APPENDIX

A RELATED WORKS

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

Generalization error bounds: The traditional way of analyzing generalization error includes the Vapnik-Chervonenkis or VC dimension (Vapnik, 2000) and the Rademacher complexity (Boucheron et al., 2005). Recently, Russo & Zou (2016) proposed using mutual information between the estimated output of an algorithm and the actual realized value of the estimates to analyze and bound the bias in data analysis, which can be regarded equivalent to the generalization error. This new approach is simpler and can handle a wider range of loss functions compared to the abovementioned methods and other methods like differential privacy, total-variation information and so on. It also paves a new way to improving generalization capability of learning algorithms from an information-theoretic aspects. Following Russo & Zou (2016), Xu & Raginsky (2017) derived upper bounds on generalization error of learning algorithms with mutual information between the input dataset and the output hypothesis, which formalizes the intuition that less information that a learning algorithm can extract from training dataset leads to less overfitting. Later Pensia et al. (2018) derived generalization error bounds for noisy and iterative algorithms and the key contribution is to bound the mutual information between input data and output hypothesis. Negrea et al. (2019) improved mutual information bounds for Stochastic Gradient Langevin Dynamics (SGLD) via data-dependent estimates compared to distribution-dependent bounds.

However, one major shortcoming of the 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 & 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 & Zakynthinou (2020) proposed using Conditional Mutual Information (CMI) to bound the generalization error, which can still preserve the chain rule property. Bu et al. (2020) provided a tightened upper bound based on the *individual* mutual information (IMI) between the *individual* data sample and the output. Wu et al. (2020) extended Bu et al. (2020)'s result to the transfer learning problems and characterized the upper bound based on IMI and KL-divergence. In a similar manner, Jose & Simeone (2020) provided a tightened bound on transfer generalization error based on the Jensen-Shannon divergence.

B PROOF OF THEOREM 1

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

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 4. *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*

$$
\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&\leq \frac{w}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_-^{*-1}(I_{\theta^{(t-1)}}(\theta_t; Z_i))\right] \\
&\quad + \frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_-^{*-1}\left(I_{\theta^{(t-1)}}(\theta_t; X'_i, \hat{Y}'_i) + D_{\theta^{(t-1)}}(P_{X'_i,\hat{Y}'_i}\|P_Z)\right)\right], \qquad (29) \\
&- \mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&\leq \frac{w}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_+^{*-1}(I_{\theta^{(t-1)}}(\theta_t; Z_i))\right] \\
&\quad + \frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm} \mathbb{E}_{\theta^{(t-1)}}\left[\psi_+^{*-1}\left(I_{\theta^{(t-1)}}(\theta_t; X'_i, \hat{Y}'_i) + D_{\theta^{(t-1)}}(P_{X'_i,\hat{Y}'_i}\|P_Z)\right)\right], \qquad (30)
\end{aligned}
$$

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 KL-divergence between any two distributions P and Q on $\mathcal{X}$:

$$
D(P\|Q) = \sup_{g\in\mathcal{G}}\{\mathbb{E}_{X\sim P}[g(X)] - \log \mathbb{E}_{X\sim Q}[e^{g(X)}]\} \qquad (31)
$$

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

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

$$
\begin{aligned}
&\mathrm{gen}_t(P_Z, P_X, \{P_{\theta_k|S_1,S_u}\}_{k=0}^t, \{f_{\theta_k}\}_{k=0}^{t-1}) \\
&= w\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)]] - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta_t,Z_i}[l(\theta_t, Z_i)]\right) \\
&\quad + (1-w)\left(\mathbb{E}_{\theta_t}[\mathbb{E}_Z[l(\theta_t, Z)]] - \frac{1}{m}\sum_{i=(t-1)m+1}^{tm}\mathbb{E}_{\theta_t,X'_i,\hat{Y}'_i}[l(\theta_t, (X'_i, \hat{Y}'_i))]\right) \qquad (32)
\end{aligned}
$$

$$= \frac{w}{n}\sum_{i=1}^{n}\left(\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})] - \mathbb{E}_{\theta_t,Z_i}[l(\theta_t,Z_i)]\right)$$
$$+ \frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm}\left(\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}[l(\theta_t,(X_i',\hat{Y}_i'))]\right) \tag{33}$$
$$= \frac{w}{n}\sum_{i=1}^{n}\mathbb{E}_{\theta^{(t-1)}}\left[\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\left[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\right] - \mathbb{E}_{\theta_t,Z_i}\left[l(\theta_t,Z_i)|\theta^{(t-1)}\right]\right]$$
$$+ \frac{1-w}{m}\sum_{i=(t-1)m+1}^{tm}\mathbb{E}_{\theta^{(t-1)}}\left[\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\left[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\right] - \mathbb{E}_{\theta_t,X_i',\hat{Y}_i'}\left[l(\theta_t,(X_i',\hat{Y}_i'))|\theta^{(t-1)}\right]\right]. \tag{34}$$

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

$$\begin{aligned}
&I_{\theta^{(t-1)}}(\theta_t;Z) = D(P_{\theta_t,Z|\theta^{(t-1)}}\|P_{\theta_t|\theta^{(t-1)}}\otimes P_Z)\\
&\geq \mathbb{E}_{\theta_t,Z}[\lambda\check{l}(\theta_t,Z)|\theta^{(t-1)}] - \log\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[e^{\lambda\check{l}(\tilde{\theta}_t,\tilde{Z})}|\theta^{(t-1)}] && (35)\\
&= \mathbb{E}_{\theta_t,Z}[\lambda\check{l}(\theta_t,Z)|\theta^{(t-1)}] - \log\mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\mathbb{E}_{\tilde{Z}}\left[e^{\lambda\check{l}(\tilde{\theta}_t,\tilde{Z})}\right] && (36)\\
&= \mathbb{E}_{\theta_t,Z}[\lambda\check{l}(\theta_t,Z)|\theta^{(t-1)}] - \log\mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\left[\exp\left(\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t)\right)\right] && (37)\\
&\geq \lambda\mathbb{E}_{\theta_t,Z}[l(\theta_t,Z) - \mathbb{E}_Z[l(\theta_t,Z)]|\theta^{(t-1)}] - \log\mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\left[\exp(\psi_+(\lambda,\tilde{\theta}_t))\right] && (38)\\
&\geq \lambda\mathbb{E}_{\theta_t,Z}[l(\theta_t,Z) - \mathbb{E}_Z[l(\theta_t,Z)]|\theta^{(t-1)}] - \psi_+(\lambda) && (39)\\
&= \lambda\big(\mathbb{E}_{\theta_t,Z}[l(\theta_t,Z)|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}]\big) - \psi_+(\lambda). && (40)
\end{aligned}$$

where (37) follows from the definition of $\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t)$ in (52), (38) 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 (39) follows since $\psi_+(\lambda) = \sup_{\tilde{\theta}_t}\psi_+(\lambda,\tilde{\theta}_t)$. Thus, we have

$$\begin{aligned}
&\mathbb{E}_{\theta_t,Z}[l(\theta_t,Z)|\theta^{(t-1)}] - \mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}]\\
&\leq \inf_{\lambda\in[0,b_+)}\frac{I_{\theta^{(t-1)}}(\theta_t;Z)+\psi_+(\lambda)}{\lambda} && (41)\\
&= \psi_+^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t;Z)\big). && (42)
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}] - \mathbb{E}_{\theta_t,Z}[l(\theta_t,Z)|\theta^{(t-1)}]\\
&\leq \inf_{\lambda\in[0,-b_-)}\frac{I_{\theta^{(t-1)}}(\theta_t;Z)+\psi_-(\lambda)}{\lambda} && (43)\\
&= \psi_-^{*-1}\big(I_{\theta^{(t-1)}}(\theta_t;Z)\big). && (44)
\end{aligned}$$

