Supplementary Materials - Proofs

A. Proof of Theorem 1

Theorem 1 (Generalization error of ERM)**.** *Assume that the cumulant generating function of the random variable $\ell(W,Z)-\mathbb{E}\{\ell(W,Z)\}$ is upper bounded by $\psi(\lambda)$ in the interval (b_-,b_+) under the product distribution $P_W\otimes\mu'$ for some $b_-<0$ and $b_+>0$. Then for any $\beta>0$, the expectation of the generalization error is upper bounded as*

$$\mathbb{E}_{WSS'}\{\text{gen}(W_{\mathsf{ERM}},S,S')\}\le\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\psi_-^{*-1}(I(W_{\mathsf{ERM}};Z_i))+\frac{(1-\alpha)}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\psi_-^{*-1}(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))$$

$$-\mathbb{E}_{WSS'}\{\text{gen}(W_{\mathsf{ERM}},S,S')\}\le\frac{\alpha}{\beta n}\sum_{i=1}^{\beta n}\psi_+^{*-1}(I(W_{\mathsf{ERM}};Z_i))+\frac{(1-\alpha)}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\psi_+^{*-1}(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))$$

where we define

$$\psi_-^{*-1}(x):=\inf_{\lambda\in[0,-b_-)}\frac{x+\psi(-\lambda)}{\lambda}$$
$$\psi_+^{*-1}(x):=\inf_{\lambda\in[0,b_+)}\frac{x+\psi(\lambda)}{\lambda}$$

Proof. We use W to denote W_{ERM} in the proof to simplify notations. First rewrite expectation of the generalization error of the ERM algorithm as

$$\begin{aligned}\mathbb{E}_{WSS'}\{L_{\mu'}(W)-\hat{L}_\alpha(W)\}&=\mathbb{E}_{WSS'}\{L_{\mu'}(W)-(1-\alpha)\hat{L}(W,S)-\alpha\hat{L}(W,S')\}\\&=\mathbb{E}_{WSS'}\{(1-\alpha)L_{\mu'}(W)+\alpha L_{\mu'}(W)-\frac{1}{n}\sum_{i=\beta n+1}^{n}\frac{1-\alpha}{1-\beta}\ell(W,Z_i)-\frac{1}{n}\sum_{i=1}^{\beta n}\frac{\alpha}{\beta}\ell(W,Z_i)\}\\&=\frac{1}{n}\mathbb{E}_{WSS'}\{\sum_{i=1}^{\beta n}\frac{\alpha}{\beta}(L_{\mu'}(W)-\ell(W,Z_i))+\sum_{i=\beta n+1}^{n}\frac{1-\alpha}{1-\beta}(L_{\mu'}(W)-\ell(W,Z_i))\}\\&=\frac{1}{n}\frac{\alpha}{\beta}\sum_{i=1}^{\beta n}\mathbb{E}_{WZ_i}\{(L_{\mu'}(W)-\ell(W,Z_i))\}+\frac{1}{n}\frac{1-\alpha}{1-\beta}\sum_{i=\beta n+1}^{n}\mathbb{E}_{WZ_i}\{L_{\mu'}(W)-\ell(W,Z_i))\}\end{aligned}$$

where the joint distribution $P_{WSS'}(w,s,s')$ on (W,S,S') is given by $P_S(s)P_{S'}(s')P_{W|SS'}(w|s,s')$

Recall that the variational representation of the KL divergence between two distributions P and Q defined over $\mathcal{X}$ is given as (see, e. g. [1])

$$D(P||Q)=\sup_f\{\mathbb{E}_P\{f(X)\}-\log\mathbb{E}_Q\{e^{f(x)}\}\}\tag{1}$$

where the supremum is taken over all measurable functions such that $\mathbb{E}_Q\{e^{f(x)}\}$ exists.

For each $i=1,\dots,n$, define the joint distribution $P_{WZ_i}(w,z_i)$ between an individual sample Z_i and the hypothesis W as induced by $P_{WSS'}(w,z^n)$ by marginalizing all samples other than z_i, and let P_W be the marginal distribution on W induced from $P_{WSS'}$.

