Appendix

A. Proof for Lemmas

A.1. Proof for Lemma 1

Proof. We try to build the connection between noisy distribution $\widetilde{\mathcal{D}}$ and the underlying Bayes optimal distribution $\mathcal{D}^*$ by the noise rates e_+ and e_-. The primary difference from the proof of Lemma 2 in [24] is the usage of $\mathcal{D}^*$. Note:

$$
\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(f(X),\widetilde{Y})] \\
=&\mathbb{E}_{\mathcal{D}^*}\left[\sum_{j\in\{-1,+1\}}\mathbb{P}(\widetilde{Y}=j|X,Y^*)\ell(f(X),j)\right] \\
=&\mathbb{E}_{\mathcal{D}^*}\left[\sum_{j\in\{-1,+1\}}\mathbb{P}(\widetilde{Y}=j|Y^*)\ell(f(X),j)\right] \\
=&\sum_{i\in\{-1,+1\}}\mathbb{P}(Y^*=i)\mathbb{E}_{\mathcal{D}^*|Y^*=i}[\mathbb{P}(\widetilde{Y}=+1|Y^*=i)\ell(f(X),+1)+\mathbb{P}(\widetilde{Y}=-1|Y^*=i)\ell(f(X),-1)] \\
=&\mathbb{P}(Y^*=+1)\mathbb{E}_{\mathcal{D}^*|Y^*=+1}[(1-e_+)\ell(f(X),+1)+e_+\ell(f(X),-1)] \\
&+\mathbb{P}(Y^*=-1)\mathbb{E}_{\mathcal{D}^*|Y^*=-1}[(1-e_-)\ell(f(X),-1)+e_-\ell(f(X),+1)].
\end{aligned}
$$

Similarly, following the proof of Lemma 2 in [24], we can prove this lemma. □

A.2. Proof for Lemma 2

Peer Loss on the Bayes Optimal Distribution Recall our goal is to learn a classifier f from the noisy distribution $\widetilde{\mathcal{D}}$ which also minimizes the loss on the corresponding Bayes optimal distribution D^*, i.e. $\mathbb{E}[\mathbb{1}(f(X),Y^*)],(X,Y^*)\sim\mathcal{D}^*$. Before considering the case with label noise, we need to prove peer loss functions induce the Bayes optimal classifier when minimizing the 0-1 loss on $\mathcal{D}^*$ as in Lemma 2.

Lemma 2. *Given the Bayes optimal distribution $\mathcal{D}^*$, the optimal peer classifier defined below:*

$$f^*_{peer}=\arg\min_f\ \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{PL}(f(X),Y^*)]$$

also minimizes $\mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X),Y^*)]$.

See the proof below. It has been shown in [24] that Lemma 2 holds for the clean distribution $\mathcal{D}$ when the clean dataset is class-balanced, i.e. $\mathbb{P}(Y=-1)=\mathbb{P}(Y=+1)=0.5$. For the Bayes optimal distribution $\mathcal{D}^*$, as shown in Lemma 2, there is *no requirement* for the prior $p^*:=\mathbb{P}(Y^*=+1)$.

Proof. Recall Y^* is the Bayes optimal label defined as

$$Y^*|X:=\arg\max_Y\ \mathbb{P}(Y|X),(X,Y)\sim\mathcal{D}.$$

We need to prove that the "optimal peer classifier" defined below:

$$f^*_{\text{peer}}=\arg\min_f\ \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f(X),Y^*)]$$

is the same as the Bayes optimal classifier f^*. To see this, suppose the claim is wrong. Denote by (notations ϵ_+ and ϵ_- are defined only for this proof):

$$\epsilon_+:=\mathbb{P}(f^*_{\text{peer}}(X)=-1|f^*(X)=+1),\ \ \epsilon_-:=\mathbb{P}(f^*_{\text{peer}}(X)=+1|f^*(X)=-1)$$