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

$$\begin{aligned}
&I_{\theta^{(t-1)}}(\theta_t;X',\hat{Y}') + D_{\theta^{(t-1)}}(P_{X',\hat{Y}'}\|P_Z)\\
&= D_{\theta^{(t-1)}}(P_{\theta_t,X',\hat{Y}'}\|P_{\theta_t}\otimes P_{X',\hat{Y}'}) + D_{\theta^{(t-1)}}(P_{\theta_t}\otimes P_{X',\hat{Y}'}\|P_{\theta_t}\otimes P_Z) && (45)\\
&\geq \mathbb{E}_{\theta_t,X',\hat{Y}'}[\lambda l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}] - \log\mathbb{E}_{\theta_t}\left[\mathbb{E}_{X',\hat{Y}'}[e^{\lambda l(\theta_t,(X',\hat{Y}'))}|\theta^{(t-1)}]|\theta^{(t-1)}\right]\\
&\quad + \mathbb{E}_{\theta_t}\left[\mathbb{E}_{X',\hat{Y}'}[\lambda l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}]|\theta^{(t-1)}\right] - \log\mathbb{E}_{\theta_t}\left[\mathbb{E}_Z[e^{\lambda l(\theta_t,Z)}]|\theta^{(t-1)}\right] && (46)
\end{aligned}$$

$$\geq \mathbb{E}_{\theta_t,X',\hat{Y}'}[\lambda l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}] - \log \mathbb{E}_{\theta_t}\left[\mathbb{E}_Z[e^{\lambda l(\theta_t,Z)}]|\theta^{(t-1)}\right] \tag{47}$$

$$= \lambda\Big(\mathbb{E}_{\theta_t,X',\hat{Y}'}[\lambda l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}] - \mathbb{E}_{\theta_t}\left[\mathbb{E}_Z[l(\theta_t,Z)]|\theta^{(t-1)}\right]\Big)$$
$$- \log \mathbb{E}_{\tilde{\theta}_t|\theta^{(t-1)}}\left[\exp\left(\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t)\right)\right] \tag{48}$$

$$\geq \lambda\Big(\mathbb{E}_{\theta_t,X',\hat{Y}'}[\lambda l(\theta_t,(X',\hat{Y}'))] - \mathbb{E}_{\theta_t}\left[\mathbb{E}_Z[l(\theta_t,Z)]\right]\Big) - \psi_+(\lambda), \tag{49}$$

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

$$\mathbb{E}_{\theta_t,X'_i,\hat{Y}'_i}\left[l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}\right] - \mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\left[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\right]$$
$$\leq \psi_+^{*-1}\left(I_{\theta^{(t-1)}}(\theta_t;X',\hat{Y}') + D_{\theta^{(t-1)}}(P_{X',\hat{Y}'}\|P_Z)\right) \tag{50}$$

and

$$\mathbb{E}_{\tilde{\theta}_t,\tilde{Z}}\left[l(\tilde{\theta}_t,\tilde{Z})|\theta^{(t-1)}\right] - \mathbb{E}_{\theta_t,X'_i,\hat{Y}'_i}\left[l(\theta_t,(X',\hat{Y}'))|\theta^{(t-1)}\right]$$
$$\leq \psi_-^{*-1}\left(I_{\theta^{(t-1)}}(\theta_t;X',\hat{Y}') + D_{\theta^{(t-1)}}(P_{X',\hat{Y}'}\|P_Z)\right). \tag{51}$$

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

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

$$\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t) := \log \mathbb{E}_{\tilde{Z}}[e^{\lambda(l(\tilde{\theta}_t,\tilde{Z})-\mathbb{E}_{\tilde{Z}}[l(\tilde{\theta}_t,\tilde{Z})])}]. \tag{52}$$

When the loss function $l(\theta,Z) \sim \mathsf{subG}(R)$ under $Z \sim P_Z$ for any $\theta \in \Theta$, we have $\Lambda_{l(\tilde{\theta}_t,\tilde{Z})}(\lambda,\tilde{\theta}_t) \leq \frac{R^2\lambda^2}{2}$ for all $\lambda \in \mathbb{R}$. Then we can let $\psi_-(\lambda,\tilde{\theta}_t) = \psi_+(\lambda,\tilde{\theta}_t) = \frac{R^2\lambda^2}{2}$ for all $\tilde{\theta}_t \in \Theta$. Hence, $\psi_+(\lambda) = \psi_-(\lambda) = \sup_{\tilde{\theta}_t\in\Theta}\frac{R^2\lambda^2}{2} = \frac{R^2\lambda^2}{2}$ and $\psi_+^{*-1}(y) = \psi_-^{*-1}(y) = \sqrt{2R^2y}$ for any $y \geq 0$. Finally, Theorem 1 can then be directly obtained from Theorem 4.

C Proof of Theorem 2

Theorem 2 can be proved iteratively by applying Theorem 1. For simplicity, in the following proofs, we abbreviate $\mathrm{gen}_t(P_{\mathbf{Z}},P_{\mathbf{X}},\{P_{\boldsymbol{\theta}_k|S_l,S_u}\}_{k=0}^t,\{f_{\boldsymbol{\theta}_k}\}_{k=0}^{t-1})$ 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 $c \in \mathbb{R}_+$,

$$\Pr(\boldsymbol{\theta}_0 \in \Theta_{\boldsymbol{\mu},c}) = \Pr(\|\boldsymbol{\theta}_0 - \boldsymbol{\mu}\|_\infty \leq c) = \left(1 - 2\Phi\Big(\frac{-\sqrt{n}c}{\sigma}\Big)\right)^d =: 1 - \delta_{\sqrt{n}c,d}. \tag{53}$$

   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 $l(\boldsymbol{\theta}_0,(\mathbf{X},Y)) \sim \mathsf{subG}((c_2-c_1)/2)$ under $(\mathbf{X},Y) \sim P_{\mathbf{X},Y}$ is given as follows

$$\Pr\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda,\tilde{\boldsymbol{\theta}}_0) \leq \frac{\lambda^2(c_2-c_1)^2}{8}\right) \tag{54}$$

$$\geq \Pr\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda,\tilde{\boldsymbol{\theta}}_0) \leq \frac{\lambda^2(c_2-c_1)^2}{8} \text{ and } \tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}\right) \tag{55}$$

$$= \Pr(\tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}) \Pr\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_0,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda,\tilde{\boldsymbol{\theta}}_0) \leq \frac{\lambda^2(c_2-c_1)^2}{8}\Big|\tilde{\boldsymbol{\theta}}_0 \in \Theta_{\boldsymbol{\mu},c}\right) \tag{56}$$

$$= (1-\delta_{\sqrt{n}c,d})(1-\delta_{r,d}), \tag{57}$$

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

Fix some $d \in \mathbb{N}$, $\epsilon > 0$ and $\delta \in (0,1)$. There exists $n_0(d,\delta) \in \mathbb{N}$, $r_0(d,\delta) \in \mathbb{R}_+$ such that for all $n > n_0, r > r_0$, $\delta_{\sqrt{n}c,d} < \frac{\delta}{3}$, $\delta_{r,d} < \frac{\delta}{3}$, and then with probability at least $1-\delta$, the absolute generalization error can be upper bounded as follows

$$|\mathrm{gen}_0| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{(c_2-c_1)^2}{2}I(\boldsymbol{\theta}_0;\mathbf{X}_i,Y_i)}. \tag{58}$$

Then mutual information can be calculated as follows

$$\begin{aligned}
I(\boldsymbol{\theta}_0;\mathbf{X}_i,Y_i) &= h(\boldsymbol{\theta}_0) - h(\boldsymbol{\theta}_0|\mathbf{X}_i,Y_i) && (59)\\
&= h\left(\frac{1}{n}\sum_{j=1}^{n}Y_j\mathbf{X}_j\right) - h\left(\frac{1}{n}\sum_{j=1}^{n}Y_j\mathbf{X}_j\middle|\mathbf{X}_i,Y_i\right) && (60)\\
&= \frac{d}{2}\log\left(\frac{2\pi e\sigma^2}{n}\right) - h\left(\frac{1}{n}\sum_{j\in[n],j\neq i}Y_j\mathbf{X}_j\right) && (61)\\
&= \frac{d}{2}\log\left(\frac{2\pi e\sigma^2}{n}\right) - \frac{d}{2}\log\left(\frac{2\pi e(n-1)\sigma^2}{n^2}\right) && (62)\\
&= \frac{d}{2}\log\frac{n}{n-1}. && (63)
\end{aligned}$$

Thus we can get (23).