We first show the first inequality in the Theorem. For any $i=1,\dots,\beta n$, let $P=P_{WZ_i}$, $Q=P_W\otimes\mu'$ in (1), and define $f:=\lambda\ell(W,Z_i)$ for some λ. The representation in (1) implies that

$$\mathbb{E}_{WZ_i}\{\lambda\ell(W,Z_i)\}\le D(P_{WZ_i}||P_W\otimes\mu')+\log\mathbb{E}\{e^{\lambda\ell(W,Z_i)}\}$$

where the expectation on the RHS is taken w. r. t. the distribution $P_W\otimes\mu'$. By the assumption that

$$\log\mathbb{E}\{e^{\lambda(\ell(W,Z_i)-\mathbb{E}\{\ell(W,Z_i)\})}\}\le\psi(\lambda)$$

for some $\lambda\in[b_-,0]$ under the distribution $P_W\otimes\mu'$, we have

$$\mathbb{E}_{WZ_i}\{\lambda(\ell(W,Z_i)-\mathbb{E}_{WZ_i\sim P_W\otimes\mu'}\{\ell(W,Z_i)\})\}\le D(P_{WZ_i}||P_W\otimes\mu')+\psi(\lambda)$$

which is equivalent to

$$\begin{aligned}\mathbb{E}_{WZ_i}\{L_{\mu'}(W)-\ell(W,Z_i)\} &\leq -\frac{1}{\lambda}\left(D(P_{WZ_i}||P_W\otimes\mu')+\psi(\lambda)\right)\right)\\ &= -\frac{1}{\lambda}\left(I(W;Z_i)+D(P_{Z_i}||\mu')+\psi(\lambda)\right)\right)\\ &= -\frac{1}{\lambda}\left(I(W;Z_i)+\psi(\lambda)\right)\right)\end{aligned}$$

as $P_{Z_i}=\mu'$ for $i=1,\ldots,\beta n$. The best upper bound is obtained by minimizing the RHS, giving

$$\mathbb{E}_{WZ_i}\{L_{\mu'}(W)-\ell(W,Z_i)\} \leq \min_{\lambda\in[0,-b_-]}\frac{1}{\lambda}(I(W;Z_i)+\psi(-\lambda)) = \psi^{*-1}(I(W;Z_i)) \tag{2}$$

For $i=\beta n+1,\ldots,n$, using the same argument we can show that

$$\mathbb{E}_{WZ_i}\{L_{\mu'}(W)-\ell(W,Z_i)\} \leq \psi^{*-1}(I(W;Z_i)+D(\mu||\mu')) \tag{3}$$

Summing over i using the upper bounds in (2) and (3), we obtain the first inequality in the theorem.

The second inequality is shown in the same way by using the fact that the cumulant generating function is upper bounded by $\psi(\lambda)$ in $[0,b_+)$. □

B. Proof of Theorem 2

Theorem 2 (Excess risk of ERM)**.** *Assume that for any $w\in\mathcal{W}$, the loss function $\ell(w,Z)$ is r^2-subgaussian under the distribution $P_W\otimes\mu'$. Then for any $\epsilon>0$ and $\delta>0$, there exists an n_0 (depending on δ and ϵ) such that for all $n\geq n_0$, the following inequality holds with probability at least $1-\delta$ (over the randomness of samples and the learning algorithm),*

$$\begin{aligned}L_{\mu'}(W_{\mathsf{ERM}})-L_{\mu'}(w^*) \leq & \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i)}+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))}\\ &+\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}}+(1-\alpha)d_{\mathcal{W}}(\mu,\mu')+\epsilon\end{aligned} \tag{4}$$

Furthermore in the case when $\beta=0$ (no samples from the distribution μ'), the inequality becomes

$$L_{\mu'}(W_{\mathsf{ERM}})-L_{\mu'}(w^*) \leq \sqrt{\frac{2r^2\log\frac{2}{\delta}}{n}}+|L_\mu(w^*)-L_{\mu'}(w^*)|+\frac{\sqrt{2r^2}}{n}\sum_{i=1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu||\mu'))}+\epsilon$$

The following lemma is used to prove the theorem.