and denote by $p^* := \mathbb{P}(f^*(X) = +1)$. Then

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f^*_{\text{peer}}(X), Y^*)] \\
&= \mathbb{P}(f^*_{\text{peer}}(X) \neq Y^*) - p^* \cdot \mathbb{P}(f^*_{\text{peer}}(X) \neq +1) - (1-p^*) \cdot \mathbb{P}(f^*_{\text{peer}}(X) \neq -1) \\
&= p^* \cdot \epsilon_+ + (1-p^*) \cdot \epsilon_- - p^* \cdot \mathbb{P}(f^*_{\text{peer}}(X) \neq +1) - (1-p^*) \cdot \mathbb{P}(f^*_{\text{peer}}(X) \neq -1) \\
&= p^* \cdot \epsilon_+ + (1-p^*) \cdot \epsilon_- \\
&- p^* \cdot \big(\mathbb{P}(f^*_{\text{peer}}(X) \neq +1 | f^*(X) \neq +1)\mathbb{P}(f^*(X) \neq +1) + \mathbb{P}(f^*_{\text{peer}}(X) \neq +1 | f^*(X) \neq -1)\mathbb{P}(f^*(X) \neq -1)\big) \\
&- (1-p^*) \cdot \big(\mathbb{P}(f^*_{\text{peer}}(X) \neq -1 | f^*(X) \neq +1)\mathbb{P}(f^*(X) \neq +1) + \mathbb{P}(f^*_{\text{peer}}(X) \neq -1 | f^*(X) \neq -1)\mathbb{P}(f^*(X) \neq -1)\big) \\
&= p^* \cdot \epsilon_+ + (1-p^*) \cdot \epsilon_- \\
&- p^* \cdot \mathbb{P}(f^*(X) \neq +1)(1-\epsilon_-) - p^* \cdot \mathbb{P}(f^*(X) \neq -1) \cdot \epsilon_+ \\
&- (1-p^*) \cdot \mathbb{P}(f^*(X) \neq -1)(1-\epsilon_+) - (1-p^*) \cdot \mathbb{P}(f^*(X) \neq +1) \cdot \epsilon_- \\
&= 0 - p^* \cdot \mathbb{P}(f^*(X) \neq +1) - (1-p^*) \cdot \mathbb{P}(f^*(X) \neq -1) \\
&+ p^*(\epsilon_+ + \mathbb{P}(f^*(X) \neq +1)\epsilon_- - \mathbb{P}(f^*(X) \neq -1)\epsilon_+) \\
&+ (1-p^*)(\epsilon_- + \mathbb{P}(f^*(X) \neq -1)\epsilon_+ - \mathbb{P}(f^*(X) \neq +1)\epsilon_-) \\
&> 0 - p^* \cdot \mathbb{P}(f^*(X) \neq +1) - (1-p^*) \cdot \mathbb{P}(f^*(X) \neq -1) \\
&= \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f^*(X), Y^*)]
\end{aligned}
$$

contradicting the optimality of f^*_{peer}. Thus our claim is proved. □

B. Proof for Theorems

B.1. Proof for Theorem 1

Proof. The covariance $\text{Cov}(\cdot, \cdot)$ in this proof is taken over the Bayes optimal distribution $\mathcal{D}^*$. The following proof is built on the result of Theorem 2, i.e. Eq. (2). First note

$$
\begin{aligned}
\text{Cov}(Z_1(X), \mathbb{1}(f_1(X), Y^*) - \mathbb{1}(f_2(X), Y^*)) &= \mathbb{E}[(Z_1(X) - \mathbb{E}[Z_1(X)]) \cdot (\mathbb{1}(f_1(X), Y^*) - \mathbb{1}(f_2(X), Y^*))] \\
&\leq \mathbb{E}[|(Z_1(X) - \mathbb{E}[Z_1(X)]|] \\
&\leq \mathbb{E}|e_+(X) - \mathbb{E}[e_+(X)]| + \mathbb{E}|e_-(X) - \mathbb{E}[e_-(X)]|
\end{aligned}
$$