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

$$\hat{Y}_i' = \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}_i'). \tag{64}$$

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 & Gulcu (2021), we can decompose it as follows:

$$\begin{aligned}
\boldsymbol{\theta}_0 &= \boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\xi} && (65)\\
&= \boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}(\xi_0\boldsymbol{\mu} + \boldsymbol{\mu}^\perp) && (66)\\
&= \left(1 + \frac{\sigma}{\sqrt{n}}\xi_0\right)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^\perp, && (67)
\end{aligned}$$

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 (19), the decomposition of $\bar{\boldsymbol{\theta}}_0$ in (20) and α, β. Since $\mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}_i') = \mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')$, 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

$$\mathbf{g}_i = g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{\upsilon} + \mathbf{g}_i^\perp, \tag{68}$$

where $g_{0,i}, g_i \sim \mathcal{N}(0,1)$, $\mathbf{g}_i^\perp \sim \mathcal{N}(0, \mathbf{I}_d - \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\upsilon}\boldsymbol{\upsilon}^\top)$, $\mathbf{g}_i^\perp \perp \boldsymbol{\mu}$, $\mathbf{g}_i^\perp \perp \boldsymbol{\upsilon}$ 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

$$\mathbf{X}_i' = \boldsymbol{\mu} + \sigma\mathbf{g}_i = \boldsymbol{\mu} + \sigma(g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{\upsilon} + \mathbf{g}_i^\perp). \tag{69}$$

Then we have

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i' &= (\alpha\boldsymbol{\mu} + \beta\boldsymbol{\upsilon})^\top(\boldsymbol{\mu} + \sigma\mathbf{g}_i) && (70)\\
&= \alpha + \sigma(\alpha\boldsymbol{\mu} + \beta\boldsymbol{\upsilon})^\top(g_{0,i}\boldsymbol{\mu} + g_i\boldsymbol{\upsilon} + \mathbf{g}_i^\perp) && (71)\\
&= \alpha + \sigma(\alpha g_{0,i} + \beta g_i) && (72)\\
&=: \alpha + \sigma h_i. && (73)
\end{aligned}$$

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

$$\mathbf{X}_i' = -\boldsymbol{\mu} + \sigma \mathbf{g}_i \tag{74}$$

and

$$\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i' = -\alpha + \sigma h_i. \tag{75}$$

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

$$\begin{aligned}
\Pr(\hat{Y}_i' = 1) &= \Pr\left(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i' > 0\right) && (76)\\
&= \frac{1}{2}\Pr\left(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i' > 0 | Y_i' = 1\right) + \frac{1}{2}\Pr\left(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i' > 0 | Y_i' = -1\right) && (77)\\
&= \frac{1}{2}\mathbb{E}_\alpha\big[\Pr(\alpha + \sigma h_i > 0)\big] + \frac{1}{2}\mathbb{E}_\alpha\big[\Pr\big(-\alpha + \sigma h_i > 0\big)\big] && (78)\\
&= \frac{1}{2}\mathbb{E}_\alpha\left[\mathrm{Q}\left(\frac{-\alpha}{\sigma}\right)\right] + \frac{1}{2}\mathbb{E}_\alpha\left[\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right] = \frac{1}{2}. && (79)
\end{aligned}$$

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

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

$$\boldsymbol{\theta}_1 = \frac{1}{m}\sum_{i=1}^m \hat{Y}_i' \mathbf{X}_i' = \frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}_i')\mathbf{X}_i' = \frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')\mathbf{X}_i'. \tag{80}$$

(a) 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:

$$\begin{aligned}
\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} &= \mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' | \xi_0, \boldsymbol{\mu}^\perp] && (81)\\
&= \mathbb{E}_{Y_j'}\big[\,\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j']\,\big] && (82)\\
&= \frac{1}{2}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = -1] + \frac{1}{2}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = 1]. && (83)
\end{aligned}$$

Different from (68), here we decompose the Gaussian random vector $\mathbf{g}_j \sim \mathcal{N}(0, \mathbf{I}_d)$ in another way

$$\mathbf{g}_j = \tilde{g}_j \bar{\boldsymbol{\theta}}_0 + \tilde{\mathbf{g}}_j^\perp, \tag{84}$$

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 $\bar{\boldsymbol{\theta}}_0\mathbf{X}_j'$ as

$$\begin{aligned}
\mathbf{X}_j' &= Y_j'\boldsymbol{\mu} + \sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0 + \sigma\tilde{\mathbf{g}}_j^\perp, \text{ and} && (85)\\
\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j' &= Y_j'\alpha + \sigma\tilde{g}_j. && (86)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j' \mid \xi_0, \boldsymbol{\mu}^\perp, Y_j' = -1]\\
&= \mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)(-\boldsymbol{\mu} + \sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0 + \sigma\tilde{\mathbf{g}}^\perp) \mid \xi_0, \boldsymbol{\mu}^\perp] && (87)\\
&= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{g}_j|\xi_0, \boldsymbol{\mu}^\perp]\bar{\boldsymbol{\theta}}_0\\
&\quad + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{\mathbf{g}}^\perp|\xi_0, \boldsymbol{\mu}^\perp] && (88)\\
&= -\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)|\xi_0, \boldsymbol{\mu}^\perp]\boldsymbol{\mu} + \sigma\mathbb{E}[\mathrm{sgn}(-\alpha + \sigma\tilde{g}_j)\tilde{g}_j|\xi_0, \boldsymbol{\mu}^\perp]\bar{\boldsymbol{\theta}}_0, && (89)
\end{aligned}$$

where (89) 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 (89), recall $\tilde{g}_j \sim \mathcal{N}(0,1)$ and we have

$$-\mathbb{E}[\mathrm{sgn}(-\alpha+\sigma\tilde{g}_j)|\xi_0,\boldsymbol{\mu}^{\perp}]\boldsymbol{\mu} = \left(1-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right)\boldsymbol{\mu}. \tag{90}$$

For the second term in (89), we have

$$\begin{aligned}
&\mathbb{E}[\mathrm{sgn}(-\alpha+\sigma\tilde{g}_j)\tilde{g}_j|\xi_0,\boldsymbol{\mu}^{\perp}]\bar{\boldsymbol{\theta}}_0 \\
&= \left(-\int_{-\infty}^{\frac{\alpha}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-g^2}{2}} g \,\mathrm{d}g + \int_{\frac{\alpha}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-g^2}{2}} g \,\mathrm{d}g\right)\bar{\boldsymbol{\theta}}_0 \qquad (91)\\
&= \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\bar{\boldsymbol{\theta}}_0. \qquad (92)
\end{aligned}$$

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

$$\mathbb{E}\big[\,\mathrm{sgn}\,(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_j')\mathbf{X}_j' \mid \xi_0,\boldsymbol{\mu}^{\perp},Y_j'=-1\big] = \left(1-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right)\boldsymbol{\mu} + \frac{2\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\bar{\boldsymbol{\theta}}_0, \tag{93}$$

and similarly,

$$\mathbb{E}\big[\,\mathrm{sgn}\,(\bar{\boldsymbol{\theta}}_0^{\top}\mathbf{X}_j')\mathbf{X}_j' \mid \xi_0,\boldsymbol{\mu}^{\perp},Y_j'=1\big] = \left(2\mathrm{Q}\left(\frac{-\alpha}{\sigma}\right)-1\right)\boldsymbol{\mu} + \frac{2\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\bar{\boldsymbol{\theta}}_0. \tag{94}$$

Thus, recall $\bar{\boldsymbol{\theta}}_0 = \alpha\boldsymbol{\mu} + \beta\boldsymbol{v}$ and $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}}$ is given by

$$\begin{aligned}
\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}} &= \mathbb{E}[\mathrm{sgn}(\boldsymbol{\theta}_0^{\top}\mathbf{X}_j')\mathbf{X}_j'|\xi_0,\boldsymbol{\mu}^{\perp}] \\
&= \left(1-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)\right)\boldsymbol{\mu} + \frac{2\sigma}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\bar{\boldsymbol{\theta}}_0 \qquad (95)\\
&= \left(1-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\boldsymbol{\mu} + \frac{2\sigma\beta}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\boldsymbol{v}. \qquad (96)
\end{aligned}$$

