Journal of Machine Learning Research 23 (2022) 1-52 Submitted 5/22; Revised 8/22; Published 8/22

Information-Theoretic Characterization of the Generalization Error for Iterative Semi-Supervised Learning

Haiyun He HAIYUN.HE@U.NUS.EDU
Department of Electrical and Computer Engineering,
National University of Singapore,
117583 Singapore

Hanshu Yan HANSHU.YAN@U.NUS.EDU
Department of Electrical and Computer Engineering,
National University of Singapore,
117583 Singapore

Vincent Y. F. Tan VTAN@NUS.EDU.SG
Department of Mathematics,
Department of Electrical and Computer Engineering,
Institute of Operations Research and Analytics,
National University of Singapore,
119076, Singapore

Editor: Aarti Singh

Abstract

Using information-theoretic principles, we consider the generalization error (gen-error) of iterative semi-supervised learning (SSL) algorithms that iteratively generate pseudo-labels for a large amount of unlabelled data to progressively refine the model parameters. In contrast to most previous works that *bound* the gen-error, we provide an *exact* expression for the gen-error and particularize it to the binary Gaussian mixture model. Our theoretical results suggest that when the class conditional variances are not too large, the gen-error decreases with the number of iterations, but quickly saturates. On the flip side, if the class conditional variances (and so amount of overlap between the classes) are large, the gen-error increases with the number of iterations. To mitigate this undesirable effect, we show that regularization can reduce the gen-error. The theoretical results are corroborated by extensive experiments on the MNIST and CIFAR datasets in which we notice that for easy-to-distinguish classes, the gen-error improves after several pseudo-labelling iterations, but saturates afterwards, and for more difficult-to-distinguish classes, regularization improves the generalization performance.

Keywords: Generalization error, Semi-supervised learning, Pseudo-label, Information theory, Binary Gaussian mixture.

1. Introduction

In real-life machine learning applications, it is relatively easy and inexpensive 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 and

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, 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 (gen-error)*, or generalization bias, 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 (Russo and Zou, 2016; Xu and Raginsky, 2017; Kawaguchi et al., 2017). In SSL problems, the unlabelled data are expected to improve the generalization performance in a certain manner and thus, it is a worthy endeavor to investigate the behaviour theoretically. Although there exist many works studying the gen-error for supervised learning problems, the gen-error of SSL algorithms is yet to be explored.

1.1 Related Works

The extensive literature review is categorized into three aspects.

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 and Karakoulas (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) partitioned 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 and 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. Zhu (2020) considered the general SSL setting by assuming the loss function to be β-exponentially concave or the 0-1 loss, and used a Bayesian method for prediction instead of empirical risk minimization which we consider. The author presented an upper bound for the excess risk and the learning rate in terms of the number of labelled and unlabelled data examples. 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. Oymak and Gülcü (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. Recently, Aminian et al. (2022) considered the scenario where the labelled and unlabelled data are not generated from the same distribution and these distributions may change over time, exhibiting so-called covariate shifts. They provided an upper bound for the gen-error and proposed the Covariate-shift SSL (CSSL) method which outperforms some previous SSL algorithms under this setting. 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 involves using the Vapnik–Chervonenkis or VC dimension (Vapnik, 2000) and the Rademacher complexity (Boucheron et al., 2005). Recently, Russo and Zou (2016) proposed using the 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 such as differential privacy. It also paves a new way to improving generalization capability of learning algorithms from an information-theoretic viewpoint. Following Russo and Zou (2016), Xu and 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 aformentioned 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 and 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-, f-divergences and maximal leakage. Steinke and Zakynthinou (2020) proposed using the Conditional Mutual Information (CMI) to bound the generalization error; the CMI is useful as it possesses 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 transfer learning problems and characterized the upper bound based on IMI and KL-divergence. In a similar manner, Jose and Simeone (2021) provided a tightened bound on transfer generalization error based on the Jensen–Shannon divergence. Moreover, recently, Aminian et al. (2021) and Bu et al. (2022) recently derived the exact characterization of gen-error for supervised learning and transfer learning with the Gibbs algorithm.

Regularization is an important technique to reduce the model variance (Anzai, 2012), but there are few works that theoretically analyse the relationship between the gen-error and regularization. Moody (1992) characterized the gen-error as a function of the regularization parameter in supervised nonlinear learning systems and showed that the gen-error decreases as the parameter increases. Bousquet and Elisseeff (2002) provided a stability-based gen-error upper bound in terms of the regularization parameter in supervised learning. Mignacco et al. (2020) studied how the regularization affects the expected accuracy in high-dimensional GMM supervised classification problem.

Gaussian mixture models (GMM): The GMM is a popular, simple but non-trivial model that has been studied by many researchers. The performance of GMM classification problems depends on the data structure. The classical work of Castelli and Cover (1996) studied the classification problem in a binary mixture model with known conditional distributions but unknown mixing parameter and characterized the relative value of labelled and unlabelled data in improving the convergence rate of classification error probability. Akaho and Kappen (2000) characterized the generalization bias of general GMMs in supervised learning and discussed its dependency on data noise. Watanabe and Watanabe (2006) considered GMM in Bayesian learning and provided bounds for variational stochastic complexity. Wang and Thrampoulidis (2022) and Muthukumar et al. (2021) studied the dependence of the bGMM classification performance (using the 0-1 loss) on the structure of data covariance by considering SVM and linear interpolation.

However, all these aforementioned works do not investigate the generalization performance of SSL algorithms.

1.2 Contributions

Our main contributions are as follows.

1. In Section 3, we leverage results by Bu et al. (2020) and Wu et al. (2020) to derive an information-theoretic gen-error bound at each iteration for iterative SSL; see Theorem 2.A. Moreover, in contrast to most previous works that bound the gen-error, we derive an *exact* characterization of gen-error at each iteration for *negative log-likelihood (NLL)* loss functions (see Theorem 2.B).

2. In Section 4, we particularize Theorem 2.B to the binary Gaussian mixture model (bGMM) with in-class variance σ^2. We show that for any fixed number of data samples, there exists a critical value σ_0 such that when the data variance (representing the overlap between classes) $\sigma^2 < \sigma_0^2$, the gen-error decreases in the iteration count t and converges quickly with a sufficiently large amount of unlabelled data. When $\sigma^2 > \sigma_0^2$, the gen-error increases instead, which means using the unlabelled data does not help to reduce the gen-error across the SSL iterations. The empirical gen-error corroborates the theoretical results, which suggests that the characterization serves as a useful rule-of-thumb to understand how the gen-error changes across the SSL

iterations and it can be used to establish conditions under which unlabelled data can help in terms of generalization.

3. In Section 5, we theoretically and empirically show that for difficult-to-classify problems with large overlap between classes, regularization can effectively help to mitigate the undesirable increase of the gen-error across the SSL iterations.

4. In Section 6, we implement the pseudo-labelling procedure on the MNIST and CIFAR datasets with few labelled data and abundant unlabelled data. The experimental results corroborate the phenomena for the bGMM that the gen-error decreases quickly in the early pseudo-labelling iterations and saturates thereafter for easy-to-distinguish classes but increases for hard-to-distinguish classes. By adding ℓ_2-regularization to the hard-to-distinguish problem, we also observe improvements to the gen-error similar to that for the bGMM.

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 Θ and the loss fucntion be $l : \mathcal{Z} \times \Theta \to \mathbb{R}$, where $d \in \mathbb{N}$. We are given a labelled training dataset $S_{\mathrm{l}} = \{Z_1, \ldots, Z_n\} = \{(X_i, Y_i)\}_{i=1}^n$ drawn from $\mathcal{Z}$, where each $Z_i = (X_i, Y_i)$ is independently and identically distributed (i.i.d.) from $P_Z = P_{X,Y} \in \mathcal{P}(\mathcal{Z})$. 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 $S_{\mathrm{u}} = \{X'_1, \ldots, X'_{\tau m}\}, \tau \in \mathbb{N}$, where each X'_i is generated i.i.d. 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 1. Let $t \in [0 : \tau]$ denote the iteration count. 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 τ disjoint equal-size sub-datasets $\{S_{\mathrm{u},k}\}_{k=1}^{\tau}$, where $S_{\mathrm{u},k} = \{X'_{(k-1)m+1}, \ldots, X'_{km}\}$. In each subsequent round $t \in [1 : \tau]$, based on θ_{t-1} trained from the previous round, we use a predictor $f_{\theta_{t-1}} : \mathcal{X} \mapsto \mathcal{Y}$ to assign a *pseudo-label* $\hat{Y}'_i$ to the unlabelled sample X'_i for all $i \in \mathcal{I}_t := \{(t-1)m, (t-1)m+1, \ldots, tm\}$. Let $\hat{S}_{\mathrm{u},t} = \{(X'_i, \hat{Y}'_i)\}_{i \in \mathcal{I}_t}$ denote the t^{th} pseudo-labelled dataset. After pseudo-labelling, both the labelled data S_{l} and the pseudo-labelled data $\hat{S}_{\mathrm{u},t}$ are used to learn a new model parameter θ_t. The procedure is then repeated iteratively until the maximum number of iterations τ is reached.

This setup is a classical and widely-used model in the realm of self-training in SSL (Chapelle et al., 2006; Zhu, 2008; Zhu and Goldberg, 2009; Lee, 2013), where in each iteration, only a subset of the unlabelled data are used. Furthermore, as discussed by Arazo et al. (2020), this method is less likely to overfit to incorrect pseudo-labels, compared to using all the unlabelled data in each iteration (also see Figure 10). Under this setup of iterative SSL, during each iteration t, our *goal* is to find a model parameter $\theta_t \in \Theta$ that

$$\begin{tikzcd}
(X_i, Y_i)_{i=1}^n \arrow[r] \arrow[rrr, dashed, bend left] \arrow[rrrrr, dashed, bend left] \arrow[rrrrrrr, dashed, bend left] & \theta_0 \arrow[r] & \left(X_i', \hat{Y}_i' = f_{\theta_0}(X_i')\right)_{i=1}^m \arrow[r] & \theta_1 \arrow[r] & \left(X_i', \hat{Y}_i' = f_{\theta_1}(X_i')\right)_{i=m+1}^{2m} \arrow[r] & \theta_2 \arrow[r] & \cdots \arrow[r] & \theta_\tau \\
 & (X_i')_{i=1}^m \arrow[u] & & (X_i')_{i=m+1}^{2m} \arrow[u] & & (X_i')_{i=2m+1}^{3m} \arrow[u] & &
\end{tikzcd}$$

Figure 1: Paradigm of iterative self-training with pseudo-labelling in SSL

minimizes the population risk with respect to the underlying data distribution

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

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

$$L_{S_\mathrm{l}}(\theta_t) := \frac{1}{n} \sum_{i=1}^n l(\theta_t, Z_i),$$

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

$$L_{\hat{S}_{\mathrm{u},t}}(\theta_t) := \frac{1}{m} \sum_{i \in \mathcal{I}_t} l(\theta_t, (X_i', \hat{Y}_i')).$$

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

$$L_{S_\mathrm{l}, \hat{S}_{\mathrm{u},t}}(\theta_t) := w L_{S_\mathrm{l}}(\theta_t) + (1 - w) L_{\hat{S}_{\mathrm{u},t}}(\theta_t). \tag{1}$$

In the usual case where the algorithm minimizes the average of the empirical training losses, one should set $w = \frac{n}{n+m}$. An SSL algorithm can be characterized by a randomized map from the labelled and unlabelled training data S_l, S_u to a model parameter θ according to a conditional distribution $P_{\theta|S_\mathrm{l}, S_\mathrm{u}}$. Then at each iteration t, we can use the sequence of conditional distributions $\{P_{\theta_k|S_\mathrm{l}, S_\mathrm{u}}\}_{k=0}^t$ with $P_{\theta_0|S_\mathrm{l}, S_\mathrm{u}} = P_{\theta_0|S_\mathrm{l}}$ 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 θ_t and the empirical risk on the training data:

$$\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l}, S_\mathrm{u}}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&:= \mathbb{E}[L_{P_Z}(\theta_t) - L_{S_\mathrm{l}, \hat{S}_{\mathrm{u},t}}(\theta_t)] \\
&= w \bigg(\mathbb{E}_{\theta_t} \mathbb{E}_Z [l(\theta_t, Z)] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\theta_t, Z_i}[l(\theta_t, Z_i)] \bigg) \\
&\quad + (1 - w) \bigg(\mathbb{E}_{\theta_t} \mathbb{E}_Z [l(\theta_t, Z)] - \frac{1}{m} \sum_{i \in \mathcal{I}_t} \mathbb{E}_{\theta_t, X_i', \hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))] \bigg).
\end{aligned}$$

When $t = 0$ and $w = 1$, the definition of the generalization error reduces to that of vanilla supervised learning. Based on this definition, the expected population risk can be decomposed as

$$\mathbb{E}[L_{P_Z}(\theta_t)] = \mathbb{E}[L_{S_\mathrm{l}, \hat{S}_{\mathrm{u},t}}(\theta_t)] + \mathrm{gen}_t, \tag{2}$$

where the first term on the right-hand side of this equation is what the algorithm minimizes and reflects how well the output hypothesis fits the dataset, and the second term gen_t is used to measure the extent to which the iterative learning algorithm overfits the training data at the t-th iteration. To minimize $\mathbb{E}[L_{P_Z}(\theta_t)]$, we need both terms in (2) to be small, but there exists a natural trade-off between them. While the algorithm aims to minimize the empirical risk $\mathbb{E}[L_{S_{\mathrm{l}},\hat{S}_{\mathrm{u},t}}(\theta_t)]$, studying and controlling gen_t can also help to reduce the population risk $\mathbb{E}[L_{P_Z}(\theta_t)]$, which is the ultimate goal of learning. Instead of focusing on the total generalization error induced during the entire process, we are interested in the following questions. How does gen_t evolve as t increases? Do the unlabelled data examples in S_{u} help to improve the generalization error?

3. General Results

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 gen-error gen_t in terms of the mutual information between input data samples (either labelled or pseudo-labelled) and the output model parameter θ_t, as well as the KL-divergence between the data distribution and the joint distribution of feature vectors and pseudo-labels (cf. Theorem 2.A). Furthermore, by considering the NLL loss function (MacKay, 2003; Goodfellow et al., 2016), we derive the exact characterization for the gen-error gen_t (cf. Theorem 2.B).

Recall that for a given $R > 0$, L is an *R-sub-Gaussian random variable* (Vershynin, 2018) if its *cumulant generating function* $\Lambda_L(\lambda) := \log \mathbb{E}[\exp(\lambda(L - \mathbb{E}[L]))] \le \exp(\lambda^2 R^2/2)$ for all $\lambda \in \mathbb{R}$. If L is R-sub-Gaussian, we write this as $L \sim \mathrm{subG}(R)$. Furthermore, let us recall the following somewhat non-standard information quantities (Negrea et al., 2019; Haghifam et al., 2020).

Definition 1. *For random variables X, Y and U, define the* disintegrated mutual information *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 that the conditional mutual information $I(X;Y|U) = \mathbb{E}_U[I_U(X;Y)]$ and the conditional KL-divergence $D(P_{X|U}\|P_{Y|U}|P_U) = \mathbb{E}_U[D_U(P_X\|P_Y)]$.*

For distributions P, Q and V, define the cross-entropy *as $h(P,Q) := \mathbb{E}_P[-\log Q]$ and the* divergence between the cross-entropies *as $\Delta\mathrm{h}(P\|Q|V) := h(P,V) - h(Q,V)$.*

Let $\theta^{(t)} = (\theta_0, \ldots, \theta_t)$ for any $t \in [0:\tau]$ and $w = 1$ for $t = 0$. In iterative SSL, we can characterize the gen-error as shown in Theorem 2 by applying the law of total expectation.

Theorem 2.A (Gen-error upper bound for iterative SSL)**.** *Suppose $l(\theta, Z) \sim \mathrm{subG}(R)$ under $Z \sim P_Z$ for all $\theta \in \Theta$, then for any $t \in [0:\tau]$,*

$$\begin{aligned}
&\left|\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_{\mathrm{l}},S_{\mathrm{u}}}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1})\right| \\
&\le \frac{w}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2 I^{(i)}_{\theta^{(t-1)}}}\right] + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta^{(t-1)}}\left[\sqrt{2R^2\big(I'^{(i)}_{\theta^{(t-1)}} + D'^{(i)}_{\theta^{(t-1)}}\big)}\right],
\end{aligned}$$

where $I^{(i)}_{\theta^{(t-1)}} := I_{\theta^{(t-1)}}(\theta_t; Z_i)$, $I'^{(i)}_{\theta^{(t-1)}} := I_{\theta^{(t-1)}}(\theta_t; X'_i, \hat{Y}'_i)$, and $D'^{(i)}_{\theta^{(t-1)}} := D_{\theta^{(t-1)}}(P_{X'_i,\hat{Y}'_i}\|P_Z)$.

Theorem 2.B (Exact gen-error for iterative SSL)**.** *Consider the NLL loss function $l(\theta, Z) = -\log p_\theta(Z)$, where $p_\theta(Z)$ is the likelihood of Z under parameter θ. For any $t \in [0:\tau]$,*

$$\begin{aligned}&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) \\ &= \mathbb{E}_{\theta^{(t)}}\left[\frac{w}{n}\sum_{i=1}^{n}\Delta\mathrm{h}_{\theta_t}^{(i)} + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t}\left(\Delta\mathrm{h}_{\theta^{(t)}}^{\prime(i)} + \widetilde{\Delta\mathrm{h}}_{\theta^{(t)}}^{\prime(i)}\right)\right],\end{aligned} \tag{3}$$

where

$$\Delta\mathrm{h}_{\theta_t}^{(i)} := \Delta\mathrm{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t}), \qquad \Delta\mathrm{h}_{\theta^{(t)}}^{\prime(i)} := \Delta\mathrm{h}(P_Z \| P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} | p_{\theta_t}), \quad \textit{and}$$
$$\widetilde{\Delta\mathrm{h}}_{\theta^{(t)}}^{\prime(i)} := \Delta\mathrm{h}(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} \| P_{X_i',\hat{Y}_i'|\theta^{(t)}} | p_{\theta_t}).$$

The proof of Theorem 2.A is provided in Appendix A, in which we provide a general upper bound not only applicable to sub-Gaussian loss functions. The proof of Theorem 2.B is provided in Appendix B. Specifically, for NLL loss functions, Theorem 2.B provides an *exact* characterization of the gen-error at each iteration. This is in stark contrast to most works on information-theoretic generalization error in which only *bounds* are provided.

In contrast to Bu et al. (2020, Theorem 1) and Wu et al. (2020, Theorem 1) which pertain to supervised learning, Theorem 2 characterizes the gen-error at each iteration during the pseudo-labelling and training process. Note that the quantities in Theorem 2.B satisfy $\Delta\mathrm{h}_{\theta_t}^{(i)} = I_{\theta^{(t-1)}}^{(i)} + D(P_Z\|p_{\theta_t}) - D(P_{Z_i|\theta_t}\|p_{\theta_t})$ and $\widetilde{\Delta\mathrm{h}}_{\theta^{(t)}}^{\prime(i)} = I_{\theta^{(t-1)}}^{\prime(i)} + D_{\theta^{(t-1)}}(P_{X_i',\hat{Y}_i'}\|p_{\theta_t}) - D_{\theta^{(t-1)}}(P_{X_i',\hat{Y}_i'|\theta_t}\|p_{\theta_t})$. Thus, it is plausible that the upper bound based on $I_{\theta^{(t-1)}}^{(i)}$ and $I_{\theta^{(t-1)}}^{\prime(i)}$ in Theorem 2.A can help to understand and control the exact gen-error. Intuitively, the mutual information between the individual input data sample Z_i and the output model parameter θ_t in Theorem 2.A and the cross-entropy divergences $\Delta\mathrm{h}_{\theta_t}^{(i)}$, $\widetilde{\Delta\mathrm{h}}_{\theta^{(t)}}^{\prime(i)}$ in Theorem 2.B both measure the extent to which the algorithm is sensitive to each data example at each iteration t. The KL-divergence between the underlying P_Z and pseudo-labelled distribution $P_{X_i',\hat{Y}_i'}$ in Theorem 2.A and the cross-entropy divergence $\Delta\mathrm{h}_{\theta^{(t)}}^{\prime(i)}$ in Theorem 2.B measure how effectively the pseudo-labelling process works. As $n \to \infty$ and $m \to \infty$, we show that the mutual information (as well as $\Delta\mathrm{h}_{\theta_t}^{(i)}$ and $\widetilde{\Delta\mathrm{h}}_{\theta^{(t)}}^{\prime(i)}$) vanishes but the divergences $D_{\theta^{(t-1)}}^{\prime(i)}$ and $\Delta\mathrm{h}_{\theta^{(t)}}^{\prime(i)}$ do not, which reflects the impact of pseudo-labelling on the gen-error.

In iterative learning algorithms, 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 gen-error iteratively. In the next section, we apply the exact iterated gen-error in Theorem 2.B to a classification problem under a specific generative model—the bGMM. This simple model allows us to derive a tractable characterization on the gen-error as a function of iteration number t that we can compute numerically.

4. Main Results on bGMM

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

4.1 Iterative SSL under bGMM

Fix a unit vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and a scalar $\sigma \in \mathbb{R}_+ = (0, \infty)$. Under the bGMM with mean $\boldsymbol{\mu}$ and standard deviation (std. dev.) σ (bGMM($\boldsymbol{\mu}, \sigma$)), we assume that the distribution of any labelled data example $(\mathbf{X}, Y)$ can be specified as follows. Let $\mathcal{Y} = \{-1, +1\}$, $Y \sim P_Y = \mathrm{unif}\{-1, +1\}$, and $\mathbf{X}|Y \sim \mathcal{N}(Y\boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$, where $\mathbf{I}_d$ is the identity matrix of size $d \times d$.

The random vector $\mathbf{X}$ is distributed according to the mixture distribution

$$p_{\boldsymbol{\mu}} = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d) + \frac{1}{2}\mathcal{N}(-\boldsymbol{\mu}, \sigma^2\mathbf{I}_d).$$

In the unlabelled dataset S_{u}, each $\mathbf{X}'_i$ for $i \in [1 : \tau m]$ is drawn i.i.d. from $p_{\boldsymbol{\mu}}$.

Let $\Theta \subset \mathbb{R}^d$ such that $\boldsymbol{\mu} \in \Theta$. For any $\boldsymbol{\theta} \in \Theta$, under the bGMM($\boldsymbol{\theta}, \sigma$), the joint distribution of any pair of $(\mathbf{X}, Y) \in \mathcal{Z}$ is given by $\mathcal{N}(Y\boldsymbol{\theta}, \sigma^2\mathbf{I}_d) \otimes P_Y$. The NLL loss function can be expressed as

$$\begin{aligned} l(\boldsymbol{\theta}, (\mathbf{X}, Y)) &= -\log p_{\boldsymbol{\theta}}(\mathbf{X}, Y) = -\log \big(P_Y(Y) p_{\boldsymbol{\theta}}(\mathbf{X}|Y)\big) \\ &= -\log \frac{1}{2\sqrt{(2\pi)^d}\sigma^d} + \frac{1}{2\sigma^2}(\mathbf{X} - Y\boldsymbol{\theta})^\top(\mathbf{X} - Y\boldsymbol{\theta}). \end{aligned}$$

The population risk minimizer is given by $\arg\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\mathbf{X},Y}[l(\boldsymbol{\theta}, (\mathbf{X}, Y))] = \boldsymbol{\mu}$.

