds via Rényi-, $f$-divergences and maximal leakage. *IEEE Transactions on Information Theory*, 67(8):4986–5004, 2021. doi: 10.1109/TIT.2021.3085190.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 33:9925–9935, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic bounds on transfer generalization gap based on Jensen-Shannon divergence. *arXiv preprint arXiv:2010.09484*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master’s thesis, University of Toronto*, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, pp. 896, 2013.

Jian Li, Yong Liu, Rong Yin, and Weiping Wang. Multi-class learning using unlabeled samples: Theory and algorithm. In *IJCAI*, pp. 2880–2886, 2019.

Adrian Tovar Lopez and Varun Jog. Generalization error bounds using wasserstein distances. In 2018 *IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2018.

David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pp. 875–884. PMLR, 2020.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pp. 11013–11023, 2019.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Samet Oymak and Talha Cihad Gulcu. A theoretical characterization of semi-supervised learning with self-training for gaussian mixture models. In *International Conference on Artificial Intelligence and Statistics*, pp. 3601–3609. PMLR, 2021.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In 2018 *IEEE International Symposium on Information Theory (ISIT)*, pp. 546–550. IEEE, 2018.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240, 2016.
Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn’t. *Advances in Neural Information Processing Systems*, 21:1513–1520, 2008.

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

Isaac Triguero, Salvador García, and Francisco Herrera. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems*, 42(2):245–284, 2015.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

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

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

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

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. 2008.
Appendix

A Related works

Semi-supervised learning: There have been many existing results discussing about various methods of SSL. The book by Chapelle et al. (2006) presented a comprehensive overview of the SSL methods both theoretically and practically. Chawla & Karakolouš (2005) presented an empirical study of various SSL techniques on a variety of datasets and investigated sample-selection bias when the labelled and unlabelled data are from different distributions. Zhu (2008) classified various SSL methods into six main classes: generative models, low-density separation methods, graph-based methods, self-training and co-training. Pseudo-labelling is a technique among the self-training and co-training (Zhu & Goldberg, 2009). In self-training, the model is initially trained by the limited number of labelled data and generate pseudo-labels to the unlabelled data. Subsequently, the model is retrained with the pseudo-labelled data and repeats the process iteratively. It is a simple and effective SSL method without restrictions on the data samples (Triguero et al., 2015). A variety of works have also shown the benefits of utilizing the unlabelled data. Singh et al. (2008) developed a finite sample analysis that characterized how the unlabelled data improves the excess risk compared to the supervised learning, with respect to the number of unlabelled data and the margin between different classes. Li et al. (2019) studied multi-class classification with unlabelled data and provided a sharper generalization error bound using the notion of Rademacher complexity that yields a faster convergence rate. Carmon et al. (2019) proved that using unlabelled data can help to achieve high robust accuracy as well as high standard accuracy at the same time. Dupre et al. (2019) considered iteratively pseudo-labelling the whole unlabelled dataset with a confidence threshold and showed that the accuracy converges relatively quickly. Dymak & Gulcu (2021), in which part of our analysis hinges on, studied SSL under the binary Gaussian mixture model setup and characterized the correlation between the learned and the optimal estimators concerning the margin and the regularization factor. However, these works do not investigate how the unlabelled data affects the generalization error over the iterations.

Generalization error bounds: The traditional way of analyzing generalization error includes the Vapnik-Chervonenkis or VC dimension (Vapnik, 2000) and the Rademacher complexity (Boucheron et al., 2005). Recently, Russo & Zou (2016) proposed using mutual information between the estimated output of an algorithm and the actual realized value of the estimates to analyze and bound the bias in data analysis, which can be regarded equivalent to the generalization error. This new approach is simpler and can handle a wider range of loss functions compared to the abovementioned methods and other methods like differential privacy, total-variation information and so on. It also paves a new way to improving generalization capability of learning algorithms from an information-theoretic aspects. Following Russo & Zou (2016), Xu & Raginsky (2017) derived upper bounds on generalization error of learning algorithms with mutual information between the input dataset and the output hypothesis, which formalizes the intuition that less information that a learning algorithm can extract from training dataset leads to less overfitting. Later Pensia et al. (2018) derived generalization error bounds for noisy and iterative algorithms and the key contribution is to bound the mutual information between input data and output hypothesis. Negrea et al. (2019) improved mutual information bounds for Stochastic Gradient Langevin Dynamics (SGLD) via data-dependent estimates compared to distribution-dependent bounds. However, one major shortcoming of the abovementioned mutual information bounds is that the bounds go to infinity for (deterministic) learning algorithms without noise, e.g., Stochastic Gradient Descent (SGD). Some other works have tried to overcome this problem. Lopez & Jog (2018) derived upper bounds on the generalization error using the Wasserstein distance involving the distributions of input data and output hypothesis, which are shown to be tighter under some natural cases. Esposito et al. (2021) derived generalization error bounds via Rényi- $\phi$-divergences and maximal leakage. Steinke & Zakynthinou (2020) proposed using Conditional Mutual Information (CMI) to bound the generalization error, which can still preserve the chain rule property. Bu et al. (2020) provided a tightened upper bound based on the individual mutual information (IMI) between the individual data sample and the output. Wu et al. (2020) extended Bu et al. (2020)’s result to the transfer learning problems and characterized the upper bound based on IMI and KL-divergence. In a similar manner, Jose & Simeone (2020) provided a tightened bound on transfer generalization error based on the Jensen-Shannon divergence.

B Proof of Theorem 1

Before the proof, let us define some notation. The cumulant generating function (CGF) of a random variable $L \in \mathbb{R}$ is $\Lambda_L(\lambda) := \log \mathbb{E}_L[e^{\lambda (L - \mathbb{E}[L])}]$ for all $\lambda \in \mathbb{R}$. Note that $\Lambda_L(0) = \Lambda_L'(0) = 0$ and $\Lambda_L(\lambda)$ is convex. Then for any $L \sim \text{subG}(R)$, it means $\Lambda_L(\lambda) \leq \frac{R^2 \lambda^2}{2}$, for all $\lambda \in \mathbb{R}$. 

12
For any convex function \( \psi : [0, b] \mapsto \mathbb{R} \), its Legendre dual \( \psi^* \) is defined as \( \psi^*(x) := \sup_{\lambda \in [0, b)} \lambda x - \psi(\lambda) \) for all \( x \in [0, \infty) \). According to [Boucheron et al., 2013; Lemma 2.4], when \( \psi(0) = \psi'(0) = 0 \), \( \psi^*(x) \) is a nonnegative convex and nondecreasing function on \( [0, \infty) \). Moreover, for every \( y > 0 \), its generalized inverse function \( \psi^{*-1}(y) := \inf \{ x \geq 0 : \psi^*(x) \geq y \} \) is concave and can be rewritten as \( \psi^{*-1}(y) = \inf_{\lambda \in [0, b)} \frac{y + \psi(\lambda)}{\lambda} \).

We first introduce the following theorem that is applicable to more general loss functions.

**Theorem 4.** For any \( \hat{\theta}_t \in \Theta_t \), let \( \psi_\lambda(\hat{\theta}_t) \) and \( \psi_\lambda(\tilde{\theta}_t) \) be convex functions of \( \lambda \) and \( \psi_\lambda(0, \hat{\theta}_t) = \psi_\lambda'(0, \hat{\theta}_t) = \psi_\lambda(0, \hat{\theta}_t) = \psi_\lambda'(0, \hat{\theta}_t) = 0 \). Assume that \( \Lambda_t(\tilde{\theta}_t, \xi) \leq \psi_\lambda(\lambda, \tilde{\theta}_t) \) for all \( \lambda \in [0, b) \) and \( \Lambda_t(\tilde{\theta}_t, Z) \leq \psi_\lambda(\lambda, \tilde{\theta}_t) \) for \( \lambda \in (b_-, 0) \) under distribution \( P_{Z|\theta^{(t-1)}} = P_Z \), where \( 0 < b_+ \leq \infty \) and \(-\infty \leq b_- < 0 \). Let \( \psi_\lambda(\lambda) = \sup_{\theta_t} \psi_\lambda(\lambda, \tilde{\theta}_t) \) and \( \psi_\lambda(\lambda) = \sup_{\theta_t} \psi_\lambda(\lambda, \hat{\theta}_t) \). We have

\[
\begin{align*}
\text{gen}_t(P_Z, P_X, \{P_{\theta_t|S_i, Z_u}\}_{i=1}^{m-1}, \{f_{k_{t-1}}\}_{k=0}^n) \\
\leq \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{g(t-1)} \left[ \psi^{*-1}_\lambda(I_{g(t-1)}(\theta; Z)) \right] \\
+ \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{g(t-1)} \left[ \psi^{*-1}_\lambda(I_{g(t-1)}(\theta; X', Y')) + D_{\theta^{(t-1)}}(P_{X'|\theta^{(t-1)}}, P_P) \right],
\end{align*}
\]

\( \quad \text{(29)} \)

where \( P_{X'|\theta^{(t-1)}, Y'}(x, y) = P_{X}(x) \mathbb{I} \{ y = f_{t_{t-1}}(x) \} \) for any \( x \in X, y \in Y \) and \( \theta^{(t-1)} \in \Theta^{t-1} \), and \( P_{Z|\theta^{(t-1)}} = P_Z \).

**Proof.** Consider the Donsker–Varadhan variational representation of KL-divergence between any two distributions \( P \) and \( Q \) on \( \mathcal{X} \):

\[
D(P \| Q) = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}_{X \sim P}[g(X)] - \log \mathbb{E}_{X \sim Q}[e^{g(X)}] \right\}
\]

(31)

where the supremum is taken over the set of measurable functions in \( \mathcal{G} = \{ g : \mathcal{X} \mapsto \mathbb{R} : \mathbb{E}_{X \sim P}[e^{g(X)}] < \infty \} \).

