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

A. Proofs

A.1. Proof of Theorem 1

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

\[
g_D^*(x) = \text{sgn} \left( \frac{g_D(x)}{\|g_D(x)\|_\infty} \right)
\]

where the last equality is justified by checking the possible binary values of \( g_D^*(x) \), e.g., when \( g_D^*(x) = +1 \), \( \text{sgn}(1) = +1 \); when \( g_D^*(x) = -1 \), \( \text{sgn}(1) = -1 \).

A.2. Proof of Theorem 2 and Corollary 1

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

\[
\tilde{\eta}(x) = P(\tilde{Y} = +1, Y = +1|x) \\
+ P(\tilde{Y} = +1, Y = -1|x)
\]

Then, we have

\[
\eta(x) \geq \frac{1}{2} \implies \tilde{\eta}(x) = (1 - \rho_1(x))\eta(x) \\
+ \rho_1(x)(1 - \eta(x))
\]

and its contrapositive

\[
\tilde{\eta}(x) < \frac{1 - UB(\rho_1(x))}{2} \implies \eta(x) < \frac{1}{2} \\
\Rightarrow g_D^*(x) = -1
\]

The last step follows by Lemma 1. Similarly, we can prove

\[
\tilde{\eta}(x) > \frac{1 + UB(\rho_1(x))}{2} \implies g_D^*(x) = +1.
\]

Corollary 1 holds by replacing \( UB(\rho_1(x)) \) and \( UB(\rho_1(x)) \) by \( \rho_{1max} \) and \( -\rho_{1max} \), 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

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

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

\[
R_{D^*,L}(\hat{f}_{D^*,L}) - R_{D^*,L}(f_{D^*,L}) \\
= (R_{D^*,L}(\hat{f}_{D^*,L}) - \hat{R}_{D^*,L}(\hat{f}_{D^*,L})) \\
+ (\hat{R}_{D^*,L}(\hat{f}_{D^*,L}) - R_{D^*,L}(f_{D^*,L})) \\
\leq 2 \sup_{f \in \mathcal{F}} |\hat{R}_{D^*,L}(f) - R_{D^*,L}(f)| \\
\leq 2\mathcal{R}(L \circ \mathcal{F}) + 2b\sqrt{\frac{\log(1/\delta)}{2m}},
\]

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

A.3.2. Proposition 2

Proof. Notice that \( R_{D,L}(f) = R_{D^*,3L}(f) \), then the proof is similar with the proof of Proposition 1.

A.4. Proof of Theorem 3

Proof. \( \forall x \in \mathcal{X} \), we have

\[
\tilde{\eta}(x) = (1 - \rho_1(x))\eta(x) + (1 - \eta(x))\rho_1(x) \\
= (1 - \rho_1(x) - \rho_1(x))\eta(x) + \rho_1(x) \\
\geq \rho_1(x),
\]

where the first equality has been derived in the proof of Theorem 2 and the inequality follows by our bounded total noise assumption \( 0 \leq \rho_1(x) + \rho_{-1}(x) < 1 \). Similarly, we can prove \( \rho_1(x) \leq 1 - \tilde{\eta}(x) \).
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 \((x, y)\) is distilled if \(g_D^*(x) = y\), where \(g_D^*(x) = \arg \max_i P_D(Y = i|x)\) is the Bayes optimal classifier under \(D\). Like in the binary case, ILN can be modeled by flip rates \(\rho_y(x) = P(\hat{Y} \neq y|x, Y = y)\) and \(\rho_{\neg y}(x) = P(\hat{Y} = y|x, Y \neq y)\). Let \(\eta_y(x) = P(Y = y|x)\) and \(\tilde{\eta}_y(x) = P(\hat{Y} = y|x)\). Easy to derive that \(\tilde{\eta}_y(x) > \frac{1+UB(\rho_{\neg y}(x))}{2} \implies \eta_y(x) > \frac{1}{2} \implies (x, y)\) is distilled. Hence, distilled examples can be collected out of noisy examples by thresholding \(\tilde{\eta}_y(x)\). 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.