Under this setup, the iterative SSL procedure is shown in Figure 1, but the labelled dataset S_{l} is only used to train in the initial round $t = 0$; we discuss the reuse of S_{l} in all iterations in Corollary 10. That is, in (1), we set $w = 0$. The algorithm operates in the following steps.

- **Step 1: Initial round $t = 0$ with S_{l}:** By minimizing the empirical risk of labelled dataset S_{l}

  $$L_{S_{\mathrm{l}}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\theta}, (\mathbf{X}_i, Y_i)) \overset{\mathrm{c}}{=} \frac{1}{2\sigma^2 n}\sum_{i=1}^{n}(\mathbf{X}_i - Y_i\boldsymbol{\theta})^\top(\mathbf{X}_i - Y_i\boldsymbol{\theta}),$$

  where $\overset{\mathrm{c}}{=}$ means that both sides differ by a constant independent of $\boldsymbol{\theta}$, we obtain the minimizer

  $$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} L_{S_{\mathrm{l}}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} Y_i\mathbf{X}_i. \tag{4}$$

- **Step 2: Pseudo-label data in S_{u}:** At each iteration $t \in [1 : \tau]$, for any $i \in \mathcal{I}_t$, we use $\boldsymbol{\theta}_{t-1}$ to assign a pseudo-label for $\mathbf{X}'_i$, that is, $\hat{Y}'_i = f_{\boldsymbol{\theta}_{t-1}}(\mathbf{X}'_i) = \mathrm{sgn}(\boldsymbol{\theta}_{t-1}^\top \mathbf{X}'_i)$.

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

  $$L_{\hat{S}_{\mathrm{u},t}}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i\in\mathcal{I}_t} l(\boldsymbol{\theta}, (\mathbf{X}'_i, \hat{Y}'_i)) \overset{\mathrm{c}}{=} \frac{1}{2\sigma^2 m}\sum_{i\in\mathcal{I}_t}(\mathbf{X}'_i - \hat{Y}'_i\boldsymbol{\theta})^\top(\mathbf{X}'_i - \hat{Y}'_i\boldsymbol{\theta}), \tag{5}$$

  we obtain the new model parameter

  $$\boldsymbol{\theta}_t = \frac{1}{m}\sum_{i\in\mathcal{I}_t} \hat{Y}'_i\mathbf{X}'_i = \frac{1}{m}\sum_{i\in\mathcal{I}_t} \mathrm{sgn}(\boldsymbol{\theta}_{t-1}^\top\mathbf{X}'_i)\mathbf{X}'_i. \tag{6}$$

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

4.2 Definitions

To state our result succinctly, we first define some non-standard notations and functions. From (4), we know that $\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$ and inspired by Oymak and Gülcü (2021), we can decompose $\boldsymbol{\theta}_0$ as

$$\boldsymbol{\theta}_0 = \Big(1 + \frac{\sigma}{\sqrt{n}}\xi_0\Big)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^{\perp},$$

where $\xi_0 \sim \mathcal{N}(0, 1)$, $\boldsymbol{\mu}^{\perp} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^{\top})$, and $\boldsymbol{\mu}^{\perp}$ is perpendicular to $\boldsymbol{\mu}$ and independent of ξ_0 (the details of this decomposition are provided in Appendix C).

Given a pair of vectors $(\mathbf{a}, \mathbf{b})$, define their correlation coefficient as $\rho(\mathbf{a}, \mathbf{b}) := \frac{\langle \mathbf{a}, \mathbf{b}\rangle}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2}$. The correlation coefficient between the estimated and true parameters is

$$\alpha(\xi_0, \boldsymbol{\mu}^{\perp}) := \rho(\boldsymbol{\theta}_0, \boldsymbol{\mu}) = \frac{1 + \frac{\sigma}{\sqrt{n}}\xi_0}{\sqrt{(1 + \frac{\sigma}{\sqrt{n}}\xi_0)^2 + \frac{\sigma^2}{n}\|\boldsymbol{\mu}^{\perp}\|_2^2}}. \tag{7}$$

Let $\beta(\xi_0, \boldsymbol{\mu}^{\perp}) = \sqrt{1 - \alpha(\xi_0, \boldsymbol{\mu}^{\perp})^2}$. We abbreviate $\alpha(\xi_0, \boldsymbol{\mu}^{\perp})$ and $\beta(\xi_0, \boldsymbol{\mu}^{\perp})$ to α and β respectively in the following. Then the normalized vector $\boldsymbol{\theta}_0/\|\boldsymbol{\theta}_0\|_2$ can be decomposed as follows

$$\bar{\boldsymbol{\theta}}_0 := \frac{\boldsymbol{\theta}_0}{\|\boldsymbol{\theta}_0\|_2} = \alpha\boldsymbol{\mu} + \beta\boldsymbol{\upsilon}, \tag{8}$$

where $\boldsymbol{\upsilon} = \boldsymbol{\mu}^{\perp}/\|\boldsymbol{\mu}^{\perp}\|_2$. Let $\bar{\boldsymbol{\theta}}_0^{\perp} := (2\beta^2\boldsymbol{\mu} - 2\alpha\beta\boldsymbol{\upsilon})/\sigma$, which is a vector perpendicular to $\bar{\boldsymbol{\theta}}_0$.

Let $\mathrm{Q}(\cdot) := 1 - \Phi(\cdot)$. 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 $t + 1$:

$$F_\sigma(x) := \frac{J_\sigma(x)}{\sqrt{J_\sigma^2(x) + K_\sigma^2(x)}}, \quad \text{where} \tag{9}$$

$$J_\sigma(x) := 1 - 2\mathrm{Q}\bigg(\frac{x}{\sigma}\bigg) + \frac{2\sigma x}{\sqrt{2\pi}}\exp\bigg(-\frac{x^2}{2\sigma^2}\bigg), \quad \text{and} \tag{10}$$

$$K_\sigma(x) := \frac{2\sigma\sqrt{1 - x^2}}{\sqrt{2\pi}}\exp\bigg(-\frac{x^2}{2\sigma^2}\bigg). \tag{11}$$

The t^{th} iterate of the function F_σ is defined recursively as $F_\sigma^{(t)} := F_\sigma \circ F_\sigma^{(t-1)}$ with $F_\sigma^{(0)}(x) = x$. As shown in Figure 2, for any fixed σ, 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 α, defined in (7), is positive, $F_\sigma^{(t)}(\alpha)$ increases with t; and vice versa. Moreover, as shown in Figure 11 in Appendix C, by varying σ, we observe that a smaller σ results in a larger $|F_\sigma(x)|$.

4.3 Main Theorem

By applying the result in Theorem 2.B, the following theorem provides an exact characterization for the generalization error at each iteration t for m large enough.

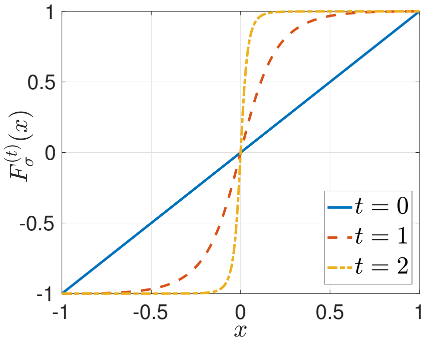


Figure 2: $F_\sigma^{(t)}(x)$ versus x for different t when $\sigma = 0.5$.

Theorem 3 (Exact gen-error for iterative SSL under bGMM)**.** *Fix any $\sigma \in \mathbb{R}_+$, $d \in \mathbb{N}$. The gen-error at $t = 0$ is given by*

$$\mathrm{gen}_0(P_{\mathbf{Z}}, P_{\mathbf{X}}, P_{\boldsymbol{\theta}_0|S_\mathrm{l},S_\mathrm{u}}) = \frac{d}{n}. \tag{12}$$

Let $\alpha = \alpha(\xi_0, \boldsymbol{\mu}^\perp)$. For each $t \in [1 : \tau]$, for almost all sample paths (i.e., almost surely),

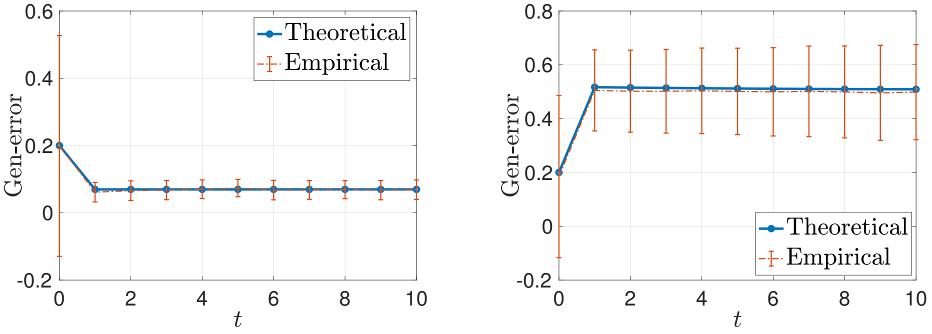
$$\begin{aligned}&\mathrm{gen}_t(P_{\mathbf{Z}}, P_{\mathbf{X}}, \{P_{\boldsymbol{\theta}_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^{t}, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})\\ &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\frac{(m-1)(J_\sigma^2(F_\sigma^{(t-1)}(\alpha)) + K_\sigma^2(F_\sigma^{(t-1)}(\alpha)))}{m\sigma^2} - \frac{J_\sigma(F_\sigma^{(t-1)}(\alpha))}{\sigma^2}\right] + o(1),\end{aligned} \tag{13}$$

where $o(1)$ is a term that vanishes as $m \to \infty$.

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

First, the gen-error at $t = 0$ corresponds to the asymptotic result of supervised maximum likelihood estimation in works by Akaho and Kappen (2000) and Aminian et al. (2021). We numerically plot the quantity in (13), $g_\sigma^{(m)}(x) := ((m-1)(J_\sigma^2(x)+K_\sigma^2(x)) - mJ_\sigma(x))/\sigma^2$, for $x \in [-1, 1]$ in Figure 12 in Appendix C, which shows that for all $\sigma_1 > \sigma_2$, $g_{\sigma_1}^{(m)}(x) > g_{\sigma_2}^{(m)}(x)$ when $x > 0$. From (7), we can see that α is close to 1 of high probability, which means that $\sigma \mapsto g_\sigma(x)$ is monotonically increasing in σ with high probability. As a result, (13) increases as σ increases. This is consistent with the intuition that when the training data have larger overlap between classes, it is more difficult to generalize well. Moreover, $F_\sigma^{(t)}(\alpha)$ is also close to 1 of high probability, and thus (13) saturates with t quickly.

Second, by ignoring the $o(1)$ term, we compare the theoretical gen_t (cf. (12) and (13)) and the empirical gen-error from the repeated synthetic experiments with $d = 2$, $n = 10$ and $m = 1000$, as shown in Figure 3. It can be seen that the theoretical gen_t matches the empirical gen-error well, which means that the characterization in (13) serves as a useful rule-of-thumb for how the gen-error changes over the SSL iterations. When the variance is small (e.g., $\sigma^2 = 0.6^2$), as shown in Figure 3(a), the gen-error decreases significantly from $t = 0$ to $t = 1$ and then quickly converges to a non-zero constant. Recall the correlation evolution function F_σ in (9). Given any pair of $(\xi_0, \boldsymbol{\mu}^\perp)$, if $\alpha(\xi_0, \boldsymbol{\mu}^\perp) > 0$, $F_\sigma^{(t)}(\alpha(\xi_0, \boldsymbol{\mu}^\perp)) > F_\sigma^{(t-1)}(\alpha(\xi_0, \boldsymbol{\mu}^\perp))$ for all $t \in [1 : \tau]$, as shown in Figure 2. This means that if the quality of the labelled data S_l is reasonably good, by using $\boldsymbol{\theta}_0$ which is learned from S_l, the generated



(a) $\sigma = 0.6$

(b) $\sigma = 3$

Figure 3: Comparison of the theoretical gen_t and the empirical gen-error at each iteration t.

pseudo-labels for the unlabelled data are largely correct. Then the subsequent parameters $\boldsymbol{\theta}_t$ for $t \geq 1$ learned from the large number of pseudo-labelled data examples can improve the generalization error. With sufficiently large amount of training data, algorithm converges at very early stage. In addition, for more general cases (e.g., non-diagonal class covariance matrices), it takes more iterations for the gen-error to reach a plateau, as shown in Figure 4.

When the variance is large (e.g. $\sigma^2 = 3^2$), as shown in Figure 3(b), the gen-error increases with iteration t. The result shows that when the overlap between different classes is large enough, using the unlabelled data may not be able to improve the generalization performance. The intuition is that at the initial iteration with a limited number of labelled data, the learned parameter $\boldsymbol{\theta}_0$ cannot pseudo-label the unlabelled data with sufficiently high accuracy. Thus, the unlabelled data is not labelled well by the pseudo-labelling operation and hence, cannot help to improve the generalization error. To gain more insight, in Figure 5, we numerically plot gen_1 versus different values of σ under the same setting. It is interesting to find that there exists a σ_0 such that for $\sigma < \sigma_0$, $\mathrm{gen}_1 < \mathrm{gen}_0$, which means the gen-error can be reduced with the help of abundant unlabelled data, while for $\sigma > \sigma_0$, using the unlabelled data can even harm the generalization performance.

Third, let us examine the effect of n, the number of labelled training samples. By expanding α, defined in (7), using a Taylor series, we have

$$\alpha = 1 - \frac{\sigma^2}{2n}\|\boldsymbol{\mu}^{\perp}\|_2^2 + o\left(\frac{1}{n}\right). \tag{14}$$

It can be seen that as n increases, α converges to 1 in probability. Suppose the dimension $d = 2$ and $\boldsymbol{\mu} = (1, 0)$. Then $\boldsymbol{\mu}^{\perp} = [0, \mu_2^{\perp}]$ where $\mu_2^{\perp} \sim \mathcal{N}(0, 1)$. By letting $m \to \infty$, the gen-error gen_1 (cf. (13)) can be rewritten as

$$\mathrm{gen}_1 \approx \int_{-\sqrt{2}}^{\sqrt{2}} \sqrt{\frac{n}{\pi\sigma^2}}\, e^{-\frac{ny^2}{\sigma^2}}\, g_\sigma^{\infty}(1-y^2)\ \mathrm{d}y,$$

where $g_\sigma^{(\infty)}(1-y^2) := (J_\sigma^2(1-y^2) + K_\sigma^2(1-y^2) - J_\sigma(1-y^2))/\sigma^2$ and thus, gen_1 is a decreasing function of n. We further deduce that for any t, gen_t is decreasing in n.

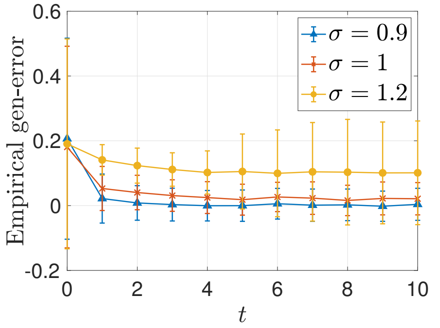


Figure 4: Empirical gen-error with covariance matrix $\sigma^2 \times [0.6, 0.3; 0.3, 0.8]$.

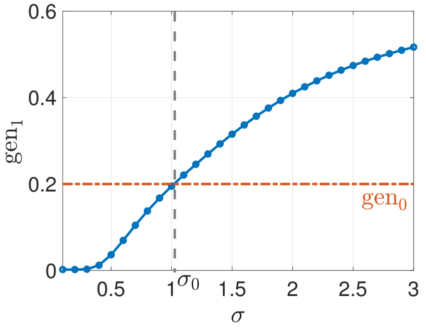


Figure 5: Theoretical gen-error at $t = 1$ versus different std. devs. $\sigma \in [0.1, 3]$.

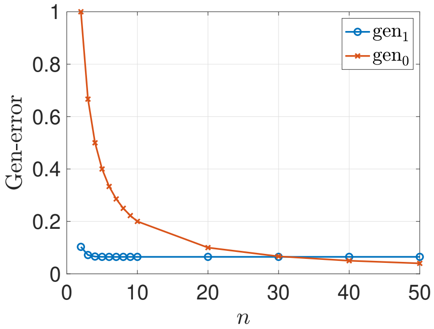


Figure 6: gen_0 vs. gen_1 for different n.

Fourth, we consider an "enhanced" scenario in which the labelled data in S_l are reused in each iteration. Set $w = \frac{n}{n+m}$ in (1). We can extend Theorem 3 to Corollary 10 provided in Appendix F. It can be seen from Figure 17 that gen_t still decreases from $t = 0$ to 1 and saturates afterwards. We find that when $\sigma = 0.6$, $n = 10$, $m = 1000$, the gen-error is almost the same as that one in Figure 3(a), which means that for large enough $\frac{m}{n}$, reusing the labelled data does not necessarily help to improve the generalization performance. Moreover, when $m = 100$, gen_t 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.

Fifth, it is natural to wonder what the effect is when m, the number of unlabelled data examples, is held fixed and n, the number of labelled data examples, increases. In Figure 6, we numerically plot $\mathrm{gen}_0 = \frac{d}{n}$ in (12) and (the theoretical) gen_1 in (13) for n ranging from 2 to 50, $m = 1000$, $\sigma = 0.6$ and $d = 2$. As n increases, gen_0 and gen_1 both decrease, which is as expected. However, when n is larger than a certain value (30 in this case), we find that gen_0 becomes smaller than gen_1. This implies that with sufficiently many labelled training data, the generalization error based on the labelled training data is already sufficiently low, and incorporating the pseudo-labelled data in fact adversely affects the generalization error. Understanding this phenomenon precisely is an interesting avenue for future work.

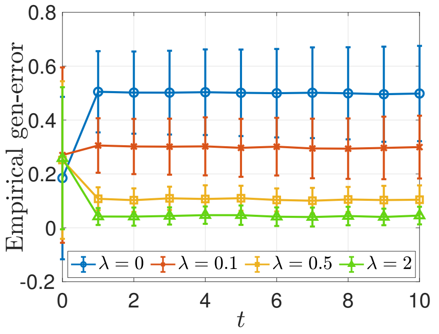


Figure 7: Empirical gen-error versus t for $\sigma = 3$ for different λ.

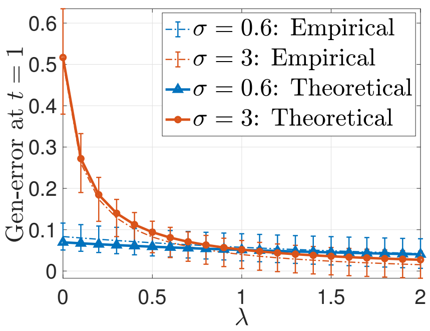


Figure 8: Theoretical and empirical gen_1 vs. λ for different σ.

Finally, to verify the validity of the gen-error upper bound in Theorem 2.A, we further apply the bound to this setup and prove that the upper bound exhibits similar behaviour of the evolution of gen-error as t increases. See Appendix D.

5. Improving the Gen-Error for Difficult Problems via Regularization

In Section 4.3, it is shown that for difficult classification problems with large class conditional variance, the gen-error increases after using pseudo-labelled data. The reason is that the learned initial parameter $\boldsymbol{\theta}_0$ can only generate low-accurate pseudo-labels and thus the pseudo-labelled data cannot help improve the generalization performance. In this section, we prove that by adding regularization to the loss function, we can mitigate the undesirable increase of gen-error across the pseudo-labelling iterations.

Since gen_0 in (12) does not depend on data variance σ^2, here we focus on subsequent iterations $t \in [1 : \tau]$. By considering the ℓ_2 regularization (i.e., adding $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ to (5)), we obtain the new parameters (cf. (6)) as follows

$$\boldsymbol{\theta}_t^{\mathrm{reg}} = \frac{\boldsymbol{\theta}_t}{1 + \sigma^2 \lambda}, \quad \forall t \in [1 : \tau].$$

The derivations are provided in Appendix H. By applying Theorem 3, the following theorem provides a characterization for the gen-error of the case with regularization at each iteration t for m large enough. Let $\mathrm{gen}_t^{\mathrm{reg}}$ denotes the gen-error of the case with regularization and we drop the fixed quantities $(P_{\mathbf{Z}}, P_{\mathbf{X}}, \{P_{\boldsymbol{\theta}_k|S_{\mathrm{l}},S_{\mathrm{u}}}\}_{k=0}^{t}, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})$ for notational simplicity.

Theorem 4 (Gen-error with regularization)**.** *Fix any $d \in \mathbb{N}$, and $\sigma, \lambda \in \mathbb{R}_+$. The gen-error at any $t \in [1 : \tau]$ is*

$$\mathrm{gen}_t^{\mathrm{reg}} = \frac{\mathrm{gen}_t}{1 + \sigma^2 \lambda}. \tag{15}$$

The proof of Theorem 4 is provided in Appendix H. From (15), we observe that as λ increases, the gen-error decreases. In Figure 7, we first empirically show that regularization can help mitigate the increase of gen-error during SSL iterations for hard-to-distinguish

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

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

Classes	horse	ship	automobile	truck	cat	dog
	→ship	→horse	→truck	→automobile	→dog	→cat
Number	17	3	61	64	93	137
Difficulty	Easy		Moderate		Challenging	

Table 2: Number of images misclassified out of 1000 (Liu and Mukhopadhyay, 2018).

classes by comparing the empirical gen-error under $\lambda = 0, 0.1, 0.5, 2$ when $\sigma = 3$ and $d = 2$. Then in Figure 8, we plot the theoretical gen-error in (15) versus λ when $t = 1$ for the cases with small and large variances, i.e., $\sigma^2 = 0.6^2$ and $\sigma^2 = 3^2$. We also compare the theoretical results with the empirical gen-errors, which turn out to corroborate the theoretical ones. For the case with smaller variance, the improvement on gen-error is barely visible as λ increases. For the case with larger variance, the decrease of the gen-error is more pronounced, which implies that ℓ_2-regularization can effectively mitigate the impact on the gen-error induced by large class overlapping and pseudo-labels with low accuracy.

The adept reader might naturally wonder why one would not set $\lambda \to \infty$ in (15), which results in the gen-error tending to zero, which presumably is a desirable phenomenon. However, ultimately, what we wish to control is the expected population risk, which, according to (2), is the sum of the expected empirical risk on the training data and the gen-error. Even if the gen-error is zero, the expected population risk might be large. Hence, as λ increases, we see a tradeoff between the gen-error and the empirical risk.

6. Experiments on Benchmark Datasets

To further illustrate that our theory is indeed behind the empirical behaviour of the iterative self-training with pseudo-labelling, in this section, we conduct experiments on real-world benchmark datasets, which demonstrates that our theoretical results on the bGMM example can also reflect the training dynamics on more realistic real-world tasks. The code to reproduce all the experiments can be found at `https://github.com/HerianHe/GenErrorSSL_2022.git`.