Lemma 1. *Assume that for any $w\in\mathcal{W}$, the loss function $\ell(w,S)$ is r^2-subgaussian under the distribution μ or μ'. With probability at least $1-\delta$, it holds that*

$$\hat{L}_\alpha(w^*)-L_\alpha(w^*) \leq \sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}} \tag{5}$$

$$L_\mu(w^*)-L_{\mu'}(w^*) \leq \sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}}+(1-\alpha)D_{\mathcal{W}}(\mu,\mu') \tag{6}$$

Proof. Using the fact that $\ell(w,S)$ is r^2-subgaussian under μ or μ' for any $w\in\mathcal{W}$. Then let $X_1,\cdots,X_{\beta n}$ be random variables that take on values

$$\frac{\alpha}{\beta}\ell(Z_i,w)$$

for $i=1,\cdots,\beta n$. Similarly let $X_{\beta n+1},\cdots,X_{\beta n}$ be random variables that take on values

$$\frac{1-\alpha}{1-\beta}\ell(Z_i,w)$$

for $i=\beta n+1,\cdots,n$. Then

$$L(w^*,S,S')=\frac{\sum_{i=1}^{n}X_i}{n}$$

It then follows from the Hoeffding's inequality [1] that

$$
\begin{aligned}
Pr[|\hat{L}_\alpha(w^*) - L_\alpha(w^*)| \geq t] &\leq 2\exp\left(\frac{-2n^2t^2}{\sum_{i=1}^n range^2(X_i)}\right) \\
&= 2\exp\left(\frac{-2n^2t^2}{\left(\beta n\frac{\alpha^2}{\beta^2} + (1-\beta)n\frac{(1-\alpha)^2}{(1-\beta)^2}\right)4r^2}\right) \\
&= 2\exp\left(\frac{-nt^2}{2\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right)r^2}\right)
\end{aligned}
$$

which leads to the upper bound in (5). To upper bound $L_\mu(w^*) - L_{\mu'}(w^*)$, we write

$$
L_\mu(w^*) - L_{\mu'}(w^*) = L_\mu(w^*) - \hat{L}(w^*, S) + \hat{L}(w^*, S') - L_{\mu'}(w^*) + \hat{L}(w^*, S) - \hat{L}(w^*, S')
$$

We can upper bound the terms $L_\mu(w^*) - \hat{L}(w^*, S)$ and $\hat{L}(w^*, S') - L_{\mu'}(w^*)$ using the same argument as we prove (5). The last difference $\hat{L}(w^*, S) - \hat{L}(w^*, S')$ is upper bounded by the empirical distance between the samples S, S'. Overall, with probability larger than $1-\delta$, we have the bound

$$
L_\mu(w^*) - L_{\mu'}(w^*) \leq \sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}} + (1-\alpha)D_{\mathcal{W}}(\mu, \mu')
$$

□

C. Proof of Corollary 3

Corollary 3. *(Generalization error bound of ERM using ϕ_1-divergence) Assume that for any $w \in \mathcal{W}$, the loss function $\ell(w, Z)$ is L_∞-norm bounded by σ under the distribution $P_W \otimes \mu'$. Then for any $\epsilon > 0$ and $\delta > 0$, there exists an n_0 (depending on δ and ϵ) such that for all $n \geq n_0$, the following inequality holds with probability at least $1-\delta$ (over the randomness of samples and the learning algorithm) that*

$$
L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*) \leq \frac{\alpha\|\sigma\|_\infty}{\beta n}\sum_{i=1}^{\beta n} I_{\phi_1}(W_{\mathsf{ERM}}; z_i) + \frac{(1-\alpha)\|\sigma\|_\infty}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\left(I_{\phi_1}(W_{\mathsf{ERM}}; z_i) + 2TV(\mu||\mu')\right) + \epsilon \tag{7}
$$

where $I_{\phi_1}(W_{\mathsf{ERM}}; z_i) = D_{\phi_1}(P_{W_{\mathsf{ERM}}, z_i}||P_{W_{\mathsf{ERM}}} \otimes P_{z_i})$ is the ϕ-divergence between the distribution $P_{W_{\mathsf{ERM}}, z_i}$ and $P_{W_{\mathsf{ERM}}} \otimes P_{z_i}$ with $\phi_1(x) = |x-1|^1$ and $TV(\mu||\mu') = \frac{1}{2}D_{\phi_1}(\mu||\mu')$ denotes the total variation distance between the distribution μ and μ'.