Similarly, one can show that

$$
\text{Cov}(Z_2(X), \mathbb{1}(f_1(X), -1) - \mathbb{1}(f_2(X), -1)) \leq \mathbb{E}|e_+(X) - \mathbb{E}[e_+(X)]| + \mathbb{E}|e_-(X) - \mathbb{E}[e_-(X)]|
$$

Now with bounded variance in the error rates, suppose:

$$
\mathbb{E}|e_+(X) - \mathbb{E}[e_+(X)]| \leq \epsilon_+, \quad \mathbb{E}|e_-(X) - \mathbb{E}[e_-(X)]| \leq \epsilon_-
$$

Note

$$
\begin{aligned}
\tilde{f}^*_{\text{peer}} &:= \arg\min_f \mathbb{E}_{\widetilde{\mathcal{D}}}\left[\mathbb{1}_{\text{PL}}(f(X), \tilde{Y})\right] \\
&= \arg\min_f \left[(1 - e_+ - e_-)\mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f(X), Y^*) + \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1))\right] \\
&= \arg\min_f \big[(1 - e_+ - e_-)\left(\mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), Y^*) - p^* \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), +1)] - (1-p^*) \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), -1)]\right) \\
&\qquad + \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1))\big].
\end{aligned}
$$

Then

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)\right] + \frac{\text{Cov}(Z_1(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1))}{1 - e_+ - e_-} \\
=&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)\right] - 0.5 \cdot \mathbb{E}_X[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), +1)] - 0.5 \cdot \mathbb{E}_X[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1)] + 0.5 \\
&+ \frac{\text{Cov}(Z_1(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1))}{1 - e_+ - e_-} \\
\leq&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)\right] - p^* \cdot \mathbb{E}_X[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), +1)] - (1 - p^*) \cdot \mathbb{E}_X[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1)] + |p^* - 0.5| + 0.5 \\
&+ \frac{\text{Cov}(Z_1(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1))}{1 - e_+ - e_-} \\
\leq&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(f^*(X), Y^*)\right] - p^* \cdot \mathbb{E}_X[\mathbb{1}(f^*(X), +1)] - (1 - p^*) \cdot \mathbb{E}_X[\mathbb{1}(f^*(X), -1)] + |p^* - 0.5| + 0.5 \\
&+ \frac{\text{Cov}(Z_1(X), \mathbb{1}(f^*(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f^*(X), -1))}{1 - e_+ - e_-} \\
\leq&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(f^*(X), Y^*)\right] + \frac{\text{Cov}(Z_1(X), \mathbb{1}(f^*(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f^*(X), -1))}{1 - e_+ - e_-} + 2|p^* - 0.5|.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}^*}\left[\mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*) - \mathbb{1}(f^*(X), Y^*)\right] \\
\leq&\frac{\text{Cov}(Z_1(X), \mathbb{1}(f^*(X), Y^*) - \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f^*(X), -1) - \mathbb{1}(\tilde{f}^*_{\text{peer}}(X), -1))}{1 - e_+ - e_-} + 2|p^* - 0.5| \\
\leq&2\frac{\mathbb{E}|e_+(X) - \mathbb{E}[e_+(X)]| + \mathbb{E}|e_-(X) - \mathbb{E}[e_-(X)]|}{1 - e_+ - e_-} + 2|p^* - 0.5| \\
\leq&\frac{2(\epsilon_+ + \epsilon_-)}{1 - e_+ - e_-} + 2|p^* - 0.5|.
\end{aligned}
$$

Noting $\mathbb{1}(f^*(X), Y^*) = 0$, we finish the proof.

