Learning from Noisy Labels with No Change to the Training Process

Supplementary Material

A. Proof of Theorem 1

Proof. We use $\langle\cdot,\cdot\rangle$ to denote the standard inner product.

$$
\begin{aligned}
&\text{regret}_D^{\mathbf{L}}[\,\widehat{h}\,] \\
&= \mathbf{E}_X\big[\langle\boldsymbol{\eta}(X),\boldsymbol{\ell}_{\widehat{h}(X)}\rangle - \min_{y\in[n]}\langle\boldsymbol{\eta}(X),\boldsymbol{\ell}_y\rangle\big] \\
&= \mathbf{E}_X\big[\max_{y\in[n]}\langle\boldsymbol{\eta}(X),\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y\rangle\big] \\
&= \mathbf{E}_X\big[\max_{y\in[n]}\langle(\mathbf{C}^\top)^{-1}\widetilde{\boldsymbol{\eta}}(X),\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y\rangle\big] \\
&= \mathbf{E}_X\big[\max_{y\in[n]}\langle\widetilde{\boldsymbol{\eta}}(X),\mathbf{C}^{-1}(\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y)\rangle\big] \quad \text{(by property of adjoint)} \\
&\le \mathbf{E}_X\big[\max_{y\in[n]}\langle\widetilde{\boldsymbol{\eta}}(X)-\widehat{\widetilde{\boldsymbol{\eta}}}(X),\mathbf{C}^{-1}(\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y)\rangle\big] \\
&\qquad \text{(since by the definition of } \widehat{h}(X),\ \langle\widehat{\widetilde{\boldsymbol{\eta}}}(X),\mathbf{C}^{-1}(\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y)\rangle \le 0\ \forall y\in[n]) \\
&\le \mathbf{E}_X\Big[\big\|\widehat{\widetilde{\boldsymbol{\eta}}}(X)-\widetilde{\boldsymbol{\eta}}(X)\big\|_2\cdot\big\|\mathbf{C}^{-1}\big\|_2\cdot\max_{y\in[n]}\big\|\boldsymbol{\ell}_{\widehat{h}(X)}-\boldsymbol{\ell}_y\big\|_2\Big] \\
&\qquad \text{(by Cauchy-Schwarz inequality)} \\
&\le 2\max_{y\in[n]}\big\|\boldsymbol{\ell}_y\big\|_2\cdot\big\|\mathbf{C}^{-1}\big\|_2\cdot\mathbf{E}_X\Big[\big\|\widehat{\widetilde{\boldsymbol{\eta}}}(X)-\widetilde{\boldsymbol{\eta}}(X)\big\|_2\Big]
\end{aligned}
$$

□

B. Proof of Theorem 4

Proof. By Theorem 1, we have

$$
\text{regret}_D^{\mathbf{L}}[\,\widehat{h}\,] \le 2\max_{y}\big\|\boldsymbol{\ell}_y\big\|_2\cdot\big\|\mathbf{C}^{-1}\big\|_2\cdot\mathbf{E}_X\Big[\big\|\widehat{\widetilde{\boldsymbol{\eta}}}(X)-\widetilde{\boldsymbol{\eta}}(X)\big\|_2\Big]. \tag{3}
$$

Then, since ψ is s-strongly proper composite with link function $\boldsymbol{\lambda}$, we have

$$
\begin{aligned}
&\text{regret}_{\widetilde{D}}^{\psi}[\,\widehat{\widetilde{\mathbf{f}}}\,] \\
&= \mathbf{E}_X\Big[\mathbf{E}_{Y|X\sim\widetilde{\boldsymbol{\eta}}(X)}\big[\psi(Y,\widehat{\widetilde{\mathbf{f}}}(X))\big] - \inf_{\mathbf{u}\in\mathbb{R}^{n-1}}\mathbf{E}_{Y|X\sim\widetilde{\boldsymbol{\eta}}(X)}\big[\psi(Y,\mathbf{u})\big]\Big] \\
&= \mathbf{E}_X\Big[\mathbf{E}_{Y|X\sim\widetilde{\boldsymbol{\eta}}(X)}\big[\psi(Y,\widehat{\widetilde{\mathbf{f}}}(X)) - \psi(Y,\boldsymbol{\lambda}(\widetilde{\boldsymbol{\eta}}(X)))\big]\Big] \\
&\qquad \text{(by definition of strongly proper composite multiclass loss)} \\
&\ge \mathbf{E}_X\Big[\frac{s}{2}\big\|\boldsymbol{\lambda}^{-1}(\widehat{\widetilde{\mathbf{f}}}(X))-\widetilde{\boldsymbol{\eta}}(X)\big\|_2^2\Big] \\
&= \frac{s}{2}\mathbf{E}_X\Big[\big\|\widehat{\widetilde{\boldsymbol{\eta}}}(X)-\widetilde{\boldsymbol{\eta}}(X)\big\|_2^2\Big]
\end{aligned} \tag{4}
$$

Combining Eqs. (3, 4), and applying Jensen's inequality (to the convex function $x\mapsto x^2$) establishes the result.