The l_∞ norm between $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}}$ and $\boldsymbol{\mu}$ can be upper bounded by

$$\begin{aligned}
&\|\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^{\perp}} - \boldsymbol{\mu}\|_{\infty} \\
&\le \sqrt{\left(-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)^2 + \frac{2\sigma^2\beta^2}{\pi}\exp\left(-\frac{2\alpha^2}{2\sigma^2}\right)} \qquad (97)\\
&< \sqrt{\left(2\Phi\left(\frac{1}{\sigma}\right)+\frac{2\sigma}{\sqrt{2\pi}}\right)^2 + \frac{2\sigma^2}{\pi}} =: \tilde{c}_1, \qquad (98)
\end{aligned}$$

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

(b) Next, we need to calculate the probability that $l(\boldsymbol{\theta}_1,(\mathbf{X},Y)) \sim \mathsf{subG}((c_2-c_1/2))$ 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 $k\in[1:d]$, 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 (85) and $\bar{\boldsymbol{\theta}}_0\mathbf{X}_i' = Y_i'\alpha + \sigma\tilde{g}_i$ in (86). 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

$$\tilde{\mathbf{g}}_i^{\perp} = \tilde{g}_{i,2}^{\perp}\mathbf{v}_2 + \ldots + \tilde{g}_{i,d}^{\perp}\mathbf{v}_d, \tag{99}$$

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 $s_1 > 0$,

$$\begin{aligned}
&\mathbb{E}_{V_{i,1}}[e^{s_1 V_{i,1}}] \\
&= \mathrm{Q}\Big(\frac{-\alpha}{\sigma}\Big)\mathbb{E}_{\tilde{g}_i}\Big[e^{s_1(\mu_{0,1}-\mu_{1,1}+\sigma\tilde{g}_i)}\Big|\tilde{g}_i > \frac{-\alpha}{\sigma}\Big] + \mathrm{Q}\Big(\frac{\alpha}{\sigma}\Big)\mathbb{E}_{\tilde{g}_i}\Big[e^{s_1(-\mu_{0,1}-\mu_{1,1}+\sigma\tilde{g}_i)}\Big|\tilde{g}_i > \frac{\alpha}{\sigma}\Big] \quad (100)\\
&= e^{s_1(\mu_{0,1}-\mu_{1,1})}e^{\frac{\sigma^2 s_1^2}{2}}\Phi\Big(\frac{\alpha}{\sigma}+\sigma s_1\Big) + e^{s_1(-\mu_{0,1}-\mu_{1,1})}e^{\frac{\sigma^2 s_1^2}{2}}\Phi\Big(\frac{-\alpha}{\sigma}+\sigma s_1\Big). \quad (101)
\end{aligned}$$

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

$$\begin{aligned}
\tilde{R}_1(s_1) &:= \frac{\mathrm{d}^2 \log \mathbb{E}_{V_{i,1}}[e^{s_1 V_{i,1}}]}{\mathrm{d}s_1^2} \quad (102)\\
&\le \sigma^2 + \frac{\text{const.}}{\big(\Phi(\frac{\alpha}{\sigma}+\sigma s_1)e^{s_k\mu_{0,k}} + \Phi(\frac{-\alpha}{\sigma}+\sigma s_1)e^{-s_k\mu_{0,k}}\big)^2} < \infty. \quad (103)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}] &= \mathbb{E}_{\sigma\tilde{g}_{i,k}^\perp, Y_i'}\Big[e^{s_k(Y_i'\mu_{0,k}-\mu_{1,k}+\sigma\tilde{g}_{i,k}^\perp)}\Big] \quad (104)\\
&= Q\Big(\frac{-\alpha}{\sigma}\Big)e^{s_k(\mu_{0,k}-\mu_{1,k})}e^{\frac{\sigma^2 s_k^2}{2}} + Q\Big(\frac{\alpha}{\sigma}\Big)e^{s_k(-\mu_{0,k}-\mu_{1,k})}e^{\frac{\sigma^2 s_k^2}{2}}, \quad (105)
\end{aligned}$$

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

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

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

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

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 $\varepsilon > 0$,

$$\sup_{s_k>0}\Big(s_k\varepsilon - \log \mathbb{E}_{V_{i,k}}[e^{s_k V_{i,k}}]\Big) \ge \sup_{s_k>0}\Big(s_k\varepsilon - \frac{\tilde{R}_k(\xi_{\mathrm{L},k})s_k^2}{2}\Big) = \frac{\varepsilon^2}{2\tilde{R}_k(\xi_{\mathrm{L},k})}. \quad (108)$$

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 $\{\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')\mathbf{X}_i'\}_{i=1}^m$ are i.i.d. random variables conditioned on $(\xi_0, \boldsymbol{\mu}^\perp)$, by applying Chernoff-Cramér inequality, we have for all $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\Big(\|\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\|_\infty > \varepsilon\Big)\\
&= \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\Big(\max_{k\in[1:d]}|\theta_{1,k}-\mu_{1,k}| > \varepsilon\Big) \quad (109)\\
&\le \sum_{k=1}^d \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\Big(|\theta_{1,k}-\mu_{1,k}| > \varepsilon\Big) \quad (110)\\
&= \sum_{k=1}^d \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\Big(\Big|\frac{1}{m}\sum_{i=1}^m V_{i,k}\Big| > \varepsilon\Big) \quad (111)\\
&\le \sum_{k=1}^d 2\exp\Big(-m\sup_{s>0}\Big(s\varepsilon - \log \mathbb{E}_{V_{i,k}}[e^{sV_{i,k}}]\Big)\Big) \quad (112)\\
&\le 2d\exp\Big(-\frac{m\varepsilon^2}{2\tilde{R}^*}\Big) \quad (113)\\
&=: \delta_{m,\varepsilon,d}, \quad (114)
\end{aligned}$$

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 (98)). We have

$$\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\big(\boldsymbol{\theta}_1 \in \Theta_{\boldsymbol{\mu},c}\big) \geq \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\big(\|\boldsymbol{\theta}_1 - \boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}\|_\infty \leq c - \tilde{c}_1\big) \geq 1 - \delta_{m,c-\tilde{c}_1,d}. \tag{115}$$

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 $l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) \sim \mathsf{subG}((c_2 - c_1)/2)$ under $(\mathbf{X}, Y) \sim P_{\mathbf{X},Y}$ is given as follows

$$\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_1,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_1) \leq \frac{\lambda^2(c_2-c_1)^2}{8}\right) \tag{116}$$

$$\geq \mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}(\tilde{\boldsymbol{\theta}}_1 \in \Theta_{\boldsymbol{\mu},c})\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\left(\Lambda_{l(\tilde{\boldsymbol{\theta}}_1,(\tilde{\mathbf{X}},\tilde{Y}))}(\lambda, \tilde{\boldsymbol{\theta}}_1) \leq \frac{\lambda^2(c_2-c_1)^2}{8}\Big|\tilde{\boldsymbol{\theta}}_1 \in \Theta_{\boldsymbol{\mu},c}\right) \tag{117}$$

$$= (1 - \delta_{m,c,d})(1 - \delta_{r,d}). \tag{118}$$

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:

$$|\mathrm{gen}_1| = |\mathbb{E}[L_{P_{\mathbf{Z}}}(\boldsymbol{\theta}_1) - L_{\hat{S}_{\mathrm{u},1}}(\boldsymbol{\theta}_1)]| \tag{119}$$

$$= \left|\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1, (\mathbf{X}'_i, \hat{Y}'_i))|\xi_0, \boldsymbol{\mu}^\perp\right]\right]\right| \tag{120}$$

$$\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\frac{(c_2-c_1)^2}{2}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1, (\mathbf{X}'_i, \hat{Y}'_i)) + D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_i,\hat{Y}'_i}\|P_{\mathbf{X},Y})\Big)}\right], \tag{121}$$