Proof. Suppose $\ell(Z_i, W)$ is L_∞-norm upper bounded by σ, the L_∞-norm of a random variable is defined as

$$
\|X\|_\infty = \inf\{M : P(X > M) = 0\}
$$

then followed by [2, Theorem 3], we have

$$
|\mathbb{E}_P \ell(Z_i, W) - \mathbb{E}_Q \ell(Z_i, W)| \leq \|\sigma\|_\infty D_{\phi_1}(P||Q) \tag{8}
$$

where $D_{\phi_1}(P||Q) = \int |dP - dQ|$ is referred to as the ϕ_1-divergence with $\phi_1(x) = |x-1|$. If $Z_i \sim \mu'$, $D_{\phi_1}(P||Q) = D_{\phi_\alpha}(P_{W,Z'}||P_W \otimes \mu') := I_{\phi_1}(Z_i; W)$. If $Z_i \sim \mu$, we have

$$
\begin{aligned}
D_{\phi_\alpha}(P||Q) &= \int_{\mathcal{W}\times\mathcal{Z}} \left|dP_{W,Z_i} - dP_W d\mu'\right| \\
&= \int_{\mathcal{W}\times\mathcal{Z}} \left|dP_{W,Z_i} - dP_W d\mu + dP_W d\mu - dP_W d\mu'\right| \\
&\leq \int_{\mathcal{W}\times\mathcal{Z}} \left|dP_{W,Z_i} - dP_W d\mu\right| + \int_{\mathcal{W}\times\mathcal{Z}} \left|dP_W d\mu - dP_W d\mu'\right| \\
&= I_{\phi_1}(W; Z_i) + 2TV(\mu||\mu')
\end{aligned}
$$

where $TV(\mu||\mu') = \frac{1}{2}D_{\phi_1}(\mu||\mu')$ denotes the total variation distance between the distribution μ and μ'. By this we can extend the mutual information measure to ϕ-divergence. □

D. *Proof of Theorem 3*

Theorem 3 (Generalization error of noisy gradient descent)**.** *Assume that $W(T)$ is obtained from noisy gradient descent algorithm at T iteration, and assume that $\ell(w,Z)$ is r^2-subgaussian over $P_W \otimes \mu'$ for any $w \in \mathcal{W}$, and the gradient is bounded, e.g., $\left\|\nabla(\ell(w(t),Z))\right\|_2 \le K_{ST}$ for any $w(t)$. then*

$$\mathbb{E}_{wSS'}\left\{\mathrm{gen}\left(W(T),S,S'\right)\right\} \le \alpha\sqrt{\frac{2r^2}{\beta n}\hat{I}(S)} + (1-\alpha)\sqrt{2r^2\left(\frac{\hat{I}(S)}{(1-\beta)n} + D\left(\mu\|\mu'\right)\right)}$$

where we define

$$\hat{I}(S) := \frac{d}{2}\sum_{t=1}^{T}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right) - \sum_{t=1}^{T} h(n_t) \tag{9}$$

The following lemma is used to prove the theorem.

Lemma 2. *For all t, if the noise $n(t) \sim \mathcal{N}(0, \sigma_t I_d)$, we have*

$$I\left(W(t);S|W(t-1)\right) \le \frac{d}{2}\log\left(1 + \frac{\eta_t^2 K_{ST}^2}{d\sigma_t^2}\right)$$
$$I\left(W(t);S'|W(t-1)\right) \le \frac{d}{2}\log\left(1 + \frac{\eta_t^2 K_{ST}^2}{d\sigma_t^2}\right)$$

Proof. From the definition of mutual entropy

$$I\left(W(t);S|W(t-1)\right) = h\left(W(t)|W(t-1)\right) - h\left(w(t)|W(t-1),S\right)$$

Each term can be bounded in the final expression. First we have

$$W(t) = W(t-1) - \eta_t(\alpha\nabla\hat{L}_\alpha(W(t-1),S') + (1-\alpha)\nabla\hat{L}_\alpha(W(t-1),S)) + n(t)$$

Note that

$$h\left(W(t) - W(t-1)|W(t-1)\right) = h\left(W(t)|W(t-1)\right)$$

since the subtraction term does not affect the entropy of a random variable. Also the perturbation $n(t)$ is independent with the gradient term, thus we can compute the upper bound of the expected squared-norm of $w(t) - w(t-1)$:

$$\begin{aligned}\mathbb{E}\left(\left\|W(t)-W(t-1)\right\|_2^2\right) &= \mathbb{E}\left(\left\|\eta_t(\alpha\nabla(\hat{L}_\alpha(W(t-1),S') + (1-\alpha)\nabla(\hat{L}_\alpha(W(t-1),S))\right\|_2^2 + \left\|n(t)\right\|_2^2\right)\\ &\le \eta_t^2(\alpha K_{ST} + (1-\alpha)K_{ST})^2 + d\sigma_t^2\\ &\le \eta_t^2 K_{ST}^2 + d\sigma_t^2\end{aligned}$$

