Appendix

The Appendix is organized as follows.

- Section A presents omitted proofs for theoretical conclusions in the main paper.
- Section B includes additional details for Figures 2 and 3.
- Section C provides additional discussions.

A. Proofs

Proof for Theorem 4

Proof. Recall the definition of τ_l:

$$\tau_l := \frac{\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l+1}\cdot(1-\alpha)^{n-l}]}{\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l}\cdot(1-\alpha)^{n-l}]} \tag{15}$$

To bound τ_l, we start with the denominator $\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^l\cdot(1-\alpha)^{n-l}]$. First $\forall\alpha\in[0,1]$, we have

$$\alpha^{l-1}\cdot(1-\alpha)^{n-l} \le \left(\frac{l-1}{n-1}\right)^{l-1}\left(1-\frac{l-1}{n-1}\right)^{n-l}.$$

The above can be easily verified by the checking the first order condition of $\log\left(\alpha^{l-1}\cdot(1-\alpha)^{n-l}\right)$. Therefore

$$\begin{aligned}
&\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^l\cdot(1-\alpha)^{n-l}]\\
\le&\left(\frac{l-1}{n-1}\right)^{l-1}\left(1-\frac{l-1}{n-1}\right)^{n-l}\cdot\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha]\\
=&\left(\frac{l-1}{n-1}\right)^{l-1}\left(1-\frac{l-1}{n-1}\right)^{n-l}\cdot\frac{1}{N}.
\end{aligned} \tag{16}$$

where the last equality is due to $\bar{\pi}^N[\alpha]=\frac{1}{N}$. Now let's look at the numerator $\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l+1}\cdot(1-\alpha)^{n-l}]$. For $\alpha\in[\frac{l-1}{n-1},\frac{l}{n}]$, we have

$$\alpha^{l+1}\cdot(1-\alpha)^{n-l}\ge\alpha\cdot\left(\frac{l-1}{n-1}\right)^{l}\cdot\left(1-\frac{l}{n}\right)^{n-l}$$

Therefore combining the bounds for both denominator and the numerator we have

$$\begin{aligned}
\tau_l \ge&\frac{\left(\frac{l-1}{n-1}\right)^{l}\cdot\left(1-\frac{l}{n}\right)^{n-l}}{\left(\frac{l-1}{n-1}\right)^{l-1}\left(1-\frac{l-1}{n-1}\right)^{n-l}}\cdot\mathbb{E}\left[N\alpha\cdot\mathbb{1}\left(\alpha\in\left[\frac{l-1}{n-1},\frac{l}{n}\right]\right)\right]\\
=&\underbrace{\frac{l-1}{n-1}}_{\text{Term 1}}\cdot\underbrace{\left(\frac{1-\frac{l}{n}}{1-\frac{l-1}{n-1}}\right)^{n-l}}_{\text{Term 2}}\cdot\underbrace{\mathsf{weight}\left(\bar{\pi}^N,\left[\frac{l-1}{n-1},\frac{l}{n}\right]\right)}_{\text{Term 3}},
\end{aligned} \tag{17}$$

where the weight is introduced by its definition. For the second term above,

$$\begin{aligned}
&\left(\frac{1-\frac{l}{n}}{1-\frac{l-1}{n-1}}\right)^{n-l} \\
=&\left(1-\frac{\frac{l}{n}-\frac{l-1}{n-1}}{1-\frac{l-1}{n-1}}\right)^{n-l} \\
=&\left(1-\frac{l(n-1)-(l-1)n}{n(n-1)-(l-1)n}\right)^{n-l} \\
=&\left(1-\frac{n-l}{n(n-l)}\right)^{n-l} \\
\geq&1-\frac{n-l}{n(n-l)}\cdot(n-l)=\frac{l}{n}.
\end{aligned} \tag{18}$$

For the third term, we will call Lemma 2.4 of Feldman (2020):

Lemma 4 (Lemma 2.4 of (Feldman, 2020))**.** *For any $0<\beta_1<\beta_2<1$, and for any $\gamma>0$,*

$$\textit{weight}\left(\bar{\pi}^N,[\beta_1,\beta_2]\right)\geq\frac{1-\delta}{1-\frac{1}{N}+\beta_2+\gamma}\cdot\textit{weight}\left(\bar{\pi}^N,\left[\frac{\beta_1}{1-\frac{1}{N}+\beta_1-\gamma},\frac{\beta_2}{1-\frac{1}{N}+\beta_2+\gamma}\right]\right), \tag{19}$$

where in above $\delta:=2\cdot e^{\frac{-\gamma^2}{2(N-1)Var(\pi)+2\gamma\pi_{\max}/3}}$, and

$$Var(\pi):=\sum_{j\in[N]}(\pi_j-\frac{1}{N})\leq\frac{\pi_{\max}}{N}.$$

Using above, we next further derive that

$$\text{weight}\left(\bar{\pi}^N,[\frac{l-1}{n-1},\frac{l}{n}]\right)\geq 0.4\cdot\text{weight}\left(\pi,\left[\frac{2}{3}\frac{l-1}{n-1},\frac{4}{3}\frac{l}{n}\right]\right)$$