Recall that \( \tilde{\theta}_t \) and \( \tilde{Z} \) are independent copies of \( \theta_t \) and \( Z \) respectively, such that \( P_{\tilde{\theta}_t, \tilde{Z}} = Q_{\theta_t} \otimes P_Z \). \( P_{\tilde{\theta}_t, \tilde{Z}|\theta^{(t-1)}} = P_{\theta_t|\theta^{(t-1)}} \otimes P_Z \). For any iterative SSL algorithm, by applying the law of total expectation, the generalization error can be rewritten as

\[
\begin{align*}
\text{gen}_t(P_Z, P_X, \{P_{\theta_t|S_i, Z_u}\}_{i=1}^{m-1}, \{f_{k_{t-1}}\}_{k=0}^n) \\
= w \left( \mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)] - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\theta_t, Z_u}[l(\theta_t, Z_u)] \right) \\
+ (1-w) \left( \mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)] - \frac{1}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta_t, X', Y'}[l(\theta_t, (X', Y'))] \right) \quad \text{(32)}
\end{align*}
\]

\[
\begin{align*}
= \frac{w}{n} \sum_{i=1}^{n} \left( \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z}) - \mathbb{E}_{\theta_t, Z_u}[l(\theta_t, Z_u)] \right) \\
+ \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \left( \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z}) - \mathbb{E}_{\theta_t, X', Y'}[l(\theta_t, (X', Y'))] \right) \quad \text{(33)}
\end{align*}
\]

\[
\begin{align*}
= \frac{w}{n} \sum_{i=1}^{n} \mathbb{E}_{g(t-1)} \left[ \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\theta_t, Z_u}[l(\theta_t, Z_u)|\theta^{(t-1)}] \right] \\
+ \frac{1-w}{m} \sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{g(t-1)} \left[ \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\theta_t, X', Y'}[l(\theta_t, (X', Y'))|\theta^{(t-1)}] \right]. \quad \text{(34)}
\end{align*}
\]
Note that \( \psi_+(\lambda) = \sup_{\tilde{\theta}_t} \psi_+(\lambda, \tilde{\theta}_t) \) and \( \psi_-(\lambda) = \sup_{\tilde{\theta}_t} \psi_-(\lambda, \tilde{\theta}_t) \) are convex, and so their Legendre duals \( \psi^*_+, \psi^*_- \), and the corresponding inverses are well-defined.

Let \( \bar{l}(\theta, z) = l(\theta, z) - \mathbb{E}_Z[l(\theta, Z)] \). We have the fact that \( \mathbb{E}_Z[l(\tilde{\theta}_t, \tilde{Z})] = 0 \) for any \( \tilde{\theta}_t \). Again, by the Donsker–Varadhan variational representation of the KL-divergence, for any fixed \( \theta^{(t-1)} \) and any \( \lambda \in [0, b_+), \) we have

\[
I_{\theta^{(t-1)}}(\theta_t; Z) = D(P_{\theta_t, Z}||P_{\theta_t^{(t-1)} \otimes P_Z}) \\
\geq \mathbb{E}_{\theta_t, Z}[\lambda(l(\theta_t, Z))|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[e^{\lambda(l(\tilde{\theta}_t, \tilde{Z}))}|\theta^{(t-1)}] \\
= \mathbb{E}_{\theta_t, Z}[\lambda(l(\theta_t, Z))|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[e^{\lambda(l(\tilde{\theta}_t, \tilde{Z}))}] \\
= \mathbb{E}_{\theta_t, Z}[\lambda(l(\theta_t, Z))|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[\exp(\Lambda_{\tilde{\theta}_t, \tilde{Z}}(\lambda, \tilde{\theta}_t))] \\
\geq \lambda \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\theta_t, Z) - \mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[\exp(\psi_+(\lambda, \tilde{\theta}_t))] \\
\geq \lambda \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\theta_t, Z) - \mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)}] - \psi_+(\lambda) \\
= \lambda(\mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\theta_t, Z)]|\theta^{(t-1)}) - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)})] - \psi_+(\lambda). 
\]

where (37) follows from the definition of \( \Lambda_{\tilde{\theta}_t, \tilde{Z}}(\lambda, \tilde{\theta}_t) \) in (52). (38) follows from the assumption that \( \Lambda_{\tilde{\theta}_t, \tilde{Z}}(\lambda, \tilde{\theta}_t) \leq \psi_+(\lambda, \tilde{\theta}_t) \) for all \( \lambda \in [0, b_+) \) and (39) follows since \( \psi_+(\lambda) = \sup_{\tilde{\theta}_t} \psi_+(\lambda, \tilde{\theta}_t) \). Thus, we have

\[
\mathbb{E}_{\theta_t, Z}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] \\
\leq \inf_{\lambda \in [0, b_+]} I_{\theta^{(t-1)}}(\theta_t; Z) + \psi_+(\lambda) \\
= \psi_+^{-1}(I_{\theta^{(t-1)}}(\theta_t; Z)). 
\]

Similarly, for \( \lambda \in (b_-, 0) \),

\[
\mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] \\
\leq \inf_{\lambda \in (b_-, 0)} I_{\theta^{(t-1)}}(\theta_t; Z) + \psi_-(\lambda) \\
= \psi_+^{-1}(I_{\theta^{(t-1)}}(\theta_t; Z)). 
\]

By applying the same techniques, for any pair of pseudo-labelled random variables \( (X', \tilde{Y}') \) used at iteration \( t \) and any \( \lambda \in [0, b_+) \), we have

\[
I_{\theta^{(t-1)}}(\theta_t; X', \tilde{Y}') + D_{\theta^{(t-1)}}(P_{X', \tilde{Y}'}||P_Z) \\
= D_{\theta^{(t-1)}}(P_{\theta_t, X', \tilde{Y}'}||P_{\theta_t} \otimes P_{X', \tilde{Y}'}, P_{\theta_t} \otimes P_Z) \\
\geq \mathbb{E}_{\theta_t, X', \tilde{Y}'}[\Lambda(\theta_t, (X', \tilde{Y}'))|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t} [\mathbb{E}_{X', \tilde{Y}'}[e^{\Lambda(\theta_t, (X', \tilde{Y}'))}|\theta^{(t-1)}]] \\
+ \mathbb{E}_{\theta_t} [\mathbb{E}_{X', \tilde{Y}'}[\Lambda(\theta_t, (X', \tilde{Y}'))]|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t} [\mathbb{E}_Z[e^{\lambda(l(\theta_t, Z))}|\theta^{(t-1)}]] \\
\geq \mathbb{E}_{\theta_t, X', \tilde{Y}'}[\Lambda(\theta_t, (X', \tilde{Y}'))|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t} [\mathbb{E}_Z[e^{\lambda(l(\theta_t, Z))}|\theta^{(t-1)}]] \\
- \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[\exp(\Lambda_{\tilde{\theta}_t, \tilde{Z}}(\lambda, \tilde{\theta}_t))] \\
\geq \lambda \mathbb{E}_{\theta_t, X', \tilde{Y}'}[\Lambda(\theta_t, (X', \tilde{Y}'))] - \mathbb{E}_{\theta_t} [\mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)})] - \psi_+(\lambda), 
\]

where (47) follows from the Jensen’s inequality. Thus, we get

\[
\mathbb{E}_{\theta_t, X', \tilde{Y}'}[l(\theta_t, (X', \tilde{Y}'))|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] \\
\leq \psi_+^{-1}(I_{\theta^{(t-1)}}(\theta_t; X', \tilde{Y}') + D_{\theta^{(t-1)}}(P_{X', \tilde{Y}'}||P_Z)) 
\]

and

\[
\mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\theta_t, X', \tilde{Y}'}[l(\theta_t, (X', \tilde{Y}'))|\theta^{(t-1)}] \\
\leq \psi_+^{-1}(I_{\theta^{(t-1)}}(\theta_t; X', \tilde{Y}') + D_{\theta^{(t-1)}}(P_{X', \tilde{Y}'}||P_Z)). 
\]

The proof is completed by applying inequalities (42), (44), (50) and (51) to the expansion of \( \psi_t \) in (34).

\[\square\]
Let $\tilde{\theta}_t$ and $\tilde{Z}$ be independent copies of $\theta_t$ and $Z$ respectively, such that $P_{\tilde{\theta}_t,\tilde{Z}} = Q_{\theta_t} \otimes P_Z$, where $Q_{\theta_t}$ is the marginal distribution of $\theta_t$. For any fixed $\tilde{\theta}_t \in \Theta$, let the cumulant generating function (CGF) of $l(\tilde{\theta}_t, \tilde{Z})$ be

$$
\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t) := \log E_{\tilde{Z}}[e^{\lambda l(\tilde{\theta}_t, \tilde{Z}) - E_{\tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})]}].
$$

(52)

When the loss function $l(\theta, Z) \sim \text{sub-G}(R)$ under $Z \sim P_Z$ for any $\theta \in \Theta$, we have $\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t) \leq \frac{R^2 \lambda^2}{2}$ for all $\lambda \in \mathbb{R}$. Then we can let $\psi_{-}(\lambda, \tilde{\theta}_t) = \psi_{+}(\lambda, \tilde{\theta}_t) = \frac{R^2 \lambda^2}{2}$ for all $\tilde{\theta}_t \in \Theta$. Hence, $\psi_{+}(\lambda) = \psi_{-}(\lambda) = \sup_{\tilde{\theta}_t \in \Theta} \frac{R^2 \lambda^2}{2}$ and $\psi_{-}^{-1}(y) = \psi_{+}^{-1}(y) = \sqrt{2R^2} y$ for any $y \geq 0$. Finally, Theorem $1$ can then be directly obtained from Theorem $1$.

### C Proof of Theorem 2

Theorem 2 can be proved iteratively by applying Theorem 1. For simplicity, in the following proofs, we abbreviate $\text{gen}_t(P_Z, P_{X}, \{P_{\theta_t|s_t, s_{n-t}}\}_{k=0}^{t-1}, \{\theta_{k=0}^{t-1}\})$ as $\text{gen}_t$.

#### 1. Initial round $t = 0$:

Since $Y_t X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2 I_d)$, we have $\theta_0 \sim \mathcal{N}(\mu, \frac{\sigma^2}{n} I_d)$ and for some constant $c \in \mathbb{R}_+$,

$$
\Pr(\theta_0 \in \Theta_{\mu,c}) = \Pr(||\theta_0 - \mu||_\infty \leq c) = \left(1 - 2 \Phi\left(\frac{-\sqrt{nc}}{\sigma}\right)\right)^d = 1 - \delta_{\sqrt{nc},d}.
$$

(53)

By choosing $c$ large enough, $\delta_{\sqrt{nc},d}$ can be made arbitrarily small. Consider $\tilde{\theta}_0$ and $(\tilde{X}, \tilde{Y})$ as independent copies of $\theta_0 \sim Q_{\theta_0}$ and $(X, Y) \sim P_{X,Y} = P_\gamma \otimes \mathcal{N}(\mu, \sigma^2 I_d)$, respectively, such that $P_{\tilde{\theta}_0, \tilde{X}, \tilde{Y}} = Q_{\theta_0} \otimes P_{X,Y}$. Then the probability that $l(\theta_0, (X, Y)) \sim \text{sub-G}(c^2 - c_1)/2$ under $(X, Y) \sim P_{X,Y}$ is given as follows

$$
\Pr\left(\Lambda_{l(\tilde{\theta}_0, \tilde{X}, \tilde{Y})}(\lambda, \tilde{\theta}_0) \leq \frac{\lambda^2(c^2 - c_1)^2}{8}\right)
\geq \Pr\left(\Lambda_{l(\tilde{\theta}_0, \tilde{X}, \tilde{Y})}(\lambda, \tilde{\theta}_0) \leq \frac{\lambda^2(c^2 - c_1)^2}{8} \quad \text{and} \quad \tilde{\theta}_0 \in \Theta_{\mu,c}\right)
= \Pr(\tilde{\theta}_0 \in \Theta_{\mu,c}) \Pr\left(\Lambda_{l(\tilde{\theta}_0, \tilde{X}, \tilde{Y})}(\lambda, \tilde{\theta}_0) \leq \frac{\lambda^2(c^2 - c_1)^2}{8} \mid \tilde{\theta}_0 \in \Theta_{\mu,c}\right)
= \left(1 - \delta_{\sqrt{nc},d}\right)^d (1 - c_{r,d}),
$$

(54)

(55)

(56)

(57)

where the last equality follows from (14) and (53).