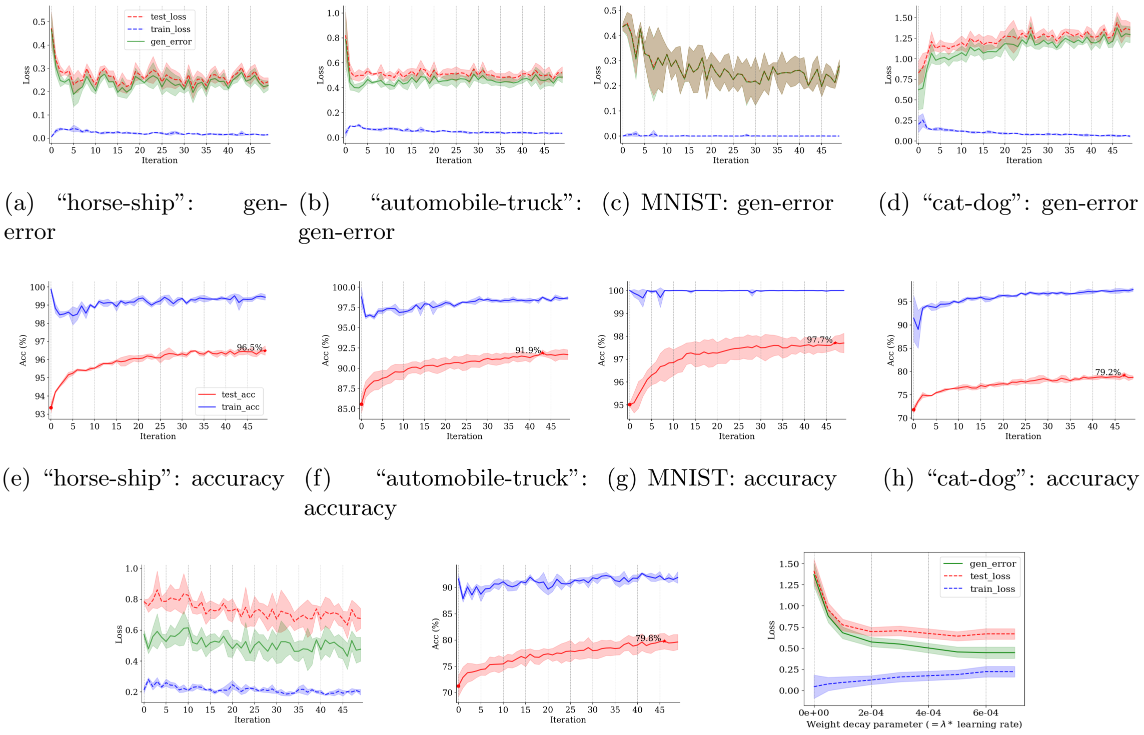
Recall that in the bGMM example, a higher standard deviation σ represents a higher in-class variance, larger class-overlap, and consequently higher difficulty in classification. By a whitening argument, this also holds for bGMMs with non-isotropic covariance matrices. In our experiments on real-world data, we use the difficulty level of classification to mimic different in-class variances of bGMM. We pick two easy-to-distinguish class pairs ("automobile" and "truck", "horse" and "ship") from the CIFAR-10 dataset (Krizhevsky, 2009) as an analogy to bGMM with small in-class variance, and one difficult-to-distinguish class pair ("cat" and "dog") from the same dataset as an analogy to bGMM with large in-class variance. Furthermore, to extend the analogy to multi-class classification, we conduct experiments on the 10-class MNIST dataset to gain more intuition.

We train deep neural networks (DNNs) via an iterative self-learning strategy (under the same setting as Figure 1) to perform binary and multi-class classification. In the first iteration, we only use a few labelled data examples to initialize the DNN with a sufficient number of epochs. In the subsequent iterations, we first sample a subset of unlabelled data and generate pseudo-labels for them via the model trained from 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", and "cat" and "dog" from the CIFAR10 (Krizhevsky, 2009) dataset. In this dataset, each class has 5000 images for training and 1000 images for testing. For each selected pair of classes, we manually divide the 10000 training images into two sets: 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), and use stochastic gradient descent (SGD) optimizer to minimize the cross-entropy loss. In PyTorch, the cross-entropy loss is defined as the negative logarithm of the output softmax probability corresponding to the true class, which is analogous to the NLL of the data under the parameters of the neural network. For the task with ℓ_2-regularization, we train the neural network by setting different weight decay parameters (equivalent to $\lambda\times$ learning rate). In each pseudo-labelling iteration, we sample 2500 unlabelled images. The complete training procedure lasts for 50 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 hand-written 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 without regularization.

Experimental observations: We perform each experiment 3 times and report the average test and training (cross entropy) losses, the gen-error, and test and training accuracies in Figures 9. To illustrate the difficulty level of classification for each pair, we first calculated the mean and variance of the RGB (i.e., the red-green-blue color values) values of the images to show the difference of the images between the two classes. 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 positive and negative classes). Furthermore, in Table 2, we quote the commonly-used confusion matrix for the CIFAR10 dataset in (Liu and Mukhopadhyay, 2018, Fig. 7), which quantifies how many out of 1000 images of each class



(a) “horse-ship”: gen-error

(b) “automobile-truck”: gen-error

(c) MNIST: gen-error

(d) “cat-dog”: gen-error

(e) “horse-ship”: accuracy

(f) “automobile-truck”: accuracy

(g) MNIST: accuracy

(h) “cat-dog”: accuracy

(i) “cat-dog”: gen-error with weight decay set to 5×10^{-4}.

(j) “cat-dog”: accuracy with weight decay set to 5×10^{-4}.

(k) “cat-dog”: gen-error after convergence versus weight decay.

Figure 9: (a)–(c) & (e)–(g): easier-to-distinguish classes “horse” vs. “ship” and “automobile” vs. “truck”; (d),(h), (i)–(k): harder-to-distinguish classes “cat” vs. “dog”.

are misclassified to any other class. It is obvious that fewer misclassified images indicates lower classification difficulty, which corresponds with Table 1. These two tables provide an indication of the level of difficulty to distinguish different pairs of classes.

In Figures 9(a)–9(c), for easier-to-distinguish classes (based on the high classification accuracy and low loss, as well as Tables 1 and 2), the gen-error appears to have relatively large reduction in the early training iterations and then fluctuates around a constant value afterwards. For example, in Figure 9(a), the gen-error converges to around 0.25 after 5 iterations; in Figure 9(b), it converges to around 0.45 after 5 iterations. For multi-classification of MNIST in Figure 9(c), the gen-error also converges to around 0.25 after 20 iterations. These results corroborate the theoretical and empirical analyses in the bGMM case with small variance, which again verifies that Theorem 3 and Corollary 10 can shed light on the empirical gen-error 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. In Figures 9(e), 9(f) and 9(g), 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 Figures 9(d) and 9(h), we perform another binary classification experiment on the harder-to-distinguish pair, “cat” and “dog” (based on low accuracy at the initial point

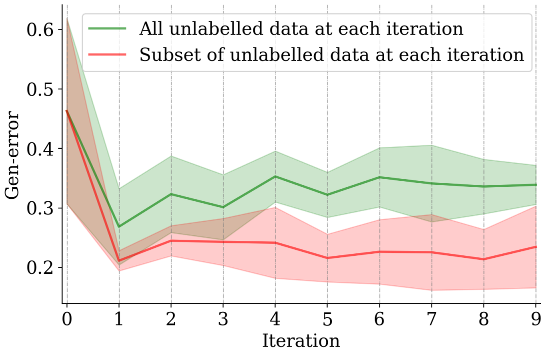


Figure 10: Comparison of the gen-error for the "horse" and "ship" classification task

as well as Tables 1 and 2). We observe that the gen-error (and the test loss) does not decrease across the self-training iterations even though the test accuracy increases. This again corroborates the result in Figure 3(b) for the bGMM with large variance. The fact that both the test loss and test accuracy appear to increase with the iteration is, in fact, not contradictory. To intuitively explain this, in binary classification using the softmax (hence, logistic) function to predict the output classes, suppose the learned probability of a data example belonging to its true class is $p \in (1/2, 1]$, 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 increases, the test loss may also increase at the same time; this represents our lack of confidence.

We further investigate the effect of ℓ_2-regularization on "cat-dog" classification. In Figures 9(i) and 9(j), we show that by setting the weight decay parameter to be 0.0005, the increase of gen-error for the "cat-dog" classification task can be mitigated and the test accuracy is improved by 0.6% as well; compare this to Figures 9(d) and 9(h). In Figure 9(k), we plot the average gen-error over the last 10 iterations versus the weight decay parameter; this is shown to decrease as the weight decay increases (compared to Figure 8). In summary, our above observations correspond to that for the bGMM, namely that the unlabelled data do not always help to improve the gen-error but adding regularization can help to compensate for the undesirable impact.

Furthermore, we study the effect of reusing *all* the unlabelled data at each iteration. Under the same experimental setup as above, we conduct an additional experiment on the "horse-ship" pair in the CIFAR-10 dataset by using *all* $9,500$ unlabelled images in *each* iteration. The self-training procedure lasts for 10 iterations. Figure 10 compares the gen-error of this additional experiment with the one of the same experiment in Figure 9(a). We find that when using the unlabelled data all at once, as the pseudo-labelling iteration increases, the gen-error is even higher than that for our original setup. This can possibly be attributed to overfitting.

7. Concluding Remarks and Future Work

In this paper, we have analyzed the gen-error of iterative SSL algorithms that pseudo-label large amounts of unlabelled data to progressively refine the parameters of a given model. We particularized the general bounds and exact expressions on the gen-error for the bGMM to gain some theoretical insight into the problem. These were then corroborated by experiments on benchmark datasets. The theoretical analyses and experimental results reinforce the main message of this paper—namely, that in the low-class-overlap or easy-to-classify scenario, pseudo-labelling can help to reduce the gen-error. On the other hand, for the high-class-overlap or difficult-to-classify scenario, pseudo-labelling can in fact hurt. Thus, the key takeaway from our paper is that practitioners should be judicious in adopting pseudo-labelling techniques, for they may degrade the overall performance.

There are three avenues for future research. First, our analytical results are only applicable to the bGMM. This yields valuable insights, but the model is admittedly restrictive. Generalizing our analyses to other statistical models for classification such as logistic regression will be instructive. Secondly, our work focuses on the gen-error. Often bounds on the *population risk* are desired as the population risk is the key determinant of the performance of classification algorithms. Bounding the population risk in the SSL setting would thus be interesting. Finally, analyzing other families of SSL algorithms beyond those that utilize pseudo-labelling would provide a clearer theoretical picture about the utility of SSL.

Acknowledgements

We sincerely thank the reviewers and action editor for their meticulous reading and insightful comments that led to a significantly improved version of the paper.

This research/project is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-018), by an NRF Fellowship (A-0005077-01-00), and by Singapore Ministry of Education (MOE) AcRF Tier 1 Grants (A-0009042-01-00 and A-8000189-01-00).

Appendices

A. Proof of Theorem 2.A

We commence with some notation. For any convex function $\psi : [0, b) \mapsto \mathbb{R}$, its *Legendre dual* ψ^* 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 $y \geq 0$, 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 5. *For any $\tilde{\theta}_t \in \Theta$, let $\psi_-(\lambda, \tilde{\theta}_t)$ and $\psi_+(\lambda, \tilde{\theta}_t)$ be convex functions of λ and $\psi_+(0, \tilde{\theta}_t) = \psi'_+(0, \tilde{\theta}_t) = \psi_-(0, \tilde{\theta}_t) = \psi'_-(0, \tilde{\theta}_t) = 0$. Assume that $\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t) \leq \psi_+(\lambda, \tilde{\theta}_t)$ for all $\lambda \in [0, b_+)$ and $\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t) \leq \psi_-(\lambda, \tilde{\theta}_t)$ for $\lambda \in (b_-, 0]$ under distribution $P_{\tilde{Z}|\theta^{(t-1)}} = P_Z$, where $0 < b_+ \leq \infty$ and $-\infty \leq b_- < 0$. Let $\psi_+(\lambda) = \sup_{\tilde{\theta}_t} \psi_+(\lambda, \tilde{\theta}_t)$ and $\psi_-(\lambda) =$*

$\sup_{\tilde{\theta}_t} \psi_-(\lambda, \tilde{\theta}_t)$. *We have*

$$\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&\le \frac{w}{n}\sum_{i=1}^n \mathbb{E}_{\theta^{(t-1)}}\left[\psi_-^{*-1}(I_{\theta^{(t-1)}}(\theta_t; Z_i))\right] \\
&\quad + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_-^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i',\hat{Y}_i'}\|P_Z)\big)\right],
\end{aligned}$$

and

$$\begin{aligned}
&-\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&\le \frac{w}{n}\sum_{i=1}^n \mathbb{E}_{\theta^{(t-1)}}\left[\psi_+^{*-1}(I_{\theta^{(t-1)}}(\theta_t; Z_i))\right] \\
&\quad + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_+^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t; X_i', \hat{Y}_i') + D_{\theta^{(t-1)}}(P_{X_i',\hat{Y}_i'}\|P_Z)\big)\right],
\end{aligned}$$

where $P_{X_i',\hat{Y}_i'|\theta^{(t-1)}}(x, y|\hat{\theta}^{(t-1)}) = P_X(x)\mathbb{1}\{y = f_{\hat{\theta}_{t-1}}(x)\}$ *for any* $x \in \mathcal{X}$*,* $y \in \mathcal{Y}$ *and* $\hat{\theta}^{(t-1)} \in \Theta^{t-1}$*, and* $P_{Z|\theta^{(t-1)}} = P_Z$*.*

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

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

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

Recall that $\tilde{\theta}_t$ and $\tilde{Z}$ are independent copies of θ_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

$$\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&= w\bigg(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)]] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta_t,Z_i}[l(\theta_t, Z_i)]\bigg) \\
&\quad + (1-w)\bigg(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)]] - \frac{1}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\bigg) \\
&= \frac{w}{n}\sum_{i=1}^n \bigg(\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})] - \mathbb{E}_{\theta_t,Z_i}[l(\theta_t, Z_i)]\bigg) \\
&\quad + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t} \bigg(\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}[l(\theta_t, (X_i', \hat{Y}_i'))]\bigg) \\
&= \frac{w}{n}\sum_{i=1}^n \mathbb{E}_{\theta^{(t-1)}}\Big[\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}\big] - \mathbb{E}_{\theta_t,Z_i}\big[l(\theta_t, Z_i)|\theta^{(t-1)}\big]\Big] \\
&\quad + \frac{1-w}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta^{(t-1)}}\Big[\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}\big] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}\big[l(\theta_t, (X_i', \hat{Y}_i'))|\theta^{(t-1)}\big]\Big]. \quad (16)
\end{aligned}$$

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 ψ_-^*, ψ_+^*, and the corresponding inverses are well-defined.

Let $\check{l}(\theta, z) = l(\theta, z) - \mathbb{E}_Z[l(\theta, Z)]$. We have the fact that $\mathbb{E}_{\tilde{Z}}[\check{l}(\tilde{\theta}_t, \tilde{Z})] = 0$ 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

$$\begin{aligned}
&I_{\theta^{(t-1)}}(\theta_t; Z) = D(P_{\theta_t, Z|\theta^{(t-1)}} \| P_{\theta_t|\theta^{(t-1)}} \otimes P_Z) \\
&\geq \mathbb{E}_{\theta_t, Z}[\lambda \check{l}(\theta_t, Z)|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[e^{\lambda \check{l}(\tilde{\theta}_t, \tilde{Z})}|\theta^{(t-1)}] \\
&= \mathbb{E}_{\theta_t, Z}[\lambda \check{l}(\theta_t, Z)|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}} \mathbb{E}_{\tilde{Z}}\big[e^{\lambda \check{l}(\tilde{\theta}_t, \tilde{Z})}\big] \\
&= \mathbb{E}_{\theta_t, Z}[\lambda \check{l}(\theta_t, Z)|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\big[\exp\big(\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t)\big)\big] && (17) \\
&\geq \lambda \mathbb{E}_{\theta_t, Z}[l(\theta_t, Z) - \mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)}] - \log \mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\big[\exp(\psi_+(\lambda, \tilde{\theta}_t))\big] && (18) \\
&\geq \lambda \mathbb{E}_{\theta_t, Z}[l(\theta_t, Z) - \mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)}] - \psi_+(\lambda) && (19) \\
&= \lambda\big(\mathbb{E}_{\theta_t, Z}[l(\theta_t, Z)|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}]\big) - \psi_+(\lambda).
\end{aligned}$$

where (17) follows from the definition of $\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t)$ in (25), (18) follows from the assumption that $\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t) \leq \psi_+(\lambda, \tilde{\theta}_t)$ for all $\lambda \in [0, b_+)$, and (19) follows because $\psi_+(\lambda) = \sup_{\tilde{\theta}_t} \psi_+(\lambda, \tilde{\theta}_t)$. Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_t, Z}[l(\theta_t, Z)|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] \\
&\leq \inf_{\lambda \in [0, b_+)} \frac{I_{\theta^{(t-1)}}(\theta_t; Z) + \psi_+(\lambda)}{\lambda} = \psi_+^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t; Z)\big). && (20)
\end{aligned}$$

Similarly, for $\lambda \in (b_-, 0]$,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\theta}_t, \tilde{Z}}[l(\tilde{\theta}_t, \tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\theta_t, Z}[l(\theta_t, Z)|\theta^{(t-1)}] \\
&\leq \inf_{\lambda \in [0, -b_-)} \frac{I_{\theta^{(t-1)}}(\theta_t; Z) + \psi_-(\lambda)}{\lambda} = \psi_-^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t; Z)\big). && (21)
\end{aligned}$$

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

$$\begin{aligned}
&I_{\theta^{(t-1)}}(\theta_t; X', \hat{Y}') + D_{\theta^{(t-1)}}(P_{X', \hat{Y}'} \| P_Z) \\
&= D_{\theta^{(t-1)}}(P_{\theta_t, X', \hat{Y}'} \| P_{\theta_t} \otimes P_{X', \hat{Y}'}) + D_{\theta^{(t-1)}}(P_{\theta_t} \otimes P_{X', \hat{Y}'} \| P_{\theta_t} \otimes P_Z) \\
&\geq \mathbb{E}_{\theta_t, X', \hat{Y}'}[\lambda l(\theta_t, (X', \hat{Y}'))|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t}\big[\mathbb{E}_{X', \hat{Y}'}[e^{\lambda l(\theta_t, (X', \hat{Y}'))}|\theta^{(t-1)}]|\theta^{(t-1)}\big] \\
&\quad + \mathbb{E}_{\theta_t}\big[\mathbb{E}_{X', \hat{Y}'}[\lambda l(\theta_t, (X', \hat{Y}'))|\theta^{(t-1)}]|\theta^{(t-1)}\big] - \log \mathbb{E}_{\theta_t}\big[\mathbb{E}_Z[e^{\lambda l(\theta_t, Z)}]|\theta^{(t-1)}\big] \\
&\geq \mathbb{E}_{\theta_t, X', \hat{Y}'}[\lambda l(\theta_t, (X', \hat{Y}'))|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t}\big[\mathbb{E}_Z[e^{\lambda l(\theta_t, Z)}]|\theta^{(t-1)}\big] && (22) \\
&= \lambda\Big(\mathbb{E}_{\theta_t, X', \hat{Y}'}[\lambda l(\theta_t, (X', \hat{Y}'))|\theta^{(t-1)}] - \mathbb{E}_{\theta_t}\big[\mathbb{E}_Z[l(\theta_t, Z)]|\theta^{(t-1)}\big]\Big) \\
&\quad - \log \mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\big[\exp\big(\Lambda_{l(\tilde{\theta}_t, \tilde{Z})}(\lambda, \tilde{\theta}_t)\big)\big] \\
&\geq \lambda\Big(\mathbb{E}_{\theta_t, X', \hat{Y}'}[\lambda l(\theta_t, (X', \hat{Y}'))] - \mathbb{E}_{\theta_t}\big[\mathbb{E}_Z[l(\theta_t, Z)]\big]\Big) - \psi_+(\lambda),
\end{aligned}$$

where (22) follows from Jensen's inequality. Thus, we get

$$\begin{aligned}&\mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}\big[l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}\big] - \mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\big]\\ &\le \psi_+^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t;X',\hat{Y}') + D_{\theta^{(t-1)}}(P_{X',\hat{Y}'}\|P_Z)\big)\end{aligned} \tag{23}$$

and

$$\begin{aligned}&\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\big] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}\big[l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}\big]\\ &\le \psi_-^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t;X',\hat{Y}') + D_{\theta^{(t-1)}}(P_{X',\hat{Y}'}\|P_Z)\big).\end{aligned} \tag{24}$$

The proof is completed by applying inequalities (20), (21), (23) and (24) to the expansion of gen_t in (16). ■

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

$$\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t) := \log \mathbb{E}_{\tilde{Z}}[e^{\lambda(l(\tilde{\theta}_t,\tilde{Z})-\mathbb{E}_{\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})])}]. \tag{25}$$

When the loss function $l(\theta,Z) \sim \text{subG}(R)$ under $Z \sim P_Z$ for any $\theta \in \Theta$, we have $\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t) \le \frac{R^2\lambda^2}{2}$ 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, $\psi_+(\lambda) = \psi_-(\lambda) = \sup_{\tilde{\theta}_t\in\Theta} \frac{R^2\lambda^2}{2} = \frac{R^2\lambda^2}{2}$ and $\psi_+^{*-1}(y) = \psi_-^{*-1}(y) = \sqrt{2R^2y}$ for any $y \ge 0$. Finally, Theorem 2.A can then be directly obtained from Theorem 5.

B. Proof of Theorem 2.B

Recall that $\tilde{\theta}_t$ and $\tilde{Z}$ are independent copies of θ_t and Z respectively, such that $P_{\tilde{\theta}_t,\tilde{Z}} = Q_{\theta_t} \otimes P_Z$, $P_{\tilde{\theta}_t,\tilde{Z}|\theta^{(t-1)}} = P_{\theta_t|\theta^{(t-1)}} \otimes P_Z$.