To see this, using above Lemma 4, for any $\gamma>0$,

$$\text{weight}\left(\bar{\pi}^N,\left[\frac{l-1}{n-1},\frac{l}{n}\right]\right)\geq\frac{1-\delta}{1-\frac{1}{N}+\frac{l}{n}+\gamma}\cdot\text{weight}\left(\pi,\left[\frac{\frac{l-1}{n-1}}{1-\frac{1}{N}+\frac{l-1}{n-1}-\gamma},\frac{\frac{l}{n}}{1-\frac{1}{N}+\frac{l}{n}+\gamma}\right]\right)$$

Let $\gamma=\frac{1}{2}$, for sufficiently large $n(\gg l),N$,

$$1-\frac{1}{N}+\frac{l-1}{n-1}-\gamma\leq\frac{2}{3}$$

Similarly

$$1-\frac{1}{N}+\frac{l}{n}+\gamma\geq\frac{4}{3}$$

Easy to show that $\delta\leq 2e^{-1/10\pi_{max}}$, and when $\pi_{max}\leq\frac{1}{20}$, we have $\delta\leq 2e^{-2}\leq 0.3$. Therefore

$$\frac{1-\delta}{1-\frac{1}{N}+\frac{l}{n}+\gamma}\geq\frac{0.7}{7/4}=0.4. \tag{20}$$

To summarize

$$\text{weight}\left(\bar{\pi}^N,[\frac{l-1}{n-1},\frac{l}{n}]\right)\geq 0.4\cdot\text{weight}\left(\pi,\left[\frac{2}{3}\frac{l-1}{n-1},\frac{4}{3}\frac{l}{n}\right]\right)$$

Combining Term 2 and 3 we complete the proof. □

Proof for Proposition 5

Proof. The major difference from proving Theorem 4 is due to the reasoning of the numerator: $\mathbb{E}_{\alpha\sim\bar{\pi}^N}[\alpha^{l+1}\cdot(1-\alpha)^{n-l}]$. Now for $\alpha\in[\frac{l-1}{1.1(n-1)},\frac{l-1}{n-1}]$, we have

$$\alpha^{l+1}\cdot(1-\alpha)^{n-l}\geq\alpha\cdot\left(\frac{l-1}{1.1(n-1)}\right)^{l}\cdot\left(1-\frac{l-1}{n-1}\right)^{n-l}$$

Therefore, using Eqn. (16), and the definition of τ_l we have

$$\begin{aligned}\tau_l\geq&\frac{\left(\frac{l-1}{1.1(n-1)}\right)^{l}\cdot\left(1-\frac{l-1}{n-1}\right)^{n-l}}{\left(\frac{l-1}{n-1}\right)^{l-1}\left(1-\frac{l-1}{n-1}\right)^{n-l}}\cdot\mathbb{E}\left[N\alpha\cdot\mathbb{1}\left(\alpha\in\left[\frac{l-1}{1.1(n-1)},\frac{l-1}{n-1}\right]\right)\right]\\=&\frac{l-1}{n-1}\cdot\frac{1}{1.1^l}\cdot\mathsf{weight}\left(\bar{\pi}^N,\left[\frac{l-1}{1.1(n-1)},\frac{l-1}{n-1}\right]\right).\end{aligned}\tag{21}$$

Again, using Lemma 2.4 of Feldman (2020), for the weight term, we further derive that for any $\gamma>0$,

$$\mathsf{weight}\left(\bar{\pi}^N,\left[\frac{l-1}{1.1(n-1)},\frac{l-1}{n-1}\right]\right)\geq\frac{1-\delta}{1-\frac{1}{N}+\frac{l-1}{n-1}+\gamma}\cdot\mathsf{weight}\left(\pi,\left[\frac{\frac{l-1}{1.1(n-1)}}{1-\frac{1}{N}+\frac{l-1}{1.1(n-1)}-\gamma},\frac{\frac{l-1}{n-1}}{1-\frac{1}{N}+\frac{l-1}{n-1}+\gamma}\right]\right)$$

Let $\gamma=\frac{1}{2}$, for sufficiently large $n(\gg l),N$,

$$1-\frac{1}{N}+\frac{l-1}{1.1(n-1)}-\gamma\leq\frac{2}{3}$$

Similarly

$$1-\frac{1}{N}+\frac{l-1}{1.1(n-1)}+\gamma\geq\frac{4}{3}$$

Similarly we show that $\delta\leq 2e^{-1/10\pi_{max}}$, and when $\pi_{max}\leq\frac{1}{20}$ we have $\delta\leq 2e^{-2}\leq 0.3$. Therefore

$$\frac{1-\delta}{1-\frac{1}{N}+\frac{l-1}{n-1}+\gamma}\geq\frac{0.7}{7/4}=0.4.\tag{22}$$

