Learning with Bounded Instance- and Label-Dependent Label Noise Supplementary Material

A. Proofs

A.1. Proof of Theorem 1

Proof. $\forall \mathbf{x} \in \text{supp}(P_{D^*}(\mathbf{X})) = \text{supp}(P_D(\mathbf{X}))$, we have

$$\begin{aligned} g^*_{D^*}(\mathbf{x}) =& \text{sgn}\left(P_{D^*}(Y=+1|\mathbf{x}) - \frac{1}{2}\right) \\ =& \text{sgn}\left(\mathbb{1}[g^*_D(\mathbf{x}) = +1] - \frac{1}{2}\right) \\ =& g^*_D(\mathbf{x}), \end{aligned}$$

where the last equality is justified by checking the possible binary values of $g^*_D(\mathbf{x})$, e.g., when $g^*_D(\mathbf{x}) = +1$, $\text{sgn}(\mathbb{1}[g^*_D(\mathbf{x}) = +1] - \frac{1}{2}) = +1$; when $g^*_D(\mathbf{x}) = -1$, $\text{sgn}(\mathbb{1}[g^*_D(\mathbf{x}) = +1] - \frac{1}{2}) = -1$. □

A.2. Proof of Theorem 2 and Corollary 1

Proof. $\forall \mathbf{x} \in \mathcal{X}$, $\tilde{\eta}(\mathbf{x})$ can be rewritten as

$$\begin{aligned} \tilde{\eta}(\mathbf{x}) =& P(\widetilde{Y} = +1, Y = +1|\mathbf{x}) \\ &+ P(\widetilde{Y} = +1, Y = -1|\mathbf{x}) \\ =& P(\widetilde{Y} = +1|Y = +1, \mathbf{x})P(Y = +1|\mathbf{x}) \\ &+ P(\widetilde{Y} = +1|Y = -1, \mathbf{x})P(Y = -1|\mathbf{x}) \\ =& (1 - \rho_{+1}(\mathbf{x}))\eta(\mathbf{x}) + \rho_{-1}(\mathbf{x})(1 - \eta(\mathbf{x})) \end{aligned}$$

Then, we have

$$\begin{aligned} \eta(\mathbf{x}) \geq \frac{1}{2} \implies \tilde{\eta}(\mathbf{x}) =& (1 - \rho_{+1}(\mathbf{x}))\eta(\mathbf{x}) \\ &+ \rho_{-1}(\mathbf{x})(1 - \eta(\mathbf{x})) \\ \geq& (1 - \rho_{+1}(\mathbf{x}))\eta(\mathbf{x}) \\ \geq& \frac{1 - UB(\rho_{+1}(\mathbf{x}))}{2} \end{aligned}$$

and its contrapositive

$$\begin{aligned} \tilde{\eta}(\mathbf{x}) < \frac{1 - UB(\rho_{+1}(\mathbf{x}))}{2} &\implies \eta(\mathbf{x}) < \frac{1}{2} \\ &\implies g^*_D(\mathbf{x}) = -1 \end{aligned}$$

The last step follows by Lemma 1. Similarly, we can prove $\tilde{\eta}(\mathbf{x}) > \frac{1+UB(\rho_{-1}(\mathbf{x}))}{2} \implies g^*_D(\mathbf{x}) = +1$.

Corollary 1 holds by replacing $UB(\rho_{+1}(\mathbf{x}))$ and $UB(\rho_{-1}(\mathbf{x}))$ by $\rho_{+1\max}$ and $\rho_{-1\max}$, respectively. □

A.3. Proof of Propositions 1 and 2

A.3.1. Proposition 1

Proof. The following Lemma holds because of the basic Rademacher bound on the maximal deviation between the expected and empirical risks (Bartlett & Mendelson, 2002).