□

B.2. Proof for Theorem 2

Proof. The covariance $\text{Cov}(\cdot, \cdot)$ in this proof is taken over the Bayes optimal distribution $\mathcal{D}^*$. Recall

$$
e_+(X) := \mathbb{P}(\widetilde{Y} = -1 | Y^* = +1, X), e_-(X) := \mathbb{P}(\widetilde{Y} = +1 | Y^* = -1, X)
$$

and

$$
e_+ := \mathbb{E}_X[e_+(X)], \;\; e_- := \mathbb{E}_X[e_-(X)]
$$

We first have the following equality:

$$
\begin{aligned}
\mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbb{1}_{\text{PL}}(f(X), \tilde{Y})] = &\mathbb{E}_{\mathcal{D}^*}[(1 - e_+(X) - e_-(X))\mathbb{1}(f(X), Y^*))] && \text{(Term-A)} \\
&+ \mathbb{E}_X[e_+(X)\mathbb{1}(f(X), -1) + e_-(X)\mathbb{1}(f(X), +1)] && \text{(Term-B)} \\
&- (1 - e_+ - e_-) \cdot \mathbb{E}_{D^*}[\mathbb{1}(f(X), Y^*_p))] && \text{(Term-C)} \\
&- \mathbb{E}_X[e_+ \cdot \mathbb{1}(f(X), -1) + e_- \cdot \mathbb{1}(f(X), +1)] && \text{(Term-D)}
\end{aligned}
$$

Term-B can be transformed to:

$$
\begin{aligned}
&\mathbb{E}_X[e_+(X) \cdot \mathbb{1}(f(X), -1) + e_-(X) \cdot \mathbb{1}(f(X), +1)] \\
&= \mathbb{E}_X[e_+(X) \cdot \mathbb{1}(f(X), -1) + e_-(X) \cdot (1 - \mathbb{1}(f(X), -1))] \\
&= \mathbb{E}_X[(e_+(X) - e_-(X)) \cdot \mathbb{1}(f(X), -1) + e_-(X)].
\end{aligned}
$$

Similarly, Term-D turns to

$$\mathbb{E}_X[e_+ \cdot \mathbb{1}(f(X), -1) + e_- \cdot \mathbb{1}(f(X), +1)] = (e_+ - e_-) \cdot \mathbb{E}_X[\mathbb{1}(f(X), -1)] + e_-.$$

Define two random variables

$$Z_1(X) := 1 - e_+(X) - e_-(X),\ Z_2(X) = e_+(X) - e_-(X).$$

Then Term-A becomes

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}^*}[(1 - e_+(X) - e_-(X))\mathbb{1}(f(X), Y^*))] \\
&= \mathbb{E}[Z_1(X)] \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), Y^*))] + \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*)) \\
&= (1 - e_+ - e_-) \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), Y^*))] + \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*))
\end{aligned}$$

Similarly, Term-B can be further transformed to

$$\begin{aligned}
&\mathbb{E}_X[(e_+(X) - e_-(X)) \cdot \mathbb{1}(f(X), -1) + e_-(X)] \\
&= \mathbb{E}[Z_2(X)]\mathbb{E}_X[\mathbb{1}(f(X), -1)] + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1)) + e_- \\
&= (e_+ - e_-)\mathbb{E}_X[\mathbb{1}(f(X), -1)] + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1)) + e_-
\end{aligned}$$

Combining the above results, we have

$$\begin{aligned}
\mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbb{1}_{\text{PL}}(f(X), \tilde{Y})] &= (1 - e_+ - e_-) \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), Y^*))] \\
&+ (e_+ - e_-)\mathbb{E}_X[\mathbb{1}(f(X), -1)] + e_- \\
&- (1 - e_+ - e_-) \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}(f(X), Y_p^*))] \\
&- (e_+ - e_-) \cdot \mathbb{E}_X[\mathbb{1}(f(X), -1)] - e_- \\
&+ \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y)) + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1)) \\
&= (1 - e_+ - e_-)\mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f(X), Y^*)] \\
&+ \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*)) + \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1))
\end{aligned}$$