□

C. Proof of Lemma 3

Proof. We will show for all $\mathbf{p} \in \Delta_n$ and $\mathbf{u} \in \mathbb{R}^{n-1}$,

$$\mathbf{E}_{Y\sim\mathbf{p}}\Big[\psi_{\text{mlog}}(Y,\mathbf{u}) - \psi_{\text{mlog}}(Y,\boldsymbol{\lambda}_{\text{mlog}}(\mathbf{p}))\Big] \geq \frac{1}{2}\big\|\boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u}) - \mathbf{p}\big\|_2^2 .$$

Fix $\mathbf{p} \in \Delta_n$ and $\mathbf{u} \in \mathbb{R}^{n-1}$. Then

$$\begin{aligned}
&\mathbf{E}_{Y\sim\mathbf{p}}\Big[\psi_{\text{mlog}}(Y,\mathbf{u}) - \psi_{\text{mlog}}(Y,\boldsymbol{\lambda}_{\text{mlog}}(\mathbf{p}))\Big] \\
&= -\sum_{i\in[n]} p_i \ln\left((\boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u}))_i\right) + \sum_{i\in[n]} p_i \ln(p_i) \\
&= \sum_{i\in[n]} p_i \ln\left(\frac{p_i}{(\boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u}))_i}\right) \\
&= D_{KL}(\mathbf{p}||\boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u})) \quad \text{by the definition of Kullback-Leibler divergence} \\
&\geq \frac{1}{2}\big\|\mathbf{p} - \boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u})\big\|_1^2 \quad \text{using Pinsker's inequality and properties of total variation distance} \\
&\geq \frac{1}{2}\big\|\mathbf{p} - \boldsymbol{\lambda}_{\text{mlog}}^{-1}(\mathbf{u})\big\|_2^2 .
\end{aligned}$$

□

D. Proof of Theorem 5

Proof. **Part 1 (Sufficiency).**

Suppose $\mathbf{C}$ satisfies the given sufficient condition, i.e. that

$$\gamma_{\widetilde{y},\widetilde{y}} > \gamma_{y,\widetilde{y}} \;\; \forall y \neq \widetilde{y} .$$

We will show that

$$\operatorname*{argmax}_x \eta_y(x) = \operatorname*{argmax}_x \widetilde{\eta}_y(x) \;\; \forall y \in [n] ;$$

the claim will then follow.

Fix any class $y \in [n]$.

First, suppose $x' \in \operatorname{argmax}_x \eta_y(x)$. Then by assumption (A), it must be the case that $\eta_y(x') = 1$, i.e. that $\boldsymbol{\eta}(x') = \mathbf{e}_y$. This gives

$$\widetilde{\eta}_y(x') = (\mathbf{C}^\top \boldsymbol{\eta}(x'))_y = (\mathbf{C}^\top \mathbf{e}_y)_y = \gamma_{y,y} .$$

Now for any $x \in \mathcal{X}$, we have

$$\widetilde{\eta}_y(x) = (\mathbf{C}^\top \boldsymbol{\eta}(x))_y = \sum_{y'=1}^{n} \gamma_{y',y}\eta_{y'}(x) \leq \sum_{y'=1}^{n} \gamma_{y,y}\eta_{y'}(x) = \gamma_{y,y} = \widetilde{\eta}_y(x') .$$

Thus $x' \in \operatorname{argmax}_x \widetilde{\eta}_y(x)$. This establishes $\operatorname{argmax}_x \eta_y(x) \subseteq \operatorname{argmax}_x \widetilde{\eta}_y(x)$.

Conversely, suppose $x' \in \operatorname{argmax}_x \widetilde{\eta}_y(x) = \operatorname{argmax}_x (\mathbf{C}^\top \boldsymbol{\eta}(x))_y$. This means

$$\sum_{y'=1}^{n} \gamma_{y',y}\eta_{y'}(x') \geq \sum_{y'=1}^{n} \gamma_{y',y}\eta_{y'}(x) \quad \forall x \in \mathcal{X} .$$

By assumption (A), there exists $\bar{x}^y \in \mathcal{X}$ such that $\boldsymbol{\eta}(\bar{x}^y) = \mathbf{e}_y$. Applying the above inequality to $x = \bar{x}^y$, we have

$$\sum_{y'=1}^{n} \gamma_{y',y}\eta_{y'}(x') \geq \sum_{y'=1}^{n} \gamma_{y',y}\eta_{y'}(\bar{x}^y) = \gamma_{y,y} .$$