Fix some $d \in \mathbb{N}$, $\epsilon > 0$ and $\delta \in (0,1)$. There exists $n_0(d, \delta) \in \mathbb{N}, r_0(d, \delta) \in \mathbb{R}_+$ such that for all $n > n_0, \epsilon > r_0, \delta_{\sqrt{nc},d} < \frac{\epsilon}{4}, \delta_{r,d} < \frac{\epsilon}{4}$, and then with probability at least $1 - \delta$, the absolute generalization error can be upper bounded as follows

$$
|\text{gen}_0| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{(c^2 - c_1)^2}{2 I(\theta_0; X_i, Y_i)}}.
$$

(58)

Then mutual information can be calculated as follows

$$
I(\theta_0; X_i, Y_i) = h(\theta_0) - h(\theta_0|X_i, Y_i)
= h\left(\frac{1}{n} \sum_{j=1}^{n} Y_j X_j\right) - h\left(\frac{1}{n} \sum_{j=1}^{n} Y_j X_j \mid X_i, Y_i\right)
= \frac{d}{2} \log \left(\frac{2\pi e \sigma^2}{n}\right) - h\left(\frac{1}{n} \sum_{j \in [n], j \neq i} Y_j X_j\right)
= \frac{d}{2} \log \left(\frac{2\pi e \sigma^2}{n}\right) - \frac{d}{2} \log \left(\frac{2\pi e (n-1) \sigma^2}{n^2}\right)
= \frac{d}{2} \log \frac{n}{n - 1}.
$$

(59)

(60)

(61)

(62)

(63)

Thus we can get (23).
2. Pseudo-label using $\theta_0$: For any $i \in [1 : m]$ and $X'_i \in S_u$, the pseudo-label is given by

$$\hat{Y}'_i = \text{sgn}(\theta_0^\top X'_i).$$

(64)

Given any pair of $(\xi_0, \mu^\perp)$, $\theta_0$ is fixed and $\{\hat{Y}'_i\}_{i=1}^m$ are conditionally i.i.d. from $P_{Y'_{\xi_0,\mu^\perp}} \in \mathcal{P}(Y)$. Recall the pseudo-labelled dataset is defined as $S_{u,1} = \{(X'_i, Y'_i)\}_{i=1}^m$.

Since $\theta_0 \sim \mathcal{N}(\mu, \frac{\sigma^2}{n} I_d)$, inspired by Oymak & Guler (2021), we can decompose it as follows:

$$\theta_0 = \mu + \frac{\sigma}{\sqrt{n}} \xi,$$

(65)

$$= \mu + \frac{\sigma}{\sqrt{n}} (\xi_0 \mu + \mu^\perp),$$

(66)

$$= \left(1 + \frac{\sigma}{\sqrt{n}} \xi_0\right) \mu + \frac{\sigma}{\sqrt{n}} \mu^\perp,$$

(67)

where $\xi \sim \mathcal{N}(0, I_d)$, $\xi_0 \sim \mathcal{N}(0, 1)$, $\mu^\perp \perp \mu, \mu^\perp \sim \mathcal{N}(0, I_d - \mu \mu^\top)$ and $\mu^\perp$ is independent of $\xi_0$.

Recall the correlation between $\theta_0$ and $\mu$ given in (19), the decomposition of $\theta_0$ in (20) and $\alpha, \beta$. Since $\text{sgn}(\theta_0^\top X'_i) = \text{sgn}(\theta_i^\top X'_i)$, in the following we can analyze the normalized parameter $\theta_0$ instead.

Given any $(\xi_0, \mu^\perp)$, $\alpha$ is fixed, and for any $i \in \mathbb{N}$, let us define a Gaussian noise vector $g_i \sim \mathcal{N}(0, I_d)$ and decompose it as follows

$$g_i = g_{0,i} \mu + g_i \mathbf{v} + g_i^\perp,$$

(68)

where $g_{0,i}, g_i \sim \mathcal{N}(0, 1)$, $g_i^\perp \sim \mathcal{N}(0, I_d - \mu \mu^\top - vv^\top), g_i^\perp \perp \mu, g_i^\perp \perp v$ and $g_{0,i}, g_i, g_i^\perp$ are mutually independent.

For any sample $X'_i \sim \mathcal{N}(\mu, \sigma^2 I_d)$, we can decompose it as

$$X'_i = \mu + \sigma g_i = \mu + \sigma (g_{0,i} \mu + g_i \mathbf{v} + g_i^\perp).$$

(69)

Then we have

$$\theta_0^\top X'_i = (\alpha \mu + \beta \mathbf{v})^\top (\mu + \sigma g_i),$$

(70)

$$= \alpha + \sigma (\alpha \mu + \beta \mathbf{v})^\top (g_{0,i} \mu + g_i \mathbf{v} + g_i^\perp),$$

(71)

$$= \alpha + \sigma (a g_{0,i} + \beta g_i),$$

(72)

$$= \alpha + \sigma h_i.$$  

(73)

Note that $h_i \sim \mathcal{N}(0, 1)$ for any $\alpha \in [-1, 1]$. Similarly, for any sample $X'_i \sim \mathcal{N}(-\mu, \sigma^2 I_d)$, we have

$$X'_i = -\mu + \sigma g_i,$$

(74)

and

$$\theta_0^\top X'_i = -\alpha + \sigma h_i.$$  

(75)

Denote the true label of $X'_i$ as $Y'_i$ and $P_{Y'_i} = P_Y \sim \text{unif}\{\{-1, +1\}\}$. The probability that the pseudo-label $\hat{Y}'_i$ is equal to 1 is given by

$$\Pr(\hat{Y}'_i = 1) = \Pr(\theta_0^\top X'_i > 0)$$

(76)

$$= \frac{1}{2} \Pr(\theta_0^\top X'_i > 0 | Y'_i = 1) + \frac{1}{2} \Pr(\theta_0^\top X'_i > 0 | Y'_i = -1)$$

(77)

$$= \frac{1}{2} \mathbb{E}_\alpha \left[ \Pr(\alpha + \sigma h_i > 0) \right] + \frac{1}{2} \mathbb{E}_\alpha \left[ \Pr\left(-\alpha + \sigma h_i > 0\right) \right]$$

(78)

$$= \frac{1}{2} \mathbb{E}_\alpha \left[ Q\left(-\frac{\alpha}{\sigma}\right) \right] + \frac{1}{2} \mathbb{E}_\alpha \left[ Q\left(\frac{\alpha}{\sigma}\right) \right] = \frac{1}{2}. $$

(79)

We also have $\Pr(\hat{Y}'_i = -1) = 1 - \Pr(\hat{Y}'_i = 1) = 1/2$, and so $P_{Y'_i} = P_Y$.

3. Iteration $t = 1$: Recall (18) and the new model parameter learned from the pseudo-labelled dataset $S_{u,1}$ is given by

$$\theta_1 = \frac{1}{m} \sum_{i=1}^m \hat{Y}'_i X'_i = \frac{1}{m} \sum_{i=1}^m \text{sgn}(\theta_0^\top X'_i) X'_i = \frac{1}{m} \sum_{i=1}^m \text{sgn}(\theta_0^\top X'_i) X'_i.$$

(80)
(a) First let us calculate the conditional expectation of $\theta_1$ given $\theta_0$.

Given any $(\xi_0, \mu^\perp)$, for any $j \in [1 : m]$, let $\mu_{\xi_0}^{\xi_0, \mu^\perp} := \mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j) X_j^\top | \xi_0, \mu^\perp]$ and $\mathbb{P}_{\xi_0, \mu^\perp}$ denotes the probability measure under the parameters $(\xi_0, \mu^\perp)$.

The expectation $\mu_{\xi_0}^{\xi_0, \mu^\perp}$ can be calculated as follows:

$$
\mu_{\xi_0}^{\xi_0, \mu^\perp} = \mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j) X_j^\top | \xi_0, \mu^\perp] = \mathbb{E}_Y[\mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j) X_j^\top | \xi_0, \mu^\perp, Y_j] \mid \xi_0, \mu^\perp, Y_j = -1] + \frac{1}{2} \mathbb{E}[\mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j) X_j^\top | \xi_0, \mu^\perp, Y_j = 1].
$$

Different from (68), here we decompose the Gaussian random vector $g_j \sim \mathcal{N}(0, L_d)$ in another way

$$
g_j = \tilde{g}_j \tilde{\theta}_0 + \tilde{g}_j^\perp,
$$

where $\tilde{g}_j \sim \mathcal{N}(0, 1)$, $\tilde{g}_j^\perp \sim \mathcal{N}(0, L_d - \tilde{\theta}_0 \tilde{\theta}_0^\top)$. $\tilde{g}_j$ and $\tilde{g}_j^\perp$ are mutually independent and $\tilde{g}_j^\perp \perp \tilde{\theta}_0$. Then we decompose $X_j^\top$ and $\tilde{\theta}_0 X_j^\top$ as

$$
X_j^\top = Y_j^\top \mu + \sigma \tilde{g}_j \tilde{\theta}_0 + \sigma \tilde{g}_j^\perp, \text{ and } \tilde{\theta}_0 X_j^\top = Y_j^\top \alpha + \sigma \tilde{g}_j.
$$

Then we have

$$
\mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j) X_j^\top | \xi_0, \mu^\perp, Y_j = -1] = \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j)(-\mu + \sigma \tilde{g}_j \tilde{\theta}_0 + \sigma \tilde{g}_j^\perp) | \xi_0, \mu^\perp]
$$

$$
= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j)|\xi_0, \mu^\perp|\mu + \sigma \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j) \tilde{\theta}_0 | \xi_0, \mu^\perp] + \sigma \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j) \tilde{g}_j^\perp | \xi_0, \mu^\perp]
$$

$$
= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j)|\xi_0, \mu^\perp|\mu + \sigma \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j) \tilde{g}_j | \xi_0, \mu^\perp] \tilde{\theta}_0,
$$

where (89) follows since $\tilde{g}_j^\perp$ is independent of $\tilde{g}_j$ and $\mathbb{E}[\tilde{g}_j^\perp] = 0$.