□

B.3. Proof for Theorem 3

Proof. From Theorem 2, we know

$$\begin{aligned}
&\mathbb{E}_{\widetilde{\mathcal{D}}}[\mathbb{1}_{\text{PL}}(f(X), \widetilde{Y})] - \text{Cov}(Z_1(X), \mathbb{1}(f(X), Y^*)) - \text{Cov}(Z_2(X), \mathbb{1}(f(X), -1))\Big] \\
=&(1 - e_- - e_+) \cdot \mathbb{E}_{\mathcal{D}^*}[\mathbb{1}_{\text{PL}}(f(X), Y^*)].
\end{aligned}$$

With Lemma 2, we can finish the proof. □

B.4. Proof for Theorem 4

Proof. Recall $\tau \in [0, 1]$ is the expected ratio (a.k.a. probability) of correct examples in $\hat{D}^\tau$, i.e. $\tau = \mathbb{E}[\mathbb{1}\{(X, \hat{Y}) \in \hat{D}^\tau | (X, Y^*) \in D^*\}] = \mathbb{P}((X, \hat{Y}) \sim \hat{\mathcal{D}}^\tau | (X, Y^*) \sim \mathcal{D}^*)$. With $\hat{D}^\tau$, the classifier learned by minimizing the 0-1 CAL loss is

$$\tilde{f}^*_{\text{CAL-}\tau} := \arg\min_f \mathbb{E}_{\widetilde{\mathcal{D}}}\Big[\mathbb{1}_{\text{PL}}(f(X), \widetilde{Y})] - \text{Cov}_{\hat{\mathcal{D}}^\tau}(Z_1(X), \mathbb{1}(f(X), \hat{Y})) - \text{Cov}_{\hat{\mathcal{D}}^\tau}(Z_2(X), \mathbb{1}(f(X), -1))\Big].$$

Note

$$\begin{aligned}
\text{Cov}_{\hat{\mathcal{D}}^\tau}(Z_1(X), \mathbb{1}(f(X), Y)) &= \mathbb{E}_{\hat{\mathcal{D}}^\tau}\left[\left(Z_1(X) - \mathbb{E}_{\hat{\mathcal{D}}^\tau}[Z_1(X)]\right)\left(\mathbb{1}(f(X), Y) - \mathbb{E}_{\hat{\mathcal{D}}^\tau}[\mathbb{1}(f(X), Y)]\right)\right] \\
&= \mathbb{E}_{\hat{\mathcal{D}}^\tau}\left[\left(Z_1(X) - \mathbb{E}_{\hat{\mathcal{D}}^\tau}[Z_1(X)]\right)\mathbb{1}(f(X), Y)\right] \\
&= \mathbb{P}((X, Y) \in D^* | (X, Y) \in \hat{D}^\tau)\mathbb{E}_{\hat{\mathcal{D}}^\tau}\left[\left(Z_1(X) - \mathbb{E}_{\hat{\mathcal{D}}^\tau}[Z_1(X)]\right)\mathbb{1}(f(X), Y) | (X, Y) \in D^*\right] \\
&+ \mathbb{P}((X, Y) \notin D^* | (X, Y) \in \hat{D}^\tau)\mathbb{E}_{\hat{\mathcal{D}}^\tau}\left[\left(Z_1(X) - \mathbb{E}_{\hat{\mathcal{D}}^\tau}[Z_1(X)]\right)\mathbb{1}(f(X), Y) | (X, Y) \notin D^*\right].
\end{aligned}$$