Lemma A1. *For any $\delta > 0$, with probability at least $1 - \delta$, we have*

$$\sup_{f \in \mathcal{F}} \left|\widehat{R}_{D^*,L}(f) - R_{D^*,L}(f)\right| \leq \mathfrak{R}(L \circ \mathcal{F}) + b\sqrt{\frac{log(1/\delta)}{2m}}.$$

Then, with probability at least $1 - \delta$, we have

$$\begin{aligned} & R_{D^*,L}(\hat{f}_{D^*,L}) - R_{D^*,L}(f^*_{D^*,L}) \\ =& (R_{D^*,L}(\hat{f}_{D^*,L}) - \widehat{R}_{D^*,L}(\hat{f}_{D^*,L})) \\ &+ (\widehat{R}_{D^*,L}(f^*_{D^*,L}) - R_{D^*,L}(f^*_{D^*,L})) \\ &+ (\widehat{R}_{D^*,L}(\hat{f}_{D^*,L}) - \widehat{R}_{D^*,L}(f^*_{D^*,L})) \\ \leq& 2 \sup_{f \in \mathcal{F}} \left|\widehat{R}_{D^*,L}(f) - R_{D^*,L}(f)\right| \\ \leq& 2\mathfrak{R}(L \circ \mathcal{F}) + 2b\sqrt{\frac{log(1/\delta)}{2m}}, \end{aligned}$$

where the first inequality holds because $\hat{f}_{D^*,L} = \arg\min_{f \in \mathcal{F}} \widehat{R}_{D^*,L}(f)$ and the second inequality follows by Lemma A1. □

A.3.2. Proposition 2

Proof. Notice that $R_{D,L}(f) = R_{D^*,\beta L}(f)$, then the proof is similar with the proof of Proposition 1. □

A.4. Proof of Theorem 3

Proof. $\forall \mathbf{x} \in \mathcal{X}$, we have

$$\begin{aligned} \tilde{\eta}(\mathbf{x}) =& (1 - \rho_{+1}(\mathbf{x}))\eta(\mathbf{x}) + (1 - \eta(\mathbf{x}))\rho_{-1}(\mathbf{x}) \\ =& (1 - \rho_{+1}(\mathbf{x}) - \rho_{-1}(\mathbf{x}))\eta(\mathbf{x}) + \rho_{-1}(\mathbf{x}) \\ \geq& \rho_{-1}(\mathbf{x}), \end{aligned}$$

where the first equality has been derived in the proof of Theorem 2 and the inequality follows by our bounded total noise assumption $0 \leq \rho_{+1}(\mathbf{x}) + \rho_{-1}(\mathbf{x}) < 1$. Similarly, we can prove $\rho_{+1}(\mathbf{x}) \leq 1 - \tilde{\eta}(\mathbf{x})$. □

B. Extension to the Multiclass Classification

By the one-vs.-all strategy, our algorithm can be easily adapted for multiclass classification. In the multi-class case, our Theorem 1 still holds and keeps the idea of learning with distilled examples justified. An example $(\mathbf{x}, y)$ is distilled if $g_D^*(\mathbf{x}) = y$, where $g_D^*(\mathbf{x}) = \arg\max_i P_D(Y = i|\mathbf{x})$ is the Bayes optimal classifier under D. Like in the binary case, ILN can be modeled by flip rates $\rho_y(\mathbf{x}) = P(\widetilde{Y} \neq y|\mathbf{x}, Y = y)$ and $\rho_{-y}(\mathbf{x}) = P(\widetilde{Y} = y|\mathbf{x}, Y \neq y)$. Let $\eta_y(\mathbf{x}) = P(Y = y|\mathbf{x})$ and $\tilde{\eta}_y(\mathbf{x}) = P(\widetilde{Y} = y|\mathbf{x})$. Easy to derive that $\tilde{\eta}_y(\mathbf{x}) > \frac{1+UB(\rho_{-y}(\mathbf{x}))}{2} \implies \eta_y(\mathbf{x}) > \frac{1}{2} \implies (\mathbf{x}, y)$ is distilled. Hence, distilled examples can be collected out of noisy examples by thresholding $\tilde{\eta}_y(\mathbf{x})$. Other parts of our algorithm can be performed without special adaptations.

References

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research (JMLR)*, 3(Nov):463–482, 2002.