To summarize

$$\mathsf{weight}\left(\bar{\pi}^N,\left[\frac{l-1}{1.1(n-1)},\frac{l-1}{n-1}\right]\right)\geq\mathsf{weight}\left(\pi,\left[0.7\cdot\frac{l-1}{n-1},\frac{4}{3}\frac{l-1}{n-1}\right]\right)$$

Putting this above bound back to Eqn. (21) we complete the proof. □

Proof for Theorem 6

Proof. This is simply because when h memorizes the noisy labels for x such that $\mathbb{P}_{h\sim\mathcal{A}(S')}[h(x)=k]=\tilde{\mathbb{P}}[\tilde{y}=k|x]$, we will have:

$$\begin{aligned}&\mathsf{err}_l^+(\mathcal{P},\mathcal{A},x|S)\\=&\tau_l\cdot\mathbb{P}_{h\sim\mathcal{A}(S')}[h(x)\neq y]\\=&\tau_l\cdot\sum_{k\neq y}\tilde{\mathbb{P}}[\tilde{y}=k|x].\end{aligned}$$

Plugging the lower bound we prepared for τ_l earlier (Theorem 4) we proved the claim. □

Proof for Lemma 2

Proof. Using the definition of $\mathbf{y}_{\text{LC}}$, and using the knowledge of Eqn. (10) we know that

$$\mathbf{y}_{\text{LC}}[1] = \frac{1}{1-e_+(x)-e_-(x)}\Big((1-e_+(x))\tilde{\mathbb{P}}[\tilde{y}=-1|x] - e_+(x)\tilde{\mathbb{P}}[\tilde{y}=+1|x]\Big)$$
$$\mathbf{y}_{\text{LC}}[2] = \frac{1}{1-e_+(x)-e_-(x)}\Big((1-e_-(x))\tilde{\mathbb{P}}[\tilde{y}=+1|x] - e_-(x)\tilde{\mathbb{P}}[\tilde{y}=-1|x]\Big)$$

Therefore

$$\mathbf{y}_{\text{LC}}[1] + \mathbf{y}_{\text{LC}}[2] = \frac{1}{1-e_+(x)-e_-(x)}\Big((1-e_+(x)-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x] + (1-e_+(x)-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x]\Big) = 1.$$

□

Proof for Theorem 7

Proof. Due to symmetricity, we consider $y=+1$. Using the definition of $\mathbf{y}_{\text{LC}}$, and using the knowledge of Eqn. (10) we know that

$$\mathbb{P}[y_{\text{LC}}=+1|x] = \frac{(1-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x] - e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x]}{1-e_+(x)-e_-(x)}$$

Recall we denote by y_{LC} the random variable drawn according to $\mathbf{y}_{\text{LC}}$. Next we show:

$$\tilde{\mathbb{P}}[\tilde{y}=+1|x] > \tilde{\mathbb{P}}[\tilde{y}=-1|x] \Leftrightarrow \mathbb{P}[y_{\text{LC}}=+1|x] > \tilde{\mathbb{P}}[\tilde{y}=+1|x]. \tag{23}$$

This is equivalent to the following comparison (when $e_+(x)+e_-(x)<1$):

RHS of Eqn. (23)
$\Leftrightarrow \frac{(1-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x] - e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x]}{1-e_+(x)-e_-(x)} > \tilde{\mathbb{P}}[\tilde{y}=+1|x]$
$\Leftrightarrow (1-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x] - e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x] > (1-e_+(x)-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x]$
$\Leftrightarrow e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x] > e_+\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x]$
$\Leftrightarrow \tilde{\mathbb{P}}[\tilde{y}=+1|x] > \tilde{\mathbb{P}}[\tilde{y}=-1|x]$
$\Leftrightarrow$ LHS of Eqn. (23)

Note that because we consider a non-trivial case $\tilde{\mathbb{P}}[\tilde{y}\neq y|x]>0$, we have $\tilde{\mathbb{P}}[\tilde{y}=+1|x]<1$. Therefore the above derivation holds even if we capped $\mathbb{P}[y_{\text{LC}}=+1|x] = \frac{(1-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x]-e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x]}{1-e_+(x)-e_-(x)}$ at 1 to make it a valid probability measure.

