APPENDIX TO "CAN LABEL-NOISE TRANSITION MATRIX HELP TO IMPROVE SAMPLE SELECTION AND LABEL CORRECTION?"

APPENDIX A

In this section, we show all the proofs.

Theorem 1. *Let* $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ *be two examples such that* $\arg\max_{i\in\{0,1\}} P(Y=i|\boldsymbol{x}_1) = \arg\max_{j\in\{0,1\}} P(\tilde{Y}=j|\boldsymbol{x}_1) = 1$, $\arg\max_{i\in\{0,1\}} P(Y=i|\boldsymbol{x}_2) = \arg\max_{j\in\{0,1\}} P(\tilde{Y}=j|\boldsymbol{x}_2) = 0$, *and* $P(Y=0|\boldsymbol{x}_2) = P(Y=1|\boldsymbol{x}_1)$*. If* $P(\tilde{Y}=1|Y=0) - P(\tilde{Y}=0|Y=1) > 0$*, then* $\min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_2), i) > \min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_1), i)$.

Proof.

$$\begin{aligned}
&P(\tilde{Y}=0|\boldsymbol{x}_2) - P(\tilde{Y}=1|\boldsymbol{x}_1)\\
=&P(\tilde{Y}=0|Y=0)P(Y=0|\boldsymbol{x}_2) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_2)\\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + P(\tilde{Y}=1|Y=1)P(Y=1|\boldsymbol{x}_1)]\\
=&(1-P(\tilde{Y}=1|Y=0))P(Y=0|\boldsymbol{x}_2) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_2)\\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + (1-P(\tilde{Y}=0|Y=1))P(Y=1|\boldsymbol{x}_1)]\\
=&(1-P(\tilde{Y}=1|Y=0))P(Y=0|\boldsymbol{x}_2) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_2)\\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + (1-P(\tilde{Y}=0|Y=1))P(Y=1|\boldsymbol{x}_1)]\\
=&P(Y=0|\boldsymbol{x}_2) - P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_2) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_2)\\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1)]\\
=&P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=1|Y=0)P(Y=1|\boldsymbol{x}_1) + P(\tilde{Y}=0|Y=1)(1-P(Y=1|\boldsymbol{x}_1))\\
-&[P(\tilde{Y}=1|Y=0)(1-P(Y=1|\boldsymbol{x}_1)) + P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1)]\\
=&P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=1|Y=0)P(Y=1|\boldsymbol{x}_1) + P(\tilde{Y}=0|Y=1) - P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1)\\
-&[P(\tilde{Y}=1|Y=0) - P(\tilde{Y}=1|Y=0)P(Y=1|\boldsymbol{x}_1) + P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1)]\\
=&P(\tilde{Y}=0|Y=1) - P(\tilde{Y}=1|Y=0) < 0.
\end{aligned} \tag{1}$$

Note that f^* is an optimal hypothesis which perfectly learns the noisy class posterior distribution. By employing the cross-entropy loss on f^*, we have

$$\ell(f^*(X),\tilde{Y}) = -\tilde{Y}\log(f^*(X)) - (1-\tilde{Y})\log(1-f^*(X)) = -\log(P(\tilde{Y}|X)), \tag{2}$$

which is a non-increasing function. Therefore, the largest noisy class posterior has the minimum loss. Because $\arg\max_{j\in\{0,1\}} P(\tilde{Y}=j|\boldsymbol{x}_2) = 0$, $\arg\max_{i\in\{0,1\}} P(\tilde{Y}=i|\boldsymbol{x}_1) = 1$, and $P(\tilde{Y}=0|\boldsymbol{x}_2) >$

$P(\tilde{Y}=1|\boldsymbol{x}_1)$ by Eq. (1), then

$$\max(P(\tilde{Y}=0|\boldsymbol{x}_2), P(\tilde{Y}=1|\boldsymbol{x}_2), P(\tilde{Y}=0|\boldsymbol{x}_1), P(\tilde{Y}=1|\boldsymbol{x}_1)) = P(\tilde{Y}=1|\boldsymbol{x}_1),$$

which implies that the minimum loss among those four noisy class posteriors is $\ell(f^*(X=\boldsymbol{x}_1), \tilde{Y}=1)$. Therefore $\min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_2), i) > \min_{i\in\{0,1\}} \ell(f^*(\boldsymbol{x}_1), i)$ holds, which completes the proof. □