For the first term in (89), recall $\tilde{g}_j \sim \mathcal{N}(0, 1)$ and we have

$$
-\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j)|\xi_0, \mu^\perp|\mu = \left(1 - 2Q\left(\frac{\alpha}{\sigma}\right)\right)\mu.
$$

For the second term in (89), we have

$$
\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma \tilde{g}_j) \tilde{\theta}_0 | \xi_0, \mu^\perp] = \left(-\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} g \, dg + \int_{\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} g \, dg\right) \tilde{\theta}_0
$$

$$
= \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right) \tilde{\theta}_0.
$$

By combining (90) and (92), we have

$$
\mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j^\top) X_j^\top | \xi_0, \mu^\perp, Y_j = -1] = \left(1 - 2Q\left(\frac{\alpha}{\sigma}\right)\right) \mu + \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right) \tilde{\theta}_0.
$$

and similarly,

$$
\mathbb{E}[\mathrm{sgn}(\theta_0^\top X_j^\top) X_j^\top | \xi_0, \mu^\perp, Y_j = 1] = \left(2Q\left(-\frac{\alpha}{\sigma}\right) - 1\right) \mu + \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right) \tilde{\theta}_0.
$$

Thus, recall $\tilde{\theta}_0 = \alpha \mu + \beta \nu$ and $\mu_{\xi_0}^{\xi_0, \mu^\perp}$ is given by

$$
\mu_{\xi_0}^{\xi_0, \mu^\perp} = \mathbb{E}[\mathrm{sgn}(\tilde{\theta}_0^\top X_j^\top) X_j^\top | \xi_0, \mu^\perp]
$$

$$
= \left(1 - 2Q\left(\frac{\alpha}{\sigma}\right)\right) \mu + \frac{2\sigma}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right) \tilde{\theta}_0
$$

$$
= \left(1 - 2Q\left(\frac{\alpha}{\sigma}\right) + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right) \mu + \frac{2\sigma \beta}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right) \nu.
$$
The $l_\infty$ norm between $\mu_1^{\tilde{\epsilon},\mu^+}$ and $\mu$ can be upper bounded by

$$
\|\mu_1^{\tilde{\epsilon},\mu^+} - \mu\|_\infty \\
\leq \sqrt{\left(-2Q\left(\frac{\alpha}{\sigma}\right) + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)^2 + \frac{2\sigma^2 \beta^2}{\pi} \exp\left(-\frac{2\alpha^2}{2\sigma^2}\right)} \quad (97)
$$

$$
< \sqrt{\left(2\Phi\left(\frac{1}{\sigma}\right) + \frac{2\sigma}{\sqrt{2\pi}}\right)^2 + \frac{2\sigma^2}{\pi}} =: \tilde{c}_1, \quad (98)
$$

where $\tilde{c}_1$ is a constant only dependent on $\sigma$.

(b) Next, we need to calculate the probability that $l(\theta_1,(X,Y)) \sim \text{subG}(c_2 - c_1/2)$ under $(X,Y) \sim P_X \cdot \mathcal{N}$.

Let $V_i = \text{sgn}(\bar{\theta}_1 X_i') X_i' - \mu_1^{\tilde{\epsilon},\mu^+}$. For any $k \in [1:d]$, let $V_{i,k}, \theta_{1,k}, \mu_{1,k}$ denote the $k$-th components of $V_i$, $\theta_1$ and $\mu_1^{\tilde{\epsilon},\mu^+}$, respectively. Recall the decompositions $X_i' = Y_i' \mu + \sigma \tilde{g}_0 + \sigma \tilde{g}_i$ in (85) and $\bar{\theta}_0 X_i' = Y_i' \alpha + \sigma \tilde{g}_i$ in (86). Suppose the basis of $\mathbb{R}^d$ is denoted by $B = \{v_1, \ldots, v_d\}$ and let $v_1 = \theta_0$. Then we have

$$
\tilde{g}_i^\perp = \tilde{g}_{i,2} v_2 + \cdots + \tilde{g}_{i,d} v_d, \quad (99)
$$

where $\tilde{g}_{i,k} \sim \mathcal{N}(0,1)$ for any $k \in [2:d]$ and $\{\tilde{g}_{i,k}\}_{k=2}^d$ are mutually independent. We also let $\mu = (\mu_{0,1}, \ldots, \mu_{0,d})$.

Given any $(\xi_0,\mu^+)$, the moment generating function (MGF) of $V_{i,1}$ is given as follows: for any $s_1 > 0$,

$$
\mathbb{E}_{V_i,1}[e^{s_1 V_{i,1}}] = Q\left(-\frac{\alpha}{\sigma}\right) \mathbb{E}_{\tilde{g}_1} \left[e^{s_1(\mu_0,1-\mu_1,1+\sigma \tilde{g}_1)} | \tilde{g}_1 > -\frac{\alpha}{\sigma}\right] + Q\left(\frac{\alpha}{\sigma}\right) \mathbb{E}_{\tilde{g}_1} \left[e^{s_1(-\mu_0,1-\mu_1,1+\sigma \tilde{g}_1)} | \tilde{g}_1 > \frac{\alpha}{\sigma}\right] \quad (100)
$$

$$
= e^{s_1(\mu_0,1-\mu_1,1)} e^{\frac{s_1^2}{2} \Phi\left(\frac{\alpha}{\sigma} + \sigma s_1\right)} + e^{s_1(-\mu_0,1-\mu_1,1)} e^{\frac{s_1^2}{2} \Phi\left(-\frac{\alpha}{\sigma} + \sigma s_1\right)}. \quad (101)
$$

The final equality follows from the fact that the MGF of a zero-mean univariate Gaussian truncated to $(a,b)$ is $e^{ax^2/2} \left[\frac{\Phi(b-a) - \Phi(a-x)}{\Phi(b) - \Phi(a)}\right]$. The second derivative of log $\mathbb{E}_{V_i,1}[e^{s_1 V_{i,1}}]$ is given as

$$
\tilde{R}_{i}(s_1) := \frac{d^2 \log \mathbb{E}_{V_i,1}[e^{s_1 V_{i,1}}]}{ds_1^2} \quad (102)
$$

$$
\leq \sigma^2 + \frac{\sigma}{Q\left(\frac{\alpha}{\sigma} + \sigma s_1\right) e^{s_1 \mu_{0,k}}} + \frac{\sigma}{Q\left(\frac{\alpha}{\sigma} + \sigma s_1\right) e^{-s_1 \mu_{0,k}}} < \infty. \quad (103)
$$

For $k \in [2:d]$ and any $s_k > 0$, the MGF of $V_{i,k}$ is given as

$$
\mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}] = \mathbb{E}_{\tilde{g}_{i,k}} e^{s_k V_{i,1} - \mu_{1,k} \tilde{g}_{i,k}} \left[\mathcal{N}\left(\mu_{0,k} \tilde{g}_{i,k}, \sigma^2 \tilde{g}_{i,k}^2\right)\right] \quad (104)
$$

$$
= Q\left(-\frac{\alpha}{\sigma}\right) e^{s_k(\mu_{0,k}-\mu_{1,k})} e^{\frac{s_k^2}{2} \Phi\left(\frac{\alpha}{\sigma} + \sigma s_1\right)} + Q\left(\frac{\alpha}{\sigma}\right) e^{s_k(-\mu_{0,k}+\mu_{1,k})} e^{\frac{s_k^2}{2} \Phi\left(-\frac{\alpha}{\sigma} + \sigma s_1\right)}, \quad (105)
$$

and the second derivative of log $\mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}]$ is given by

$$
\tilde{R}_{k}(s_k) := \frac{d^2 \log \mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}]}{ds_k^2} = \sigma^2 + \frac{4 \mu_{0,k}^2 Q\left(-\frac{\alpha}{\sigma}\right) Q\left(\frac{\alpha}{\sigma}\right)}{(Q\left(\frac{\alpha}{\sigma}\right) e^{s_k \mu_{0,k}} + Q\left(\frac{\sigma}{\sigma}\right) e^{-s_k \mu_{0,k}})^2}. \quad (106)
$$

Fix $k \in [1:d]$. According to Taylor’s theorem, we have

$$
\log \mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}] = \tilde{R}_{k}(\xi_{L,k}) \frac{s_k^2}{2}, \quad (107)
$$

for some $\xi_{L,k} \in (0,s_k)$ and $\tilde{R}_{k}(\xi_{L,k}) < \infty$. Then the Cramér transform of log $\mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}]$ can be lower bounded as follows: for any $\varepsilon > 0$,

$$
\sup_{s_k > 0} \left( s_k \varepsilon - \log \mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}] \right) \geq \sup_{s_k > 0} \left( s_k \varepsilon - \frac{\tilde{R}_{k}(\xi_{L,k}) s_k^2}{2} \right) = \frac{\varepsilon^2}{2 \tilde{R}_{k}(\xi_{L,k})}. \quad (108)
$$
Let \( \bar{R}^* = \max_{\xi_0, \mu^+} \min_{k \in [1 : |d|]} \bar{R}_k(\xi_0, k) \), which is a finite constant only dependent on \( \sigma \). Since \( \{\text{sgn}(\theta_0^i | X_i X_i')\}_{i=1}^m \) are i.i.d. random variables conditioned on \( (\xi_0, \mu^+) \), by applying Chernoff-Cramér inequality, we have for all \( \varepsilon > 0 \)

\[
\mathbb{P}_{\xi_0, \mu^+} \left( \| \theta_1 - \mu_1 \xi_0, \mu^+ \|_\infty > \varepsilon \right) \\
= \mathbb{P}_{\xi_0, \mu^+} \left( \max_{k \in [1 : |d|]} | \theta_{1,k} - \mu_{1,k} | > \varepsilon \right) \\
\leq \sum_{k=1}^d \mathbb{P}_{\xi_0, \mu^+} \left( | \theta_{1,k} - \mu_{1,k} | > \varepsilon \right) \\
= \sum_{k=1}^d \mathbb{P}_{\xi_0, \mu^+} \left( \left| \frac{1}{m} \sum_{i=1}^m V_{i,k} \right| > \varepsilon \right) \\
\leq \sum_{k=1}^d 2 \exp \left( - m \sup_{s > 0} \left( s \varepsilon - \log \mathbb{E}_{V_{i,k}} [e^{s V_{i,k}}] \right) \right) \\
\leq 2d \exp \left( - \frac{m \varepsilon^2}{2 R^*} \right) \\
= \delta_{m, \varepsilon, d},
\]

where \( \delta_{m, \varepsilon, d} \xrightarrow{\Delta} 0 \) as \( m \to \infty \) and does not depend on \( \xi_0, \mu^+ \).