Now we derive when $\tilde{\mathbb{P}}[\tilde{y}=+1|x] > \tilde{\mathbb{P}}[\tilde{y}=-1|x]$. Recall $\tilde{y}(1),...,\tilde{y}(l)$ denote the l noisy labels for $x\in X_{S=l}$, and let $Z_1,...,Z_l$ denote the l Bernoulli random variable that $Z_k = \mathbb{1}(\tilde{y}(k)=+1), k\in[l]$. Then

$$\begin{aligned}
&\tilde{\mathbb{P}}[\tilde{y}=+1|x] > \tilde{\mathbb{P}}[\tilde{y}=-1|x] \\
\Leftrightarrow\ &\tilde{\mathbb{P}}[\tilde{y}=+1|x] > 1/2 \\
\Leftrightarrow\ &\frac{\sum_{k\in[l]} Z_k}{l} > 1/2.
\end{aligned}$$

Now we derive $\mathbb{P}\left[\frac{\sum_{k\in[l]} Z_k}{l} > 1/2\right]$. Note $\mathbb{E}\left[\frac{\sum_{k\in[l]} Z_k}{l}\right] = 1-e_+(x)$. By applying Hoeffding inequality we prove that

$$\mathbb{P}\left[\frac{\sum_{k\in[l]} Z_k}{l} \leq 1/2\right] \leq e^{-2l(1/2-e_+(x))^2}.$$

Therefore

$$\mathbb{P}\left[\frac{\sum_{k\in[l]} Z_k}{l} > 1/2\right] \geq 1 - e^{-2l(1/2-e_+(x))^2}.$$

The case with $y = -1$ is entirely symmetric so we omit the details. □

Proof for Corollary 1

Proof. Because $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S) := \tau_l \cdot \mathbb{P}_{h\sim\mathcal{A}(S')}[h(x) \neq y]$, as well as w.p. at least $1 - e^{-2l(1/2-e_{sgn(y)}(x))^2}$, memorizing $\mathbf{y}_{\text{LC}}$ returns lower error $\mathbb{P}[h(x) \neq y]$ than memorizing the noisy labels s.t. $\mathbb{P}[h(x) = k] = \tilde{\mathbb{P}}[\tilde{y} = k|x]$, the corollary is then true via the lower bound for τ_l. □

Proof for Theorem 8

Proof. Again due to symmetricity, we consider $y = +1$. Similar as argued in the proof of Theorem 7, from Eqn. (23) we know that when $\tilde{\mathbb{P}}[\tilde{y} = +1|x] < \tilde{\mathbb{P}}[\tilde{y} = -1|x]$, we have $\mathbb{P}[y_{\text{LC}} = +1|x] < \tilde{\mathbb{P}}[\tilde{y} = +1|x]$, that is memorizing $\mathbf{y}_{\text{LC}}$ is worse than memorizing $\tilde{\mathbf{y}}$. Again because we consider a non-trivial case $\tilde{\mathbb{P}}[\tilde{y} \neq y|x] < 1$, we have $\tilde{\mathbb{P}}[\tilde{y} = +1|x] > 0$. Therefore the above derivation holds even if we capped $\mathbb{P}[y_{\text{LC}} = +1|x] = \frac{(1-e_-(x))\cdot\tilde{\mathbb{P}}[\tilde{y}=+1|x]-e_+(x)\cdot\tilde{\mathbb{P}}[\tilde{y}=-1|x]}{1-e_+(x)-e_-(x)}$ at 0 to make it a valid probability measure.

Similarly define Z_k for $k \in [l]$ as in Proof for Theorem 7: the l Bernoulli random variable that $Z_k = \mathbb{1}(\tilde{y}(k) = +1), k \in [l]$. Then

$$\mathbb{P}[y_{\text{LC}} = +1|x] < \tilde{\mathbb{P}}[\tilde{y} = +1|x] \Leftrightarrow \tilde{\mathbb{P}}[\tilde{y} = +1|x] < \tilde{\mathbb{P}}[\tilde{y} = -1|x] \Leftrightarrow \frac{\sum_{k\in[l]} Z_k}{l} < 1/2$$

We bound $\mathbb{P}\left[\frac{\sum_{k\in[l]} Z_k}{l} < 1/2\right]$:

$$\begin{aligned}
&\mathbb{P}\left[\frac{\sum_{k\in[l]} Z_k}{l} < 1/2\right] \\
=&\mathbb{P}\left[\sum_{k\in[l]} 1 - Z_k \geq l/2\right] \\
=&\mathbb{P}[\text{Bin}(l, e_+(x)) \geq l/2],
\end{aligned}$$

where we use Bin to denote a Binomial random variable, and the fact that $\mathbb{E}\left[\frac{\sum_{k\in[l]}(1-Z_k)}{l}\right] = e_+(x)$. Using tail bound for Bin (e.g., Lemma 4.7.2 of (Ash, 1990)) yields:

$$\begin{aligned}
\mathbb{P}[\text{Bin}(l, e_+(x)) \geq l/2] \geq& \frac{1}{\sqrt{8\cdot\frac{1}{2}l(1-\frac{1}{2})}} \cdot e^{-l\cdot D_{\text{KL}}\left(\frac{1}{2}\|e_+(x)\right)} \\
=& \frac{1}{\sqrt{2l}} \cdot e^{-l\cdot D_{\text{KL}}\left(\frac{1}{2}\|e_+(x)\right)},
\end{aligned}$$

completing the proof. The case with $y = -1$ is entirely symmetric so we omit the details. □