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 the closed form expressions of the mutual information and KL-divergence in (121). 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

$$h_U(X) := h(P_{X|U}). \tag{122}$$

Conditioned on a certain pair of $(\xi_0, \boldsymbol{\mu}^\perp)$, the mutual information between $\boldsymbol{\theta}_1$ and $(\mathbf{X}'_i, \hat{Y}'_i)$ is given by

$$I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_j, \hat{Y}'_j)$$

$$= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) - h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{j=1}^m \hat{Y}'_i\mathbf{X}'_i\Big|\mathbf{X}'_j, \hat{Y}'_j\right) \tag{123}$$

$$= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) - h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i\in[m],i\neq j} \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) \tag{124}$$

$$= h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^m \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right)$$
$$- h_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m-1}\sum_{i\in[m],i\neq j} \mathrm{sgn}(\boldsymbol{\theta}_0^\top \mathbf{X}'_i)\mathbf{X}'_i\right) - d\log\frac{m-1}{m}. \tag{125}$$

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 $m > m_0$,

$$\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; \mathbf{X}'_i, \hat{Y}'_i) > \epsilon) \leq \delta. \tag{126}$$

- **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. (79)) regardless of the values of $(\xi_0, \boldsymbol{\mu}^\perp)$, the *disintegrated conditional KL-divergence* can be rewritten as

$$D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j,\hat{Y}'_j}\|P_{\mathbf{X},Y})$$
$$= P_{\hat{Y}'_j}(-1)D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=-1}\|P_{\mathbf{X}|Y=-1}) + P_{\hat{Y}'_j}(1)D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=1}\|P_{\mathbf{X}|Y=1}). \tag{127}$$

Recall the decomposition of a Gaussian vector $\tilde{\mathbf{g}}_j \sim \mathcal{N}(0, \mathbf{I}_d)$ in (84). Note that $\mathrm{rank}(\mathsf{Cov}(\tilde{\mathbf{g}}_j^{\perp})) = \mathrm{rank}(\mathbf{I}_d - \bar{\boldsymbol{\theta}}\bar{\boldsymbol{\theta}}^\top) = d-1$.

For any pair of labelled data sample $(\mathbf{X}, Y)$, from (85), 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

$$
\begin{aligned}
P_{\mathbf{X},Y}(\mathbf{x}, 1) &= P_Y(1) p_{\boldsymbol{\mu}}(\mathbf{x}|1) \\
&= \frac{P_Y(1)}{\sqrt{(2\pi)^d \sigma^d}} \exp\bigg(-\frac{1}{2\sigma^2}(\mathbf{x} - y\boldsymbol{\mu})^\top (\mathbf{x} - y\boldsymbol{\mu}) \bigg) && (128) \\
&= \frac{P_Y(1)}{\sqrt{(2\pi)^d \sigma^d}} \exp\bigg(-\frac{1}{2\sigma^2}(\sigma u\bar{\boldsymbol{\theta}} + \sigma\mathbf{u}^{\perp})^\top (\sigma u\bar{\boldsymbol{\theta}} + \sigma\mathbf{u}^{\perp}) \bigg) && (129) \\
&= \frac{P_Y(y)}{\sqrt{(2\pi)^d \sigma^d}} \exp\bigg(-\frac{u^2}{2} \bigg) \exp\bigg(-\frac{(\mathbf{u}^{\perp})^\top \mathbf{u}^{\perp}}{2} \bigg) && (130) \\
&= P_Y(1) p_{\tilde{g}}(u) p_{\tilde{\mathbf{g}}^{\perp}}(\mathbf{u}^{\perp}). && (131)
\end{aligned}
$$

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

$$
P_{\mathbf{X},Y}(\mathbf{x}, -1) = P_Y(-1) p_{\boldsymbol{\mu}}(\mathbf{x}|-1) = P_Y(-1) p_{\tilde{g}}(u) p_{\tilde{\mathbf{g}}^{\perp}}(\mathbf{u}^{\perp}). \tag{132}
$$

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

$$
P_{Y'_j|\hat{Y}'_j} = \frac{P_{\hat{Y}'_j|Y'_j} P_{Y'_j}}{P_{\hat{Y}'_j}} = P_{\hat{Y}'_j|Y'_j}, \tag{133}
$$

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 $\hat{Y}'_j = \mathrm{sgn}(Y'_j\alpha + \sigma\tilde{g}_j)$ (cf. (86)), we have

$$
P_{\hat{Y}'_j|Y'_j}(-1|-1) = \Pr(Y'_j\alpha + \sigma\tilde{g}_j < 0 | Y'_j = -1) = \mathrm{Q}\bigg(\frac{-\alpha}{\sigma}\bigg), \tag{134}
$$

and similarly,

$$
P_{\hat{Y}'_j|Y'_j}(1|-1) = \mathrm{Q}\bigg(\frac{\alpha}{\sigma}\bigg),\ P_{\hat{Y}'_j|Y'_j}(-1|1) = \mathrm{Q}\bigg(\frac{\alpha}{\sigma}\bigg),\ P_{\hat{Y}'_j|Y'_j}(1|1) = \mathrm{Q}\bigg(\frac{-\alpha}{\sigma}\bigg). \tag{135}
$$

Thus, we conclude that

$$
P_{Y'_j|\hat{Y}'_j}(y'_j|\hat{y}'_j) = \begin{cases} \mathrm{Q}(\frac{-\alpha}{\sigma}) & y'_j = \hat{y}'_j \\ \mathrm{Q}(\frac{\alpha}{\sigma}) & y'_j \neq \hat{y}'_j. \end{cases} \tag{136}
$$

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 (85) and (86). 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

$$
p_{\tilde{g}_j|\hat{Y}'_j, Y'_j}(u|-1,-1) = p_{\tilde{g}_j|\tilde{g}_j \leq \alpha/\sigma}(u) = \frac{\mathbb{1}\{u \leq \alpha/\sigma\} p_{\tilde{g}_j}(u)}{\Phi(\alpha/\sigma)}, \quad \forall u \in \mathbb{R}. \tag{137}
$$

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

$$
p_{\tilde{g}_j|\hat{Y}'_j, Y'_j}(u|-1,1) = p_{\tilde{g}_j|\tilde{g}_j \leq -\alpha/\sigma}(u) = \frac{\mathbb{1}\{u \leq -\alpha/\sigma\} f_{\tilde{g}_j}(u)}{\Phi(-\alpha/\sigma)}, \tag{138}
$$

$$
p_{\tilde{g}_j|\hat{Y}'_j, Y'_j}(u|1,-1) = p_{\tilde{g}_j|\tilde{g}_j > \alpha/\sigma}(u) = \frac{\mathbb{1}\{z > \alpha/\sigma\} f_{\tilde{g}_j}(u)}{\mathrm{Q}(\alpha/\sigma)}, \tag{139}
$$

$$
p_{\tilde{g}_j|\hat{Y}'_j, Y'_j}(u|1,1) = p_{\tilde{g}_j|\tilde{g}_j > -\alpha/\sigma}(u) = \frac{\mathbb{1}\{u > -\alpha/\sigma\} p_{\tilde{g}_j}(u)}{\mathrm{Q}(-\alpha/\sigma)}. \tag{140}
$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|1,1) = P_{\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,1) \tag{141}$$

$$= P_{\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,1) \tag{142}$$

$$= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|1,1)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp), \tag{143}$$

where (143) 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

$$\frac{2\boldsymbol{\mu}}{\sigma} = \frac{2\alpha\bar{\boldsymbol{\theta}}_0 + 2\beta^2\boldsymbol{\mu} - 2\alpha\beta\boldsymbol{\upsilon}}{\sigma} = \frac{2\alpha}{\sigma}\bar{\boldsymbol{\theta}}_0 + \bar{\boldsymbol{\theta}}_0^\perp, \tag{144}$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|1,-1) = P_{-\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|1,-1) \tag{145}$$

$$= P_{\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}\Big(\sigma\Big(\frac{2\boldsymbol{\mu}}{\sigma} + u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp\Big)\Big|1,-1\Big) \tag{146}$$

$$= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}\Big(u + \frac{2\alpha}{\sigma}\Big|1,-1\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp + \bar{\boldsymbol{\theta}}_0^\perp). \tag{147}$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|-1,1) = P_{\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(-\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|-1,1) \tag{148}$$