Similarly,

$$\begin{aligned}\mathrm{Cov}_{\mathcal{D}^*}(Z_1(X), \mathbb{1}(f(X),Y)) &= \mathbb{E}_{\mathcal{D}^*}\left[(Z_1(X) - \mathbb{E}_{\mathcal{D}^*}[Z_1(X)])\,\mathbb{1}(f(X),Y)\right] \\ &= \mathbb{P}((X,Y)\in\hat{D}^\tau|(X,Y)\in D^*)\mathbb{E}_{\mathcal{D}^*}\left[(Z_1(X) - \mathbb{E}_{\mathcal{D}^*}[Z_1(X)])\,\mathbb{1}(f(X),Y)|(X,Y)\in\hat{D}^\tau\right] \\ &+ \mathbb{P}((X,Y)\notin\hat{D}^\tau|(X,Y)\in D^*)\mathbb{E}_{\mathcal{D}^*}\left[(Z_1(X) - \mathbb{E}_{\mathcal{D}^*}[Z_1(X)])\,\mathbb{1}(f(X),Y)|(X,Y)\notin\hat{D}^\tau\right].\end{aligned}$$

When D^*, $\hat{D}^\tau$ and $\tilde{D}$ have the same feature set, we have

$$\begin{aligned}\mathbb{P}((X,Y)\in D^*|(X,Y)\in\hat{D}^\tau) &= \mathbb{P}((X,Y)\in\hat{D}^\tau|(X,Y)\in D^*) = \tau, \\ \mathbb{P}((X,Y)\notin D^*|(X,Y)\in\hat{D}^\tau) &= \mathbb{P}((X,Y)\notin\hat{D}^\tau|(X,Y)\in D^*) = 1-\tau.\end{aligned}$$

Therefore,

$$\mathrm{Cov}_{\hat{\mathcal{D}}^\tau}(Z_1(X), \mathbb{1}(f(X),Y)) - \mathrm{Cov}_{\mathcal{D}^*}(Z_1(X), \mathbb{1}(f(X),Y)) \le 2(1-\tau)(\epsilon_+ + \epsilon_-).$$

The rest of the proof can be accomplished by following the proof of Theorem 1. □

C. Proof for Corollaries

C.1. Proof for Corollary 1

Proof.

$$\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell_{\mathrm{PL}}(f(X),\widetilde{Y})] = \mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(f(X),\widetilde{Y})] - \mathbb{E}_{\widetilde{\mathcal{D}}_Y}\left[\mathbb{E}_{\mathcal{D}_X}[\ell(f(X_p),\widetilde{Y}_p)]\right]. \tag{5}$$

The first term in (5) is

$$\begin{aligned}&\mathbb{E}_{\widetilde{\mathcal{D}}}[\ell(f(X),\widetilde{Y})] \\ =&\mathbb{E}_{\mathcal{D}^*}\left[\sum_{j\in[K]}\mathbb{P}(\widetilde{Y}=j|X,Y^*)\ell(f(X),j)\right] \\ =&\sum_{j\in[K]}\sum_{i\in[K]}\mathbb{P}(Y^*=i)\mathbb{E}_{\mathcal{D}^*|Y^*=i}[T_{ij}(X)\ell(f(X),j)] \\ =&\sum_{j\in[K]}\sum_{i\in[K]}\mathbb{P}(Y^*=i)\left[T_{ij}\mathbb{E}_{\mathcal{D}^*|Y^*=i}[\ell(f(X),j)] + \mathrm{Cov}_{\mathcal{D}^*|Y^*=i}[T_{ij}(X),\ell(f(X),j)]\right] \\ =&\sum_{j\in[K]}\left[\mathbb{P}(Y^*=j)\left(1-\sum_{i\neq j,i\in[K]}T_{ji}\right)\mathbb{E}_{\mathcal{D}^*|Y^*=j}\left[\ell(f(X),j)\right] + \sum_{i\in[K],i\neq j}\mathbb{P}(Y^*=i)T_{ij}\mathbb{E}_{\mathcal{D}^*|Y^*=i}\left[\ell(f(X),j)\right]\right] \\ &+\sum_{j\in[K]}\sum_{i\in[K]}P(Y^*=i)\mathrm{Cov}_{\mathcal{D}^*|Y^*=i}\left[T_{ij}(X),\ell(f(X),j)\right] \\ =&\sum_{j\in[K]}\left[\mathbb{P}(Y^*=j)\left(1-\sum_{i\neq j,i\in[K]}e_i\right)\mathbb{E}_{\mathcal{D}^*|Y^*=j}\left[\ell(f(X),j)\right] + \sum_{i\in[K],i\neq j}\mathbb{P}(Y^*=i)e_j\mathbb{E}_{\mathcal{D}^*|Y^*=i}\left[\ell(f(X),j)\right]\right] \\ &+\sum_{j\in[K]}\sum_{i\in[K]}P(Y^*=i)\mathrm{Cov}_{\mathcal{D}^*|Y^*=i}\left[T_{ij}(X),\ell(f(X),j)\right] \\ =&\left(1-\sum_{i\in[K]}e_i\right)\mathbb{E}_{\mathcal{D}^*}\left[\ell(f(X),Y^*)\right] + \sum_{j\in[K]}\sum_{i\in[K]}\mathbb{P}(Y^*=i)e_j\mathbb{E}_{\mathcal{D}^*|Y^*=i}\left[\ell(f(X),j)\right] \\ &+\sum_{j\in[K]}\sum_{i\in[K]}P(Y^*=i)\mathrm{Cov}_{\mathcal{D}^*|Y^*=i}\left[T_{ij}(X),\ell(f(X),j)\right]\end{aligned}$$