Proof for Theorem 9

Proof. Note that label smoothing smooths the noisy labels in the following way:

$$\tilde{\mathbb{P}}[y_{\text{LS}} = +1|x] = (1 - a) \cdot \tilde{\mathbb{P}}[\tilde{y} = +1|x] + \frac{a}{2}. \tag{24}$$

The proof is then simple: when $\mathcal{E}_+$ is true, from the proof of Theorem 7, Eqn. (23), we know that $\mathbf{y}_{\text{LC}}$ further extremizes/increases the prediction of $+1$ that $\mathbb{P}[y_{\text{LC}} = +1|x] > \tilde{\mathbb{P}}[\tilde{y} = +1|x]$, while label smoothing y_{LS} reduces from $\tilde{\mathbb{P}}[\tilde{y} = +1|x]$ by a factor of $a\cdot\tilde{\mathbb{P}}[\tilde{y} = +1|x] - \frac{a}{2} > 0$. Therefore, memorizing smoothed label increases the error $\mathbb{P}[h(x) \neq y]$.

The above observation is reversed when the opposite event $\bar{\mathcal{E}}_+$ is instead true: in this case, $\mathbf{y}_{\text{LC}}$ further extremizes/increases the prediction of -1 that $\mathbb{P}[y_{\text{LC}} = +1|x] < \tilde{\mathbb{P}}[\tilde{y} = +1|x]$, while label smoothing y_{LS} increases from $\tilde{\mathbb{P}}[\tilde{y} = +1|x]$ by a factor of $\frac{a}{2} - a \cdot \tilde{\mathbb{P}}[\tilde{y} = +1|x] > 0$. □

Proof for Lemma 3

Proof. Denote by ℓ_{CE} the CE loss, and $\mathcal{D}_x, \mathcal{D}_{\tilde{y}}$ the marginal distribution of $x, \tilde{y}$ explicitly.

$$\begin{aligned}
&\mathbb{E}_{x\times\tilde{y}}\left[\ell_{\text{PL}}(h(x),\tilde{y})\right] \\
=&\mathbb{E}_{x\times\tilde{y}}\left[\ell_{\text{CE}}(h(x),\tilde{y})\right] - \mathbb{E}_{\mathcal{D}_{\tilde{y}}}\left[\mathbb{E}_{\mathcal{D}_x}\left[\ell_{\text{CE}}\left(h\left(x\right),\tilde{y}\right)\right]\right] \\
=&\underbrace{-\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x,\tilde{y})\log\mathbb{Q}(\tilde{y}|x)}_{\text{CE term}} + \underbrace{\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x)\mathbb{P}(\tilde{y})\log\mathbb{Q}(\tilde{y}|x)}_{\text{Peer term}},
\end{aligned} \tag{25}$$

The conditional probability $\mathbb{Q}(\tilde{y}|x)$ is defined as the prediction of the underlying neural network model, while $\mathbb{P}(x)\mathbb{P}(y)$ captures the probabilities of the marginal-product of the training distribution. We call the first sum-integration as *CE term* and the second as *peer term*.

For the CE term, we further have

$$\begin{aligned}
&-\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x,\tilde{y})\log\mathbb{Q}(\tilde{y}|x) \\
=&-\sum_{x\in X}\sum_{\tilde{y}\in Y}\left[\mathbb{P}(x,\tilde{y})\log\mathbb{Q}(\tilde{y}|x)\mathbb{P}(x) - \mathbb{P}(x,\tilde{y})\log\mathbb{P}(x,\tilde{y})\right] \\
=&-\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x,\tilde{y})\log\frac{\mathbb{Q}(x,\tilde{y})}{\mathbb{P}(x,\tilde{y})} \\
=&D_{\text{KL}}(\mathbb{Q}(x,\tilde{y})\|\mathbb{P}(x,\tilde{y}))
\end{aligned}$$

In the above, we used $\mathbb{Q}(x,\tilde{y}) = \mathbb{Q}(\tilde{y}|x)\mathbb{P}(x)$ as in classification task the model prediction does not affect the feature distribution. Similarly, for the peer term

$$\begin{aligned}
&\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x)\mathbb{P}(\tilde{y})\log\mathbb{Q}(\tilde{y}|x) \\
=&\sum_{x\in X}\sum_{\tilde{y}\in Y}\left[\mathbb{P}(x)\mathbb{P}(\tilde{y})\log\mathbb{Q}(\tilde{y}|x)\mathbb{P}(x) - \mathbb{P}(x)\mathbb{P}(\tilde{y})\log\mathbb{P}(x)\mathbb{P}(\tilde{y})\right] \\
=&\sum_{x\in X}\sum_{\tilde{y}\in Y}\mathbb{P}(x)\mathbb{P}(\tilde{y})\log\frac{\mathbb{Q}(x,\tilde{y})}{\mathbb{P}(x)\mathbb{P}(\tilde{y})} \\
=&-D_{\text{KL}}(\mathbb{Q}(x,\tilde{y})\|\mathbb{P}(x)\times\mathbb{P}(\tilde{y})).
\end{aligned}$$