$$= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}\Big(u - \frac{2\alpha}{\sigma}\Big| - 1,1\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp - \bar{\boldsymbol{\theta}}_0^\perp); \tag{149}$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j',Y_j'}(\mathbf{x}|-1,-1) = P_{-\boldsymbol{\mu}+\sigma\tilde{g}_j\bar{\boldsymbol{\theta}}_0+\sigma\tilde{\mathbf{g}}_j^\perp|\hat{Y}_j',Y_j'}(-\boldsymbol{\mu} + \sigma(u\bar{\boldsymbol{\theta}}_0 + \mathbf{u}^\perp)|-1,-1) \tag{150}$$

$$= p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|-1,-1)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp). \tag{151}$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j'=-1}(\mathbf{x}) = \sum_{y\in\{-1,+1\}} P_{\mathbf{X}_j'|\hat{Y}_j'=-1,Y_j'=y}(\mathbf{x})P_{Y_j'|\hat{Y}_j'=-1}(y) \tag{152}$$

$$\begin{aligned} &= P_{Y_j'|\hat{Y}_j'=-1}(1)p_{\tilde{g}_j|\hat{Y}_j',Y_j'}\Big(u - \frac{2\alpha}{\sigma}\Big| - 1,1\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp - \bar{\boldsymbol{\theta}}^\perp) \\ &\quad + P_{Y_j'|\hat{Y}_j'=-1}(-1)p_{\tilde{g}_j|\hat{Y}_j',Y_j'}(u|-1,-1)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp) \end{aligned} \tag{153}$$

$$= \mathbb{1}\Big\{u \le \frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}\Big(u - \frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp - \bar{\boldsymbol{\theta}}^\perp) + \mathbb{1}\Big\{u \le \frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp); \tag{154}$$

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

$$P_{\mathbf{X}_j'|\hat{Y}_j'=1}(\mathbf{x}) = \sum_{y\in\{-1,+1\}} P_{\mathbf{X}_j'|\hat{Y}_j'=1,Y_j'=y}(\mathbf{x})P_{Y_j'|\hat{Y}_j'=1}(y) \tag{155}$$

$$= \mathbb{1}\Big\{u > -\frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}\Big(u + \frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp + \bar{\boldsymbol{\theta}}^\perp) + \mathbb{1}\Big\{u > -\frac{\alpha}{\sigma}\Big\}p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp). \tag{156}$$

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 (21) 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

$$
\begin{aligned}
&D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=-1}\|P_{\mathbf{X}|Y=-1}) \\
&= \int_{\mathcal{U}_0^\perp}\int_{-\infty}^{\frac{\alpha}{\sigma}} \left(p_{\tilde{g}_j}\Big(u-\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp)+p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)\right) \\
&\qquad\qquad \times\log\left(1+\frac{p_{\tilde{g}_j}\Big(u-\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp-\bar{\boldsymbol{\theta}}_0^\perp)}{p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)}\right)\mathrm{d}u\,\mathrm{d}\mathbf{u}^\perp \qquad (157)\\
&= G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp) \qquad (158)
\end{aligned}
$$

and

$$
\begin{aligned}
&D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j|\hat{Y}'_j=1}\|P_{\mathbf{X}|Y=1}) \\
&= \int_{\mathcal{U}_0^\perp}\int_{-\frac{\alpha}{\sigma}}^{+\infty} \left(p_{\tilde{g}_j}\Big(u+\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp+\bar{\boldsymbol{\theta}}_0^\perp)+p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)\right) \\
&\qquad\qquad \times\log\left(1+\frac{p_{\tilde{g}_j}\Big(u+\frac{2\alpha}{\sigma}\Big)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp+\bar{\boldsymbol{\theta}}_0^\perp)}{p_{\tilde{g}_j}(u)p_{\tilde{\mathbf{g}}_j^\perp}(\mathbf{u}^\perp)}\right)\mathrm{d}u\,\mathrm{d}\mathbf{u}^\perp \qquad (159)\\
&= G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp), \qquad (160)
\end{aligned}
$$

where (160) follows from since $p_{\tilde{g}_j}$ and $p_{\tilde{\mathbf{g}}_j^\perp}$ are zero-mean Gaussian distributions. Then from (127), we have

$$
D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}'_j,\hat{Y}'_j}\|P_{\mathbf{X},Y}) = G_\sigma(\alpha,\xi_0,\boldsymbol{\mu}^\perp). \qquad (161)
$$

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 $1-\delta$,

$$
|\mathrm{gen}_1| \le \sqrt{\frac{(c_2-c_1)^2}{2}}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{G_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp),\xi_0,\boldsymbol{\mu}^\perp)+\epsilon}\,\right]. \qquad (162)
$$

4. **Pseudo-label using $\boldsymbol{\theta}_1$:** Let $\bar{\boldsymbol{\theta}}_1 := \boldsymbol{\theta}_1/\|\boldsymbol{\theta}_1\|_2$. For any $i\in[m+1:2m]$, the pseudo-labels are given by

$$
\hat{Y}'_i = \mathrm{sgn}(\boldsymbol{\theta}_1^\top\mathbf{X}'_i) = \mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top\mathbf{X}'_i). \qquad (163)
$$

It can be seen that the pseudo-labels $\{\hat{Y}'_i\}_{i=m+1}^{2m}$ 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 $\hat{S}_{\mathrm{u},2}=\{(\mathbf{X}'_i,\hat{Y}'_i)\}_{i=m+1}^{2m}$.

For any fixed $\bar{\boldsymbol{\theta}}_1\in\Theta$, we can decompose it as $\bar{\boldsymbol{\theta}}_1=\alpha'_1\boldsymbol{\mu}+\beta'_1\boldsymbol{v}$, where $\alpha'_1\in[-1,1]$ and $\beta'_1=\sqrt{1-(\alpha'_1)^2}$. Recall the decomposition of $\mathbf{X}'_i$ and $\bar{\boldsymbol{\theta}}_0^\top\mathbf{X}'_i$ in (69) and (73). Similarly, we have

$$
\bar{\boldsymbol{\theta}}_1^\top\mathbf{X}'_i =: Y'_i\alpha'_1+\sigma h_i^1, \qquad (164)
$$

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

$$
\begin{aligned}
P_{\hat{Y}'_i|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(1) &= P_{\hat{Y}'_i|\boldsymbol{\theta}_1}(1) = \mathbb{P}_{\boldsymbol{\theta}_1}\big(\bar{\boldsymbol{\theta}}_1^\top\mathbf{X}'_i>0\big) \qquad (165)\\
&= \frac{1}{2}\mathbb{P}_{\boldsymbol{\theta}_1}\big(\alpha'_1+\sigma h_i^1>0\big)+\frac{1}{2}\mathbb{P}_{\boldsymbol{\theta}_1}\big(\alpha'_1+\sigma h_i^1\le 0\big)=\frac{1}{2}, \qquad (166)
\end{aligned}
$$

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 $\boldsymbol{\theta}_1$.

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

$$\boldsymbol{\theta}_2 = \frac{1}{m}\sum_{i=m+1}^{2m} \hat{Y}_i'\mathbf{X}_i' = \frac{1}{m}\sum_{i=m+1}^{2m} \mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i')\mathbf{X}_i', \tag{167}$$

where $\{\mathrm{sgn}(\bar{\boldsymbol{\theta}}_1^\top \mathbf{X}_i')\mathbf{X}_i'\}_{i=m+1}^{2m}$ are conditionally i.i.d. random variables given $\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp$.

