Appendix

1 PROOFS

In this section, we provide proofs for all the lemmas and theorems in the main paper. We always assume that a class-conditional extension of the Classification Noise Process (CNP) (Angluin & Laird, 1988) maps true labels y to observed labels s such that each label in P is flipped independently with probability ρ_1 and each label in N is flipped independently with probability ρ_0 ($s \leftarrow CNP(y, \rho_1, \rho_0)$), so that $P(s = s|y = y, x) = P(s = s|y = y)$. Remember that $\rho_1 + \rho_0 < 1$ is a necessary condition of minimal information, otherwise we may learn opposite labels.

In Lemma 1, Theorem 2, Lemma 3 and Theorem 4, we assume that P and N have infinite number of examples so that they are the true, hidden distributions.

A fundamental equation we use in the proofs is the following lemma:

Lemma S1 *When g is ideal, i.e. $g(x) = g^*(x)$ and P and N have non-overlapping support, we have*

$$g(x) = (1 - \rho_1) \cdot \mathbb{1}[[y = 1]] + \rho_0 \cdot \mathbb{1}[[y = 0]] \tag{1}$$

Proof: Since $g(x) = g^*(x)$ and P and N have non-overlapping support, we have

$$\begin{aligned} g(x) =& g^*(x) = P(s = 1|x) \\ =& P(s = 1|y = 1, x) \cdot P(y = 1|x) + P(s = 1|y = 0, x) \cdot P(y = 0|x) \\ =& P(s = 1|y = 1) \cdot P(y = 1|x) + P(s = 1|y = 0) \cdot P(y = 0|x) \\ =& (1 - \rho_1) \cdot \mathbb{1}[[y = 1]] + \rho_0 \cdot \mathbb{1}[[y = 0]] \end{aligned}$$

1.1 PROOF OF LEMMA 1

Lemma 1 *When g is ideal, i.e. $g(x) = g^*(x)$ and P and N have non-overlapping support, we have*

$$\begin{cases} \tilde{P}_{y=1} = \{x \in P|s = 1\}, \tilde{N}_{y=1} = \{x \in P|s = 0\} \\ \tilde{P}_{y=0} = \{x \in N|s = 1\}, \tilde{N}_{y=0} = \{x \in N|s = 0\} \end{cases} \tag{2}$$

Proof: Firstly, we compute the threshold $LB_{y=1}$ and $UB_{y=0}$ used by $\tilde{P}_{y=1}$, $\tilde{N}_{y=1}$, $\tilde{P}_{y=0}$ and $\tilde{N}_{y=0}$. Since P and N have non-overlapping support, we have $P(y = 1|x) = \mathbb{1}[[y = 1]]$. Also using $g(x) = g^*(x)$, we have

$$\begin{aligned} LB_{y=1} =& E_{x \in \tilde{P}}[g(x)] = E_{x \in \tilde{P}}[P(s = 1|x)] \\ =& E_{x \in \tilde{P}}[P(s = 1|x, y = 1)P(y = 1|x) + P(s = 1|x, y = 0)P(y = 0|x)] \\ =& E_{x \in \tilde{P}}[P(s = 1|y = 1)P(y = 1|x) + P(s = 1|y = 0)P(y = 0|x)] \\ =& (1 - \rho_1)(1 - \pi_1) + \rho_0\pi_1 \end{aligned} \tag{3}$$

Similarly, we have

$$UB_{y=0} = (1 - \rho_1)\pi_0 + \rho_0(1 - \pi_0)$$

Since $\pi_1 = P(y = 0|s = 1)$, we have $\pi_1 \in [0, 1]$. Furthermore, we have the requirement that $\rho_1 + \rho_0 < 1$, then $\pi_1 = 1$ will lead to $\rho_1 = P(s = 0|y = 1) = 1 - P(s = 1|y = 1) = 1 - \frac{P(y=1|s=1)P(s=1)}{P(y=1)} = 1 - 0 = 1$

which violates the requirement of $\rho_1 + \rho_0 < 1$. Therefore, $\pi_1 \in [0, 1)$. Similarly, we can prove $\pi_0 \in [0, 1)$. Therefore, we see that both $LB_{y=1}$ and $UB_{y=0}$ are interpolations of $(1 - \rho_1)$ and ρ_0:

$$\rho_0 < LB_{y=1} \leq 1 - \rho_1$$
$$\rho_0 \leq UB_{y=0} < 1 - \rho_1$$

The first equality holds iff $\pi_1 = 0$ and the second equality holds iff $\pi_0 = 0$.