where in the expression above, we used the assumption that $n(t) \sim N(0,\sigma_t^2 I_d)$. Among all random variables X with a fixed expectation bound $\mathbb{E}\left\|X\right\|_2^2 < A$, then the norm distribution $Y \sim N(0, \sqrt{\frac{A}{d}}I_d)$ has the largest entropy given by:

$$h(Y) = d\log\left(\sqrt{2\pi e\sigma_Y^2}\right) = \frac{d}{2}\log\left(\frac{2\pi e A}{d}\right)$$

which indicates that:

$$h\left(W(t)|W(t-1)\right) \le \frac{d}{2}\log\left(2\pi e\frac{\eta_t^2 K_{ST}^2 + d\sigma_t^2}{d}\right)$$

By entropy power inequality [3], we have:

$$\begin{aligned}h\left(W(t)|W(t-1),S\right) =& h\left(W(t-1) + \eta_t\nabla\hat{L}_\alpha(W(t-1),S,S') + n(t)|W(t-1),S\right)\\ =& h\left(n(t) + \eta_t\alpha\nabla\hat{L}_\alpha(W(t-1),S')|W(t-1),S\right)\\ \ge& \frac{1}{2}\log(e^{2h(n(t))} + e^{2h\left(\eta_t\alpha\nabla\hat{L}_\alpha(W(t-1),S')|W(t-1),S\right)})\\ \ge& h\left(n(t)\right)\end{aligned}$$

this leads to the following desired bound for the mutual entropy $I(W(t);S|W(t-1))$:

$$h\left(W(t)|W(t-1)\right)-h\left(W(t)|S,W(t-1)\right)\leq\frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-h(n(t))$$

Similarly, we can achieve the upper bound for the mutual entropy $I(W(t);S'|W(t-1))$:

$$h\left(W(t)|W(t-1)\right)-h\left(W(t)|S',W(t-1)\right)\leq\frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-h(n(t))$$

Therefore, consider the mutual information $I(W(t);S'|W(t-1))$ and $I(W(t);S|W(t-1))$ with Gaussian noise $n(t)$, e.g., $h(n(t))=\frac{d}{2}\log 2\pi e\sigma_t^2$, we can write

$$\begin{aligned}I(W(t);S'|W(t-1))&=h\left(W(t)|W(t-1)\right)-h\left(W(t)|S',W(t-1)\right)\\&\leq\frac{d}{2}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-\frac{d}{2}\log 2\pi e\sigma_t^2\\&=\frac{d}{2}\log\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d\sigma_t^2}\\&=\frac{d}{2}\log\left(1+\frac{\eta_t^2K_{ST}^2}{d\sigma_t^2}\right)\end{aligned}$$

Similarly, we have:

$$I(W(t);S|W(t-1))\leq\frac{d}{2}\log\left(1+\frac{\eta_t^2K_{ST}^2}{d\sigma_t^2}\right)$$

□

Proof. (of Theorem 3) Use Jensen-inequality, we reach

$$\begin{aligned}\mathbb{E}_{WSS'}\left\{\text{gen}\left(W(T),S,S'\right)\right\}&\leq\frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I\left(W(T);Z_i\right)}+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{\left(I\left(W(T);Z_i\right)+D\left(\mu\|\mu'\right)\right)}\\&\leq\alpha\sqrt{\frac{2r^2}{\beta n}I(W(T);S')}+(1-\alpha)\sqrt{2r^2\left(\frac{I(W(T);S)}{(1-\beta)n}+D\left(\mu\|\mu'\right)\right)}\end{aligned}\tag{10}$$

Let $W^T=(W(1),W(2),W(3),\cdots,W(T))$, with the characteristic of the gradient descent algorithm, we can show that

$$h(W(t)|W^{(t-1)},S)=h(W(t)|W(t-1),S)\tag{11}$$

which follows from the Markov chain that $S\rightarrow W(1)\rightarrow W(2)\cdots\rightarrow W(T)$. Using lemma 2, both the mutual information $I(W(T);S)$ and $I(W(T);S')$ are bounded as:

$$\begin{aligned}I(W(T);S)\leq&I(W^T;S)\\=&I(W(1);S|W(0))+I(W(2);S|W(1))+I(W(3);S|W(2),W(1))\\&+I(W(4);S|(W(3),W(2),W(1)))+\cdots+I(W(T);S|W^{T-1})\\=&\sum_{t=1}^{T}I(W(t);S|W(t-1))\\\leq&\frac{d}{2}\sum_{t=1}^{T}\log\left(2\pi e\frac{\eta_t^2K_{ST}^2+d\sigma_t^2}{d}\right)-\sum_{t=1}^{T}h(n(t))\end{aligned}$$

where the first inequality follow from the Markov chain $S\rightarrow W^T$. □

E. Proof of excess risk upper bound under strong convex loss function

In this section, we further give the upper bound for excess risk under strong convex loss function as an compl.