The rest of proofs can be done following standard multi-class peer loss derivations [24].

□

D. More Discussions

D.1. Setting Thresholds $L_{\min}$ and $L_{\max}$

In a high level, there are two strategies for setting $L_{\min}$ and $L_{\max}$: 1) $L_{\min} < L_{\max}$ and 2) $L_{\min} = L_{\max}$.

Strategy-1: $L_{\min} < L_{\max}$: This strategy may provide a higher ratio of true Bayes optimal labels among feasible examples in $\hat{D}$ since some ambiguous examples are dropped. However, dropping examples changes the distribution of X (as well as the distribution of the unobservable Y^*), a.k.a. covariate shift [14, 6]. Importance re-weighting with weight $\gamma(X)$ is necessary for correcting the covariate shift, i.e. the weight of each feasible example $(x, \hat{y}) \in \hat{D}$ should be changed from 1 to $\gamma(x)$. Let $\mathcal{D}_X$ and $\hat{\mathcal{D}}_X$ be the marginal distributions of $\mathcal{D}$ and $\hat{\mathcal{D}}$ on X. With a particular kernel $\Phi(X)$, the optimization problem is:

$$\begin{aligned} \min_{\gamma(X)} \quad & \|\mathbb{E}_{\mathcal{D}_X}[\Phi(X)] - \mathbb{E}_{\hat{\mathcal{D}}_X}[\gamma(X)\Phi(X)]\| \\ \text{s.t.} \quad & \gamma(X) > 0 \ \text{ and } \ \mathbb{E}_{\hat{\mathcal{D}}_X}[\gamma(X)] = 1. \end{aligned} \tag{6}$$

The optimal solution is supposed to be $\gamma^*(X) = \frac{\mathbb{P}_{\mathcal{D}_X}(X)}{\mathbb{P}_{\hat{\mathcal{D}}_X}(X)}$. Note the selection of kernel $\Phi(\cdot)$ is non-trivial, especially for complicated features [7] in DNN solutions. Using this strategy, with appropriate $L_{\min}$ and $L_{\max}$ such that all the examples in $\hat{D}$ are Bayes optimal, the covariance could be guaranteed to be optimal when each example in $\hat{D}$ is re-weighted by $\gamma^*(X)$.

Strategy-2: $L_{\min} = L_{\max}$: Compared with Strategy-1, we effectively lose one degree of freedom for getting a better $\hat{D}$. However, this is not entirely harmful since $\hat{D}$ and D^* have the same feature set, indicating estimating $\gamma(X)$ is no longer necessary and $\gamma(X) = 1$ is an optimal solution for (6) with this strategy.