Choose some \( c \in (\bar{c}_1, \infty) \) (\( \bar{c}_1 \) defined in (108)). We have

\[
\mathbb{P}_{\xi_0, \mu^+} \left( \theta_1 \in \Theta_{\mu, c} \right) \geq \mathbb{P}_{\xi_0, \mu^+} \left( \| \theta_1 - \mu_1 \xi_0, \mu^+ \|_\infty \leq c - \bar{c}_1 \right) \geq 1 - \delta_{m, c - \bar{c}_1, d}.
\]

Consider \( \tilde{\theta}_1 \) as an independent copy of \( \theta_1 \) and independent of \( (\tilde{X}, \tilde{Y}) \). Then the probability that \( l(\theta_1, (X, Y)) \sim \text{subG}((c_2 - c_1)/2) \) under \( (X, Y) \sim P_{X, Y} \) is given as follows

\[
\mathbb{P}_{\xi_0, \mu^+} \left( \Lambda_l(\tilde{\theta}_1, (\tilde{X}, \tilde{Y}))(\lambda, \tilde{\theta}_1) \leq \frac{\lambda^2(c_2 - c_1)^2}{8} \right) \\
\geq \mathbb{P}_{\xi_0, \mu^+} \left( \tilde{\theta}_1 \in \Theta_{\mu, c} \right) \mathbb{P}_{\xi_0, \mu^+} \left( \Lambda_l(\tilde{\theta}_1, (\tilde{X}, \tilde{Y}))(\lambda, \tilde{\theta}_1) \leq \frac{\lambda^2(c_2 - c_1)^2}{8} | \tilde{\theta}_1 \in \Theta_{\mu, c} \right) \\
= (1 - \delta_{m, c - \bar{c}_1, d})(1 - \delta_{r, d}).
\]

Thus, for some \( c \in (\bar{c}_1, \infty) \), with probability at least \( (1 - \delta_{m, c - \bar{c}_1, d})(1 - \delta_{r, d}) \), the absolute generalization error can be upper bounded as follows:

\[
|\text{gen}_1| = |E[L_{P_{X}}(\theta_1) - L_{\hat{S}_n,1}(\theta_1)]| \\
= \left| \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_0, \mu^+} \left[ E \left[ l(\theta_1, (X, Y)) - l(\theta_1, (X'_i, Y'_i)) \| \xi_0, \mu^+ \right] \right] \right| \\
\leq \frac{1}{m} \sum_{i=1}^m \mathbb{E}_{\xi_0, \mu^+} \left[ \sqrt{\frac{(c_2 - c_1)^2}{2}} \left( I_{\xi_0, \mu^+}(\theta_1, (X'_i, Y'_i)) + D_{\xi_0, \mu^+}(P_{X'_i, Y'_i} \| P_{X, Y}) \right) \right],
\]

where \( P_{\theta_1, (X, Y)}(\xi_0, \mu^+) = Q_{\theta_1}(\xi_0, \mu^+) \otimes P_{X, Y} \) and \( Q_{\theta_1}(\xi_0, \mu^+) \) denotes the marginal distribution of \( \theta_1 \) under parameters \( (\xi_0, \mu^+) \).

In the following, we derive the closed form expressions of the mutual information and KL-divergence in (121). For any \( j \in [1 : m] \):

- **Calculate** \( I_{\xi_0, \mu^+}(\theta_1; X'_i, Y'_i) \): For arbitrary random variables \( X \) and \( U \), we define the **disintegrated conditional differential entropy** of \( X \) given \( U \) as
  \[
  h_U(X) := h(P_{X|U}).
  \]

Conditioned on a certain pair of \( (\xi_0, \mu^+) \), the mutual information between \( \theta_1 \) and \( (X'_i, Y'_i) \) is given
by
\[
I_{\xi_{0}, \mu^{+}}(\theta; X', \hat{Y}'') = h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{i=1}^{m} \text{sgn}(\theta_{0}' X'_{i}) X'_{i}\right) - h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{j=1}^{m} \bar{Y}'_{j} X'_{j} \right)
\]
\[
= h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{i=1}^{m} \text{sgn}(\theta_{0}' X'_{i}) X'_{i}\right) - h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{i \in [m], i \neq j} \text{sgn}(\theta_{0}' X'_{i}) X'_{i}\right)
\]
\[
= h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{i=1}^{m} \text{sgn}(\theta_{0}' X'_{i}) X'_{i}\right)
- h_{\xi_{0}, \mu^{+}}\left(\frac{1}{m} \sum_{i \in [m], i \neq j} \text{sgn}(\theta_{0}' X'_{i}) X'_{i}\right) - d \log \frac{m - 1}{m}.
\] (125)

As \( m \to \infty \), \( I_{\xi_{0}, \mu^{+}}(\theta; X', \hat{Y}'') \to 0 \) almost surely and hence, in probability. Thus, for any \( \epsilon > 0 \), and there exists \( m_{0}(\epsilon, d, \delta) \in \mathbb{N} \) such that for all \( m > m_{0} \),
\[
\mathbb{P}_{\xi_{0}, \mu^{+}}(I_{\xi_{0}, \mu^{+}}(\theta; X', \hat{Y}'') > \epsilon) \leq \delta.
\] (126)

**Calculate \( D_{\xi_{0}, \mu^{+}}(P_{X|Y'}, P_{Y|X}) \):** First of all, since \( P_{Y'} = P_{Y} \) (cf. [79]) regardless of the values of \( (\xi_{0}, \mu^{+}) \), the **disintegrated conditional KL-divergence** can be rewritten as
\[
D_{\xi_{0}, \mu^{+}}(P_{X|Y'}, P_{Y|X}) = P_{Y'}(1)D_{\xi_{0}, \mu^{+}}(P_{X|Y'=1} || P_{Y|X=1}) + P_{Y'}(1)D_{\xi_{0}, \mu^{+}}(P_{X|Y'=1} || P_{X|Y=1}).
\] (127)

Recall the decomposition of a Gaussian vector \( \vec{g}_{y} \sim \mathcal{N}(0, \mathbf{I}_{d}) \) in [84]. Note that \( \text{rank}(\text{Cov}(\vec{g}_{y}^\top)) = \text{rank}(\mathbf{I}_{d} - \theta \theta^\top) = d - 1 \). For any pair of labelled data sample \((X, Y)\), from [85], we similarly decompose \( X \) as \( X = Y \mu + \sigma(\bar{g}_{0} + \bar{g}^{\perp}) \), where \( \bar{g} \sim \mathcal{N}(0, 1) \) and \( \bar{g}^{\perp} \sim \mathcal{N}(0, \mathbf{I}_{d} - \theta \theta^\top) \). Let \( p_{\bar{g}} \) and \( p_{\bar{g}^{\perp}} \) denote the probability density functions of \( \bar{g} \) and \( \bar{g}^{\perp} \), respectively. For any \( x = \mu + \sigma(u \theta_{0} + u^{\perp}) \in \mathbb{R}^{d} \), the joint probability distribution at \((X, Y) = (x, 1)\) is given by
\[
P_{X,Y}(x, 1) = P_{Y}(1)p_{\mu}(x|1)
\]
\[
= \frac{P_{Y}(1)}{\sqrt{2\pi \sigma_{d}^{2}}} \exp\left(-\frac{1}{2\sigma^{2}}(x - y \mu)^{\top}(x - y \mu)\right)
\] (128)
\[
= \frac{P_{Y}(1)}{\sqrt{2\pi \sigma_{d}^{2}}} \exp\left(-\frac{1}{2\sigma^{2}}(\sigma u \theta + \sigma u^{\perp})^{\top}(\sigma u \theta + \sigma u^{\perp})\right)
\] (129)
\[
= \frac{P_{Y}(1)}{\sqrt{2\pi \sigma_{d}^{2}}} \exp\left(-\frac{u^{2}}{2}\sigma \right) \exp\left(-\frac{(u^{\perp})^{\top}u^{\perp}}{2}\right)
\] (130)
\[
= P_{Y}(1)p_{\bar{g}}(u)p_{\bar{g}^{\perp}}(u^{\perp}).
\] (131)

Similarly, for any \( x = -\mu + \sigma(u \theta_{0} + u^{\perp}) \in \mathbb{R}^{d} \), the joint probability density evaluated at \((X, Y) = (x, -1)\) is given by
\[
P_{X,Y}(x, -1) = P_{Y}(-1)p_{\mu}(x|1) = P_{Y}(-1)p_{\bar{g}}(u)p_{\bar{g}^{\perp}}(u^{\perp}).
\] (132)

Second, we have \( P_{X'|Y'} = \sum_{y \in \{-1, 1\}} P_{X|Y,y} P_{Y'|y} \). The conditional probability distribution \( P_{Y'|Y'} \) can be calculated as follows
\[
P_{Y'|Y'} = \frac{P_{Y'|Y'} P_{Y'}}{P_{Y'}} = P_{Y'|Y'},
\] (133)
where the last equality follows since \( P_{Y'}(-1) = P_{Y'}(1) = P_{Y'}(1) = 1/2 \). Since \( \bar{Y}' = \text{sgn}(Y' \alpha + \sigma \bar{g}_{y}) \) (cf. [86]), we have
\[
P_{Y'|Y'}(-1| -1) = \text{Pr}(Y' \alpha + \sigma \bar{g}_{y} < 0|Y' = -1) = Q\left(-\frac{\alpha}{\sigma}\right),
\] (134)
and similarly,

\[ P_{Y_j|Y_j}(1|1) = Q\left(\frac{\alpha}{\sigma}\right), \quad P_{Y_j|Y_j}(-1|1) = Q\left(\frac{-\alpha}{\sigma}\right). \]  

(135)

Thus, we conclude that

\[ P_{Y_j|Y_j}(y_j'|y_j) = \begin{cases} 
Q\left(\frac{-\alpha}{\sigma}\right) & y_j' = y_j' \\
Q\left(\frac{\alpha}{\sigma}\right) & y_j' \neq y_j'. 
\end{cases} \]  

(136)

To calculate the conditional probability distribution \( P_{X_j|Y_j'} Y_j' \), recall the decomposition of \( X_j' \) and \( \theta_j X_j' \) in (85) and (86). Since the event \( \{Y_j' = -1, Y_j' = -1\} \) is equivalent to \( \{\tilde{g}_j < \alpha/\sigma\} \) and \( \tilde{g}_j \sim N(0, 1) \), the conditional density of \( \tilde{g}_j \) given \( Y_j' = -1, Y_j' = -1 \) is given by

\[ p_{\tilde{g}_j|Y_j', Y_j'}(u|1, -1) = p_{\tilde{g}_j|\tilde{g}_j \leq \alpha/\sigma}(u) = \frac{u \leq \alpha/\sigma}{\Phi(\alpha/\sigma)}, \quad \forall u \in \mathbb{R}. \]  