Given any $(\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp)$, for any $j \in [m+1 : 2m]$, 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 (114), for any $\varepsilon > 0$, we have

$$\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(\|\boldsymbol{\theta}_2 - \boldsymbol{\mu}_2^{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\|_\infty > \varepsilon\big) \le \delta_{m,\varepsilon,d}. \tag{168}$$

From (98), 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 $c \in (\tilde{c}_1, \infty)$,

$$\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2 \in \Theta_{\boldsymbol{\mu},c}) \ge 1 - \delta_{m,c-\tilde{c}_1,d}. \tag{169}$$

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

$$|\mathrm{gen}_2| = |\mathbb{E}[L_{P_{\mathbf{Z}}}(\boldsymbol{\theta}_2) - L_{\hat{S}_{\mathrm{u},2}}(\boldsymbol{\theta}_2)]| \tag{170}$$

$$= \left|\frac{1}{m}\sum_{i=m+1}^{2m} \mathbb{E}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_2,(\mathbf{X},Y)) - l(\boldsymbol{\theta}_2,(\mathbf{X}_i',\hat{Y}_i'))|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp\right]\right]\right| \tag{171}$$

$$\le \sqrt{\frac{(c_2-c_1)^2}{2}} \times \frac{1}{m}\sum_{i=m+1}^{2m} \mathbb{E}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{I_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2;(\mathbf{X}_i',\hat{Y}_i')) + D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}_i',\hat{Y}_i'}\|P_{\mathbf{X},Y})}\right], \tag{172}$$

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

$$\mathbb{P}_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(I_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_2;(\mathbf{X}_i',\hat{Y}_i')) > \epsilon) \le \delta. \tag{173}$$

Recall (166) that $P_{\hat{Y}_i'|\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp} \sim \mathrm{unif}(\{-1,+1\})$. For any fixed $(\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp)$, let $\bar{\boldsymbol{\theta}}_1$ be decomposed as $\bar{\boldsymbol{\theta}}_1 = \alpha_1'(\xi_0,\boldsymbol{\mu}^\perp)\boldsymbol{\mu} + \beta_1'(\xi_0,\boldsymbol{\mu}^\perp)\boldsymbol{\upsilon}$, where $\alpha_1'(\xi_0,\boldsymbol{\mu}^\perp) \in [-1,1]$ and $\beta_1'(\xi_0,\boldsymbol{\mu}^\perp) = \sqrt{1-(\alpha_1'(\xi_0,\boldsymbol{\mu}^\perp))^2}$.

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

$$D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i',\hat{Y}_i'}\|P_{\mathbf{X},Y}\big)$$

$$= \frac{1}{2}D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i'|\hat{Y}_i'=-1}\|P_{\mathbf{X}|Y=-1}\big) + \frac{1}{2}D_{\boldsymbol{\theta}_1,\xi_0,\boldsymbol{\mu}^\perp}\big(P_{\mathbf{X}_i'|\hat{Y}_i'=1}\|P_{\mathbf{X}|Y=1}\big) \tag{174}$$

$$= G_\sigma\big(\alpha_1'(\xi_0,\boldsymbol{\mu}^\perp),\xi_0,\boldsymbol{\mu}^\perp\big), \tag{175}$$

Given any pair of $(\xi_0,\boldsymbol{\mu}^\perp)$, recall the decomposition of $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ in (96). Then the correlation between $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}$ and $\boldsymbol{\mu}$ is given by

$$\rho(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp},\boldsymbol{\mu}) = \frac{1-2\mathrm{Q}\big(\frac{\alpha}{\sigma}\big)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp(-\frac{\alpha^2}{2\sigma^2})}{\sqrt{\big(1-2\mathrm{Q}\big(\frac{\alpha}{\sigma}\big)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp(-\frac{\alpha^2}{2\sigma^2})\big)^2+\frac{2\sigma^2(1-\alpha^2)}{\pi}\exp(-\frac{\alpha^2}{\sigma^2})}} \tag{176}$$

$$= F_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp)). \tag{177}$$

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 $m > m_2$,

$$\mathbb{P}_{\boldsymbol{\theta}_1, \xi_0, \boldsymbol{\mu}^\perp}\left(\left|G_\sigma\left(\alpha_1'(\xi_0, \boldsymbol{\mu}^\perp), \xi_0, \boldsymbol{\mu}^\perp\right) - G_\sigma\left(F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp)), \xi_0, \boldsymbol{\mu}^\perp\right)\right| > \epsilon\right) \leq \delta. \tag{178}$$

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:

$$|\text{gen}_2| \leq \sqrt{\frac{(c_2 - c_1)^2}{2}} \mathbb{E}_{\xi_0, \boldsymbol{\mu}^\perp}\left[\sqrt{G_\sigma\left(F_\sigma(\alpha(\xi_0, \boldsymbol{\mu}^\perp)), \xi_0, \boldsymbol{\mu}^\perp\right) + \epsilon}\right]. \tag{179}$$

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

Remark 2 (Extra remarks about Theorem 2)**.** *In the other extreme case, when* $\alpha = \rho(\boldsymbol{\theta}_0, \boldsymbol{\mu}) = -1$ *and* $\bar{\boldsymbol{\theta}}_0 = -\boldsymbol{\mu}$*, the error probability* $\Pr(\hat{Y}_j' \neq Y_j') = 1 - \text{Q}(1/\sigma) > \frac{1}{2}$ *(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.*

D REUSING S_1 IN EACH ITERATION

If the labelled data S_1 are reused in each iteration and $w = \frac{n}{n+m}$ (cf. (5)), for each $t \in [1 : \tau]$, the learned model parameter is given by

$$\boldsymbol{\theta}_t = \frac{n}{n+m}\boldsymbol{\theta}_{t-1} + \frac{1}{n+m}\sum_{i=(t-1)m+1}^{tm} \hat{Y}_i' \mathbf{X}_i' \tag{180}$$

$$= \frac{n}{n+m}\boldsymbol{\theta}_{t-1} + \frac{1}{n+m}\sum_{i=(t-1)m+1}^{tm} \text{sgn}(\bar{\boldsymbol{\theta}}_{t-1}^\top \mathbf{X}_i')\mathbf{X}_i'. \tag{181}$$

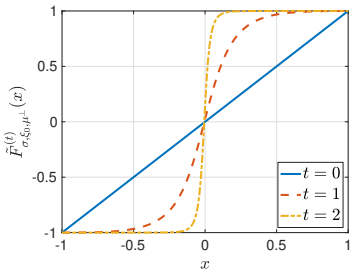

Figure 10: $\tilde{F}^{(t)}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(x)$ versus x under $t \in \{0,1,2\}$ when $\sigma = 0.5$, $\xi_0 = 0$, $\|\boldsymbol{\mu}^\perp\|_2 = 1$, $n = 10$, $m = 1000$.

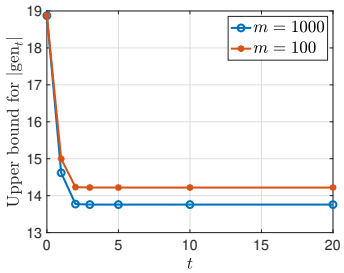

Figure 11: Upper bound for $|\text{gen}_t|$ versus t for $m = 100$ and $m = 1000$, when $n = 10$, $\sigma = 0.6$, $d = 2$, $\boldsymbol{\mu} = (1, 0)$.

Recall the definition of the function $\tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}$ in (27). 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 10, 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 $F^{(t)}_\sigma$ as t increases, which implies that the upper bound in (28) in Corollary 3 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 11, under the same setup as Figure 6(c), 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 $m = 1000$, the upper bound is almost the same as that one in Figure 6(c), which means that for large enough m/n, reusing the labelled data does not necessarily help to improve the generalization performance. Moreover, when $m = 100$, the upper bound is higher than that for $m = 1000$, which coincides with the intuition that increasing the number of unlabelled data helps to reduce the generalization error.