Corollary 4 (Excess risk of strongly convex loss function)**.** *Assume Theorem 3 holds and $\ell(W,Z)$ has $\mathcal{L}$-Lipschitz-continuous gradient such that $|\nabla\ell(w_1,Z)-\nabla\ell(w_2,Z)|\leq\mathcal{L}|w_1-w_2|$ for any w_1,w_2 with respect to any Z. Define $\kappa=\frac{\nu}{\mathcal{L}}$, setting $\eta=\frac{1}{\mathcal{L}}$, and W is arbitrarily initialized with $W(0)$ then for any $\epsilon>0$ and $\delta>0$, there exists an n_0 such that for all $n\geq n_0$, the excess risk can be bounded with probability $1-\delta$ over the randomness of samples and learning algorithm as*

$$L_{\mu'}\left(W(T)\right)-L_{\mu'}\left(w^*\right)\leq(1-\alpha)d_{\mathcal{W}}(\mu,\mu')+\alpha\sqrt{\frac{2r^2}{\beta n}\hat{I}(S)}+(1-\alpha)\sqrt{2r^2\left(\frac{\hat{I}(S)}{(1-\beta)n}+D\left(\mu\|\mu'\right)\right)}+\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}}$$
$$+K_{ST}(1-\kappa)^T\left\|W(0)-W_{\mathsf{ERM}}\right\|+K_{ST}\sum_{t=1}^{T}(1-\kappa)^{T-i}\left\|n(t)\right\|+\epsilon$$

We leverage the following proposition that

Proposition 1. *Under the given assumptions, we define $\kappa=\frac{\nu}{\mathcal{L}}\in(0,1)$, setting $\eta=\frac{1}{\mathcal{L}}$, for all $k\geq1$, we have:*

$$\begin{aligned}\hat{L}_\alpha(W(T),S,S')-\hat{L}_\alpha(w_{\mathsf{ERM}},S,S')&\leq K_{ST}\left\|W(T)-w_{\mathsf{ERM}}\right\|\\&\leq K_{ST}(1-\kappa)^T(\left\|W(0)-W_{\mathsf{ERM}}\right\|+\hat{A}_T)\end{aligned}$$

where we define $\hat{A}_T$

$$\hat{A}_T:=\sum_{t=1}^{T}(1-\kappa)^{-t}\left\|n(t)\right\|$$

We firstly claim that $\hat{L}_\alpha$ is K_{ST}-Lipschitz continuity with K_{ST} bounded gradient, then the proof follows the proposition 3 in the work [4].

Proof. (of Corollary 4) We firstly decompose the excess risk $L_{\mu'}(W(T))-L_{\mu'}(w^*)$ into five fractions as follows.

$$\begin{aligned}L_{\mu'}(W(T))-L_{\mu'}(w^*)=&\,L_{\mu'}(W(T))-\hat{L}_\alpha(W(T))+\hat{L}_\alpha(W(T))-\hat{L}_\alpha(W_{\mathsf{ERM}})\\&+\hat{L}_\alpha(W_{\mathsf{ERM}})-\hat{L}_\alpha(w^*)+\hat{L}_\alpha(w^*)-L_\alpha(w^*)+L_\alpha(w^*)-\hat{L}_{\mu'}(w^*)\end{aligned}$$

Following proposition 1, we have

$$\mathbb{E}\{L_{\mu'}(W(T))-\hat{L}_\alpha(W(T))\}\leq\frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I\left(W(T);Z_i\right)}+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{\left(I\left(W(T);Z_i\right)+D\left(\mu\|\mu'\right)\right)}\quad(12)$$