Recall the gen-error given in (16). The first term in (16) can be rewritten as

$$\begin{aligned}&\mathbb{E}_{\theta^{(t-1)}}\Big[\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\big] - \mathbb{E}_{\theta_t,Z_i}\big[l(\theta_t,Z_i)|\theta^{(t-1)}\big]\Big]\\ &= \mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\big[-\log p_{\tilde{\theta}_t}\big] - \mathbb{E}_{\theta_t,Z_i}\big[-\log p_{\theta_t}\big]\\ &= \mathbb{E}_{\theta_t}\Big[h(P_Z,p_{\theta_t}) - h(P_{Z_i|\theta_t},p_{\theta_t})\Big]\\ &= \mathbb{E}_{\theta_t}\Big[\Delta\text{h}(P_Z\|P_{Z_i|\theta_t}|p_{\theta_t})\Big].\end{aligned} \tag{26}$$

The second term in (16) can be rewritten as

$$\begin{aligned}&\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}[l(\theta_t,(X_i',\hat{Y}_i'))]\\ &= \mathbb{E}_{\theta_t,Z_i}[-\log p_{\theta_t}] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}[-\log p_{\theta_t}]\\ &= \mathbb{E}_{\theta^{(t-1)}}\Big[\mathbb{E}_{\theta_t|\theta^{(t-1)}}\big[h(P_Z,p_{\theta_t}) - h(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}},p_{\theta_t}) + h(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}},p_{\theta_t})\\ &\quad - h(P_{X_i',\hat{Y}_i'|\theta^{(t)}},p_{\theta_t})\big]\Big]\\ &= \mathbb{E}_{\theta^{(t)}}\Big[\Delta\text{h}(P_Z\|P_{X_i',\hat{Y}_i'|\theta^{(t-1)}}|p_{\theta_t}) + \Delta\text{h}(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}}\|P_{X_i',\hat{Y}_i'|\theta^{(t)}}|p_{\theta_t})\Big].\end{aligned} \tag{27}$$

By combining (26) and (27), the gen-error is finally given by

$$\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^{t}, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&= \frac{w}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta_t}\Big[\Delta\mathrm{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t})\Big] \\
&\quad + \frac{w}{m}\sum_{i\in\mathcal{I}_t} \mathbb{E}_{\theta^{(t)}}\Big[\Delta\mathrm{h}(P_Z \| P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} | p_{\theta_t}) + \Delta\mathrm{h}(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} \| P_{X_i',\hat{Y}_i'|\theta^{(t)}} | p_{\theta_t})\Big] \\
&= \mathbb{E}_{\theta^{(t)}}\Bigg[\frac{w}{n}\sum_{i=1}^{n} \Delta\mathrm{h}(P_Z \| P_{Z_i|\theta_t} | p_{\theta_t}) + \frac{w}{m}\sum_{i\in\mathcal{I}_t} \big(\Delta\mathrm{h}(P_Z \| P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} | p_{\theta_t}) \\
&\quad + \Delta\mathrm{h}(P_{X_i',\hat{Y}_i'|\theta^{(t-1)}} \| P_{X_i',\hat{Y}_i'|\theta^{(t)}} | p_{\theta_t})\big)\Bigg].
\end{aligned}$$

Theorem 2.B is thus proved.

C. Proof of Theorem 3

In the following, we abbreviate $\mathrm{gen}_t(P_\mathbf{Z}, P_\mathbf{X}, \{P_{\boldsymbol{\theta}_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^{t}, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})$ as gen_t, if there is no risk of confusion. When the labelled data are not reused in the subsequent iterations, for $t \geq 1$, $w = 0$.

- **Iteration $t = 0$:** Since $Y_i\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$, we have $\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$. The gen-error gen_0 is given by

$$\begin{aligned}
\mathrm{gen}_0 &= \mathbb{E}_{\boldsymbol{\theta}_0}\Big[\mathbb{E}_\mathbf{Z}[-\log p_{\boldsymbol{\theta}_0}(\mathbf{Z})] - \mathbb{E}_{\mathbf{Z}_i|\boldsymbol{\theta}_0}[-\log p_{\boldsymbol{\theta}_0}(\mathbf{Z}_i)]\Big] \\
&= \int Q_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})(P_\mathbf{Z}(\mathbf{z}) - P_{\mathbf{Z}_i|\boldsymbol{\theta}_0}(\mathbf{z}|\boldsymbol{\theta}))\log\frac{1}{p_{\boldsymbol{\theta}}(\mathbf{z})}\mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{2\sigma^2}\int Q_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})(P_\mathbf{Z}(\mathbf{x},y) - P_{\mathbf{Z}_i|\boldsymbol{\theta}_0}(\mathbf{x},y|\boldsymbol{\theta}))\big(\mathbf{x}^\top\mathbf{x} - 2y\boldsymbol{\theta}^\top\mathbf{x} + \boldsymbol{\theta}^\top\boldsymbol{\theta}\big)\mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}\int P_\mathbf{Z}(\mathbf{x},y)(Q_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) - P_{\boldsymbol{\theta}_0|\mathbf{Z}_i}(\boldsymbol{\theta}|\mathbf{x},y))2y\boldsymbol{\theta}^\top\mathbf{x}\ \mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}\int \Big(P_\mathbf{Z}(\mathbf{x},1)(Q_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) - P_{\boldsymbol{\theta}_0|\mathbf{Z}_i}(\boldsymbol{\theta}|\mathbf{x},1)) \\
&\qquad\qquad - P_\mathbf{Z}(\mathbf{x},-1)(Q_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) - P_{\boldsymbol{\theta}_0|\mathbf{Z}_i}(\boldsymbol{\theta}|\mathbf{x},-1))\Big)2\boldsymbol{\theta}^\top\mathbf{x}\ \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{\sigma^2}\int \Big(\frac{\boldsymbol{\mu}-\mathbf{x}}{n}P_\mathbf{Z}(\mathbf{x},1) - \frac{\boldsymbol{\mu}+\mathbf{x}}{n}P_\mathbf{Z}(\mathbf{x},-1)\Big)^\top\mathbf{x}\ \mathrm{d}\mathbf{x} \\
&= -\frac{1}{\sigma^2}\bigg(-\frac{d\sigma^2}{2n} - \frac{d\sigma^2}{2n}\bigg) \\
&= \frac{d}{n}.
\end{aligned}$$

- **Pseudo-label using $\boldsymbol{\theta}_0$:** For any $i \in [1:m]$ and $X_i' \in S_\mathrm{u}$, the pseudo-label is

$$\hat{Y}_i' = \mathrm{sgn}(\boldsymbol{\theta}_0^\top\mathbf{X}_i').$$

Given any pair of $(\xi_0, \boldsymbol{\mu}^{\perp})$, $\boldsymbol{\theta}_0$ is fixed and $\{\hat{Y}_i'\}_{i\in[1:m]}$ are conditionally i.i.d. from $P_{\hat{Y}'|\xi_0,\boldsymbol{\mu}^{\perp}} \in \mathcal{P}(\mathcal{Y})$. Recall the pseudo-labelled dataset is defined as $\hat{S}_{u,1} = \{(\mathbf{X}_i', \hat{Y}_i')\}_{i=1}^m$.

Since $\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$, inspired by Oymak and Gülcü (2021), we can decompose it as follows:

$$\boldsymbol{\theta}_0 = \boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\xi} = \boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}(\xi_0\boldsymbol{\mu} + \boldsymbol{\mu}^{\perp}) = \left(1 + \frac{\sigma}{\sqrt{n}}\xi_0\right)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^{\perp}, \tag{28}$$

where $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_d)$, $\xi_0 \sim \mathcal{N}(0,1)$, $\boldsymbol{\mu}^{\perp} \perp \boldsymbol{\mu}$, $\boldsymbol{\mu}^{\perp} \sim \mathcal{N}(0, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^{\top})$ and $\boldsymbol{\mu}^{\perp}$ is independent of ξ_0.

Recall the correlation between $\boldsymbol{\theta}_0$ and $\boldsymbol{\mu}$ given in (7), the decomposition of $\bar{\boldsymbol{\theta}}_0$ in (8) and α, β. Since $\operatorname{sgn}(\boldsymbol{\theta}_0^{\top}\mathbf{X}_i') = \operatorname{sgn}(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i')$, in the following we can analyze the normalized parameter $\bar{\boldsymbol{\theta}}_0$ instead.

Given any $(\xi_0, \boldsymbol{\mu}^{\perp})$, α is fixed, and for any $i \in \mathbb{N}$, let us define a Gaussian noise vector $\mathbf{g}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ and decompose it as follows

$$\mathbf{g}_i = g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{v} + \mathbf{g}_i^{\perp}, \tag{29}$$

where $g_{0,i}, g_i \sim \mathcal{N}(0,1)$, $\mathbf{g}_i^{\perp} \sim \mathcal{N}(0, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^{\top} - \boldsymbol{v}\boldsymbol{v}^{\top})$, $\mathbf{g}_i^{\perp} \perp \boldsymbol{\mu}$, $\mathbf{g}_i^{\perp} \perp \boldsymbol{v}$, and $g_{0,i}, g_i, \mathbf{g}_i^{\perp}$ are mutually independent.

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

$$\mathbf{X}_i' = \boldsymbol{\mu} + \sigma\mathbf{g}_i = \boldsymbol{\mu} + \sigma(g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{v} + \mathbf{g}_i^{\perp}). \tag{30}$$

Then we have

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i' &= (\alpha\boldsymbol{\mu} + \beta\boldsymbol{v})^{\top}(\boldsymbol{\mu} + \sigma\mathbf{g}_i) \\
&= \alpha + \sigma(\alpha\boldsymbol{\mu} + \beta\boldsymbol{v})^{\top}(g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{v} + \mathbf{g}_i^{\perp}) \\
&= \alpha + \sigma(\alpha g_{0,i} + \beta g_i) =: \alpha + \sigma h_i. \qquad (31)
\end{aligned}$$

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

$$\mathbf{X}_i' = -\boldsymbol{\mu} + \sigma\mathbf{g}_i$$

and

$$\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i' = -\alpha + \sigma h_i.$$

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

$$\begin{aligned}
\Pr(\hat{Y}_i' = 1) &= \Pr\left(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i' > 0\right) \\
&= \frac{1}{2}\Pr\left(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i' > 0 | Y_i' = 1\right) + \frac{1}{2}\Pr\left(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_i' > 0 | Y_i' = -1\right) \\
&= \frac{1}{2}\mathbb{E}_{\alpha}\big[\Pr(\alpha + \sigma h_i > 0)\big] + \frac{1}{2}\mathbb{E}_{\alpha}\big[\Pr\left(-\alpha + \sigma h_i > 0\right)\big] \\
&= \frac{1}{2}\mathbb{E}_{\alpha}\left[\mathrm{Q}\left(-\frac{\alpha}{\sigma}\right)\right] + \frac{1}{2}\mathbb{E}_{\alpha}\left[\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right] = \frac{1}{2}. \qquad (32)
\end{aligned}$$

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

- **Iteration $t = 1$:** Recall (6) and the new model parameter learned from the pseudo-labelled dataset $\hat{S}_{\mathrm{u},1}$ is given by

$$\boldsymbol{\theta}_1 = \frac{1}{m}\sum_{i=1}^{m} \hat{Y}_i' \mathbf{X}_i' = \frac{1}{m}\sum_{i=1}^{m} \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}_i')\mathbf{X}_i' = \frac{1}{m}\sum_{i=1}^{m} \mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')\mathbf{X}_i'. \tag{33}$$

First let us calculate the conditional expectation of $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_0$. Given any $(\xi_0, \boldsymbol{\mu}^\perp)$, for any $j \in [1 : m]$, let $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} := \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' | \xi_0, \boldsymbol{\mu}^\perp]$ and $\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}$ denotes the probability measure under the parameters $(\xi_0, \boldsymbol{\mu}^\perp)$.

The expectation $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ can be calculated as follows:

$$\begin{aligned}
\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} &= \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' | \xi_0, \boldsymbol{\mu}^\perp] \\
&= \mathbb{E}_{Y_j'}[\ \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j']\ ] \\
&= \frac{1}{2}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = -1] + \frac{1}{2}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = 1].
\end{aligned}$$

In contrast to (29), here we decompose the Gaussian random vector $\mathbf{g}_j \sim \mathcal{N}(0, \mathbf{I}_d)$ in another way

$$\mathbf{g}_j = \tilde{g}_j \bar{\boldsymbol{\theta}}_0 + \tilde{\mathbf{g}}_j^\perp, \tag{34}$$

where $\tilde{g}_j \sim \mathcal{N}(0,1)$, $\tilde{\mathbf{g}}_j^\perp \sim \mathcal{N}(0, \mathbf{I}_d - \bar{\boldsymbol{\theta}}_0\bar{\boldsymbol{\theta}}_0^\top)$, $\tilde{g}_j$ and $\tilde{\mathbf{g}}_j^\perp$ are mutually independent and $\tilde{\mathbf{g}}_j^\perp \perp \bar{\boldsymbol{\theta}}_0$.

Then we decompose $\mathbf{X}_j'$ and $\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j'$ as

$$\begin{aligned}
\mathbf{X}_j' &= Y_j'\boldsymbol{\mu} + \sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0 + \sigma\tilde{\mathbf{g}}_j^\perp, \text{ and} && (35) \\
\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j' &= Y_j'\alpha + \sigma\tilde{g}_j. && (36)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = -1] \\
&= \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)(-\boldsymbol{\mu} + \sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0 + \sigma\tilde{\mathbf{g}}^\perp) \mid \xi_0, \boldsymbol{\mu}^\perp] \\
&= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{g}_j|\xi_0, \boldsymbol{\mu}^\perp]\bar{\boldsymbol{\theta}}_0 \\
&\quad + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{\mathbf{g}}^\perp|\xi_0, \boldsymbol{\mu}^\perp] \\
&= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{g}_j|\xi_0, \boldsymbol{\mu}^\perp]\bar{\boldsymbol{\theta}}_0, && (37)
\end{aligned}$$

where (37) follows since $\tilde{\mathbf{g}}^\perp$ is independent of $\tilde{g}_j$ and $\mathbb{E}[\tilde{\mathbf{g}}^\perp] = 0$.

For the first term in (37), recall $\tilde{g}_j \sim \mathcal{N}(0,1)$ and we have

$$-\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} = \left(1 - 2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right)\boldsymbol{\mu}. \tag{38}$$

For the second term in (37), we have

$$
\begin{aligned}
&\mathbb{E}[\mathrm{sgn}(-\alpha+\sigma\tilde{g}_j)\tilde{g}_j|\xi_0,\boldsymbol{\mu}^{\perp}]\bar{\boldsymbol{\theta}}_0\\
&=\Big(-\int_{-\infty}^{\frac{\alpha}{\sigma}}\frac{1}{\sqrt{2\pi}}e^{-\frac{g^2}{2}}g\ \mathrm{d}g+\int_{\frac{\alpha}{\sigma}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{g^2}{2}}g\ \mathrm{d}g\Big)\bar{\boldsymbol{\theta}}_0=\frac{2}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\bar{\boldsymbol{\theta}}_0. \quad (39)
\end{aligned}
$$

By combining (38) and (39), we have

$$
\mathbb{E}\big[\,\mathrm{sgn}\,(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_j')\mathbf{X}_j'\mid\xi_0,\boldsymbol{\mu}^{\perp},Y_j'=-1\big]=\Big(1-2\mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big)\Big)\boldsymbol{\mu}+\frac{2\sigma}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\bar{\boldsymbol{\theta}}_0,
$$

and similarly,

$$
\mathbb{E}\big[\,\mathrm{sgn}\,(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_j')\mathbf{X}_j'\mid\xi_0,\boldsymbol{\mu}^{\perp},Y_j'=1\big]=\Big(2\mathrm{Q}\Big(-\frac{\alpha}{\sigma}\Big)-1\Big)\boldsymbol{\mu}+\frac{2\sigma}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\bar{\boldsymbol{\theta}}_0.
$$

Thus, recall J_σ and K_σ defined in (10) and (11), $\bar{\boldsymbol{\theta}}_0=\alpha\boldsymbol{\mu}+\beta\boldsymbol{\upsilon}$ and $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}}$ is given by

$$
\begin{aligned}
\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}}&=\mathbb{E}[\mathrm{sgn}(\boldsymbol{\theta}_0^{\top}\mathbf{X}_j')\mathbf{X}_j'|\xi_0,\boldsymbol{\mu}^{\perp}]\\
&=\Big(1-2\mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big)\Big)\boldsymbol{\mu}+\frac{2\sigma}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\bar{\boldsymbol{\theta}}_0\\
&=\Big(1-2\mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\Big)\boldsymbol{\mu}+\frac{2\sigma\beta}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\boldsymbol{\upsilon}\\
&=J_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^{\perp}))\boldsymbol{\mu}+K_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^{\perp}))\boldsymbol{\upsilon}. \qquad (40)
\end{aligned}
$$

From (3), the gen-error at $t=1$ is given by

$$
\begin{aligned}
\mathrm{gen}_1=\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^{\perp}}\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}\Big[&\Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta}_1}|p_{\boldsymbol{\theta}_1})\\
&+\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}|p_{\boldsymbol{\theta}_1})\Big]. \qquad (41)
\end{aligned}
$$

Next, we calculate the two Δh terms in (41) respectively.

- **Calculate** $\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}\big[\Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta}_1}|p_{\boldsymbol{\theta}_1})\big]$**:**

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}\big[\Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta}_1}|p_{\boldsymbol{\theta}_1})\big]\\
&=\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}\Big[h(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}},p_{\boldsymbol{\theta}_1})-h(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta}_1},p_{\boldsymbol{\theta}_1})\Big]\\
&=\frac{1}{2\sigma^2}\int Q_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}(\boldsymbol{\theta}|\xi_0,\boldsymbol{\mu}^{\perp})\big(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^{\perp})\\
&\qquad\qquad-P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta}_1}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^{\perp},\boldsymbol{\theta})\big)(\mathbf{x}^{\top}\mathbf{x}-2y\boldsymbol{\theta}^{\top}\mathbf{x}+\boldsymbol{\theta}^{\top}\boldsymbol{\theta})\mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta}\\
&=-\frac{1}{\sigma^2}\int P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^{\perp}}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^{\perp})\big(Q_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^{\perp}}(\boldsymbol{\theta}|\xi_0,\boldsymbol{\mu}^{\perp})\\
&\qquad\qquad-P_{\boldsymbol{\theta}_1|\mathbf{X}_i',\hat{Y}_i',\xi_0,\boldsymbol{\mu}^{\perp}}(\boldsymbol{\theta}|\mathbf{x},y,\xi_0,\boldsymbol{\mu}^{\perp})\big)(y\boldsymbol{\theta}^{\top}\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta}
\end{aligned}
$$

$$
\begin{aligned}
&= -\frac{1}{\sigma^2}\int P_{\mathbf{X}'_i,\hat{Y}'_i|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp)\bigg(\frac{\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} - y\mathbf{x}}{m}\bigg)^\top (y\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y \\
&= \frac{1}{m\sigma^2}\Big(\mathbb{E}[\mathbf{X}_i'^\top \mathbf{X}'_i|\xi_0,\boldsymbol{\mu}^\perp] - (\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\Big) \\
&= \frac{d\sigma^2 + \boldsymbol{\mu}^\top\boldsymbol{\mu} - (\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}}{m\sigma^2} \\
&= \frac{d\sigma^2 + 1 - J_\sigma^2(\alpha) - K_\sigma^2(\alpha)}{m\sigma^2}.
\end{aligned}
$$

- **Calculate** $\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}'_i,\hat{Y}'_i|\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_1})$**:** Given any $(\xi_0,\boldsymbol{\mu}^\perp)$, in the following, we drop the condition on $\xi_0,\boldsymbol{\mu}^\perp$ for notational simplicity. Since $P_{\hat{Y}'_i}(1) = P_{\hat{Y}'_i}(-1) = P_{Y'_i}(1) = P_{Y'_i}(-1) = \frac{1}{2}$ (cf. (32)), we have

$$
\begin{aligned}
&\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}'_i,\hat{Y}'_i|\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_1}) = h(P_{\mathbf{Z}},p_{\boldsymbol{\theta}_1}) - h(P_{\mathbf{X}'_i,\hat{Y}'_i|\xi_0,\boldsymbol{\mu}^\perp},p_{\boldsymbol{\theta}_1}) \\
&= \frac{1}{2}\int (P_{\mathbf{X}|Y=1}(\mathbf{x}) - P_{\mathbf{X}'_i|\hat{Y}'_i=1}(\mathbf{x}))\log\frac{1}{P_Y(1)p_{\boldsymbol{\theta}_1}(\mathbf{x}|1)}\mathrm{d}\mathbf{x} \\
&\quad + \frac{1}{2}\int (P_{\mathbf{X}|Y=-1}(\mathbf{x}) - P_{\mathbf{X}'_i|\hat{Y}'_i=-1}(\mathbf{x}))\log\frac{1}{P_Y(-1)p_{\boldsymbol{\theta}_1}(\mathbf{x}|-1)}\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\int (P_{\mathbf{X}|Y=1}(\mathbf{x}) - P_{\mathbf{X}'_i|\hat{Y}'_i=1}(\mathbf{x}))\log\frac{1}{p_{\boldsymbol{\theta}_1}(\mathbf{x}|1)}\mathrm{d}\mathbf{x} \\
&\quad + \frac{1}{2}\int (P_{\mathbf{X}|Y=-1}(\mathbf{x}) - P_{\mathbf{X}'_i|\hat{Y}'_i=-1}(\mathbf{x}))\log\frac{1}{p_{\boldsymbol{\theta}_1}(\mathbf{x}|-1)}\mathrm{d}\mathbf{x}.
\end{aligned}
$$

Since given $\boldsymbol{\theta}_1$, $p_{\boldsymbol{\theta}_1}(\mathbf{x}|\cdot)$ is a Gaussian distribution, for any $y\in\{\pm 1\}$, we have

$$
\begin{aligned}
&\frac{1}{2}\int P_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\big(P_{\mathbf{X}|Y}(\mathbf{x}|\mathbf{y}) - P_{\mathbf{X}'_i|\hat{Y}'_i}(\mathbf{x}|y)\big)\log\frac{1}{p_{\boldsymbol{\theta}}(\mathbf{x}|y)}\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= \frac{1}{4\sigma^2}\int P_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\big(P_{\mathbf{X}|Y}(\mathbf{x}|y) - P_{\mathbf{X}'_i|\hat{Y}'_i}(\mathbf{x}|y)\big)\big(\mathbf{x}^\top\mathbf{x} - 2y\boldsymbol{\theta}^\top\mathbf{x} + \boldsymbol{\theta}^\top\boldsymbol{\theta}\big)\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}\int P_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\bigg(\frac{1}{2}P_{\mathbf{X}|Y}(\mathbf{x}|y) - \frac{1}{2}P_{\mathbf{X}'_i|\hat{Y}'_i}(\mathbf{x}|y)\bigg)\big(y\boldsymbol{\theta}^\top\mathbf{x}\big)\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top\big(\boldsymbol{\mu} - \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\big) \\
&= \frac{J_\sigma^2(\alpha) + K_\sigma^2(\alpha) - J_\sigma(\alpha)}{2\sigma^2}.
\end{aligned}
$$

Thus,

$$
\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}[\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}'_i,\hat{Y}'_i|\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_1})] = \frac{J_\sigma^2(\alpha) + K_\sigma^2(\alpha) - J_\sigma(\alpha)}{\sigma^2}.
$$

Finally, the gen-error at $t = 1$ can be characterized as follows:

$$\begin{aligned}\mathrm{gen}_1 &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\Bigg[\frac{J_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp)) + K_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp)) - J_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp))}{\sigma^2} \\ &\qquad\qquad + \frac{d\sigma^2 + 1 - J_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp)) - K_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp))}{m\sigma^2}\Bigg] \\ &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\Bigg[\frac{(m-1)(J_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp)) + K_\sigma^2(\alpha(\xi_0,\boldsymbol{\mu}^\perp))) - mJ_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp)) + d\sigma^2 + 1}{m\sigma^2}\Bigg].\end{aligned}$$

- **Pseudo-label using $\boldsymbol{\theta}_1$:** Let $\bar{\boldsymbol{\theta}}_1 := \boldsymbol{\theta}_1/\|\boldsymbol{\theta}_1\|_2$. For any $i \in \mathcal{I}_2$, the pseudo-labels are given by

  $$\hat{Y}_i' = \mathrm{sgn}(\boldsymbol{\theta}_1^\top \mathbf{X}_i') = \mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i').$$

  It can be seen that the pseudo-labels $\{\hat{Y}_i'\}_{i\in\mathcal{I}_2}$ are conditionally i.i.d. given $\boldsymbol{\theta}_1$ and let us denote the conditional distribution under fixed $\boldsymbol{\theta}_1$ as $P_{\hat{Y}'|\boldsymbol{\theta}_1} \in \mathcal{P}(\mathcal{Y})$. The pseudo-labelled dataset is denoted as $\hat{S}_{\mathrm{u},2} = \{(\mathbf{X}_i', \hat{Y}_i')\}_{i\in\mathcal{I}_2}$.