(137)

Similarly, for any \( u \in \mathbb{R} \)

\[ p_{\tilde{g}_j|Y_j', Y_j'}(u|1, 1) = p_{\tilde{g}_j|\tilde{g}_j > \alpha/\sigma}(u) = \frac{u > \alpha/\sigma}{\Phi(-\alpha/\sigma)}. \]  

(138)

\[ p_{\tilde{g}_j|Y_j', Y_j'}(u|1, -1) = p_{\tilde{g}_j|\tilde{g}_j > \alpha/\sigma}(u) = \frac{u > -\alpha/\sigma}{\Phi(-\alpha/\sigma)}. \]  

(139)

\[ p_{\tilde{g}_j|Y_j', Y_j'}(u|1, 1) = p_{\tilde{g}_j|\tilde{g}_j \leq \alpha/\sigma}(u) = \frac{u \leq -\alpha/\sigma}{\Phi(-\alpha/\sigma)}. \]  

(140)

For any \( x = \mu + \sigma(u\tilde{\theta}_0 + u^\perp) \in \mathbb{R}^d \), given \( Y_j' = 1, Y_j' = 1 \), the conditional probability distribution at \( X_j' = x \) is given by

\[ P_{X_j|Y_j'}(x|1, 1) = P_{\mu + \sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(x|1, 1) \]  

(141)

\[ = P_{\sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(\sigma(u\tilde{\theta}_0 + u^\perp)|1, 1) \]  

(142)

\[ = p_{\tilde{g}_j|Y_j', Y_j'}(u|1, 1)p_{\tilde{g}_+^\perp}(u^\perp), \]  

(143)

where (143) follows since \( \tilde{g}_j \) and \( \tilde{g}_+^\perp \) are mutually independent and \( \tilde{\theta}_0 \perp \tilde{g}_+^\perp \).

Since we can decompose \( 2\mu/\sigma \) as

\[ 2\mu/\sigma = \frac{2\alpha\tilde{\theta}_0 + 2\beta^2 \mu - 2\alpha \beta \nu}{\sigma} = \frac{2\alpha}{\sigma} \tilde{\theta}_0 + \tilde{\theta}_0^+, \]  

(144)

given \( Y_j' = 1, Y_j' = 1 \), the conditional probability distribution at \( X_j' = x \) is given by

\[ P_{X_j|Y_j'}(x|1, -1) = P_{\mu + \sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(x|1, -1) \]  

(145)

\[ = P_{\sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(\sigma(u\tilde{\theta}_0 + u^\perp)|1, -1) \]  

(146)

\[ = p_{\tilde{g}_j|Y_j', Y_j'}(u + \frac{2\alpha}{\sigma}|1, -1)p_{\tilde{g}_+^\perp}(u^\perp + \tilde{\theta}_0^+). \]  

(147)

Similarly, for any \( x = -\mu + \sigma(u\tilde{\theta}_0 + u^\perp) \in \mathbb{R}^d \), given \( Y_j' = -1, Y_j' = 1 \), the conditional distribution at \( X_j' = x \) is given by

\[ P_{X_j|Y_j'}(x|-1, 1) = P_{\mu + \sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(-\mu + \sigma(u\tilde{\theta}_0 + u^\perp)|-1, 1) \]  

(148)

\[ = p_{\tilde{g}_j|Y_j', Y_j'}(u + \frac{2\alpha}{\sigma}|1, 1)p_{\tilde{g}_+^\perp}(u^\perp + \tilde{\theta}_0^+). \]  

(149)

and given \( \tilde{Y}_j' = -1, Y_j' = -1 \),

\[ P_{X_j|Y_j'}(x|-1, -1) = P_{\mu + \sigma \tilde{g}_j \theta_0 + \sigma \tilde{g}_+^\perp Y_j', Y_j'}(-\mu + \sigma(u\tilde{\theta}_0 + u^\perp)|-1, -1) \]  

(150)

\[ = p_{\tilde{g}_j|Y_j', Y_j'}(u - 1, -1)p_{\tilde{g}_+^\perp}(u^\perp). \]  

(151)
Furthermore, for any \( x = -\mu + \sigma(u\theta_0 + u^+) \in \mathbb{R}^d \), we have
\[
P_{X_j|\hat{Y}_j=-1}(x) = \sum_{y \in \{-1,+1\}} P_{X_j|\hat{Y}_j=-1,Y_j=y}(x)P_{Y_j|\hat{Y}_j=-1}(y)
\] (152)
\[
= P_{Y_j|\hat{Y}_j=-1}(1)p_{\hat{Y}_j,Y_j}(u - \frac{2\alpha}{\sigma} - 1, 1)p_{\delta_j}(u^- + \theta_j^-) + P_{Y_j|\hat{Y}_j=-1}(-1)p_{Y_j}(u^- - \theta_j^-) + \mathbb{I}\left\{u \leq \frac{\alpha}{\sigma}\right\}p_{\delta_j}(u)p_{\delta_j}(u^+); \quad (153)
\]
for any \( x = \mu + \sigma(u\theta_0 + u^+) \in \mathbb{R}^d \), we have
\[
P_{X_j|\hat{Y}_j=1}(x) = \sum_{y \in \{-1,+1\}} P_{X_j|\hat{Y}_j=1,Y_j=y}(x)P_{Y_j|\hat{Y}_j=1}(y)
\] (155)
\[
= \mathbb{I}\left\{u > -\frac{\alpha}{\sigma}\right\}p_{\delta_j}(u + \frac{2\alpha}{\sigma})p_{\delta_j}(u^+) + \mathbb{I}\left\{u > -\frac{\alpha}{\sigma}\right\}p_{\delta_j}(u)p_{\delta_j}(u^+). \quad (156)
\]
Define the set \( \mathcal{U}^+_{d_0}(\xi_0, \mu^+) := \{u^+ \in \mathbb{R}^d : u^+ \perp \theta_0\} \). We also use \( \mathcal{U}^+_{d_0} \) to represent \( \mathcal{U}^+_{d_0}(\xi_0, \mu^+) \), if there is no risk of confusion. Recall (21) and note that \( \int_{\mathcal{U}^+_{d_0}} p_{\delta_j}(u) du^+ = 1 \). Finally, the KL-divergence is given by
\[
D_{\xi_0,\mu^+}(P_{X_j|\hat{Y}_j=-1}\|P_{X|Y=-1}) = \int_{\mathcal{U}^+_{d_0}} \int_{-\infty}^{\frac{\alpha}{\sigma}} \left(p_{\delta_j}(u - \frac{2\alpha}{\sigma})p_{\delta_j}(u^+) + p_{\delta_j}(u)p_{\delta_j}(u^+)^2\right) du^+ \quad (157)
\]
\[
G_\sigma(\alpha, \xi_0, \mu^+) \quad (158)
\]
and
\[
D_{\xi_0,\mu^+}(P_{X_j|\hat{Y}_j=1}\|P_{X|Y=1}) = \int_{\mathcal{U}^+_{d_0}} \int_{-\infty}^{\frac{\alpha}{\sigma}} \left(p_{\delta_j}(u + \frac{2\alpha}{\sigma})p_{\delta_j}(u^+) + p_{\delta_j}(u)p_{\delta_j}(u^+)^2\right) du^+ \quad (159)
\]
\[
G_\sigma(\alpha, \xi_0, \mu^+) \quad (160)
\]
where (160) follows from \( p_{\delta_j} \) and \( p_{\delta_j} \) are zero-mean Gaussian distributions. Then from (127), we have
\[
D_{\xi_0,\mu^+}(P_{X_j|\hat{Y}_j|\|P_{X|Y}) = G_\sigma(\alpha, \xi_0, \mu^+). \quad (161)
\]
Thus, by combining the aforementioned results, we get the closed-form expression of the upper bound for \([\text{gen1}].\) Indeed, if we fix some \( d \in \mathbb{N}, \epsilon > 0 \) and \( \delta \in (0, 1) \), there exists \( n_0(d) \in \mathbb{N}, m_0(\epsilon, d, \delta) \in \mathbb{N}, \) \( c_0(d, \delta) \in (c_1, \infty), \) \( c_0(d, \delta) \in \mathbb{R}^+ \) such that for all \( n > n_0, m > m_0, c > c_0, \) \( r > r_0, \delta = c_0 - c_1 < \frac{d}{\delta}, \) \( \delta < \frac{d}{\delta}, \) and with probability at least \( 1 - \delta, \)
\[
|\text{gen1}| \leq \sqrt{\frac{(C_2 - c_1)^2}{2} E_{\xi_0,\mu^+} \left[ \sqrt{G_\sigma(\alpha(\xi_0, \mu^+), \xi_0, \mu^+) + \epsilon} \right]. \quad (162)
\]
4. Pseudo-label using \( \theta_1 \): Let \( \theta_1 := \theta_1/\|\theta_1\|_2. \) For any \( i \in [m + 1 : 2m] \), the pseudo-labels are given by
\[
\hat{Y}_i = \text{sgn}(\theta_1^T X_i) = \text{sgn}(\theta_1^T X_i). \quad (163)
\]
It can be seen that the pseudo-labels \( \{\hat{Y}_i\}_{i=1}^{2m} \) are conditionally i.i.d. given \( \theta_1 \) and let us denote the conditional distribution under fixed \( \theta_1 \) as \( P_{Y|\theta_1} \in \mathcal{P}(Y) \). The pseudo-labelled dataset is denoted as \( \hat{S}_{u,2} = \{(X'_i, \hat{Y}_i)\}_{i=m+1}^{2m} \).
For any fixed \( \theta_1 \in \Theta \), we can decompose it as \( \tilde{\theta}_1 = \alpha_1' \mu + \beta_1' \nu \), where \( \alpha_1' \in [-1, 1] \) and \( \beta_1' = \sqrt{1 - (\alpha_1')^2} \).

Recall the decomposition of \( X_i' \) and \( \tilde{X}_i' \) in (69) and (73). Similarly, we have

\[
\tilde{\theta}_1' X_i' = Y_i' \alpha_1' + \sigma h_1^i, \tag{164}
\]

where \( h_1^i \sim N(0, 1) \). Note that \( P_{Y_i' \mid \theta_1, \tilde{\theta}_1, \mu^\perp} = P_{Y_i' \mid \theta_1} \) and then the conditional probability \( P_{Y_i' \mid \theta_1, \tilde{\theta}_1, \mu^\perp} \) can be given by

\[
P_{Y_i' \mid \theta_1, \tilde{\theta}_1, \mu^\perp}(1) = P_{Y_i' \mid \theta_1}(1) = \mathbb{P}_{\tilde{\theta}_1}(\tilde{\theta}_1' X_i' > 0)
= \frac{1}{2}\mathbb{P}_{\tilde{\theta}_1}(\alpha_1' + \sigma h_1^i > 0) + \frac{1}{2}\mathbb{P}_{\tilde{\theta}_1}(\alpha_1' + \sigma h_1^i \leq 0) = \frac{1}{2}, \tag{165}
\]

and \( P_{Y_i' \mid \theta_1, \tilde{\theta}_1, \mu^\perp}(-1) = 1/2 \), where \( \mathbb{P}_{\tilde{\theta}_1} \) denotes the probability measure under parameter \( \theta_1 \).