Combing the above derivations for the CE and peer term we complete the proof. □

Proof for Theorem 11

Proof. Again due to symmetricity, we consider $y = +1$. It was shown in (Cheng et al., 2020a) that, when taking expectation over the data distribution, the original definition of peer loss

$$\ell_{\text{PL}}(h(x),\tilde{y}) := \ell(h(x),\tilde{y}) - \ell(h(x_{p_1}),\tilde{y}_{p_2}) \tag{26}$$

is equivalent to

$$\ell(h(x),\tilde{y}) - \tilde{\mathbb{E}}[\ell(h(x),\tilde{y}_q)], \tag{27}$$

where q is a randomly sampled index from $[n]$, and the expectation is over the randomness in $\tilde{y}_q$. The high-level intuition is that each x appears in the peer terms exactly once in expectation, so we can fix x_{p_1} be the same x. Then we will only vary p_2 (index q in our notation). The operation of taking expectation is to reduce uncertainty in the peer term.

The average empirical peer loss on x is then given by

$$\begin{aligned}
&\frac{1}{l}\sum_{i=1}^{l}\ell_{\text{PL}}(h(x),\tilde{y}(i))\\
=&\frac{1}{l}\sum_{i=1}^{l}\ell(h(x),\tilde{y})-\tilde{\mathbb{E}}[\ell(h(x),\tilde{y}_q)]\\
=&\tilde{\mathbb{E}}\left[\ell(h(x),\tilde{y})\right]-\ell(h(x),-1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=-1]\\
&\quad-\ell(h(x),+1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=+1].
\end{aligned}$$

where $\tilde{\mathbb{E}}$ denotes the empirical expectation w.r.t. the empirical distribution of $\tilde{y}|x$, and $\tilde{\mathbb{P}}[\tilde{y}_q]$s are the empirical distribution of $\tilde{y}_q$. As shown already, peer loss pushes h to predict one class confidently, therefore there are two cases: $h(x)=+1$ or $h(x)=-1$.

When $h(x)=+1$ we have

$$\begin{aligned}
&\tilde{\mathbb{E}}\left[\ell(h(x)=+1,\tilde{y})\right]-\ell(h(x),-1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=-1]\\
&\quad-\ell(h(x),+1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=+1]\\
=&\tilde{\mathbb{P}}[\tilde{y}=+1|x]\cdot\ell(h(x)=+1,+1)+\tilde{\mathbb{P}}[\tilde{y}=-1|x]\cdot\ell(h(x)=+1,-1)\\
&\quad-\ell(h(x)=+1,-1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=-1]-\ell(h(x)=+1,+1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=+1]\\
&=\left(\tilde{\mathbb{P}}[\tilde{y}=+1|x]-\tilde{\mathbb{P}}[\tilde{y}_q=+1]\right)\cdot\left(\ell(h(x)=+1,+1)-\ell(h(x)=+1,-1)\right)
\end{aligned}$$

While if $h(x)=-1$, we have

$$\begin{aligned}
&\tilde{\mathbb{E}}\left[\ell(h(x)=-1,\tilde{y})\right]-\ell(h(x),-1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=-1]\\
&\quad-\ell(h(x),+1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=+1]\\
=&\tilde{\mathbb{P}}[\tilde{y}=+1|x]\cdot\ell(h(x)=-1,+1)+\tilde{\mathbb{P}}[\tilde{y}=-1|x]\cdot\ell(h(x)=-1,-1)\\
&\quad-\ell(h(x)=-1,-1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=-1]-\ell(h(x)=-1,+1)\cdot\tilde{\mathbb{P}}[\tilde{y}_q=+1]\\
&=\left(\tilde{\mathbb{P}}[\tilde{y}=-1|x]-\tilde{\mathbb{P}}[\tilde{y}_q=-1]\right)\left(\ell(h(x)=-1,-1)-\ell(h(x)=-1,+1)\right)\\
&=-\left(\tilde{\mathbb{P}}[\tilde{y}=+1|x]-\tilde{\mathbb{P}}[\tilde{y}_q=+1]\right)\left(\ell(h(x)=-1,-1)-\ell(h(x)=-1,+1)\right)
\end{aligned}$$

For CE, we have $\ell(h(x)=+1,+1)=\ell(h(x)=-1,-1)=0$, and in Cheng et al. (2020a); Liu & Guo (2020), $\ell(h(x)=-1,+1)=\ell(h(x)=+1,-1)$ is set to be a large positive quantity. Let's denote it as C. Then $h(x)=+1$ returns a lower loss ($\left(\tilde{\mathbb{P}}[\tilde{y}=+1|x]-\tilde{\mathbb{P}}[\tilde{y}_q=+1]\right)\cdot(-C)$, a negative quantity as compared to $-\left(\tilde{\mathbb{P}}[\tilde{y}=+1|x]-\tilde{\mathbb{P}}[\tilde{y}_q=+1]\right)\cdot(-C)$ a positive one) if $\tilde{\mathbb{P}}[\tilde{y}=+1|x]-\mathbb{P}[\tilde{y}_q=+1]>0$.