  For any fixed $(\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp)$, let $\boldsymbol{\theta}_1$ be decomposed as $\boldsymbol{\theta}_1 = A_1(\xi_0,\boldsymbol{\mu}^\perp)\boldsymbol{\mu} + B_1(\xi_0,\boldsymbol{\mu}^\perp)\boldsymbol{\upsilon}$, where $A_1^2(\xi_0,\boldsymbol{\mu}^\perp) + B_1^2(\xi_0,\boldsymbol{\mu}^\perp) = \|\boldsymbol{\theta}_1\|_2^2$. In addition, let $\alpha_1(\xi_0,\boldsymbol{\mu}^\perp) := A_1/\sqrt{A_1^2 + B_1^2}$ and $\beta_1(\xi_0,\boldsymbol{\mu}^\perp) = \sqrt{1 - (\alpha_1(\xi_0,\boldsymbol{\mu}^\perp))^2}$. In the following, we use A_1, B_1, α_1, and β_1 for the above quantities if there is no risk of confusion.

  Recall the decomposition of $\mathbf{X}_i'$ and $\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i'$ in (30) and (31). Similarly, we have

  $$\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i' =: Y_i'\alpha_1 + \sigma h_i^1,$$

  where $h_i^1 \sim \mathcal{N}(0,1)$. Note that $P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} = P_{\hat{Y}_i'|\boldsymbol{\theta}_1}$ and then the conditional probability $P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}$ can be given by

  $$\begin{aligned}P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(1) &= P_{\hat{Y}_i'|\boldsymbol{\theta}_1}(1) = \mathbb{P}_{\boldsymbol{\theta}_1}\big(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i' > 0\big) \\ &= \frac{1}{2}\mathbb{P}_{\boldsymbol{\theta}_1}\big(\alpha_1 + \sigma h_i^1 > 0\big) + \frac{1}{2}\mathbb{P}_{\boldsymbol{\theta}_1}\big(\alpha_1 + \sigma h_i^1 \le 0\big) = \frac{1}{2},\end{aligned} \tag{42}$$

  and $P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(-1) = 1/2$, where $\mathbb{P}_{\boldsymbol{\theta}_1}$ denotes the probability measure under parameter $\boldsymbol{\theta}_1$.

- **Iteration $t = 2$:** Recall (6) and the new model parameter learned from the pseudo-labelled dataset $\hat{S}_{\mathrm{u},2}$ is given by

  $$\boldsymbol{\theta}_2 = \frac{1}{m}\sum_{i\in\mathcal{I}_2} \hat{Y}_i'\mathbf{X}_i' = \frac{1}{m}\sum_{i\in\mathcal{I}_2} \mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i')\mathbf{X}_i', \tag{43}$$

  where $\{\mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i')\mathbf{X}_i'\}_{i\in\mathcal{I}_2}$ are conditionally i.i.d. random variables given $\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp$.

  Similar to (40), the expectation of $\boldsymbol{\theta}_2$ conditioned on $\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp$ is given by

  $$\begin{aligned}\boldsymbol{\mu}_2^{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} &:= \mathbb{E}[\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp] = \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_j')\mathbf{X}_j'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp] \\ &= \bigg(1 - 2\mathrm{Q}\Big(\frac{\alpha_1}{\sigma}\Big) + \frac{2\sigma\alpha_1}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha_1^2}{2\sigma^2}\Big)\bigg)\boldsymbol{\mu} + \frac{2\sigma\beta_1}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha_1^2}{2\sigma^2}\Big)\boldsymbol{\upsilon} \\ &= J_\sigma(\alpha_1(\xi_0,\boldsymbol{\mu}^\perp))\boldsymbol{\mu} + K_\sigma(\alpha_1(\xi_0,\boldsymbol{\mu}^\perp))\boldsymbol{\upsilon}.\end{aligned}$$

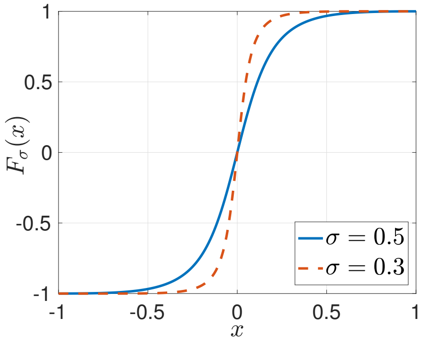


Figure 11: $F_\sigma(x)$ versus x for $\sigma \in \{0.3, 0.5\}$.

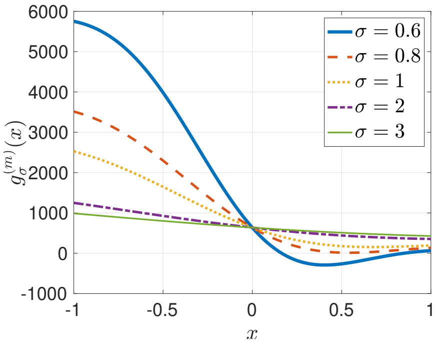


Figure 12: $g_\sigma^{(m)}(x)$ versus x for $\sigma \in \{0.6, 0.8, 1, 2, 3\}$.

As $m \to \infty$, by the strong law of large numbers, $\boldsymbol{\theta}_1|\xi_0, \boldsymbol{\mu}^\perp \to \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ almost surely. By the continuous mapping theorem, we also have $\alpha_1(\xi_0, \boldsymbol{\mu}^\perp) \to F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp))$ almost surely. Equivalently, for almost all sample paths, there exists a vanishing sequence ϵ_m (i.e., $\epsilon_m \to 0$ as $m \to \infty$) such that $|\alpha_1(\xi_0, \boldsymbol{\mu}^\perp) - F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp))| = \epsilon_m$.

The gen-error at $t = 2$ is given by

$$\begin{aligned}\mathrm{gen}_2 = \frac{1}{m}\sum_{i\in\mathcal{I}_2} \mathbb{E}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\Big[\mathbb{E}_{\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big[\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_2}) \\ + \Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\|P_{\mathbf{X}_i',\hat{Y}_i'|\boldsymbol{\theta}_2,\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_2})\big]\Big].\end{aligned}$$

By applying the same techniques in the iteration $t = 1$, we obtain the exact characterization of gen-error at $t = 2$ as follows. By the uniform continuity of J_σ , for any vanishing sequence $\epsilon_m > 0$, there exist $\epsilon_m', \epsilon_m'' > 0$ such that $\sup_{x\in[0,1]} |J_\sigma(x+\epsilon_m) - J_\sigma(x)| = \epsilon_m'$ and $\sup_{x\in[0,1]} |J_\sigma(x-\epsilon_m) - J_\sigma(x)| = \epsilon_m''$, where $\epsilon_m', \epsilon_m'' \downarrow 0$ as $\epsilon_m \downarrow 0$. The same result holds for K_σ.

Finally we can obtain the gen-error as follows. For almost all sample paths, there exists a vanishing sequence ϵ_m (i.e., $\epsilon_m \to 0$ as $m \to \infty$), such that

$$\begin{aligned}\mathrm{gen}_2 &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\Bigg[\frac{J_\sigma^2(F_\sigma(\alpha)) + K_\sigma^2(F_\sigma(\alpha)) - J_\sigma(F_\sigma(\alpha))}{\sigma^2} \\ &\qquad + \frac{d\sigma^2 + 1 - J_\sigma^2(F_\sigma(\alpha)) - K_\sigma^2(F_\sigma(\alpha))}{m\sigma^2}\Bigg] + \epsilon_m \\ &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\frac{(m-1)(J_\sigma^2(F_\sigma(\alpha)) + K_\sigma^2(F_\sigma(\alpha))) - mJ_\sigma(F_\sigma(\alpha))}{m\sigma^2}\right] + \epsilon_m',\end{aligned}$$

where $\epsilon_m' = \epsilon_m + \frac{d\sigma^2+1}{m\sigma^2}$ and α stands for $\alpha(\xi_0, \boldsymbol{\mu}^\perp)$.

- **Iteration $t \in [2:\tau]$:** By iteratively implementing the calculation, we finally obtain the characterization of gen_t as follows. For almost all sample paths, there exists a vanishing sequence $\epsilon_{m,t}$ ($\epsilon_{m,t} \to 0$ as $m \to \infty$) , such that

$$\mathrm{gen}_t = \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\frac{(m-1)(J_\sigma^2(F_\sigma^{(t-1)}(\alpha)) + K_\sigma^2(F_\sigma^{(t-1)}(\alpha))) - mJ_\sigma(F_\sigma^{(t-1)}(\alpha))}{m\sigma^2}\right] + \epsilon'_{m,t},$$

where $\epsilon'_{m,t} = \epsilon_{m,t} + \frac{d\sigma^2+1}{m\sigma^2}$ and α stands for $\alpha(\xi_0, \boldsymbol{\mu}^\perp)$.

The proof is thus completed.

Remark 6 (Numerical plots of $F_\sigma^{(t)}(\cdot)$ and $g_\sigma^{(m)}(\cdot)$)**.** *Recall* $g_\sigma^{(m)}(x) = ((m-1)(J_\sigma^2(x) + K_\sigma^2(x)) - mJ_\sigma(x))/\sigma^2$ *for any* $x \in [-1, 1]$*, which is the quantity that determines the behaviour of* (13)*. To gain more insight, we numerically plot* $F_\sigma^{(t)}(x)$ *for* $t = 0, 1, 2$ *in Figure 2 and* $F_\sigma(x)$ *under different* σ *in Figure 11. We also plot* $g_\sigma^{(m)}(x)$ *under different* σ *in Figure 12.*

D. Applying Theorem 2.A to bGMMs

In anticipation of leveraging Theorem 2.A 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 and Zou (2016); Xu and Raginsky (2017); Bu et al. (2020)), we find it convenient to show that $\mathbf{X}$ and $l(\boldsymbol{\theta}, (\mathbf{X}, Y))$ are bounded w.h.p.. By defining the ℓ_∞ ball $\mathcal{B}_r^y := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - y\boldsymbol{\mu}\|_\infty \le r\}$, we see that

$$\Pr(\mathbf{X} \in \mathcal{B}_r^Y) = \left(1 - 2\Phi\left(-\frac{r}{\sigma}\right)\right)^d =: 1 - \delta_{r,d},$$

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

To show that $\boldsymbol{\theta}$ is bounded with high probability, define the set $\Theta_{\boldsymbol{\mu},c} := \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_\infty \le c\}$ for some $c > 0$. For any $\boldsymbol{\theta} \in \Theta_{\boldsymbol{\mu},c}$, we have

$$\min_{(\mathbf{x},y)\in\mathcal{Z}} l(\boldsymbol{\theta}, (\mathbf{x}, y)) = \log(2\sqrt{(2\pi)^d}\sigma^d) =: c_1, \quad \text{and}$$

$$\max_{\mathbf{x}\in\mathcal{B}_r^y, y\in\mathcal{Y}} l(\boldsymbol{\theta}, (\mathbf{x}, y)) \le \log(2\sqrt{(2\pi)^d}\sigma^d) + \frac{d(c+r)^2}{2\sigma^2} =: c_2.$$

For any $(\mathbf{X}, Y)$ from the bGMM$(\boldsymbol{\mu}, \sigma)$ and any $\boldsymbol{\theta} \in \Theta_{\boldsymbol{\mu},c}$, the probability that $l(\boldsymbol{\theta}, (\mathbf{X}, Y))$ belongs to the interval $[c_1, c_2]$ (c_1, c_2 depend on $\delta_{r,d}$) can be lower bounded by

$$\Pr\big(l(\boldsymbol{\theta}, (\mathbf{X}, Y)) \in [c_1, c_2]\big) \ge 1 - \delta_{r,d}.$$

Thus, according to Hoeffding's lemma, with probability at least $1 - \delta_{r,d}$, $l(\boldsymbol{\theta}, (\mathbf{X}, Y)) \sim$ subG$((c_2 - c_1)/2)$ under $(\mathbf{X}, Y) \sim \mathcal{N}(Y\boldsymbol{\mu}, \sigma^2\mathbf{I}_d) \otimes P_Y$ for all $\boldsymbol{\theta} \in \Theta_{\boldsymbol{\mu},c}$, i.e., for all $\lambda \in \mathbb{R}$,

$$\begin{aligned}&\mathbb{E}_{\mathbf{X},Y}\big[\exp\big(\lambda\big(l(\boldsymbol{\theta}, (\mathbf{X}, Y)) - \mathbb{E}_{\mathbf{X},Y}[l(\boldsymbol{\theta}, (\mathbf{X}, Y))]\big)\big)\big]\\ &\le \exp\big(\lambda^2(c_2 - c_1)^2/8\big).\end{aligned} \tag{44}$$

Recall the definition of α in (7) and the decomposition of $\bar{\boldsymbol{\theta}}_0$ in (8). 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 follows:

$$G_\sigma(\alpha, \xi_0, \boldsymbol{\mu}^\perp) := D\big(p'_{\tilde{g},\tilde{\mathbf{g}}^\perp}\big\|p_{\tilde{g}} \otimes p_{\tilde{\mathbf{g}}^\perp}\big), \tag{45}$$

where

$$p'_{\tilde{g},\tilde{\mathbf{g}}^{\perp}} = \Phi\Big(-\frac{\alpha}{\sigma}\Big) p_{\tilde{g}+\frac{2\alpha}{\sigma}|\tilde{g}\le -\frac{\alpha}{\sigma}} \otimes p_{\tilde{\mathbf{g}}^{\perp}+\bar{\boldsymbol{\theta}}_0^{\perp}} + \Phi\Big(\frac{\alpha}{\sigma}\Big) p_{\tilde{g}|\tilde{g}\le \frac{\alpha}{\sigma}} \otimes p_{\tilde{\mathbf{g}}^{\perp}},$$

$\tilde{g} \sim \mathcal{N}(0,1)$, $\tilde{\mathbf{g}}^{\perp} \sim \mathcal{N}(0, \mathbf{I}_d - \bar{\boldsymbol{\theta}}_0 \bar{\boldsymbol{\theta}}_0^{\top})$, $\tilde{\mathbf{g}}^{\perp}$ is independent of $\tilde{g}$ and perpendicular to $\bar{\boldsymbol{\theta}}_0$. Note that $p_{\tilde{g}+\frac{2\alpha}{\sigma}|\tilde{g}\le -\frac{\alpha}{\sigma}}$ is the Gaussian probability density function with mean $\frac{2\alpha}{\sigma}$ and variance 1 *truncated* to the interval $(-\infty, -\frac{\alpha}{\sigma})$, and similarly for $p_{\tilde{g}|\tilde{g}\le\frac{\alpha}{\sigma}}$. In general, when $G_\sigma(\alpha, \xi_0, \boldsymbol{\mu}^{\perp})$ is small, so is the generalization error.

By applying the result in Theorem 2.A, the following theorem provides an upper bound for the generalization error at each iteration t for m large enough.

Theorem 7. *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*

$$\big|\mathrm{gen}_0(P_{\mathbf{Z}}, P_{\mathbf{X}}, P_{\boldsymbol{\theta}_0|S_{\mathrm{l}},S_{\mathrm{u}}})\big| \le \sqrt{\frac{(c_2-c_1)^2 d}{4}\log\frac{n}{n-1}}. \tag{46}$$

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

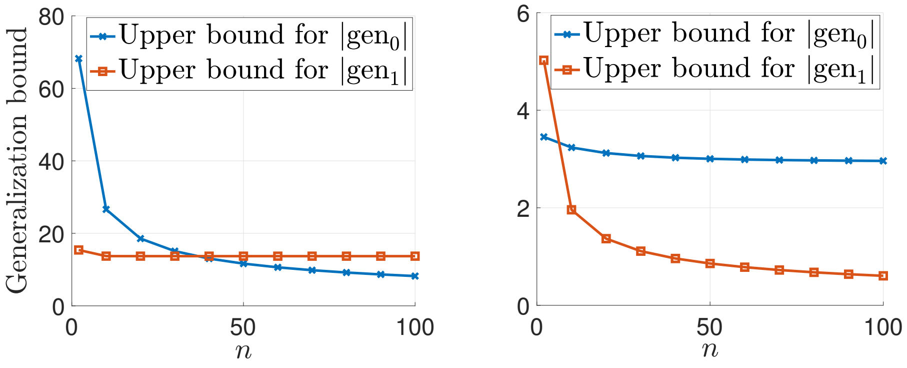
$$\begin{aligned}&\big|\mathrm{gen}_t(P_{\mathbf{Z}}, P_{\mathbf{X}}, \{P_{\boldsymbol{\theta}_k|S_{\mathrm{l}},S_{\mathrm{u}}}\}_{k=0}^{t}, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})\big| \\ &\le \frac{c_2-c_1}{\sqrt{2}} \mathbb{E}_{\xi_0,\boldsymbol{\mu}^{\perp}}\left[\sqrt{G_\sigma\big(F_\sigma^{(t-1)}(\alpha(\xi_0,\boldsymbol{\mu}^{\perp})), \xi_0, \boldsymbol{\mu}^{\perp}\big) + \epsilon}\right].\end{aligned} \tag{47}$$

The proof of Theorem 7 is provided in Appendix E. Several remarks about Theorem 7 are as follows.

First, we compare the upper bounds for $|\mathrm{gen}_0|$ and $|\mathrm{gen}_1|$, as shown in Figures 13(a) and 13(b). For any fixed σ, when n is sufficiently small, the upper bound for $|\mathrm{gen}_0|$ is greater than that for $|\mathrm{gen}_1|$. As n increases, the upper bound for $|\mathrm{gen}_1|$ surpasses that of $|\mathrm{gen}_0|$, as shown in Figure 13(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 $\boldsymbol{\theta}_t$ well and additional pseudo-labelled data does not help to further boost the generalization performance. Furthermore, by comparing Figures 13(a) and 13(b), we can see that for smaller σ, the improvement from $|\mathrm{gen}_0|$ to $|\mathrm{gen}_1|$ is more pronounced. The intuition is that when σ 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. Recall the Taylor expansion of α in (14). It can be seen that as n increases, α converges to 1 in probability. Suppose the dimension $d = 2$ and $\boldsymbol{\mu} = (1, 0)$. Then $\boldsymbol{\mu}^{\perp} = [0, \mu_2^{\perp}]$ where $\mu_2^{\perp} \sim \mathcal{N}(0,1)$. The upper bound for the absolute generalization error at $t = 1$ can be rewritten as

$$|\mathrm{gen}_1| \lessapprox \frac{c_2-c_1}{\sqrt{2}} \int_{-\sqrt{2}}^{\sqrt{2}} \sqrt{\frac{n}{\pi\sigma^2}}\, e^{-\frac{ny^2}{\sigma^2}} \sqrt{G_\sigma(1-y^2)}\, \mathrm{d}y,$$

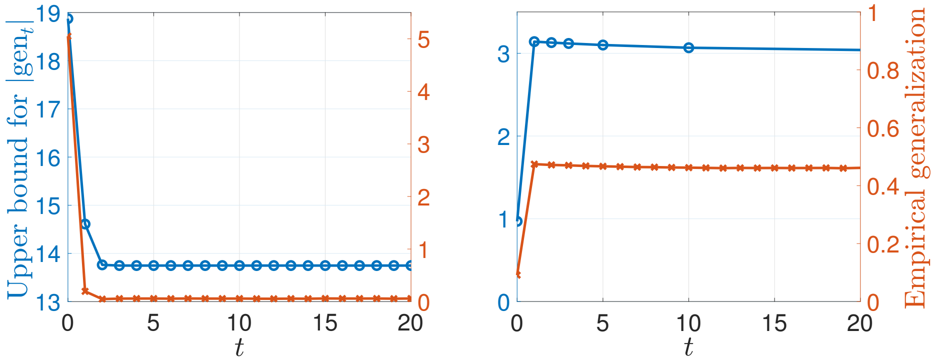
which is a decreasing function of n, as shown in Figures 13(a) and 13(b).



(a) $\sigma = 0.6$

(b) $\sigma = 3$

Figure 13: Upper bounds for generalization error at $t = 0$ and $t = 1$ under different σ when $d = 2$ and $\boldsymbol{\mu} = (1, 0)$.



(a) $\sigma = 0.6$

(b) $\sigma = 3$

Figure 14: The comparison between the upper bound for $|\text{gen}_t|$ and the empirical generalization error at each iteration t. The upper bounds are both for $d = 2$.

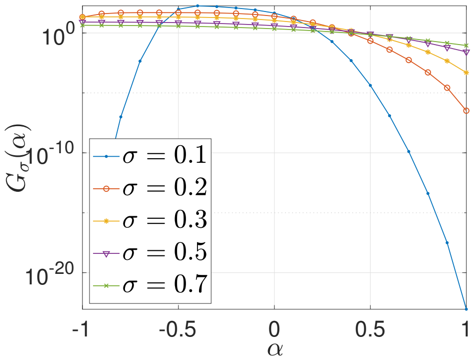


Figure 15: $G_\sigma(\alpha)$ versus α for different σ.

Second, in Figure 14(a), we plot the theoretical upper bound in (47) by ignoring ϵ. Unfortunately it is computationally difficult to numerically calculate the bound in (47) 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 $|\mathrm{gen}_t|$ decreases as t increases and finally converges to a non-zero constant. The gap between the upper bounds for $|\mathrm{gen}_t|$ and for $|\mathrm{gen}_{t+1}|$ 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 $n = 10$, $d = 50$, $\boldsymbol{\mu} = (1, 0, \ldots, 0)$ 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 (47) serves as a useful rule-of-thumb for how the generalization error changes over iterations. In Figure 14(b), we plot the theoretical bound and result from the empirical simulation based on the toy example for $d = 2$ but larger n and σ. This figure shows that when we increase n and σ, 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 σ is larger, the data samples have larger variance and the classes have a larger overlap, and thus, the initial parameter $\boldsymbol{\theta}_0$ learned by the labelled data cannot produce pseudo-labels with sufficiently high accuracy. Thus, the pseudo-labelled data cannot help to improve the generalization error significantly.