5. **Iteration** \( t = 2 \): Recall (18) and the new model parameter learned from the pseudo-labelled dataset \( \hat{S}_{u, 2} \) is given by

\[
\theta_2 = \frac{1}{m} \sum_{i=m+1}^{2m} \tilde{Y}_i' X_i' = \frac{1}{m} \sum_{i=m+1}^{2m} \text{sgn}(\tilde{\theta}_1' X_i') X_i', \tag{167}
\]

where \( \{\text{sgn}(\tilde{\theta}_1' X_i') X_i'\}_{m+1}^{2m} \) are conditionally i.i.d. random variables given \( \theta_1, \xi_0, \mu^\perp \).

Given any \( (\theta_1, \xi_0, \mu^\perp) \), for any \( j \in [m + 1 : 2m] \), let \( \mu_2^{\theta_1, \xi_0, \mu^\perp} := \mathbb{E}[\text{sgn}(\tilde{\theta}_1' X_j') X_j' \mid \theta_1, \xi_0, \mu^\perp] \) and \( \mathbb{P}_{\theta_1, \xi_0, \mu^\perp} \) denotes the probability measure under the parameters \( \theta_1, \xi_0, \mu^\perp \). Following the similar steps that derive (114), for any \( \varepsilon > 0 \), we have

\[
\mathbb{P}_{\theta_1, \xi_0, \mu^\perp}(\|\theta_2 - \mu_2^{\theta_1, \xi_0, \mu^\perp}\|_\infty > \varepsilon) \leq \delta_{m, \varepsilon, d}. \tag{168}
\]

From (98), no matter what \( \theta_1 \) is, we always have \( \|\mu_2^{\theta_1, \xi_0, \mu^\perp} - \mu_{\xi_0, \mu^\perp}\| \leq \tilde{c}_1 \). Then, for some \( c \in (\tilde{c}_1, \infty) \),

\[
\mathbb{P}_{\theta_1, \xi_0, \mu^\perp}(\theta_2 \in \Theta_{\mu, c}) \geq 1 - \delta_{m, c - \tilde{c}_1, d}. \tag{169}
\]

With probability at least \((1 - \delta_{m, c - \tilde{c}_1, d})(1 - \delta_{r, d})\), the absolute generalization error can be upper bounded as follows:

\[
|\text{gen}_2| = \|\mathbb{E}[L_{P_1}(\theta_2)] - L_{\hat{S}_{u, 2}}(\theta_2)\|
= \frac{1}{m} \sum_{i=m+1}^{2m} \mathbb{E}_{\theta_1, \xi_0, \mu^\perp} \left[ \mathbb{E} \left[ l(\theta_2, (X, Y)) - l(\theta_2, (X_i', Y_i')) \mid \theta_1, \xi_0, \mu^\perp \right] \right]
\leq \sqrt{\left(\frac{c_2 - c_1}{2}\right)^2} \times \frac{1}{m} \sum_{i=m+1}^{2m} \mathbb{E}_{\theta_1, \xi_0, \mu^\perp} \left[ \mathbb{E} \left[ l(\theta_2, (X_i', Y_i')) + D_{\theta_1, \xi_0, \mu^\perp}(P_{\theta_2, X_i', Y_i'} || P_{X, Y}) \right] \right], \tag{170}
\]

where \( \mathbb{P}_{\theta_2, X_i', Y_i'}(\theta_2, (X_i', Y_i')) = \mathbb{P}_{\theta_2, X_i', Y_i'} \otimes P_{X, Y} \).

Similar to (120), for any \( \varepsilon > 0 \) and \( \delta \in (0, 1) \), there exists \( m_1(\varepsilon, d, \delta) \) such that for all \( m > m_1 \),

\[
\mathbb{P}_{\theta_1, \xi_0, \mu^\perp}(l(\theta_2, (X_i', Y_i')) > \varepsilon) < \delta. \tag{171}
\]

Recall (166) that \( P_{Y_i' \mid \theta_1, \tilde{\theta}_1, \mu^\perp} \approx \text{unif}([-1, 1]) \). For any fixed \( (\theta_1, \xi_0, \mu^\perp) \), let \( \tilde{\theta}_1 \) be decomposed as \( \tilde{\theta}_1 = \alpha_1'(\xi_0, \mu^\perp) \mu + \beta_1'(\xi_0, \mu^\perp) \nu \), where \( \alpha_1'(\xi_0, \mu^\perp) \in [-1, 1] \) and \( \beta_1'(\xi_0, \mu^\perp) = \sqrt{1 - (\alpha_1'(\xi_0, \mu^\perp))^2} \).

By following the similar steps in the first iteration, the disintegrated conditional KL-divergence between pseudo-labelled distribution and true distribution is given by

\[
D_{\theta_1, \xi_0, \mu^\perp}(P_{X_i', Y_i'} || P_{X, Y})
= \frac{1}{2} D_{\theta_1, \xi_0, \mu^\perp}(P_{X_i' \mid Y_i' = -1} || P_{X \mid Y = -1}) + \frac{1}{2} D_{\theta_1, \xi_0, \mu^\perp}(P_{X_i' \mid Y_i' = 1} || P_{X \mid Y = 1}) \tag{172}
\]

\[
= G_\sigma(\alpha_1'(\xi_0, \mu^\perp), \xi_0, \mu^\perp), \tag{173}
\]

\[
= G_\sigma(\alpha_1'(\xi_0, \mu^\perp), \xi_0, \mu^\perp), \tag{174}
\]

\[
= G_\sigma(\alpha_1'(\xi_0, \mu^\perp), \xi_0, \mu^\perp), \tag{175}
\]
Given any pair of \((\xi_0, \mu^+)\), recall the decomposition of \(\mu_1^{\xi_0, \mu^+}\) in \([96]\). Then the correlation between \(\mu_1^{\xi_0, \mu^+}\) and \(\mu\) is given by
\[
\rho(\mu_1^{\xi_0, \mu^+}, \mu) = \frac{1}{\sqrt{(1 - 2Q(\alpha)) + \frac{2\sigma^2}{\sqrt{2\pi}} \exp(-\frac{\alpha^2}{2\sigma^2})}} \Big(\frac{1}{2} + \frac{2\sigma^2}{\sqrt{2\pi}} \exp(-\frac{\alpha^2}{2\sigma^2})\Big) \exp(-\frac{\alpha^2}{2\sigma^2})
\]

(176)

By the strong law of large numbers, we have \(\alpha'_{\delta}(\xi_0, \mu^+) \xrightarrow{a.s.} F_\sigma(\alpha(\xi_0, \mu^+))\) as \(m \to \infty\). Then for any \(\epsilon > 0\) and \(\delta \in (0, 1)\), there exists \(m_2(\epsilon, d, \delta)\) such that for all \(m > m_2\),
\[
\Pr_{\theta, \xi_0, \mu^+} \left( |G_\sigma(\alpha'_{\delta}(\xi_0, \mu^+), \xi_0, \mu^+) - G_\sigma(F_\sigma(\alpha(\xi_0, \mu^+)), \xi_0, \mu^+)| > \epsilon \right) \leq \delta.
\]

(178)

Therefore, fix some \(d \in \mathbb{N}, \epsilon > 0\) and \(\delta \in (0, 1)\). There exists \(n_0(\epsilon, d) \in \mathbb{N}, m_3(\epsilon, d, \delta) \in \mathbb{N}, c_0(\delta, d) \in (c_1, \infty), r_0(\epsilon, d) \in \mathbb{R}_+\) such that for all \(n > n_0, m > m_3, c > c_0, r > r_0, \delta_m, \epsilon, d < \frac{\epsilon}{4}, \delta_r, c, d < \frac{\epsilon}{4}\), and then with probability at least \(1 - \delta\), the absolute generalization error at \(t = 2\) can be upper bounded as follows:
\[
|\text{gen}_2| \leq \sqrt{\frac{(C_2 - C_1)^2}{2}} \mathbb{E}_{\xi_0, \mu^+} \left[ \left| G_\sigma(F_\sigma(\alpha(\xi_0, \mu^+)), \xi_0, \mu^+) + \epsilon \right| \right].
\]

(179)

6. Any iteration \(t \in [3 : \tau]\): By similarly repeating the calculation in iteration \(t = 2\), we obtain the upper bound for \(|\text{gen}_{t+1}|\) in \([24]\).

Remark 2: In the other extreme case, when \(\alpha = \rho(\theta_0, \mu) = -1\) and \(\theta_0 = -\mu\), the error probability \(\Pr(Y'_j \neq Y'_j) = 1 - Q(1/\sigma) > \frac{1}{2}\) (for all \(\sigma > 0\)) and \(D_{\alpha, \mu^+}(P_{X'_j, Y'_j}||P_{X, Y}) < \infty\), so in this other extreme (flipped) scenario, we have more mistakes than correct pseudo-labels. The reason why \(D_{\alpha, \mu^+}(P_{X'_j, Y'_j}||P_{X, Y})\) is finite is that when \(P_{X, Y}(x, y)\) is small, it means that \(x\) is far from both \(-\mu\) and \(\mu\), and then \(P_{X}(x)\) is also small. Thus, \(P_{X'_j, Y'_j}(x, y) = P_{Y'_j||X'_j}(y|x)P_{X}(x)\) is also small.

D  Reusing \(S_t\) in Each Iteration

If the labelled data \(S_t\) are reused in each iteration and \(w = \frac{n}{n+m}\) (cf. \([1]\)), for each \(t \in [1 : \tau]\), the learned model parameter is given by
\[
\theta_t = \frac{n}{n+m}\theta_{t-1} + \frac{1}{n+m} \sum_{i=(t-1)m+1}^{tm} Y'_iX'_i
\]

(180)
\[
= \frac{n}{n+m}\theta_{t-1} + \frac{1}{n+m} \sum_{i=(t-1)m+1}^{tm} \text{sgn}(\theta'_{t-1}X'_i)X'_i
\]

(181)

Figure 10: \(\tilde{F}^{(t)}_{\sigma, \xi_0, \mu^+}(x)\) versus \(x\) under \(t \in \{0, 1, 2\}\) when \(\sigma = 0.5, \xi_0 = 0, ||\mu^+||_2 = 1, n = 10, m = 1000\).