Then use proposition 1, we reach

$$\hat{L}_\alpha(W(T))-\hat{L}_\alpha(W_{\mathsf{ERM}})\leq K_{ST}(1-\kappa)^T(\left\|W(0)-W_{\mathsf{ERM}}\right\|+\sum_{t=1}^{T}(1-\kappa)^{-t}\left\|n(t)\right\|)\quad(13)$$

$\hat{L}_\alpha(w^*)-L_\alpha(w^*)+L_\alpha(w^*)-\hat{L}_{\mu'}(w^*)$ can be bounded with Theorem 2 for any $w^*\in\mathcal{W}$ with probability at least $1-\delta$ that

$$\hat{L}_\alpha(w^*)-L_\alpha(w^*)+L_\alpha(w^*)-\hat{L}_{\mu'}(w^*)\leq\sqrt{\frac{\alpha^2}{\beta}+\frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}}+(1-\alpha)d_{\mathcal{W}}(\mu,\mu')\quad(14)$$

With the property $\hat{L}_\alpha(W_{\mathsf{ERM}})-\hat{L}_\alpha(w^*)<0$, we combine the inequality 12, 13 and 14 and claim the result. □

F. Experiments Settings

To apply Corollary 2 for generalization error, we firstly write

$$|\mathbb{E}_{WSS'}\{\text{gen}(W_{\mathsf{ERM}},S,S')\}|\leq\frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i)}+\frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i)+D(\mu\|\mu'))}$$

To evaluate the mutual information $I(W_{\mathsf{ERM}}, Z_i)$ efficiently, we follow the work [5] by repeatedly generating W_{ERM} and Z_i. For other parameters we set

$$\begin{aligned}
n &= n_t + n_s \\
\alpha = \beta &= \frac{n_t}{n_t + n_s} \\
D(\mu(X,Y)\|\mu'(X',Y')) &= D(\mathbb{P}_X\|\mathbb{P}_{X'}) + \mathbb{E}_{X\sim\mathbb{P}_X}\{D(\mathbb{P}_{Y|X=x}\|\mathbb{P}_{Y'|X=x})\} \\
r^2 &= \frac{(\max_{Z\in S, w\in\mathcal{W}} \ell(w,Z) - \min_{Z\in S, w\in\mathcal{W}} \ell(w,Z))^2}{4}
\end{aligned}$$

To evaluate the KL-divergence between the source and target, the first term $D(\mathbb{P}_X\|\mathbb{P}_{X'})$ can be theoretically calculated using the parameters of Gaussian distributions. The latter term denotes the expected KL-divergence over $\mathbb{P}_X$ between two Bernoulli distributions, which can be evaluated by generating abundant samples from the source domain.

To apply Theorem 2 for excess risk, we write

$$\begin{aligned}
\mathbb{E}\{L_{\mu'}(W_{\mathsf{ERM}}) - L_{\mu'}(w^*)\} \leq & \frac{\alpha\sqrt{2r^2}}{\beta n}\sum_{i=1}^{\beta n}\sqrt{I(W_{\mathsf{ERM}};Z_i)} + \frac{(1-\alpha)\sqrt{2r^2}}{(1-\beta)n}\sum_{i=\beta n+1}^{n}\sqrt{(I(W_{\mathsf{ERM}};Z_i) + D(\mu||\mu'))} \\
& + \sqrt{\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{(1-\beta)}}\sqrt{\frac{2r^2\ln\frac{2}{\delta}}{n}} + (1-\alpha)d_{\mathcal{W}}(\mu,\mu')
\end{aligned} \tag{15}$$

We give a data-dependent estimation for the integral probability metric term $d_{\mathcal{W}}(\mu,\mu')$ as

$$\hat{d}_{\mathcal{W}}(\mu,\mu') = \sup_{w\in\mathcal{W}}|\hat{L}(w,S) - \hat{L}(w,S')|.$$

Here we restrict the solution space as $\mathcal{W} := \{w : |w| < 3\}$ and $\delta = 0.01$.

References

[1] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, Feb. 2013.

[2] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *2017 IEEE International Symposium on Information Theory (ISIT)*, Jun. 2017, pp. 1475–1479.

[3] A. Dembo, T. M. Cover, and J. A. Thomas, "Information theoretic inequalities," *IEEE Transactions on Information theory*, vol. 37, no. 6, pp. 1501–1518, 1991.

[4] M. Schmidt, N. L. Roux, and F. R. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in neural information processing systems*, 2011, pp. 1458–1466.

[5] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal processing*, vol. 16, no. 3, pp. 233–248, 1989.