Remark 8 (Numerical plot of $G_\sigma(\cdot)$)**.** *To gain more insight, we plot $G_\sigma(\alpha, \xi_0, \boldsymbol{\mu}^\perp)$ when $d = 2$ and $\boldsymbol{\mu} = (1, 0)$ in Figure 15. Under these settings, $G_\sigma(\alpha, \xi_0, \boldsymbol{\mu}^\perp)$ depends only on α and hence, we can rewrite it as $G_\sigma(\alpha)$. As shown in Figure 15 in Appendix D, for all $\sigma_1 > \sigma_2$, there exists an $\alpha_0 \in [-1, 1]$ such that for all $\alpha \geq \alpha_0 = \alpha_0(\sigma_1, \sigma_2)$, $G_{\sigma_1}(\alpha) > G_{\sigma_2}(\alpha)$. From (7), we can see that α is close to 1 of high probability, which means that $\sigma \mapsto G_\sigma(\alpha)$ is monotonically increasing in σ with high probability. As a result, $\mathbb{E}_\alpha[\sqrt{G_\sigma(\alpha)}]$ increases as σ increases. This is consistent with the intuition that when the training data has larger in-class variance, it is more difficult to generalize well.*

Remark 9 (Discussion on Theorem 7)**.** *As $n \to \infty$, $\boldsymbol{\theta}_0 \to \boldsymbol{\mu}$ and $\alpha = \rho(\boldsymbol{\theta}_0, \boldsymbol{\mu}) \to 1$ 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 $\Pr(\hat{Y}'_j \neq Y'_j) \to \mathrm{Q}(1/\sigma) > 0$ (which is the Bayes error rate) and the disintegrated KL-divergence $D_{\xi_0, \boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j, \hat{Y}'_j} \| P_{\mathbf{X}, Y})$ between the estimated and underlying distributions cannot converge to 0.*

In the other extreme case, when $\alpha = \rho(\boldsymbol{\theta}_0, \boldsymbol{\mu}) = -1$ and $\bar{\boldsymbol{\theta}}_0 = -\boldsymbol{\mu}$, the error probability $\Pr(\hat{Y}'_j \neq Y'_j) = 1 - \mathrm{Q}(1/\sigma) > \frac{1}{2}$ (for all $\sigma > 0$) and $D_{\xi_0, \boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j, \hat{Y}'_j} \| P_{\mathbf{X}, Y}) < \infty$, so in this other extreme (flipped) scenario, we have more mistakes than correct pseudo-labels. The reason why $D_{\alpha, \boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j, \hat{Y}'_j} \| P_{\mathbf{X}, Y})$ is finite is that when $P_{\mathbf{X}, Y}(\mathbf{x}, y)$ is small, it means that $\mathbf{x}$ is far from both $-\boldsymbol{\mu}$ and $\boldsymbol{\mu}$, and then $P_{\mathbf{X}}(\mathbf{x})$ is also small. Thus, $P_{\mathbf{X}'_j, \hat{Y}'_j}(\mathbf{x}, y) = P_{\hat{Y}'_j | \mathbf{X}'_j}(y|\mathbf{x}) P_{\mathbf{X}}(\mathbf{x})$ is also small.

E. Proof of Theorem 7

For simplicity, in the following, we abbreviate $\mathrm{gen}_t(P_{\mathbf{Z}}, P_{\mathbf{X}}, \{P_{\boldsymbol{\theta}_k|S_1,S_{\mathrm{u}}}\}_{k=0}^{t}, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})$ as gen_t.

1. **Initial round $t = 0$:** Since $Y_i\mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_d)$, we have $\boldsymbol{\theta}_0 \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}_d)$ and for some constant $c \in \mathbb{R}_+$,

$$\Pr(\boldsymbol{\theta}_0 \in \Theta_{\boldsymbol{\mu},c}) = \Pr(\|\boldsymbol{\theta}_0 - \boldsymbol{\mu}\|_\infty \le c) = \bigg(1 - 2\Phi\Big(-\frac{\sqrt{n}c}{\sigma}\Big)\bigg)^d =: 1 - \delta_{\sqrt{n}c,d}. \quad (48)$$

By choosing c large enough, $\delta_{\sqrt{n}c,d}$ can be made arbitrarily small. Consider $\tilde{\boldsymbol{\theta}}_0$ and $(\tilde{\mathbf{X}}, \tilde{Y})$ as independent copies of $\boldsymbol{\theta}_0 \sim Q_{\boldsymbol{\theta}_0}$ and $(\mathbf{X}, Y) \sim P_{\mathbf{X},Y} = P_Y \otimes \mathcal{N}(Y\boldsymbol{\mu}, \sigma^2 I_d)$, respectively, such that $P_{\tilde{\theta}_0,\tilde{\mathbf{X}},\tilde{Y}} = Q_{\boldsymbol{\theta}_0} \otimes P_{\mathbf{X},Y}$. Then the probability that $l(\boldsymbol{\theta}_0, (\mathbf{X}, Y)) \sim \mathrm{subG}((c_2 - c_1)/2)$ under $(\mathbf{X}, Y) \sim P_{\mathbf{X},Y}$ is given as follows

$$\begin{aligned}
&\Pr\bigg(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_0) \le \frac{\lambda^2(c_2 - c_1)^2}{8}\bigg) \\
&\ge \Pr\bigg(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_0) \le \frac{\lambda^2(c_2 - c_1)^2}{8} \text{ and } \tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}\bigg) \\
&= \Pr(\tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}) \Pr\bigg(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_0) \le \frac{\lambda^2(c_2 - c_1)^2}{8}\bigg|\tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}\bigg) \\
&= (1 - \delta_{\sqrt{n}c,d})(1 - \delta_{r,d}), \qquad (49)
\end{aligned}$$

where (49) follows from (44) and (48).

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 $n > n_0, r > r_0$, $\delta_{\sqrt{n}c,d} < \frac{\delta}{3}$, $\delta_{r,d} < \frac{\delta}{3}$, and then with probability at least $1 - \delta$, the absolute generalization error can be upper bounded as follows

$$|\mathrm{gen}_0| \le \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{(c_2 - c_1)^2}{2}I(\boldsymbol{\theta}_0; \mathbf{X}_i, Y_i)}.$$

Then mutual information can be calculated as follows

$$\begin{aligned}
I(\boldsymbol{\theta}_0; \mathbf{X}_i, Y_i) &= h(\boldsymbol{\theta}_0) - h(\boldsymbol{\theta}_0|\mathbf{X}_i, Y_i) \\
&= h\bigg(\frac{1}{n}\sum_{j=1}^{n} Y_j\mathbf{X}_j\bigg) - h\bigg(\frac{1}{n}\sum_{j=1}^{n} Y_j\mathbf{X}_j\bigg|\mathbf{X}_i, Y_i\bigg) \\
&= \frac{d}{2}\log\bigg(\frac{2\pi e\sigma^2}{n}\bigg) - h\bigg(\frac{1}{n}\sum_{j\in[n],j\ne i} Y_j\mathbf{X}_j\bigg) \\
&= \frac{d}{2}\log\bigg(\frac{2\pi e\sigma^2}{n}\bigg) - \frac{d}{2}\log\bigg(\frac{2\pi e(n-1)\sigma^2}{n^2}\bigg) \\
&= \frac{d}{2}\log\frac{n}{n-1}.
\end{aligned}$$

Thus we obtain (46).

2. **Pseudo-label using $\boldsymbol{\theta}_0$:** The same as those in Appendix C.

3. **Iteration $t = 1$:** Recall $\boldsymbol{\theta}_1$ in (33), the new model parameter learned from the pseudo-labelled dataset $\hat{S}_{\mathrm{u},1}$.

(a) Recall the condition expectation of $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ in (40). The l_∞ norm between $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ and $\boldsymbol{\mu}$ can be upper bounded by

$$\begin{aligned}\|\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} - \boldsymbol{\mu}\|_\infty &\le \sqrt{\bigg(- 2\mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big) + \frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\Big(-\frac{\alpha^2}{2\sigma^2}\Big)\bigg)^2 + \frac{2\sigma^2\beta^2}{\pi}\exp\Big(-\frac{2\alpha^2}{2\sigma^2}\Big)} \\ &< \sqrt{\bigg(2\Phi\Big(\frac{1}{\sigma}\Big) + \frac{2\sigma}{\sqrt{2\pi}}\bigg)^2 + \frac{2\sigma^2}{\pi}} =: \tilde{c}_1, \qquad (50)\end{aligned}$$

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

(b) Next, we need to calculate the probability that $l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) \sim \mathrm{subG}((c_2 - c_1/2))$ under $(\mathbf{X}, Y) \sim P_{\mathbf{X},Y}$.

Let $\mathbf{V}_i = \mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')\mathbf{X}_i' - \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$. For any $k \in [d]$, let $V_{i,k}$, $\theta_{1,k}$, $\mu_{1,k}$ denote the k-th components of $\mathbf{V}_i$, $\boldsymbol{\theta}_1$ and $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$, respectively. Recall the decompositions $\mathbf{X}_i' = Y_i'\boldsymbol{\mu} + \sigma\tilde{g}_i\bar{\boldsymbol{\theta}}_0 + \sigma\tilde{\mathbf{g}}_i^\perp$ in (35) and $\bar{\boldsymbol{\theta}}_0\mathbf{X}_i' = Y_i'\alpha + \sigma\tilde{g}_i$ in (36). Suppose the basis of $\mathbb{R}^d$ is denoted by $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_d\}$ and let $\mathbf{v}_1 = \bar{\boldsymbol{\theta}}_0$. Then we have

$$\tilde{\mathbf{g}}_i^\perp = \tilde{g}_{i,2}^\perp\mathbf{v}_2 + \ldots + \tilde{g}_{i,d}^\perp\mathbf{v}_d,$$

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

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

$$\begin{aligned}&\mathbb{E}_{V_{i,1}}[e^{s_1V_{i,1}}] \\ &= \mathrm{Q}\Big(-\frac{\alpha}{\sigma}\Big)\mathbb{E}_{\tilde{g}_i}\Big[e^{s_1(\mu_{0,1}-\mu_{1,1}+\sigma\tilde{g}_i)}\Big|\tilde{g}_i > -\frac{\alpha}{\sigma}\Big] + \mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big)\mathbb{E}_{\tilde{g}_i}\Big[e^{s_1(-\mu_{0,1}-\mu_{1,1}+\sigma\tilde{g}_i)}\Big|\tilde{g}_i > \frac{\alpha}{\sigma}\Big] \\ &= e^{s_1(\mu_{0,1}-\mu_{1,1})}e^{\frac{\sigma^2s_1^2}{2}}\Phi\Big(\frac{\alpha}{\sigma} + \sigma s_1\Big) + e^{s_1(-\mu_{0,1}-\mu_{1,1})}e^{\frac{\sigma^2s_1^2}{2}}\Phi\Big(-\frac{\alpha}{\sigma} + \sigma s_1\Big).\end{aligned}$$

The final equality follows from the fact that the MGF of a zero-mean univariate Gaussian truncated to (a, b) is $e^{\sigma^2s^2/2}\Big[\frac{\Phi(b-\sigma s)-\Phi(a-\sigma s)}{\Phi(b)-\Phi(a)}\Big]$. The second derivative of $\log\mathbb{E}_{V_{i,1}}[e^{s_1V_{i,1}}]$ is given as

$$\begin{aligned}\tilde{R}_1(s_1) &:= \frac{\mathrm{d}^2\log\mathbb{E}_{V_{i,1}}[e^{s_1V_{i,1}}]}{\mathrm{d}s_1^2} \\ &\le \sigma^2 + \frac{\mathrm{const.}}{\big(\Phi(\frac{\alpha}{\sigma} + \sigma s_1)e^{s_k\mu_{0,k}} + \Phi(\frac{-\alpha}{\sigma} + \sigma s_1)e^{-s_k\mu_{0,k}}\big)^2} < \infty.\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}] &= \mathbb{E}_{\sigma \tilde{g}_{i,k}^{\perp}, Y_i'}\Big[e^{s_k(Y_i'\mu_{0,k} - \mu_{1,k} + \sigma \tilde{g}_{i,k}^{\perp})}\Big] \\ &= Q\Big(-\frac{\alpha}{\sigma}\Big) e^{s_k(\mu_{0,k} - \mu_{1,k})} e^{\frac{\sigma^2 s_k^2}{2}} + Q\Big(\frac{\alpha}{\sigma}\Big) e^{s_k(-\mu_{0,k} - \mu_{1,k})} e^{\frac{\sigma^2 s_k^2}{2}},\end{aligned}$$

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

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

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

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

for some $\xi_{\mathrm{L},k} \in (0, s_k)$ and $\tilde{R}_k(\xi_{\mathrm{L},k}) < \infty$. 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 $\varepsilon > 0$,

$$\sup_{s_k > 0} \Big(s_k \varepsilon - \log \mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}]\Big) \geq \sup_{s_k > 0} \Big(s_k \varepsilon - \frac{\tilde{R}_k(\xi_{\mathrm{L},k}) s_k^2}{2}\Big) = \frac{\varepsilon^2}{2\tilde{R}_k(\xi_{\mathrm{L},k})}.$$

Let $\tilde{R}^* = \max_{\xi_0, \boldsymbol{\mu}^{\perp}} \min_{k \in [1:d]} \tilde{R}_k(\xi_{\mathrm{L},k})$, which is a finite constant only dependent on σ. Since $\{\mathrm{sgn}(\boldsymbol{\theta}_0^{\top} \mathbf{X}_i') \mathbf{X}_i'\}_{i=1}^m$ are i.i.d. random variables conditioned on $(\xi_0, \boldsymbol{\mu}^{\perp})$, by applying Chernoff-Cramér inequality, we have for all $\varepsilon > 0$

$$\begin{aligned}&\mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}\Big(\|\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1^{\xi_0, \boldsymbol{\mu}^{\perp}}\|_{\infty} > \varepsilon\Big) \\ &= \mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}\Big(\max_{k \in [1:d]} |\theta_{1,k} - \mu_{1,k}| > \varepsilon\Big) \\ &\leq \sum_{k=1}^{d} \mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}\Big(|\theta_{1,k} - \mu_{1,k}| > \varepsilon\Big) \\ &= \sum_{k=1}^{d} \mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}\Big(\Big|\frac{1}{m}\sum_{i=1}^{m} V_{i,k}\Big| > \varepsilon\Big) \\ &\leq \sum_{k=1}^{d} 2\exp\Big(-m \sup_{s > 0}\Big(s\varepsilon - \log \mathbb{E}_{V_{i,k}}[e^{sV_{i,k}}]\Big)\Big) \\ &\leq 2d \exp\Big(-\frac{m\varepsilon^2}{2\tilde{R}^*}\Big) \\ &=: \delta_{m,\varepsilon,d},\end{aligned} \tag{51}$$

where $\delta_{m,\varepsilon,d} \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$ and does not depend on $\xi_0, \boldsymbol{\mu}^{\perp}$.

Choose some $c \in (\tilde{c}_1, \infty)$ ($\tilde{c}_1$ defined in (50)). We have

$$\mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}(\boldsymbol{\theta}_1 \in \Theta_{\boldsymbol{\mu}, c}) \geq \mathbb{P}_{\xi_0, \boldsymbol{\mu}^{\perp}}(\|\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1^{\xi_0, \boldsymbol{\mu}^{\perp}}\|_{\infty} \leq c - \tilde{c}_1) \geq 1 - \delta_{m, c - \tilde{c}_1, d}.$$

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

$$\begin{aligned}
&\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_1,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_1) \le \frac{\lambda^2(c_2 - c_1)^2}{8}\right) \\
&\ge \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}(\tilde{\boldsymbol{\theta}}_1 \in \Theta_{\boldsymbol{\mu},c})\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_1,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_1) \le \frac{\lambda^2(c_2 - c_1)^2}{8}\middle|\tilde{\boldsymbol{\theta}}_1 \in \Theta_{\boldsymbol{\mu},c}\right) \\
&= (1 - \delta_{m,c,d})(1 - \delta_{r,d}).
\end{aligned}$$

Thus, for some $c \in (\tilde{c}_1, \infty)$, 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:

$$\begin{aligned}
|\mathrm{gen}_1| &= |\mathbb{E}[L_{P_{\mathbf{Z}}}(\boldsymbol{\theta}_1) - L_{\hat{S}_{\mathrm{u},1}}(\boldsymbol{\theta}_1)]| \\
&= \left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1, (\mathbf{X}'_i, \hat{Y}'_i))|\xi_0, \boldsymbol{\mu}^\perp\right]\right]\right| \\
&\le \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\frac{(c_2 - c_1)^2}{2}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1, (\mathbf{X}'_i, \hat{Y}'_i)) + D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_i,\hat{Y}'_i}\|P_{\mathbf{X},Y})\Big)}\,\right], \quad (52)
\end{aligned}$$

where $P_{\boldsymbol{\theta}_1,(\mathbf{X},Y)|\xi_0,\boldsymbol{\mu}^\perp} = Q_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp} \otimes P_{\mathbf{X},Y}$ and $Q_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}$ denotes the marginal distribution of $\boldsymbol{\theta}_1$ under parameters $(\xi_0, \boldsymbol{\mu}^\perp)$.

In the following, we derive closed-form expressions of the mutual information and KL-divergence in (52). For any $j \in [1 : m]$:

- **Calculate $I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_j, \hat{Y}'_j)$:** For arbitrary random variables X and U, we define the *disintegrated conditional differential entropy* of X given U as

$$h_U(X) := h(P_{X|U}).$$

  This is a $\sigma(U)$-measurable random variable. Conditioned on a certain pair $(\xi_0, \boldsymbol{\mu}^\perp)$, the mutual information between $\boldsymbol{\theta}_1$ and $(\mathbf{X}'_i, \hat{Y}'_i)$ is

$$\begin{aligned}
&I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_i, \hat{Y}'_i) \\
&= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) - h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{j=1}^m \hat{Y}'_j\mathbf{X}'_j\middle|\mathbf{X}'_i, \hat{Y}'_i\right) \\
&= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) - h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{j\in[m],j\neq i} \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_j)\mathbf{X}'_j\right) \\
&= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) \\
&\quad - h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m-1}\sum_{j\in[m],j\neq i} \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_j)\mathbf{X}'_j\right) - d\log\frac{m-1}{m}.
\end{aligned}$$

  As $m \to \infty$, $I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_i, \hat{Y}'_i) \to 0$ 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 $m > m_0$,

$$\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_i, \hat{Y}'_i) > \epsilon) \le \delta. \quad (53)$$

- **Calculate $D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j,\hat{Y}'_j}\|P_{\mathbf{X},Y})$:** First of all, since $P_{\hat{Y}'_j}=P_Y$ (cf. (32)) regardless of the values of $(\xi_0,\boldsymbol{\mu}^\perp)$, the *disintegrated conditional KL-divergence* can be rewritten as

$$\begin{aligned}&D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j,\hat{Y}'_j}\|P_{\mathbf{X},Y})\\&=P_{\hat{Y}'_j}(-1)D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=-1}\|P_{\mathbf{X}|Y=-1})+P_{\hat{Y}'_j}(1)D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=1}\|P_{\mathbf{X}|Y=1}).\end{aligned}\tag{54}$$

Recall the decomposition of a Gaussian vector $\tilde{\mathbf{g}}_j\sim\mathcal{N}(0,\mathbf{I}_d)$ in (34). Note that $\operatorname{rank}(\mathsf{Cov}(\tilde{\mathbf{g}}_j^\perp))=\operatorname{rank}(\mathbf{I}_d-\bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}^\top)=d-1$.
For any pair of labelled data sample $(\mathbf{X},Y)$, from (35), we similarly decompose $\mathbf{X}$ as $\mathbf{X}=Y\boldsymbol{\mu}+\sigma(\tilde{g}\bar{\boldsymbol{\theta}}_0+\tilde{\mathbf{g}}^\perp)$, where $\tilde{g}\sim\mathcal{N}(0,1)$ and $\tilde{\mathbf{g}}^\perp\sim\mathcal{N}(0,\mathbf{I}_d-\bar{\boldsymbol{\theta}}_0\bar{\boldsymbol{\theta}}_0^\top)$. Let $p_{\tilde{g}}$ and $p_{\tilde{\mathbf{g}}^\perp}$ denote the probability density functions of $\tilde{g}$ and $\tilde{\mathbf{g}}^\perp$, respectively. For any $\mathbf{x}=\boldsymbol{\mu}+\sigma(u\bar{\boldsymbol{\theta}}_0+\mathbf{u}^\perp)\in\mathbb{R}^d$, the joint probability distribution at $(\mathbf{X},Y)=(\mathbf{x},1)$ is given by

$$\begin{aligned}P_{\mathbf{X},Y}(\mathbf{x},1)&=P_Y(1)p_{\boldsymbol{\mu}}(\mathbf{x}|1)\\&=\frac{P_Y(1)}{\sqrt{(2\pi)^d}\sigma^d}\exp\bigg(-\frac{1}{2\sigma^2}(\mathbf{x}-y\boldsymbol{\mu})^\top(\mathbf{x}-y\boldsymbol{\mu})\bigg)\\&=\frac{P_Y(1)}{\sqrt{(2\pi)^d}\sigma^d}\exp\bigg(-\frac{1}{2\sigma^2}(\sigma u\bar{\boldsymbol{\theta}}_0+\sigma\mathbf{u}^\perp)^\top(\sigma u\bar{\boldsymbol{\theta}}_0+\sigma\mathbf{u}^\perp)\bigg)\\&=\frac{P_Y(y)}{\sqrt{(2\pi)^d}\sigma^d}\exp\bigg(-\frac{u^2}{2}\bigg)\exp\bigg(-\frac{(\mathbf{u}^\perp)^\top\mathbf{u}^\perp}{2}\bigg)\\&=P_Y(1)p_{\tilde{g}}(u)p_{\tilde{\mathbf{g}}^\perp}(\mathbf{u}^\perp).\end{aligned}$$

Similarly, for any $\mathbf{x}=-\boldsymbol{\mu}+\sigma(u\bar{\boldsymbol{\theta}}_0+\mathbf{u}^\perp)\in\mathbb{R}^d$, the joint probability density evaluated at $(X,Y)=(\mathbf{x},-1)$ is given by

$$P_{\mathbf{X},Y}(\mathbf{x},-1)=P_Y(-1)p_{\boldsymbol{\mu}}(\mathbf{x}|-1)=P_Y(-1)p_{\tilde{g}}(u)p_{\tilde{\mathbf{g}}^\perp}(\mathbf{u}^\perp).$$

Second, we have $P_{\mathbf{X}'_j|\hat{Y}'_j}=\sum_{y\in\{-1,+1\}}P_{\mathbf{X}'_j|\hat{Y}'_j,Y'_j=y}P_{Y'_j=y|\hat{Y}'_j}$. The conditional probability distribution $P_{Y'_j|\hat{Y}'_j}$ can be calculated as follows

$$P_{Y'_j|\hat{Y}'_j}=\frac{P_{\hat{Y}'_j|Y'_j}P_{Y'_j}}{P_{\hat{Y}'_j}}=P_{\hat{Y}'_j|Y'_j},$$

where the last equality follows since $P_{Y'_j}(-1)=P_{Y'_j}(1)=P_{\hat{Y}'_j}(-1)=P_{\hat{Y}'_j}(1)=1/2$. Since $\hat{Y}'_j=\operatorname{sgn}(Y'_j\alpha+\sigma\tilde{g}_j)$ (cf. (36)), we have

$$P_{\hat{Y}'_j|Y'_j}(-1|-1)=\Pr(Y'_j\alpha+\sigma\tilde{g}_j<0|Y'_j=-1)=\mathrm{Q}\bigg(-\frac{\alpha}{\sigma}\bigg),$$

and similarly,

$$P_{\hat{Y}'_j|Y'_j}(1|-1)=\mathrm{Q}\bigg(\frac{\alpha}{\sigma}\bigg),\quad P_{\hat{Y}'_j|Y'_j}(-1|1)=\mathrm{Q}\bigg(\frac{\alpha}{\sigma}\bigg),\quad P_{\hat{Y}'_j|Y'_j}(1|1)=\mathrm{Q}\bigg(-\frac{\alpha}{\sigma}\bigg).$$