Next we derive the probability of having $\tilde{\mathbb{P}}[\tilde{y}=+1|x]>\tilde{\mathbb{P}}[\tilde{y}_q=+1]$. When n is sufficiently large, $\tilde{\mathbb{P}}[\tilde{y}_q=+1]\approx p_+\cdot(1-e_+)+p_-\cdot e_-<\mathbb{P}[\tilde{y}=+1|x]=1-e_+$. Define Z_l as in Proof for Theorem 7: the l Bernoulli random variable that $Z_k=\mathbb{1}(\tilde{y}(k)=+1), k\in[l]$. Then using Hoeffding bound we prove that

$$\begin{aligned}
&\mathbb{P}\left[\tilde{\mathbb{P}}[\tilde{y}=+1|x]>p_+\cdot(1-e_+)+p_-\cdot e_-\right]\\
=&1-\mathbb{P}\left[\frac{\sum_{k\in[l]}Z_k}{l}\le p_+\cdot(1-e_+)+p_-\cdot e_-\right]\\
\ge&1-e^{\frac{-2l}{\left(p_+\cdot(1-e_+)+p_-\cdot e_--(1-e_+)\right)^2}}\\
\ge&1-e^{\frac{-2l}{p_-^2\cdot(1-e_+-e_-)^2}},
\end{aligned}$$

completing the proof. Again the case with $y=-1$ is symmetric. □

Proof for Corollary 2

Proof. Because we have shown that w.p. at least $1 - e^{\frac{-2l}{p_{sgn(-y)}^2 \cdot (1-e_+-e_-)^2}}$, training with ℓ_{PL} on $x \in X_{S=l}$ returns $h(x) = y$, results a 0 individual excessive generalization error $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S) = 0$. On the other hand, Theorem 6 informs us that $\mathsf{err}_l^+(\mathcal{P}, \mathcal{A}, x|S)$ is in the following order when memorizing the noisy labels:

$$\Omega\left(\frac{l^2}{n^2}\mathsf{weight}\left(\pi, \left[\frac{2}{3}\frac{l-1}{n-1}, \frac{4}{3}\frac{l}{n}\right]\right) \cdot \sum_{k\neq y} \tilde{\mathbb{P}}[\tilde{y} = k|x]\right) \tag{28}$$

Taking the difference (Eqn.(28) -0) we proved the claim. □

Proof for Theorem 12

Proof. Consider $y = +1$. As argued in the proof of Theorem 11, when $\tilde{\mathbb{P}}[\tilde{y} = +1|x] < \mathbb{P}[\tilde{y}_q = +1]$, we have $h(x) = -1$ returns a smaller loss. So the output $h(x)$ predicts -1, the wrong label. Now we show when $\tilde{\mathbb{P}}[\tilde{y} = +1|x] < \tilde{\mathbb{P}}[\tilde{y}_q = +1]$. Define Z_l as in Proof for Theorem 7: the l Bernoulli random variable that $Z_k = \mathbb{1}(\tilde{y}(k) = +1), k \in [l]$. Then

$$\begin{aligned}
&\mathbb{P}\left[\tilde{\mathbb{P}}[\tilde{y} = +1|x] < \tilde{\mathbb{P}}[\tilde{y}_q = +1]\right] \\
=&\mathbb{P}\left[\tilde{\mathbb{P}}[\tilde{y} = +1|x] < p_+(1-e_+) + p_- \cdot e_-\right] \\
=&\mathbb{P}\left[\sum_{k\in[l]} 1 - Z_k \geq l(p_+ \cdot e_+ + p_- \cdot (1-e_-))\right] \\
=&\mathbb{P}\left[\text{Bin}(l, e_+) \geq l\left(p_+ \cdot e_+ + p_- \cdot (1-e_-)\right)\right].
\end{aligned}$$

Again using the tail bound for Bin, we prove that

$$\begin{aligned}
&\mathbb{P}\left[\text{Bin}(l, e_+) \geq l(p_+ \cdot e_+ + p_- \cdot (1-e_-))\right] \\
\geq&\frac{e^{-l \cdot D_{\text{KL}}(\frac{1}{2}\|e_+)}}{\sqrt{8l\left(p_+ \cdot e_+ + p_- \cdot (1-e_-)\right) \cdot \left(p_+ \cdot (1-e_+) + p_- \cdot e_-\right)}}.
\end{aligned}$$

Note that $(p_+ \cdot e_+ + p_- \cdot (1-e_-)) + (p_+ \cdot (1-e_+) + p_- \cdot e_-) = 1$, and each term $(p_+ \cdot e_+ + p_- \cdot (1-e_-))$ and $(p_+ \cdot (1-e_+) + p_- \cdot e_-)$ is positive. Therefore

$$(p_+ \cdot e_+ + p_- \cdot (1-e_-)) \cdot (p_+ \cdot (1-e_+) + p_- \cdot e_-) \leq \frac{1}{4}.$$