Moreover, we have

$$\sum_{y'=1}^{n} \gamma_{y',y} \eta_{y'}(x') \leq \gamma_{y,y} \,.$$

Combining the above two inequalities, we get

$$\sum_{y'=1}^{n} \gamma_{y',y} \eta_{y'}(x') = \gamma_{y,y} \,.$$

Since $\gamma_{y',y} < \gamma_{y,y}$ for all $y' \neq y$, this means we must have $\boldsymbol{\eta}(x') = \mathbf{e}_y$. Thus, $x' \in \operatorname{argmax}_x \eta_y(x)$. This establishes $\operatorname{argmax}_x \widetilde{\eta}_y(x) \subseteq \operatorname{argmax}_x \eta_y(x)$.

Part 2 (Necessity).

Suppose that $\mathbf{C}$ fails to satisfy the given necessary condition, i.e. that there exist $y \neq \widetilde{y}$ such that

$$\gamma_{\widetilde{y},\widetilde{y}} < \gamma_{y,\widetilde{y}} \,.$$

We will show that $\operatorname{argmax}_x \eta_{\widetilde{y}}(x) \neq \operatorname{argmax}_x \widetilde{\eta}_{\widetilde{y}}(x)$.

We give a proof by contradiction. In particular, let if possible $\operatorname{argmax}_x \eta_{\widetilde{y}}(x) = \operatorname{argmax}_x \widetilde{\eta}_{\widetilde{y}}(x) = \operatorname{argmax}_x (\mathbf{C}^\top \boldsymbol{\eta}(x))_{\widetilde{y}}$.

By assumption (A), there exists $\bar{x}^{\widetilde{y}} \in \mathcal{X}$ such that $\boldsymbol{\eta}(\bar{x}^{\widetilde{y}}) = \mathbf{e}_{\widetilde{y}}$, so this means $\bar{x}^{\widetilde{y}} \in \operatorname{argmax}_x \eta_{\widetilde{y}}(x) = \operatorname{argmax}_x \widetilde{\eta}_{\widetilde{y}}(x) = \operatorname{argmax}_x (\mathbf{C}^\top \boldsymbol{\eta}(x))_{\widetilde{y}}$. This means

$$\gamma_{\widetilde{y},\widetilde{y}} = \sum_{y'=1}^{n} \gamma_{y',\widetilde{y}} \eta_{y'}(\bar{x}^{\widetilde{y}}) \geq \sum_{y'=1}^{n} \gamma_{y',\widetilde{y}} \eta_{y'}(x) \quad \forall x \in \mathcal{X} \,.$$

But by assumption (A), we can also find $\bar{x}^{y} \in \mathcal{X}$ such that $\boldsymbol{\eta}(\bar{x}^{y}) = \mathbf{e}_y$. Applying the above inequality to $x = \bar{x}^y$ then gives

$$\gamma_{\widetilde{y},\widetilde{y}} \geq \sum_{y'=1}^{n} \gamma_{y',\widetilde{y}} \eta_{y'}(\bar{x}^{y}) = \gamma_{y,\widetilde{y}} \,,$$

contradicting our assumption. Therefore, we must have $\operatorname{argmax}_x \eta_{\widetilde{y}}(x) \neq \operatorname{argmax}_x \widetilde{\eta}_{\widetilde{y}}(x)$. □

E. Additional Experimental Details

Table 3. Details of MNIST and CIFAR10 data sets.

Data set	# train	# test	# classes (n)	# features (d)
MNIST	60,000	10,000	10	784
CIFAR10	50,000	10,000	10	3072

For MNIST, the asymmetric noise matrix $\mathbf{C}^{\text{MNIST}(\gamma)}$ includes the following label noise transitions: $2 \to 7$, $3 \to 8$, $5 \leftrightarrow 6$, $7 \to 1$. Following Patrini et al. (2017), features were normalized to $[0, 1]$, and two fully connected hidden layers of size 128 were trained, with ReLU activation and dropout rate 0.2.[13]

For CIFAR10, the asymmetric noise matrix $\mathbf{C}^{\text{CIFAR10}(\gamma)}$ includes the following label noise transitions: `Truck` $\to$ `Automobile`, `Bird` $\to$ `Airplane`, `Deer` $\to$ `Horse`, `Cat` $\leftrightarrow$ `Dog`. Again following Patrini et al. (2017), per-pixel mean subtraction and data augmentation were performed, and a 14-layer residual network (ResNet) (He et al., 2016) was trained.[14]

[13]Batch size was 32. AdaGrad (Duchi et al., 2010) was run for 40 epochs with default parameters.

[14]Batch size was 32. SGD was run for 120 epochs with momentum 0.9 and learning rate set to 0.1 initially and divided by 10 after 40 and 80 epochs; weight decay was 10^{-4}.