E PROOF OF COROLLARY 3

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

At $t = 1$, from (65) and (96), the expectation $\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} = \mathbb{E}[\boldsymbol{\theta}_1|\xi_0, \boldsymbol{\mu}^\perp]$ is rewritten as

$$\begin{aligned}
\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp} &= \frac{n}{n+m}\boldsymbol{\theta}_0 + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}[\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_j')\mathbf{X}_j'|\xi_0, \boldsymbol{\mu}^\perp] \\
&= \frac{n}{n+m}\left(\left(1+\frac{\sigma}{\sqrt{n}}\xi_0\right)\boldsymbol{\mu} + \frac{\sigma}{\sqrt{n}}\boldsymbol{\mu}^\perp\right) \\
&\quad + \frac{m}{n+m}\left(\left(1-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\boldsymbol{\mu} + \frac{2\sigma\beta}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\boldsymbol{v}\right) \qquad (182) \\
&= \left(1+\frac{\sqrt{n}\sigma\xi_0}{n+m} + \frac{m}{n+m}\left(-2\mathrm{Q}\left(\frac{\alpha}{\sigma}\right)+\frac{2\sigma\alpha}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\right)\boldsymbol{\mu} \\
&\quad + \left(\frac{\sqrt{n}\sigma\|\boldsymbol{\mu}^\perp\|_2}{n+m} + \frac{m}{n+m}\frac{2\sigma\beta}{\sqrt{2\pi}}\exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\right)\boldsymbol{v}. \qquad (183)
\end{aligned}$$

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

$$\rho(\boldsymbol{\mu}_1^{\xi_0,\boldsymbol{\mu}^\perp}, \boldsymbol{\mu}) = \tilde{F}_{\sigma,\xi_0,\boldsymbol{\mu}^\perp}(\alpha). \qquad (184)$$

Let $\boldsymbol{\theta}_1' = \frac{1}{m}\sum_{i=1}^{m}\mathrm{sgn}(\bar{\boldsymbol{\theta}}_0^\top \mathbf{X}_i')\mathbf{X}_i'$. For some $c \in (\tilde{c}_1, \infty)$, from (53) and (114),

$$\begin{aligned}
&\Pr\left(\|\boldsymbol{\theta}_1 - \boldsymbol{\mu}\|_\infty > c\right) \\
&\le \Pr\left(\frac{n}{n+m}\|\boldsymbol{\theta}_0 - \boldsymbol{\mu}\|_\infty + \frac{m}{n+m}\|\boldsymbol{\theta}_1' - \boldsymbol{\mu}\|_\infty > c\right) \qquad (185) \\
&\le \Pr\left(\|\boldsymbol{\theta}_0 - \boldsymbol{\mu}\|_\infty > c\right) + \Pr\left(\|\boldsymbol{\theta}_1' - \boldsymbol{\mu}\|_\infty > c\right) \qquad (186) \\
&\le \delta_{\sqrt{n}c,d} + \delta_{m,c-\tilde{c}_1,d} \qquad (187)
\end{aligned}$$

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

$$\begin{aligned}
&|\mathrm{gen}_1| \\
&= \left|\frac{1}{n+m}\sum_{i=1}^{n}\mathbb{E}\Big[l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1, (\mathbf{X}_i, Y_i))\Big]\right. \\
&\quad \left. + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\mathbb{E}\left[l(\boldsymbol{\theta}_1, (\mathbf{X}, Y)) - l(\boldsymbol{\theta}_1, (\mathbf{X}_i', \hat{Y}_i'))|\xi_0, \boldsymbol{\mu}^\perp\right]\right]\right| \qquad (188) \\
&\le \frac{1}{n+m}\sum_{i=1}^{n}\sqrt{\frac{(c_2-c_1)^2}{2}I(\boldsymbol{\theta}_1; (\mathbf{X}_i, Y_i))} \\
&\quad + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\frac{(c_2-c_1)^2}{2}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1; (\mathbf{X}_i', \hat{Y}_i')) + D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}_i',\hat{Y}_i'}\|P_{\mathbf{X},Y})\Big)}\right] \qquad (189)
\end{aligned}$$

$$\leq \frac{1}{n+m}\sum_{i=1}^{n}\sqrt{\frac{(c_2-c_1)^2}{2}I(\boldsymbol{\theta}_0;(\mathbf{X}_i,Y_i))} + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\frac{(c_2-c_1)^2}{2}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1;(\mathbf{X}_i',\hat{Y}_i')) + D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}_i',\hat{Y}_i'}\|P_{\mathbf{X},Y})\Big)}\right] \tag{190}$$

$$= \frac{1}{n+m}\sum_{i=1}^{n}\sqrt{\frac{(c_2-c_1)^2 d}{4}\log\frac{n}{n-1}} + \frac{1}{n+m}\sum_{i=1}^{m}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\frac{(c_2-c_1)^2}{2}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1;(\mathbf{X}_i',\hat{Y}_i')) + D_{\xi_0,\boldsymbol{\mu}^\perp}(P_{\mathbf{X}_i',\hat{Y}_i'}\|P_{\mathbf{X},Y})\Big)}\right], \tag{191}$$

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 $(\xi_0,\boldsymbol{\mu}^\perp)$, and (190) follows since $(\mathbf{X}_i,Y_i)-\boldsymbol{\theta}_0-\boldsymbol{\theta}_1$ forms a Markov chain.

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

$$I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1;(\mathbf{X}_i',\hat{Y}_i')) = I_{\xi_0,\boldsymbol{\mu}^\perp}\left(\frac{1}{m}\sum_{i=1}^{m}\operatorname{sgn}(\bar{\boldsymbol{\theta}}_{t-1}^\top\mathbf{X}_i')\mathbf{X}_i';(\mathbf{X}_i',\hat{Y}_i')\right), \tag{192}$$

from (126), for any $\epsilon>0$ and any $\delta\in(0,1)$, there exists $m_1'(\epsilon,\delta,d)\in\mathbb{N}$ such that for all $m>m_1'(\epsilon,\delta,d)$,

$$\mathbb{P}_{\xi_0,\boldsymbol{\mu}^\perp}\Big(I_{\xi_0,\boldsymbol{\mu}^\perp}(\boldsymbol{\theta}_1;\mathbf{X}_i',\hat{Y}_i')>\epsilon\Big)\leq 1-\delta. \tag{193}$$

Therefore, fix $d\in\mathbb{N}$, any $\epsilon>0$ and any $\delta\in(0,1)$, and there exists $n_0(d,\delta)\in\mathbb{N}$, $m_4(\epsilon,d,\delta)\in\mathbb{N}$, $c_0(d,\delta)\in(\tilde{c}_1,\infty)$, $r_0(d,\delta)\in\mathbb{R}_+$ such that for all $n>n_0, m>m_3, c>c_0, r>r_0$, $\delta_{\sqrt{n}c,d}<\frac{\delta}{6}$, $\delta_{m,c-\tilde{c}_1,d}<\frac{\delta}{6}$, $\delta_{r,d}<\frac{\delta}{3}$, and with probability at least $1-\delta$, the absolute generalization error $|\mathrm{gen}_1|$ can be upper bounded as follows:

$$|\mathrm{gen}_1| \leq w\sqrt{\frac{(c_2-c_1)^2 d}{4}\log\frac{n}{n-1}} + (1-w)\sqrt{\frac{(c_2-c_1)^2}{2}}\mathbb{E}_{\xi_0,\boldsymbol{\mu}^\perp}\left[\sqrt{\Big(G_\sigma(\alpha(\xi_0,\boldsymbol{\mu}^\perp),\xi_0,\boldsymbol{\mu}^\perp)+\epsilon\Big)}\right]. \tag{194}$$

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

F Additional Experiments

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

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

In Table 1, we display the RGB means and variances of the test data in six classes taken from the CIFAR10 dataset. We observe that the RGB variances of each pair are almost 0 (and small compared to the RGB-mean l_2 distances), and thus, the RGB-mean l_2 distance is indicative of the difficulty of

the classification task. Indeed, a smaller RGB-mean l_2 distance implies a higher overlap of the two classes and consequently, greater difficulty in distinguishing them. Therefore, the "cat-dog" pair, which is more difficult to disambiguate compared to the "horse-ship" and "automobile-truck" pairs, is analogous to the bGMM with large variance (i.e. large overlap between the $\{\pm 1\}$ classes).

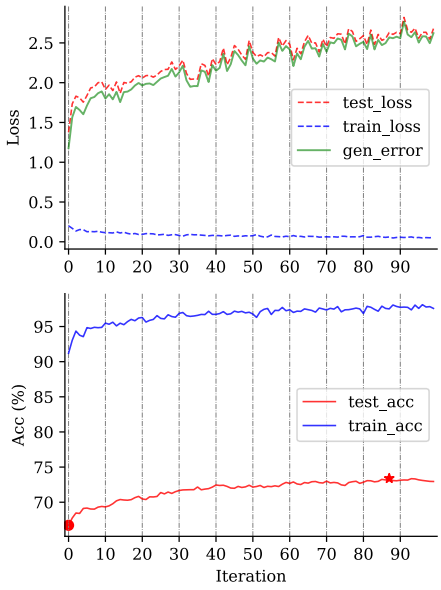

Figure 12: Binary classification on "cat" and "dog" from the CIFAR10 dataset.

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

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

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