Thus, we conclude that

$$P_{Y_j'|\hat{Y}_j'}(y_j'|\hat{y}_j') = \begin{cases} \mathrm{Q}(-\frac{\alpha}{\sigma}) & y_j' = \hat{y}_j' \\ \mathrm{Q}(\frac{\alpha}{\sigma}) & y_j' \neq \hat{y}_j'. \end{cases}$$

To calculate the conditional probability distribution $P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}$, recall the decomposition of $\mathbf{X}_j'$ and $\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j'$ in (35) and (36). Since the event $\{\hat{Y}_j' = -1, Y_j' = -1\}$ is equivalent to $\{\tilde{g}_j < \alpha/\sigma\}$ and $\tilde{g}_j \sim \mathcal{N}(0,1)$, the conditional density of $\tilde{g}_j$ given $\hat{Y}_j' = -1, Y_j' = -1$ is given by

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

Similarly, for any $u \in \mathbb{R}$

$$\begin{aligned} p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|-1,1) &= p_{\tilde{g}_j|\tilde{g}_j \le -\alpha/\sigma}(u) = \frac{\mathbb{1}\{u \le -\alpha/\sigma\} f_{\tilde{g}_j}(u)}{\Phi(-\alpha/\sigma)}, \\ p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|1,-1) &= p_{\tilde{g}_j|\tilde{g}_j > \alpha/\sigma}(u) = \frac{\mathbb{1}\{z > \alpha/\sigma\} f_{\tilde{g}_j}(u)}{\mathrm{Q}(\alpha/\sigma)}, \\ p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|1,1) &= p_{\tilde{g}_j|\tilde{g}_j > -\alpha/\sigma}(u) = \frac{\mathbb{1}\{u > -\alpha/\sigma\} p_{\tilde{g}_j}(u)}{\mathrm{Q}(-\alpha/\sigma)}. \end{aligned}$$

For any $\mathbf{x} = \boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp) \in \mathbb{R}^d$, given $\hat{Y}_j' = 1, Y_j' = 1$, the conditional probability distribution at $\mathbf{X}_j' = \mathbf{x}$ is given by

$$\begin{aligned} P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|1,1) &= P_{\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,1) \\ &= P_{\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,1) \\ &= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|1,1) p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp), \end{aligned} \tag{55}$$

where (55) follows since $\tilde{g}_j$ and $\tilde{\mathbf{g}}_j^\perp$ are mutually independent and $\bar{\boldsymbol{\theta}}_0 \perp \tilde{\mathbf{g}}_j^\perp$.
Since we can decompose $2\boldsymbol{\mu}/\sigma$ as

$$\frac{2\boldsymbol{\mu}}{\sigma} = \frac{2\alpha\bar{\boldsymbol{\theta}}_0 + 2\beta^2\boldsymbol{\mu} - 2\alpha\beta\boldsymbol{\upsilon}}{\sigma} = \frac{2\alpha}{\sigma}\bar{\boldsymbol{\theta}}_0 + \bar{\boldsymbol{\theta}}_0^\perp,$$

given $\hat{Y}_j' = 1, Y_j' = -1$, the conditional probability distribution at $\mathbf{X}_j' = \mathbf{x}$ is given by

$$\begin{aligned} P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|1,-1) &= P_{-\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,-1) \\ &= P_{\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}\Big(\sigma\Big(\frac{2\boldsymbol{\mu}}{\sigma} + u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp\Big)\Big|1,-1\Big) \\ &= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}\Big(u + \frac{2\alpha}{\sigma}\Big|1,-1\Big) p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp + \bar{\boldsymbol{\theta}}_0^\perp). \end{aligned}$$

Similarly, for any $\mathbf{x} = -\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp) \in \mathbb{R}^d$, given $\hat{Y}'_j = -1, Y'_j = 1$, the conditional distribution at $\mathbf{X}'_j = \mathbf{x}$ is given by

$$\begin{aligned}
P_{\mathbf{X}'_j|\hat{Y}'_j,Y'_j}(\mathbf{x}|-1,1) &= P_{\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}'_j,Y'_j}(-\boldsymbol{\mu}+\sigma(u\bar{\boldsymbol{\theta}}_0+\mathbf{u}^\perp)|-1,1) \\
&= p_{\tilde{g}_j|\hat{Y}'_j,Y'_j}\Big(u-\frac{2\alpha}{\sigma}\Big|-1,1\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp);
\end{aligned}$$

and given $\hat{Y}'_j = -1, Y'_j = -1$,

$$\begin{aligned}
P_{\mathbf{X}'_j|\hat{Y}'_j,Y'_j}(\mathbf{x}|-1,-1) &= P_{-\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}'_j,Y'_j}(-\boldsymbol{\mu}+\sigma(u\bar{\boldsymbol{\theta}}_0+\mathbf{u}^\perp)|-1,-1) \\
&= p_{\tilde{g}_j|\hat{Y}'_j,Y'_j}(u|-1,-1)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp).
\end{aligned}$$

Furthermore, for any $\mathbf{x} = -\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp) \in \mathbb{R}^d$, we have

$$\begin{aligned}
P_{\mathbf{X}'_j|\hat{Y}'_j=-1}(\mathbf{x}) &= \sum_{y\in\{-1,+1\}} P_{\mathbf{X}'_j|\hat{Y}'_j=-1,Y'_j=y}(\mathbf{x})P_{Y'_j|\hat{Y}'_j=-1}(y) \\
&= P_{Y'_j|\hat{Y}'_j=-1}(1)p_{\tilde{g}_j|\hat{Y}'_j,Y'_j}\Big(u-\frac{2\alpha}{\sigma}\Big|-1,1\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp) \\
&\quad + P_{Y'_j|\hat{Y}'_j=-1}(-1)p_{\tilde{g}_j|\hat{Y}'_j,Y'_j}(u|-1,-1)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp) \\
&= \mathbb{1}\Big\{u \le \frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}\Big(u-\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp) + \mathbb{1}\Big\{u \le \frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp);
\end{aligned}$$

for any $\mathbf{x} = \boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp) \in \mathbb{R}^d$, we have

$$\begin{aligned}
P_{\mathbf{X}'_j|\hat{Y}'_j=1}(\mathbf{x}) &= \sum_{y\in\{-1,+1\}} P_{\mathbf{X}'_j|\hat{Y}'_j=1,Y'_j=y}(\mathbf{x})P_{Y'_j|\hat{Y}'_j=1}(y) \\
&= \mathbb{1}\Big\{u > -\frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}\Big(u+\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp+\bar{\boldsymbol{\theta}}_0^\perp) + \mathbb{1}\Big\{u > -\frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp).
\end{aligned}$$

Define the set $\mathcal{U}_0^\perp(\xi_0, \boldsymbol{\mu}^\perp) := \{\mathbf{u}^\perp \in \mathbb{R}^d : \mathbf{u}^\perp \perp \boldsymbol{\theta}_0\}$. We also use $\mathcal{U}_0^\perp$ to represent $\mathcal{U}_0^\perp(\xi_0, \boldsymbol{\mu}^\perp)$, if there is no risk of confusion. Recall (45) and note that $\int_{\mathcal{U}_0^\perp} p_{\tilde{\mathbf{g}}^\perp}(\mathbf{u}^\perp)\mathrm{d}\mathbf{u}^\perp = 1$. Finally, the KL-divergence is given by

$$\begin{aligned}
&D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=-1}\|P_{\mathbf{X}|Y=-1}) \\
&= \int_{\mathcal{U}_0^\perp}\int_{-\infty}^{\frac{\alpha}{\sigma}} \Big(p_{\tilde{g}_j}\Big(u-\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp) + p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)\Big) \\
&\qquad \times \log\left(1 + \frac{p_{\tilde{g}_j}\Big(u-\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp)}{p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)}\right)\mathrm{d}u\,\mathrm{d}\mathbf{u}^\perp = G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp)
\end{aligned}$$

and

$$
\begin{aligned}
&D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=1}\|P_{\mathbf{X}|Y=1}) \\
&= \int_{\mathcal{U}_0^\perp}\int_{-\frac{\alpha}{\sigma}}^{+\infty}\left(p_{\tilde{g}_j}\left(u+\frac{2\alpha}{\sigma}\right)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp+\bar{\boldsymbol{\theta}}_0^\perp)+p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)\right) \\
&\qquad \times \log\left(1+\frac{p_{\tilde{g}_j}\left(u+\frac{2\alpha}{\sigma}\right)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp+\bar{\boldsymbol{\theta}}_0^\perp)}{p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)}\right)\mathrm{d}u\,\mathrm{d}\mathbf{u}^\perp = G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp), \qquad (56)
\end{aligned}
$$

where (56) follows from since $p_{\tilde{g}_j}$ and $p_{\tilde{\mathbf{g}}_j^\perp}$ are zero-mean Gaussian distributions. Then from (54), we have

$$
D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j,\hat{Y}'_j}\|P_{\mathbf{X},Y}) = G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp).
$$

Thus, by combining the aforementioned results, we get the closed-form expression of the upper bound for $|\mathrm{gen}_1|$. Indeed, if we fix some $d\in\mathbb{N}$, $\epsilon>0$ and $\delta\in(0,1)$, there exists $n_0(d,\delta)\in\mathbb{N}$, $m_0(\epsilon,d,\delta)\in\mathbb{N}$, $c_0(d,\delta)\in(\tilde{c}_1,\infty)$, $r_0(d,\delta)\in\mathbb{R}_+$ such that for all $n>n_0, m>m_0, c>c_0, r>r_0$, $\delta_{m,c-\tilde{c}_1,d}<\frac{\delta}{3}$, $\delta_{r,d}<\frac{\delta}{3}$, and with probability at least $1-\delta$,

$$
|\mathrm{gen}_1| \le \sqrt{\frac{(c_2-c_1)^2}{2}}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{G_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp),\xi_0,\boldsymbol{\mu}^\perp)+\epsilon}\;\right].
$$

4. **Pseudo-label using $\boldsymbol{\theta}_1$:** The same as those in Appendix C.
5. **Iteration $t=2$:** Recall $\boldsymbol{\theta}_2$ in (43), the new model parameter learned from the pseudo-labelled dataset $\hat{S}_{\mathrm{u},2}$.

   Given any $(\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp)$, for any $j\in\mathcal{I}_2$, let $\boldsymbol{\mu}_2^{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} := \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top\mathbf{X}'_j)\mathbf{X}'_j|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp]$ and $\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}$ denotes the probability measure under the parameters $\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp$. Following the similar steps that derive (51), for any $\varepsilon>0$, we have

$$
\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\left(\|\boldsymbol{\theta}_2-\boldsymbol{\mu}_2^{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\|_\infty>\varepsilon\right)\le\delta_{m,\varepsilon,d}.
$$

   From (50), no matter what $\boldsymbol{\theta}_1$ is, we always have $\|\boldsymbol{\mu}_2^{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}-\boldsymbol{\mu}^{\xi_0,\boldsymbol{\mu}^\perp}\|\le\tilde{c}_1$. Then, for some $c\in(\tilde{c}_1,\infty)$,

$$
\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2\in\Theta_{\boldsymbol{\mu},c})\ge 1-\delta_{m,c-\tilde{c}_1,d}.
$$

   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:

$$
\begin{aligned}
&|\mathrm{gen}_2| = |\mathbb{E}[L_{P_{\mathbf{Z}}}(\boldsymbol{\theta}_2)-L_{\hat{S}_{\mathrm{u},2}}(\boldsymbol{\theta}_2)]| \\
&= \left|\frac{1}{m}\sum_{i\in\mathcal{I}_2}\mathbb{E}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_2,(\mathbf{X},Y))-l(\boldsymbol{\theta}_2,(\mathbf{X}'_i,\hat{Y}'_i))|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp\right]\right]\right| \\
&\le \sqrt{\frac{(c_2-c_1)^2}{2}}\frac{1}{m}\sum_{i\in\mathcal{I}_2}\mathbb{E}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{I_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2;(\mathbf{X}'_i,\hat{Y}'_i))+D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_i,\hat{Y}'_i}\|P_{\mathbf{X},Y})}\right],
\end{aligned}
$$

where $P_{\boldsymbol{\theta}_2,\mathbf{X},Y|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} = P_{\boldsymbol{\theta}_2|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} \otimes P_{\mathbf{X},Y}$.

Similar to (53), for any $\epsilon > 0$ and $\delta \in (0,1)$, there exists $m_1(\epsilon, d, \delta)$ such that for all $m > m_1$,

$$\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(I_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2; (\mathbf{X}_i', \hat{Y}_i')) > \epsilon) \le \delta.$$

Recall (42) that $P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} \sim \mathrm{unif}(\{-1,+1\})$. For any fixed $(\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp)$, recall $\bar{\boldsymbol{\theta}}_1$ can be decomposed as $\bar{\boldsymbol{\theta}}_1 = \alpha_1(\xi_0, \boldsymbol{\mu}^\perp)\boldsymbol{\mu} + \beta_1(\xi_0, \boldsymbol{\mu}^\perp)\boldsymbol{v}$.

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

$$\begin{aligned}
&D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i',\hat{Y}_i'} \| P_{\mathbf{X},Y}\big) \\
&= \frac{1}{2} D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i'|\hat{Y}_i'=-1} \| P_{\mathbf{X}|Y=-1}\big) + \frac{1}{2} D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i'|\hat{Y}_i'=1} \| P_{\mathbf{X}|Y=1}\big) \\
&= G_\sigma\big(\alpha_1(\xi_0, \boldsymbol{\mu}^\perp), \xi_0, \boldsymbol{\mu}^\perp\big).
\end{aligned}$$

Given any pair of $(\xi_0, \boldsymbol{\mu}^\perp)$, recall the decomposition of $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ in (40). Then the correlation between $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ and $\boldsymbol{\mu}$ is given by

$$\begin{aligned}
\rho(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}, \boldsymbol{\mu}) &= \frac{1 - 2\mathrm{Q}\big(\frac{\alpha}{\sigma}\big) + \frac{2\sigma\alpha}{\sqrt{2\pi}} \exp(-\frac{\alpha^2}{2\sigma^2})}{\sqrt{\big(1 - 2\mathrm{Q}\big(\frac{\alpha}{\sigma}\big) + \frac{2\sigma\alpha}{\sqrt{2\pi}} \exp(-\frac{\alpha^2}{2\sigma^2})\big)^2 + \frac{2\sigma^2(1-\alpha^2)}{\pi} \exp(-\frac{\alpha^2}{\sigma^2})}} \\
&= F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp)).
\end{aligned}$$

By the strong law of large numbers, we have $\alpha_1(\xi_0, \boldsymbol{\mu}^\perp) \xrightarrow{\text{a.s.}} F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp))$ 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 $m > m_2$,

$$\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\Big(\Big|G_\sigma\big(\alpha_1(\xi_0, \boldsymbol{\mu}^\perp), \xi_0, \boldsymbol{\mu}^\perp\big) - G_\sigma\Big(F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp)), \xi_0, \boldsymbol{\mu}^\perp\Big)\Big| > \epsilon\Big) \le \delta.$$

Therefore, fix some $d \in \mathbb{N}$, $\epsilon > 0$ and $\delta \in (0,1)$. There exists $n_0(d,\delta) \in \mathbb{N}$, $m_3(\epsilon, d, \delta) \in \mathbb{N}$, $c_0(d,\delta) \in (\tilde{c}_1, \infty)$, $r_0(d,\delta) \in \mathbb{R}_+$ such that for all $n > n_0, m > m_3, c > c_0, r > r_0$, $\delta_{m,c-\tilde{c}_1,d} < \frac{\delta}{3}$, $\delta_{r,d} < \frac{\delta}{3}$, and then with probability at least $1 - \delta$, the absolute generalization error at $t = 2$ can be upper bounded as follows:

$$|\mathrm{gen}_2| \le \frac{c_2 - c_1}{\sqrt{2}} \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{G_\sigma\Big(F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp)), \xi_0, \boldsymbol{\mu}^\perp\Big) + \epsilon}\right].$$

6. **Any iteration** $t \in [3:\tau]$**:** By similarly repeating the calculation in iteration $t = 2$, we obtain the upper bound for $|\mathrm{gen}_t|$ in (47).

F. Reusing S_{l} in Each Iteration

If the labelled data S_{l} 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

$$\begin{aligned}\boldsymbol{\theta}'_t &= \frac{n}{n+m}\boldsymbol{\theta}_0 + \frac{1}{n+m}\sum_{i \in \mathcal{I}_t} \hat{Y}'_i \mathbf{X}'_i \\ &= \frac{n}{n+m}\boldsymbol{\theta}_0 + \frac{1}{n+m}\sum_{i \in \mathcal{I}_t} \operatorname{sgn}(\bar{\boldsymbol{\theta}}'^{\top}_{t-1}\mathbf{X}'_i)\mathbf{X}'_i.\end{aligned}$$

Similarly to F_σ, let us define the *enhanced correlation evolution function* $\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp} : [-1,1] \to [-1,1]$ as follows:

$$\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(x) = \left(1 + \frac{\left(w\frac{\sigma\|\boldsymbol{\mu}^\perp\|_2}{n} + (1-w)\left(\frac{2\sigma\sqrt{1-x^2}}{\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2})\right)\right)^2}{\left(w(1+\frac{\sigma}{\sqrt{n}}\xi_0) + (1-w)(1-2\mathrm{Q}\left(\frac{x}{\sigma}\right) + \frac{2\sigma x}{\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2}))\right)^2}\right)^{-\frac{1}{2}}. \tag{57}$$

From Theorem 7, we can obtain similar characterization for gen_t.

Corollary 10. *Fix any $\sigma \in \mathbb{R}_+$, $d \in \mathbb{N}$ and $\alpha = \alpha(\xi_0, \boldsymbol{\mu}^\perp)$. For almost all sample paths,*

$$\begin{aligned}&\mathrm{gen}_t - o(1) \\ &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\frac{m(m-1)(J_\sigma^2(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha)) + K_\sigma^2(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha))) - m(m-n)J_\sigma(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha)) - nm}{(n+m)^2\sigma^2}\right].\end{aligned} \tag{58}$$

The proof of Corollary 10 is provided in Appendix G.

Recall the definition of the function $\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ in (57). Let the t-th iterate of $\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ be denoted as $\tilde{F}^{(t)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ with initial condition $\tilde{F}^{(0)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(x) = x$. As shown in Figure 16, we can see that for any fixed $(\sigma, \xi_0, \boldsymbol{\mu}^\perp)$, $\tilde{F}^{(t)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ has a similar behaviour as $F_\sigma^{(t)}$ as t increases, which implies that the gen-error in (58) in Corollary 10 also decreases as t increases. As a result, $\tilde{F}^{(t)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ represents the improvement of the model parameter $\boldsymbol{\theta}_t$ over the iterations.

As shown in Figure 17, under the same setup as Figure 14(a), when the labelled data S_{l} are reused in each iteration,

the upper bound for $|\mathrm{gen}_t|$ is also a decreasing function of t. When $m = 1000$, gen_t is almost the same as the gen-error when the labelled data are not reused in the subsequent iterations, which means that for large enough m/n, reusing the labelled data does not necessarily help to improve the generalization performance. Moreover, when $m = 100$, gen_t 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.

G. Proof of Corollary 10

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

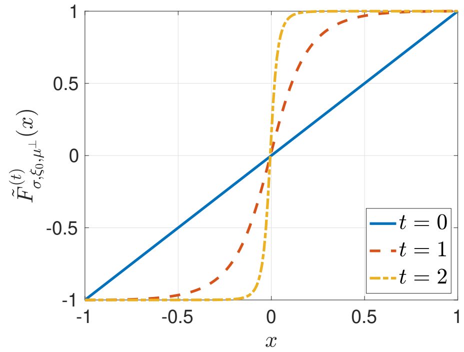


Figure 16: $\tilde{F}^{(t)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(x)$ versus x for $t \in \{0, 1, 2\}$ when $\sigma = 0.5$, $\xi_0 = 0$, $\|\boldsymbol{\mu}^\perp\|_2 = 1$, $n = 10$, and $m = 1000$.

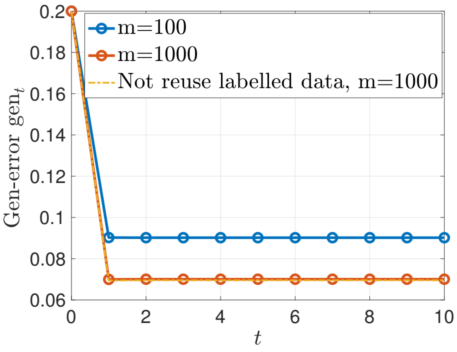


Figure 17: gen_t versus t for $m = 100$ and $m = 1000$, when $n = 10$, $\sigma = 0.6$, $d = 2$, and $\boldsymbol{\mu} = (1, 0)$.