Figure 11: Upper bound for \(|\text{gen}_{t}|\) versus \(t\) for \(m = 100\) and \(m = 1000\), when \(n = 10, \sigma = 0.6, d = 2, \mu = (1, 0)\).

Recall the definition of the function \(\tilde{F}^{(t)}_{\sigma, \xi_0, \mu^+}\) in \([27]\). Let the \(t\)-th iterate of \(\tilde{F}^{(t)}_{\sigma, \xi_0, \mu^+}\) be denoted as \(\tilde{F}^{(t)}_{\sigma, \xi_0, \mu^+}\) with initial condition \(\tilde{F}^{(0)}_{\sigma, \xi_0, \mu^+}(x) = x\). As shown in Figure 10, we can see that for any fixed
Following the similar steps in Appendix C, we first derive the upper bound for $\sigma, \xi$ in Corollary 3 also decreases as $t$ increases. As a result, $\tilde{F}^{(t)}_{\sigma, \xi, \mu^+}$ represents the improvement of the model parameter $\theta_t$ over the iterations.

As shown in Figure 11 under the same setup as Figure 6(c) when the labelled data $S_t$ are reused in each iteration, the upper bound for $|\text{gen}_1|$ is also a decreasing function of $t$. When $m = 1000$, the upper bound is almost the same as that one in Figure 6(c), which means that for large enough $m/n$, reusing the labelled data does not necessarily help to improve the generalization performance. Moreover, when $m = 100$, the upper bound is higher than that for $m = 1000$, which coincides with the intuition that increasing the number of unlabelled data helps to reduce the generalization error.

**E Proof of Corollary 3**

Following the similar steps in Appendix C, we first derive the upper bound for $|\text{gen}_1|$ as follows.

At $t = 1$, from (65) and (96), the expectation $\mu_{1, \mu^+} = E[\theta_1 | \xi, \mu^+]$ is rewritten as

$$\mu_{1, \mu^+} = \frac{n}{n + m} \sum_{i=1}^{m} E[\text{sgn}(\theta_0^T X_i^T) X_i | \xi, \mu]$$

$$= \frac{n}{n + m} \left( 1 + \frac{\sigma}{\sqrt{n}} \xi \right) \mu + \frac{\sigma}{\sqrt{n}} \mu_{1, \mu^+} + \frac{m}{n + m} \left( 1 - 2Q \frac{\alpha}{\sigma} + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp \left( -\frac{\alpha^2}{2\sigma^2} \right) \right) \mu + \frac{m}{n + m} \left( -2Q \frac{\alpha}{\sigma} + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp \left( -\frac{\alpha^2}{2\sigma^2} \right) \right) \mu$$

$$= \left( 1 + \frac{\sqrt{n} \xi}{n + m} + \frac{m}{n + m} \left( -2Q \frac{\alpha}{\sigma} + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp \left( -\frac{\alpha^2}{2\sigma^2} \right) \right) \right) \mu + \frac{m}{n + m} \sum_{i=1}^{m} \left( 1 - 2Q \frac{\alpha}{\sigma} + \frac{2\sigma \alpha}{\sqrt{2\pi}} \exp \left( -\frac{\alpha^2}{2\sigma^2} \right) \right) \mu$$

Then the correlation between $\mu_{1, \mu^+}$ and $\mu$ is given by

$$\rho(\mu_{1, \mu^+}, \mu) = \tilde{F}_{\sigma, \xi, \mu^+}(\alpha).$$

Let $\theta_1 = \frac{1}{m} \sum_{i=1}^{m} \text{sgn}(\theta_0^T X_i) X_i$. For some $c \in (\bar{c}, \infty)$, from (53) and (114),

$$\Pr \left( \|\theta_1 - \mu\|_\infty > c \right) \leq \Pr \left( \frac{n}{n + m} \|\theta_0 - \mu\|_\infty + \frac{m}{n + m} \|\theta_1 - \mu\|_\infty > c \right)$$

$$\leq \Pr \left( \|\theta_0 - \mu\|_\infty > c \right) + \Pr \left( \|\theta_1 - \mu\|_\infty > c \right)$$

$$\leq \delta_{\sqrt{m}, d} + \delta_{c, e - c, d}$$

Thus, from Theorem 1 for some $c \in (\bar{c}, \infty)$, with probability at least $(1 - \delta_{\sqrt{m}, d} - \delta_{c, e - c, d})(1 - \delta_{r, d})$, the absolute generalization error can be upper bounded as follows:

$$|\text{gen}_1|$$

$$= \left| \frac{1}{n + m} \sum_{i=1}^{n} E \left[ l(\theta_1, (X, Y)) - l(\theta_1, (X_i, Y_i)) \right] \right|$$

$$\leq \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_1; (X_i, Y_i))}$$

$$\leq \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_0; (X_i, Y_i))}$$

$$\leq \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_1; (X_i^T, Y_i^T))}$$

$$+ \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_0; (X_i^T, Y_i^T))}$$

$$+ \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_1; (X_i^T, Y_i^T))}$$

$$+ \frac{1}{n + m} \sum_{i=1}^{n} \sqrt{\frac{(c_2 - c_1)^2}{2} I(\theta_0; (X_i^T, Y_i^T))}$$

$$\leq 1$$
where $P_{\theta_1}(X,Y)\xi_0,\mu^+ = Q_{\theta_1}(\xi_0,\mu^+) \otimes P_{X,Y}$ and $Q_{\theta_1}(\xi_0,\mu^+)$ denotes the marginal distribution of $\theta_1$ under parameters $(\xi_0,\mu^+)$, and (190) follows since $(X_i,Y_i) = \theta_0 - \theta_1$ forms a Markov chain.

In (191), the KL-divergence is already given in (161) and the disintegrated conditional mutual information can be calculated as follows. Since we have

$$I_{\xi_0,\mu^{+}}(\theta_1; (X'_i, Y'_i)) = I_{\xi_0,\mu^{+}}\left(\frac{1}{m} \sum_{i=1}^{m} \text{sgn}(\theta_{I-1}^T X'_i) X'_i; (X'_i, Y'_i)\right),$$

(192)

from (126), for any $\epsilon > 0$ and any $\delta \in (0,1)$, there exists $m'_1(\epsilon, \delta, d) \in \mathbb{N}$ such that for all $m > m'_1(\epsilon, \delta, d)$,

$$\mathbb{P}_{\xi_0,\mu^{+}}\left(I_{\xi_0,\mu^{+}}(\theta_1; X'_i, Y'_i) > \epsilon\right) \leq 1 - \delta.$$

(193)

Therefore, fix $d \in \mathbb{N}$, any $\epsilon > 0$ and any $\delta \in (0,1)$, and there exists $n_0(d, \delta) \in \mathbb{N}$, $m_1(\epsilon, d, \delta) \in \mathbb{N}$, $c_0(d, \delta) \in (c_1, \infty)$, $r_0(d, \delta) \in \mathbb{R}$, such that for all $n > n_0, m > m_1, c > c_0, r > r_0, \delta < c_0 - c_1, d < \delta, \delta_{d} < \delta$, and with probability at least $1 - \delta$, the absolute generalization error $|\text{gen}_1|$ can be upper bounded as follows:

$$|\text{gen}_1| \leq w \sqrt{\frac{(c_2 - c_1)^2d}{4} \log\frac{n}{n-1} + (1 - w) \sqrt{\frac{(c_2 - c_1)^2d}{2} \mathbb{E}_{\xi_0,\mu^{+}} \left[\sqrt{\left(G_{(\alpha(\xi_0,\mu^{+}), \xi_0,\mu^{+})} + \epsilon\right)^2}\right]}}.$$

(194)

For $t \geq 2$, the only difference from the derivation in Appendix C is the correlation function $\tilde{F}_{\sigma,\xi_0,\mu^{+}}(\cdot)$ (compared to (177)). Thus by replacing $F_\sigma(\cdot)$ with $\tilde{F}_{\sigma,\xi_0,\mu^{+}}(\cdot)$, we obtain the upper bound in (28), completing the proof of Corollary 3.

### Additional Experiments

**Table 1:** The $l_2$ distances between the RGB-mean and RGB-variance of different pairs of classes from the CIFAR10 dataset.

| Classes         | RGB-mean $l_2$ distance | RGB-variance $l_2$ distance | Difficulty |
|-----------------|-------------------------|-----------------------------|------------|
| horse-ship      | 0.0180                  | 3.90e-05                    | Easy       |
| automobile-truck| 0.0038                  | 7.06e-05                    | Moderate   |
| cat-dog         | 0.0007                  | 4.95e-05                    | Challenging|

In Table 1 we display the RGB means and variances of the test data in six classes taken from the CIFAR10 dataset. We observe that the RGB variances of each pair are almost 0 (and small compared to the RGB-mean $l_2$ distances), and thus, the RGB-mean $l_2$ distance is indicative of the difficulty of the classification task. Indeed, a smaller RGB-mean $l_2$ distance implies a higher overlap of the two classes and consequently, greater difficulty in distinguishing them. Therefore, the “cat-dog” pair, which is more difficult to disambiguate compared to the “horse-ship” and “automobile-truck” pairs, is analogous to the bGMM with large variance (i.e. large overlap between the $\{\pm 1\}$ classes).

Under the same experimental settings as in Section 5, we perform another classification experiment on the “cat-dog” pair (from the CIFAR10 dataset). As shown in Figure 12, the test accuracy at the initial point when only labelled data are used is about 65% and the test loss is about 1.4; these are much worse than the performances of the classification tasks as shown in Figures 7 and 8 which means that the two classes are more challenging to classify.

It can be observed from Figure 12 that although the training loss decreases and the test and training accuracies increase as the iteration count increases (which are expected), the test loss and the generalization error both increase. The fact that both the test loss and test accuracy appear to increase with $t$ is, in fact, not contradictory. To intuitively explain this, in binary classification using the softmax (hence, logistic) function to predict the output classes, the learned probability of a data example belonging to
its true class is $p \in [0, 1]$ and if $p \in (1/2, 1]$, the classification is correct. In other words, the accuracy is 100%. However, when $p$ (i.e., the classification confidence) decreases towards $(1/2)^+$, the corresponding decision margin $2p - 1$ (Cao et al., 2019) also decreases and the test loss $-\log p$ increases commensurately. Thus, when the decision margin is small, even though the test accuracy may increase as the iteration counter $t$ increases, the test loss (representing our lack of confidence) may also increase at the same time.

In summary, for the classification task involving the “cat” and “dog” classes, our above observations correspond to that for the bGMM in Figure 6(d), namely that the unlabelled data does not help to improve the generalization error when the classification task is challenging and the initialization with the labelled data $S_l$ does not already result in a relatively high accuracy.