Theorem 2. *When* $P(\tilde{Y}=1|Y=0) - P(\tilde{Y}=0|Y=1) > 0$*, if an example* x_1 *such that* $0.5 < P(Y=0|\boldsymbol{x}_1) < \frac{(1-2P(\tilde{Y}=0|Y=1))}{(1-2P(\tilde{Y}=1|Y=0))} P(Y=1|\boldsymbol{x}_1)$*, then* $P(\tilde{Y}=1|\boldsymbol{x}_1) > 0.5$*.*

Proof.

$$\begin{aligned}
&P(\tilde{Y}=0|\boldsymbol{x}_1) - P(\tilde{Y}=1|\boldsymbol{x}_1) \\
=&P(\tilde{Y}=0|Y=0)P(Y=0|\boldsymbol{x}_1) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1) \\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + P(\tilde{Y}=1|Y=1)P(Y=1|\boldsymbol{x}_1)] \\
=&(1-P(\tilde{Y}=1|Y=0))P(Y=0|\boldsymbol{x}_1) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1) \\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + (1-P(\tilde{Y}=0|Y=1))P(Y=1|\boldsymbol{x}_1)] \\
=&P(Y=0|\boldsymbol{x}_1) - P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1) \\
-&[P(\tilde{Y}=1|Y=0)P(Y=0|\boldsymbol{x}_1) + P(Y=1|\boldsymbol{x}_1) - P(\tilde{Y}=0|Y=1)P(Y=1|\boldsymbol{x}_1)] \\
=&(1-2P(\tilde{Y}=1|Y=0))P(Y=0|\boldsymbol{x}_1) + (2P(\tilde{Y}=0|Y=1)-1)P(Y=1|\boldsymbol{x}_1). \qquad (3)
\end{aligned}$$

Let $P(Y=0|\boldsymbol{x}_1) < \frac{(1-2P(\tilde{Y}=0|Y=1))}{(1-2P(\tilde{Y}=1|Y=0))} P(Y=1|\boldsymbol{x}_1)$, by combining with Eq. (3), we have

$$\begin{aligned}
&P(\tilde{Y}=0|\boldsymbol{x}_1) - P(\tilde{Y}=1|\boldsymbol{x}_1) \\
<&(1-2P(\tilde{Y}=1|Y=0))\frac{(1-2P(\tilde{Y}=0|Y=1))}{(1-2P(\tilde{Y}=1|Y=0))}P(Y=1|\boldsymbol{x}_1) + (2P(\tilde{Y}=0|Y=1)-1)P(Y=1|\boldsymbol{x}_1) \\
<&(1-2P(\tilde{Y}=0|Y=1))P(Y=1|\boldsymbol{x}_1) + (2P(\tilde{Y}=0|Y=1)-1)P(Y=1|\boldsymbol{x}_1) < 0, \qquad (4)
\end{aligned}$$

which implies that $P(\tilde{Y}=1|\boldsymbol{x}_1) > 0.5$. Let the Bayes label on the clean class-posterior distribution of $\boldsymbol{x}_1$ be 0[1], then $0.5 < P(Y=0|\boldsymbol{x}_1) < \frac{(1-2P(\tilde{Y}=0|Y=1))}{(1-2P(\tilde{Y}=1|Y=0))} P(Y=1|\boldsymbol{x}_1)$, which completes the proof. □

REFERENCES

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

---

[1] The Bayes label is the label with the largest class posterior. For example, the Bayes label on the clean class-posterior distribution Y^* of a instance $\boldsymbol{x}$ is defined as $Y^* = \arg\max_{i\in\{0,1\}} P(Y=i|x)$Mohri et al. (2018)