At $t = 1$, from (28) and (40), the expectation $\boldsymbol{\mu}_1'^{\xi_0,\boldsymbol{\mu}^\perp} := \mathbb{E}[\boldsymbol{\theta}_1'|\xi_0, \boldsymbol{\mu}^\perp]$ is rewritten as

$$\begin{aligned}
\boldsymbol{\mu}_1'^{\xi_0,\boldsymbol{\mu}^\perp} &= \frac{n}{n+m}\boldsymbol{\theta}_0 + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j'|\xi_0, \boldsymbol{\mu}^\perp] \\
&= \frac{n}{n+m}\left(\left(1 + \frac{\sigma}{\sqrt{n}}\xi_0\right)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^\perp\right) \\
&\quad + \frac{m}{n+m}\left(\left(1 - 2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right) + \frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\boldsymbol{\mu} + \frac{2\sigma\beta}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\boldsymbol{\upsilon}\right) \\
&= \left(1 + \frac{\sqrt{n}\sigma\xi_0}{n+m} + \frac{m}{n+m}\left(-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right) + \frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\right)\boldsymbol{\mu} \\
&\quad + \left(\frac{\sqrt{n}\sigma\|\boldsymbol{\mu}^\perp\|_2}{n+m} + \frac{m}{n+m}\frac{2\sigma\beta}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\boldsymbol{\upsilon}.
\end{aligned}$$

Then the correlation between $\boldsymbol{\mu}_1'^{\xi_0,\boldsymbol{\mu}^\perp}$ and $\boldsymbol{\mu}$ is given by

$$\rho(\boldsymbol{\mu}_1'^{\xi_0,\boldsymbol{\mu}^\perp}, \boldsymbol{\mu}) = \tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha).$$

The gen-error gen_1 is given by

$$\begin{aligned}
\mathrm{gen}_1 &= \frac{1}{n+m}\sum_{i=1}^{n}\mathbb{E}\Big[l(\boldsymbol{\theta}_1', (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1, (\mathbf{X}_i, Y_i))\Big] \\
&\quad + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_1', (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1', (\mathbf{X}_i', \hat{Y}_i'))|\xi_0, \boldsymbol{\mu}^\perp\right]\right] \\
&= \frac{1}{n+m}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}_1'}\left[\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{Z}_i|\boldsymbol{\theta}_1'}|p_{\boldsymbol{\theta}_1'})\right]
\end{aligned}$$

$$+ \frac{1}{n+m} \sum_{i=1}^{m} \sum_{i=1}^{n} \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp} \mathbb{E}_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp} \Big[\Delta \mathrm{h}(P_{\mathbf{Z}} \| P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp} | p_{\boldsymbol{\theta}_1'})$$
$$+ \Delta \mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp} \| P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1'} | p_{\boldsymbol{\theta}_1}) \Big].$$

- **Calculate $\mathbb{E}_{\boldsymbol{\theta}_1'}\Big[\Delta \mathrm{h}(P_{\mathbf{Z}} \| P_{\mathbf{Z}_i|\boldsymbol{\theta}_1'} | p_{\boldsymbol{\theta}_1'})\Big]$:**

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_1'}\Big[\Delta \mathrm{h}(P_{\mathbf{Z}} \| P_{\mathbf{Z}_i|\boldsymbol{\theta}_1'} | p_{\boldsymbol{\theta}_1'})\Big] \\
&= \int Q_{\boldsymbol{\theta}_1'}(\boldsymbol{\theta}) (P_{\mathbf{Z}}(\mathbf{z}) - P_{\mathbf{Z}_i|\boldsymbol{\theta}_1'}(\mathbf{z}|\boldsymbol{\theta})) \log \frac{1}{p_{\boldsymbol{\theta}}(\mathbf{z})} \mathrm{d}\mathbf{z}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2} \int \Big(P_{\mathbf{Z}}(\mathbf{x},1)(Q_{\boldsymbol{\theta}_1'}(\boldsymbol{\theta}) - P_{\boldsymbol{\theta}_1'|\mathbf{Z}_i}(\boldsymbol{\theta}|\mathbf{x},1)) \\
&\qquad - P_{\mathbf{Z}}(\mathbf{x},-1)(Q_{\boldsymbol{\theta}_1'}(\boldsymbol{\theta}) - P_{\boldsymbol{\theta}_1'|\mathbf{Z}_i}(\boldsymbol{\theta}|\mathbf{x},-1)) \Big) 2\boldsymbol{\theta}^\top \mathbf{x} \; \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{\sigma^2} \int \Big(\frac{\boldsymbol{\mu}-\mathbf{x}}{n+m} P_{\mathbf{Z}}(\mathbf{x},1) - \frac{\boldsymbol{\mu}+\mathbf{x}}{n+m} P_{\mathbf{Z}}(\mathbf{x},-1) \Big)^\top \mathbf{x} \; \mathrm{d}\mathbf{x} \\
&= -\frac{1}{\sigma^2} \bigg(\frac{\boldsymbol{\mu}^\top\boldsymbol{\mu} - \boldsymbol{\mu}^\top\boldsymbol{\mu} - d\sigma^2}{2(n+m)} - \frac{-\boldsymbol{\mu}^\top\boldsymbol{\mu} + \boldsymbol{\mu}^\top\boldsymbol{\mu} + d\sigma^2}{2(n+m)} \bigg) \\
&= \frac{d}{n+m}.
\end{aligned}$$

- **Calculate $\mathbb{E}_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}\big[\Delta \mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp} \| P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1'} | p_{\boldsymbol{\theta}_1})\big]$:**

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}\big[\Delta \mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp} \| P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1'} | p_{\boldsymbol{\theta}_1})\big] \\
&= -\frac{1}{\sigma^2} \int P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp) \big(Q_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}|\xi_0,\boldsymbol{\mu}^\perp) \\
&\qquad\qquad - P_{\boldsymbol{\theta}_1'|\mathbf{X}_i',\hat{Y}_i',\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}|\mathbf{x},y,\xi_0,\boldsymbol{\mu}^\perp) \big) (y\boldsymbol{\theta}^\top\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{\sigma^2} \int P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp) \bigg(\frac{\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} - y\mathbf{x}}{n+m} \bigg)^\top (y\mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}y \\
&= \frac{1}{(n+m)\sigma^2} \Big(\mathbb{E}[\mathbf{X}_i'^\top \mathbf{X}_i' | \xi_0, \boldsymbol{\mu}^\perp] - (\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} \Big) \\
&= \frac{d\sigma^2 + \boldsymbol{\mu}^\top\boldsymbol{\mu} - (\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}}{(n+m)\sigma^2} \\
&= \frac{d\sigma^2 + 1 - J_\sigma^2(\alpha) - K_\sigma^2(\alpha)}{(n+m)\sigma^2}.
\end{aligned}$$

- **Calculate $\mathbb{E}_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}[\Delta \mathrm{h}(P_{\mathbf{Z}} \| P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp} | p_{\boldsymbol{\theta}_1'})]$:** Since given any fixed $\boldsymbol{\theta}_1$, $p_{\boldsymbol{\theta}_1'}(\mathbf{x}|\cdot)$ is a Gaussian distribution, for any $y \in \{\pm 1\}$, we have

$$\frac{1}{2} \int P_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}) \big(P_{\mathbf{X}|Y}(\mathbf{x}|\mathbf{y}) - P_{\mathbf{X}_i'|\hat{Y}_i'}(\mathbf{x}|y) \big) \log \frac{1}{p_{\boldsymbol{\theta}}(\mathbf{x}|y)} \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}$$

$$
\begin{aligned}
&= \frac{1}{4\sigma^2}\int P_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\big(P_{\mathbf{X}|Y}(\mathbf{x}|y) - P_{\mathbf{X}_i'|\hat{Y}_i'}(\mathbf{x}|y)\big)\big(\mathbf{x}^\top\mathbf{x} - 2y\boldsymbol{\theta}^\top\mathbf{x} + \boldsymbol{\theta}^\top\boldsymbol{\theta}\big)\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}\int P_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\bigg(\frac{1}{2}P_{\mathbf{X}|Y}(\mathbf{x}|y) - \frac{1}{2}P_{\mathbf{X}_i'|\hat{Y}_i'}(\mathbf{x}|y)\bigg)\big(y\boldsymbol{\theta}^\top\mathbf{x}\big)\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta} \\
&= -\frac{1}{2\sigma^2}(\boldsymbol{\mu}_1'^{\xi_0,\boldsymbol{\mu}^\perp})^\top\big(\boldsymbol{\mu} - \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\big) \\
&= \frac{m(J_\sigma^2(\alpha) + K_\sigma^2(\alpha)) - (m-n)J_\sigma(\alpha) - n}{2\sigma^2(n+m)},
\end{aligned}
$$

and then

$$
\mathbb{E}_{\boldsymbol{\theta}_1'|\xi_0,\boldsymbol{\mu}^\perp}[\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_1'})] = \frac{m(J_\sigma^2(\alpha) + K_\sigma^2(\alpha)) - (m-n)J_\sigma(\alpha) - n}{\sigma^2(n+m)}.
$$

Therefore, the gen-error at $t = 1$ can be exactly characterized as follows

$$
\begin{aligned}
\mathrm{gen}_1 &= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\bigg[\frac{nd}{(n+m)^2\sigma^2} \\
&\qquad + \frac{m}{n+m}\frac{d\sigma^2 + 1 - J_\sigma^2(\alpha) - K_\sigma^2(\alpha) + m(J_\sigma^2(\alpha) + K_\sigma^2(\alpha)) - (m-n)J_\sigma(\alpha) - n}{(n+m)\sigma^2}\bigg] \\
&= \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\bigg[\frac{m(m-1)(J_\sigma^2(\alpha) + K_\sigma^2(\alpha)) - m(m-n)J_\sigma(\alpha) - nm + m(1+d\sigma^2) + nd}{(n+m)^2\sigma^2}\bigg].
\end{aligned}
$$

For $t \geq 2$, similar to the derivation in Appendix C, by iteratively implementing the calculation, we only need to replace α with the correlation evolution function $\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\cdot)$ (cf. (57)) and then the gen-error for any $t \geq 1$ is characterized as follows. For almost all sample paths, there exists a vanishing sequence ϵ_m ($\epsilon_m \to 0$ as $m \to \infty$) , such that

$$
\begin{aligned}
\mathrm{gen}_t = \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\bigg[&\frac{m(m-1)(J_\sigma^2(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha)) + K_\sigma^2(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha)))}{(n+m)^2\sigma^2} \\
&+ \frac{-m(m-n)J_\sigma(\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha)) - nm}{(n+m)^2\sigma^2}\bigg] + \epsilon_m',
\end{aligned}
$$

where $\epsilon_m' = \epsilon_m + \frac{m(1+d\sigma^2)+nd}{(n+m)^2\sigma^2} \to 0$ as $m \to \infty$ and α stands for $\alpha(\xi_0, \boldsymbol{\mu}^\perp)$.

The proof of Corollary 10 is thus completed.

H. Proof of Theorem 4

Theorem 4 can be proved similarly from the proof of Theorem 7. For simplicity, in the following proofs, we abbrviate $\mathrm{gen}_t(P_{\mathbf{Z}}, P_{\mathbf{X}}, \{P_{\boldsymbol{\theta}_k|S_\mathrm{l},S_\mathrm{u}}\}_{k=0}^t, \{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})$ as gen_t. With ℓ_2-regularization, the algorithm operates in the following steps. Let $\lambda \in \mathbb{R}_+$ be the regularization parameter.

- **Step 1: Initial round $t = 0$ with S_l:** By minimizing the regularized empirical risk of labelled dataset S_l

$$
L_{S_\mathrm{l}}^{\mathrm{reg}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n l(\boldsymbol{\theta}, (\mathbf{X}_i, Y_i)) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2 \overset{\mathrm{c}}{=} \frac{1}{2\sigma^2 n}\sum_{i=1}^n(\mathbf{X}_i - Y_i\boldsymbol{\theta})^\top(\mathbf{X}_i - Y_i\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2,
$$

where $\overset{\mathrm{c}}{=}$ means that both sides differ by a constant independent of $\boldsymbol{\theta}$, we obtain the minimizer

$$\begin{aligned}\boldsymbol{\theta}_0^{\mathrm{reg}} &= \arg\min_{\boldsymbol{\theta}\in\Theta} L_{S_\mathrm{l}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i\mathbf{X}_i}{1+\sigma^2\lambda} \\ &= \frac{\boldsymbol{\theta}_0}{1+\sigma^2\lambda} \sim \mathcal{N}\biggl(\frac{\boldsymbol{\mu}}{1+\sigma^2\lambda}, \frac{\sigma^2}{n(1+\sigma^2\lambda)^2}\mathbf{I}_d\biggr). \end{aligned} \tag{59}$$

- **Step 2: Pseudo-label data in S_u:** At each iteration $t \in [1:\tau]$, for any $i \in \mathcal{I}_t$, we use $\boldsymbol{\theta}_{t-1}^{\mathrm{reg}}$ to assign a pseudo-label for $\mathbf{X}_i'$, that is, $\hat{Y}_i' = f_{\boldsymbol{\theta}_{t-1}^{\mathrm{reg}}}(\mathbf{X}_i') = \mathrm{sgn}(\mathbf{X}_i'^{\top}\boldsymbol{\theta}_{t-1}^{\mathrm{reg}})$.

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

$$\begin{aligned} L_{\hat{S}_{\mathrm{u},t}}(\boldsymbol{\theta}) &= \frac{1}{m}\sum_{i\in\mathcal{I}_t} l(\boldsymbol{\theta},(\mathbf{X}_i',\hat{Y}_i')) + \frac{\lambda}{2m}\|\boldsymbol{\theta}\|_2^2 \\ &\overset{\mathrm{c}}{=} \frac{1}{2\sigma^2 m}\sum_{i\in\mathcal{I}_t}(\mathbf{X}_i' - \hat{Y}_i'\boldsymbol{\theta})^\top(\mathbf{X}_i' - \hat{Y}_i'\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2, \end{aligned}$$

we obtain the new model parameter

$$\boldsymbol{\theta}_t^{\mathrm{reg}} = \frac{1}{m}\sum_{i\in\mathcal{I}_t}\frac{\hat{Y}_i'\mathbf{X}_i'}{1+\sigma^2\lambda} = \frac{1}{m}\sum_{i\in\mathcal{I}_t}\frac{\mathrm{sgn}(\mathbf{X}_i'^{\top}\boldsymbol{\theta}_{t-1}^{\mathrm{reg}})\mathbf{X}_i'}{1+\sigma^2\lambda}. \tag{60}$$

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

1. **Characterization of** $|\mathrm{gen}_1|$**:**

From (59), we still have $\rho(\boldsymbol{\theta}_0^{\mathrm{reg}}, \boldsymbol{\mu}) = \alpha(\xi_0, \boldsymbol{\mu}^\perp)$ and

$$\bar{\boldsymbol{\theta}}_0^{\mathrm{reg}} := \frac{\boldsymbol{\theta}_0^{\mathrm{reg}}}{\|\boldsymbol{\theta}_0^{\mathrm{reg}}\|_2^2} = \alpha\boldsymbol{\mu} + \beta\boldsymbol{v} = \bar{\boldsymbol{\theta}}_0. \tag{61}$$

From (61), we can rewrite (60) as follows

$$\boldsymbol{\theta}_1^{\mathrm{reg}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\mathrm{sgn}(\mathbf{X}_i'^{\top}\bar{\boldsymbol{\theta}}_1^{\mathrm{reg}})\mathbf{X}_i'}{1+\sigma^2\lambda} = \frac{1}{m}\sum_{i=1}^{m}\frac{\mathrm{sgn}(\mathbf{X}_i'^{\top}\bar{\boldsymbol{\theta}}_0)\mathbf{X}_i'}{1+\sigma^2\lambda} = \frac{\boldsymbol{\theta}_1}{1+\sigma^2\lambda}.$$

Thus, the expectation of $\boldsymbol{\theta}_1^{\mathrm{reg}}$ conditioned on $(\xi_0, \boldsymbol{\mu}^\perp)$ is given by

$$\begin{aligned} \boldsymbol{\mu}_1^{\mathrm{reg}|\xi_0,\boldsymbol{\mu}^\perp} &:= \frac{1}{1+\sigma^2\lambda}\mathbb{E}[\mathrm{sgn}(\mathbf{X}_j'^{\top}\boldsymbol{\theta}_0^{\mathrm{reg}})\mathbf{X}_j'|\xi_0,\boldsymbol{\mu}^\perp] \\ &= \frac{1}{1+\sigma^2\lambda}\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} \\ &= \frac{1}{1+\sigma^2\lambda}\biggl(\biggl(1 - 2\mathrm{Q}\Bigl(\frac{\alpha}{\sigma}\Bigr) + \frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\Bigl(-\frac{\alpha^2}{2\sigma^2}\Bigr)\biggr)\boldsymbol{\mu} + \frac{2\sigma\beta}{\sqrt{2\pi}}\exp\Bigl(-\frac{\alpha^2}{2\sigma^2}\Bigr)\boldsymbol{v}\biggr) \\ &= \frac{J_\sigma(\alpha)\boldsymbol{\mu} + K_\sigma(\alpha)\boldsymbol{v}}{1+\sigma^2\lambda}. \end{aligned}$$

Recall gen_1 given in (41). In the case with regularization, the gen-error $\mathrm{gen}_1^{\mathrm{reg}}$ has the same definition as gen_1. To derive $\mathrm{gen}_1^{\mathrm{reg}}$, we need to calculate the following two terms.

- **Calculate** $\mathbb{E}_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}\big[\Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1^{\mathrm{reg}}}|p_{\boldsymbol{\theta}_1^{\mathrm{reg}}})\big]$**:**

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}\big[\Delta\mathrm{h}(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1^{\mathrm{reg}}}|p_{\boldsymbol{\theta}_1^{\mathrm{reg}}})\big]\\
&= \frac{1}{2\sigma^2}\int Q_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}|\xi_0,\boldsymbol{\mu}^\perp)\big(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp)\\
&\qquad\qquad - P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta}_1^{\mathrm{reg}}}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp,\boldsymbol{\theta})\big)(\mathbf{x}^\top\mathbf{x}-2y\boldsymbol{\theta}^\top\mathbf{x}+\boldsymbol{\theta}^\top\boldsymbol{\theta})\mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta}\\
&= -\frac{1}{\sigma^2}\int P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp)\big(Q_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}|\xi_0,\boldsymbol{\mu}^\perp)\\
&\qquad\qquad - P_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\mathbf{X}_i',\hat{Y}_i',\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}|\mathbf{x},y,\xi_0,\boldsymbol{\mu}^\perp)\big)(y\boldsymbol{\theta}^\top\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y\mathrm{d}\boldsymbol{\theta}\\
&= -\frac{1}{\sigma^2}\int P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}(\mathbf{x},y|\xi_0,\boldsymbol{\mu}^\perp)\bigg(\frac{\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}-y\mathbf{x}}{m(1+\sigma^2\lambda)}\bigg)^\top(y\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}y\\
&= \frac{1}{m\sigma^2(1+\sigma^2\lambda)}\Big(\mathbb{E}[\mathbf{X}_i'^\top\mathbf{X}_i'|\xi_0,\boldsymbol{\mu}^\perp]-(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\Big)\\
&= \frac{d\sigma^2+\boldsymbol{\mu}^\top\boldsymbol{\mu}-(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})^\top\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}}{m\sigma^2(1+\sigma^2\lambda)}\\
&= \frac{d\sigma^2+1-J_\sigma^2(\alpha)-K_\sigma^2(\alpha)}{m\sigma^2(1+\sigma^2\lambda)}.
\end{aligned}
$$

- **Calculate** $\mathbb{E}_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}\big[h(P_{\mathbf{Z}},p_{\boldsymbol{\theta}_1^{\mathrm{reg}}})-h(P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp},p_{\boldsymbol{\theta}_1^{\mathrm{reg}}})\big]$**:** Given any $(\xi_0,\boldsymbol{\mu}^\perp)$, in the following, we drop the condition on $\xi_0,\boldsymbol{\mu}^\perp$ for notational simplicity. Since given $\boldsymbol{\theta}_1$, $p_{\boldsymbol{\theta}_1}(\mathbf{x}|\cdot)$ is a Gaussian distribution, for any $y\in\{\pm1\}$, we have

$$
\begin{aligned}
&\frac{1}{2}\int P_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\big(P_{\mathbf{X}|Y}(\mathbf{x}|\mathbf{y})-P_{\mathbf{X}_i'|\hat{Y}_i'}(\mathbf{x}|y)\big)\log\frac{1}{p_{\boldsymbol{\theta}}(\mathbf{x}|y)}\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\\
&= -\frac{1}{2\sigma^2}\int P_{\boldsymbol{\theta}_1^{\mathrm{reg}}|\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta})\bigg(\frac{1}{2}P_{\mathbf{X}|Y}(\mathbf{x}|y)-\frac{1}{2}P_{\mathbf{X}_i'|\hat{Y}_i'}(\mathbf{x}|y)\bigg)(y\boldsymbol{\theta}^\top\mathbf{x})\mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\\
&= -\frac{1}{2\sigma^2}(\boldsymbol{\mu}_1^{\mathrm{reg}|\xi_0,\boldsymbol{\mu}^\perp})^\top(\boldsymbol{\mu}-\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp})\\
&= \frac{J_\sigma^2(\alpha)+K_\sigma^2(\alpha)-J_\sigma(\alpha)}{2\sigma^2(1+\sigma^2\lambda)}.
\end{aligned}
$$

  Thus, we have

$$
\mathbb{E}_{\boldsymbol{\theta}_1|\xi_0,\boldsymbol{\mu}^\perp}[\Delta\mathrm{h}(P_{\mathbf{Z}}\|P_{\mathbf{X}_i',\hat{Y}_i'|\xi_0,\boldsymbol{\mu}^\perp}|p_{\boldsymbol{\theta}_1})]=\frac{J_\sigma^2(\alpha)+K_\sigma^2(\alpha)-J_\sigma(\alpha)}{\sigma^2(1+\sigma^2\lambda)}.
$$

Finally, the gen-error at $t=1$ can be characterized as follows:

$$
\mathrm{gen}_1^{\mathrm{reg}}=\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\bigg[\frac{J_\sigma^2(\alpha)+K_\sigma^2(\alpha)-J_\sigma(\alpha)}{\sigma^2(1+\sigma^2\lambda)}+\frac{d\sigma^2+1-J_\sigma^2(\alpha)-K_\sigma^2(\alpha)}{m\sigma^2(1+\sigma^2\lambda)}\bigg]=\frac{\mathrm{gen}_1}{1+\sigma^2\lambda},
$$

where α stands for $\alpha(\xi_0,\boldsymbol{\mu}^\perp)$.

2. **Iteration $t \in [2 : \tau]$:** Since $\boldsymbol{\theta}_t^{\mathrm{reg}} = \frac{\boldsymbol{\theta}_t}{1+\sigma^2\lambda}$, by iteratively applying the same techniques in iteration $t = 1$ and in Appendix C, the gen-error at any $t \in [2 : \tau]$ can be characterized as follows
$$\mathrm{gen}_t^{\mathrm{reg}} = \frac{\mathrm{gen}_t}{1+\sigma^2\lambda}.$$

Theorem 4 is thus proved.

References

Shotaro Akaho and Hilbert J Kappen. Nonmonotonic generalization bias of Gaussian mixture models. *Neural Computation*, 12(6):1411–1427, 2000.

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

Gholamali Aminian, Mahed Abroshan, Mohammad Mahdi Khalili, Laura Toni, and Miguel Rodrigues. An information-theoretical approach to semi-supervised learning under covariate-shift. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 7433–7449. PMLR, 28–30 Mar 2022.

Yuichiro Anzai. *Pattern Recognition and Machine Learning.* Elsevier, 2012.

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

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.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

Yuheng Bu, Gholamali Aminian, Laura Toni, Gregory W. Wornell, and Miguel Rodrigues. Characterizing and understanding the generalization error of transfer learning with Gibbs algorithm. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 8673–8699. PMLR, 28–30 Mar 2022.

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*, pages 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*, pages 11192–11203, 2019.

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

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

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.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning.* MIT press, 2016.

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*, pages 770–778, 2016.

Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic bounds on transfer generalization gap based on Jensen–Shannon divergence. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 1461–1465. IEEE, 2021.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 2017.

Alex 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. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 896, 2013.

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

Qun Liu and Supratik Mukhopadhyay. Unsupervised learning using pretrained CNN and associative memory bank. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 01–08. IEEE, 2018.

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

David J. C. MacKay. *Information Theory, Inference and Learning Algorithms.* Cambridge University Press, 2003.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.

John Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. *Advances in Neural Information Processing Systems*, 4, 1992.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research*, 22(222):1–69, 2021.

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*, pages 11013–11023, 2019.

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

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

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR.

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 Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3437–3452. PMLR, 09–12 Jul 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.

Vladimir 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.

Ke Wang and Christos Thrampoulidis. Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, 4(1):260–284, 2022.

Kazuho Watanabe and Sumio Watanabe. Stochastic complexities of Gaussian mixtures in variational Bayesian approximation. *The Journal of Machine Learning Research*, 7: 625–644, 2006.

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

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

Jingge Zhu. Semi-supervised learning: the case when unlabeled data is equally useful. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124, pages 709–718. PMLR, 03–06 Aug 2020.

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

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