Therefore we conclude that

$$\frac{e^{-l \cdot D_{\text{KL}}(\frac{1}{2}\|e_+)}}{\sqrt{8l\left(p_+ \cdot e_+ + p_- \cdot (1-e_-)\right) \cdot \left(p_+ \cdot (1-e_+) + p_- \cdot e_-\right)}} \geq \frac{1}{\sqrt{2l}} e^{-l \cdot D_{\text{KL}}(\frac{1}{2}\|e_+)},$$

completing the proof. □

B. Figures

More examples for Figure 2

In Figure 5, we provide one additional figure for training with 40% random label noise for comparison.

Details for generating Figure 3

The training uses ResNet-34 as the backbone with the following setups: mini-batch size (64), optimizer (SGD), initial learning rate (0.1), momentum (0.9), weight decay (0.0005), number of epochs (100) and learning rate decay (0.1 at 50 epochs). Standard data augmentation is applied to each dataset.

The generation of instance-dependent label noise follows the steps in (Xia et al., 2020; Cheng et al., 2020a):

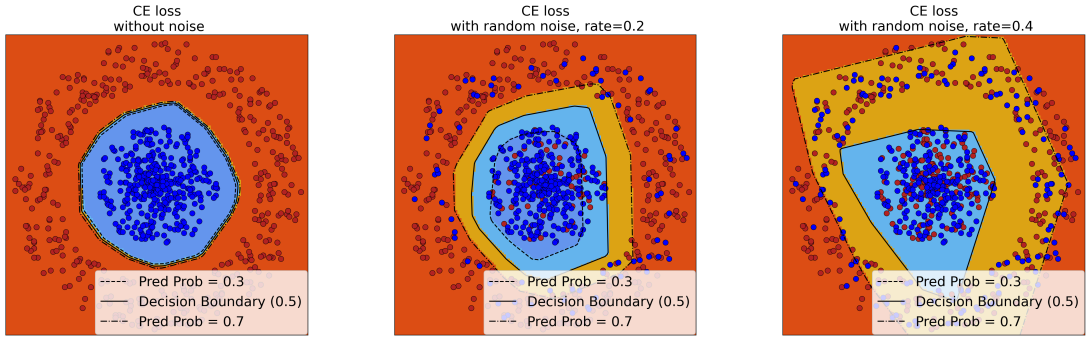

Figure 5. A 2D example illustrating the memorization of noisy labels. Left panel: Training with clean labels. Middle panel: Training with random 20% noisy labels. Right panel: Training with random 40% noisy labels.

- Define a noise rate (the global flipping rate) as ε.
- For each instance x, sample a q from the truncated normal distribution $\mathbf{N}(\varepsilon, 0.1^2, [0, 1])$.
- Sample parameters W from the standard normal distribution. Compute $x \times W$.
- Then the final noise rate is a function of both q and $x \cdot W$.

C. Discussion

Sample cleaning Sample cleaning promotes the procedure of identifying possible wrong labels and removing them from training. If the identification is done correctly, the above procedure is effectively pushing $\tilde{\mathbb{P}}(\tilde{y})$ to the direction of the correct label, achieving a better generalization performance. From the other perspective, removing noisy labels helps the h to de-memorize the noisy ones.

The ability to identify the wrong labels has been more or less analyzed in the literature, but certainly would enjoy a more thorough and in-depth investigation. We view this as an important theoretical question for the community. Similar to the previous observations, we would like to caution the existence of rare examples with small l - the previously introduced approaches have been shown to have a non-negligible chance of failing in such cases. We imagine this is probably true for sample cleaning, when the majority of a small number of noisy labels are in fact misleading.

Understanding the difficulty of labeling A better understanding of the possibility of handling noisy labels calls for immediate effort for understanding the probability of observing a wrong label for different instances. The salient challenge in doing so is again due to the missing of ground truth supervision information. One promising direction to explore is to leverage inference models (Liu et al., 2012) to infer the hidden difficulty factors in the generation of noisy labels.

Hybrid & decoupled training Our results also point out a promising direction to treat samples from different regimes differently. While the existing approaches seem to be comfortable with the highly frequent instances (large l), the rare instances might deserve different handling.

One thing we observed is rare samples suffer from a small l and insufficient label information. While dropping rare samples is clearly hurting the generalization power, a better alternative would be to collecting multiple labels for these instances to boost the classifier's confidence in evaluating these instances.

Dataset effort Last but not least, a concrete understanding of the effects of instance-dependent label noise would require a high-quality dataset that contains real human-level noise patterns. While there have been some recent efforts (Xiao et al., 2015; Jiang et al., 2020), most of the evaluations and studies stay with the synthetic ones.