Strategy selection When we can get a high-quality $\hat{D}$ by fine-tuning $L_{\min}$ and $L_{\max}$ or $\hat{D}$ is already provided from other sources, we may solve the optimization problem in (6) to find the optimal weight $\gamma(X)$. However, considering the fact that estimating $\gamma(X)$ introduces extra computation and potentially extra errors, we focus on Strategy-2 in this paper. Using Strategy-2 also reduces the effort on tuning hyperparameters. Besides, the proposed CAL loss is tolerant of an imperfect $\hat{D}$ (shown theoretically in Section 4.3).

D.2. Generation of Instance-Dependent Label Noise

Pseudo codes for generate instance-based label noise are provided in Algorithm 2. This algorithm follows the state-of-the-art method [40]. Define the overall noise rate as η.

Algorithm 2: Generating Instance-Dependent Label Noise

Input: Clean examples $(x_n, y_n)_{n=1}^N$; Noise rate: η; Number of classes: K; Shape of each feature x_n: $S \times 1$.

Sample instance flip rates q_n from the truncated normal distribution $\mathcal{N}(\eta, 0.1^2, [0, 1])$; *// mean η, variance 0.1^2, range $[0, 1]$*

Sample $W \in \mathcal{R}^{S \times K}$ from the standard normal distribution $\mathcal{N}(0, 1^2)$;

for $n \in [N]$ **do**

  $p = x_n^\top W$ *// Generate instance dependent flip rates. The size of p is $1 \times K$.*

  $p_{y_n} = -\infty$ *// Only consider entries that are different from the true label*

  $p = q_n \cdot \text{softmax}(p)$ *// Let q_n be the probability of getting a wrong label*

  $p_{y_n} = 1 - q_n$ *// Keep clean w.p. $1 - q_n$*

  Randomly choose a label from the label space as noisy label $\tilde{y}_n$ according to p;

end

Output: Noisy examples $\{(x_i, \tilde{y}_n), n \in [N]\}$.

Note Algorithm 2 cannot ensure $T_{ii}(X) > T_{ij}(X)$ when $\eta > 0.5$. To generate an informative dataset, we set $0.9 \cdot T_{ii}(X)$ as the upper bound of $T_{ij}(X)$ and distribute the remaining probability to other classes.

D.3. Performance without Data Augmentations

For a fair comparison with the recent work on instance-dependent label noise [40], we adopt the same data augmentations as [40] and re-produce their results using the same noise file as we employed in Table 1. Each noise rate is tested 5 times with a different generation matrix W (defined in Algorithm 2). Table 4 shows the advantages of our second-order approach.

Table 4. Performance comparisons without data augmentations

Method	$\eta = 0.2$	$\eta = 0.4$
PTD-R-V[40]	69.62 ± 3.35	64.73 ± 3.64
CAL	75.52 ± 3.94	70.30 ± 2.96

D.4. More Implementation Details on Clothing1M

Construct $\hat{D}$ We first train the network for 120 epochs on 1 million noisy training images using the method in [5]. The batch-size is set to 32. The initial learning rate is set as 0.01 and reduced by a factor of 10 at 30, 60, 90 epochs. We sample 1000 mini-batches from the training data for each epoch while ensuring the (noisy) labels are balanced. Mixup [48] is adopted for data augmentations. Hyperparameter β is set to 0 at first 80 epochs, and linearly increased to 0.4 for next 20 epochs and kept as 0.4 for the rest of the epochs. We construct $\hat{D}$ with the best model.

Train with CAL We change the loss to the CAL loss after getting $\hat{D}$ and continue training the model (without mixup) with an initial learning rate of 10^{-5} for 120 epochs (reduced by a factor of 10 at 30, 60, 90 epochs). We also tested re-train the model with $\hat{D}$ and get an accuracy of 73.56. A randomly-collected balanced dataset with $18,976$ noisy examples in each class is employed in training with CAL. Examples that are not in this balanced dataset are removed from $\hat{D}$ for ease of implementation.