Using Lemma S1, we know that under the condition of $g(x) = g^*(x)$ and non-overlapping support, $g(x) = (1 - \rho_1) \cdot \mathbb{1}[[y = 1]] + \rho_0 \cdot \mathbb{1}[[y = 0]]$. In other words,

$$g(x) \geq LB_{y=1} \Leftrightarrow x \in P$$
$$g(x) \leq UB_{y=0} \Leftrightarrow x \in N$$

Since

$$\begin{cases} \tilde{P}_{y=1} = \{x \in \tilde{P} | g(x) \geq LB_{y=1}\} \\ \tilde{N}_{y=1} = \{x \in \tilde{N} | g(x) \geq LB_{y=1}\} \\ \tilde{P}_{y=0} = \{x \in \tilde{P} | g(x) \leq UB_{y=0}\} \\ \tilde{N}_{y=0} = \{x \in \tilde{N} | g(x) \leq UB_{y=0}\} \end{cases}$$

where $\tilde{P} = \{x | s = 1\}$ and $\tilde{N} = \{x | s = 0\}$, we have

$$\begin{cases} \tilde{P}_{y=1} = \{x \in P | s = 1\}, \tilde{N}_{y=1} = \{x \in P | s = 0\} \\ \tilde{P}_{y=0} = \{x \in N | s = 1\}, \tilde{N}_{y=0} = \{x \in N | s = 0\} \end{cases}$$

1.2 PROOF OF LEMMA 2

We restate Theorem 2 here:

Theorem 2 *When g is ideal, i.e. $g(x) = g^*(x)$ and P and N have non-overlapping support, we have*

$$\hat{\rho}_1^{conf} = \rho_1, \hat{\rho}_0^{conf} = \rho_0$$

Proof: Using the definition of $\hat{\rho}_1^{conf}$ in the main paper:

$$\hat{\rho}_1^{conf} = \frac{|\tilde{N}_{y=1}|}{|\tilde{N}_{y=1}| + |\tilde{P}_{y=1}|}, \ \hat{\rho}_0^{conf} = \frac{|\tilde{P}_{y=0}|}{|\tilde{P}_{y=0}| + |\tilde{N}_{y=0}|}$$

Since $g(x) = g^*(x)$ and P and N have non-overlapping support, using Lemma 1, we know

$$\begin{cases} \tilde{P}_{y=1} = \{x \in P | s = 1\}, \tilde{N}_{y=1} = \{x \in P | s = 0\} \\ \tilde{P}_{y=0} = \{x \in N | s = 1\}, \tilde{N}_{y=0} = \{x \in N | s = 0\} \end{cases}$$

Since $\rho_1 = P(s = 0|y = 1)$ and $\rho_0 = P(s = 1|y = 0)$, we immediately have

$$\hat{\rho}_1^{conf} = \frac{|\{x \in P | s = 0\}|}{|P|} = \rho_1, \ \hat{\rho}_0^{conf} = \frac{|\{x \in N | s = 1\}|}{|N|} = \rho_0$$

1.3 PROOF OF LEMMA 3

We rewrite Lemma 3 below:

Lemma 3 *When g is unassuming, i.e., $\Delta g(x) := g(x) - g^*(x)$ can be nonzero, and P and N can have overlapping support, we have*

$$\begin{cases} LB_{y=1} = LB^*_{y=1} + E_{x\in\tilde{P}}[\Delta g(x)] - \frac{(1-\rho_1-\rho_0)^2}{p_{s1}}\Delta p_o \\ UB_{y=0} = UB^*_{y=0} + E_{x\in\tilde{N}}[\Delta g(x)] + \frac{(1-\rho_1-\rho_0)^2}{1-p_{s1}}\Delta p_o \\ \hat{\rho}_1^{conf} = \rho_1 + \frac{1-\rho_1-\rho_0}{|P|-|\Delta P_1|+|\Delta N_1|}|\Delta N_1| \\ \hat{\rho}_0^{conf} = \rho_0 + \frac{1-\rho_1-\rho_0}{|N|-|\Delta N_0|+|\Delta P_0|}|\Delta P_0| \end{cases} \tag{4}$$

where

$$\begin{cases} \Delta p_o := \frac{|P\cap N|}{|P\cup N|} \\ \Delta P_1 = \{x \in P | g(x) < LB_{y=1}\} \\ \Delta N_1 = \{x \in N | g(x) \geq LB_{y=1}\} \\ \Delta P_0 = \{x \in P | g(x) \leq UB_{y=0}\} \\ \Delta N_0 = \{x \in N | g(x) > UB_{y=0}\} \end{cases} \tag{5}$$

Proof: We first calculate $LB_{y=1}$ and $UB_{y=0}$ under unassuming condition, then calculate $\hat{\rho}_i^{conf}$, $i = 0, 1$ under unassuming condition.

Note that Δp_o can also be expressed as

$$\Delta p_o := \frac{|P \cap N|}{|P \cup N|} = P(\hat{y} = 1, y = 0) = P(\hat{y} = 0, y = 1)$$

Here $P(\hat{y} = 1, y = 0) \equiv P(\hat{y} = 1|y = 0)P(y = 0)$, where $P(\hat{y} = 1|y = 0)$ means for a perfect classifier $f^*(x) = P(y = 1|x)$, the expected probability that it will label a $y = 0$ example as positive ($\hat{y} = 1$).

(1) $LB_{y=1}$ and $UB_{y=0}$ under unassuming condition

Firstly, we calculate $LB_{y=1}$ and $UB_{y=0}$ with perfect probability estimation $g^*(x)$, but the support may overlap. Secondly, we allow the probability estimation to be imperfect, superimposed onto the overlapping support condition, and calculate $LB_{y=1}$ and $UB_{y=0}$.

I. Calculating $LB_{y=1}$ and $UB_{y=0}$ when $g(x) = g^*(x)$ and support may overlap

With overlapping support, we no longer have $P(y = 1|x) = \mathbb{1}[[y = 1]]$. Instead, we have

$$\begin{aligned} LB_{y=1} =& E_{x\in\tilde{P}}[g^*(x)] = E_{x\in\tilde{P}}[P(s = 1|x)] \\ =& E_{x\in\tilde{P}}[P(s = 1|x, y = 1)P(y = 1|x) + P(s = 1|x, y = 0)P(y = 0|x)] \\ =& E_{x\in\tilde{P}}[P(s = 1|y = 1)P(y = 1|x) + P(s = 1|y = 0)P(y = 0|x)] \\ =& (1 - \rho_1) \cdot E_{x\in\tilde{P}}[P(y = 1|x)] + \rho_0 \cdot E_{x\in\tilde{P}}[P(y = 0|x)] \\ =& (1 - \rho_1) \cdot P(\hat{y} = 1|s = 1) + \rho_0 \cdot P(\hat{y} = 0|s = 1) \end{aligned}$$

Here $P(\hat{y}=1|s=1)$ can be calculated using Δp_o:

$$\begin{aligned}
P(\hat{y}=1|s=1) &= \frac{P(\hat{y}=1,s=1)}{P(s=1)} \\
&= \frac{P(\hat{y}=1,y=1,s=1)+P(\hat{y}=1,y=0,s=1)}{P(s=1)} \\
&= \frac{P(s=1|y=1)P(\hat{y}=1,y=1)+P(s=1|y=0)P(\hat{y}=1,y=0)}{P(s=1)} \\
&= \frac{(1-\rho_1)(p_{y1}-\Delta p_o)+\rho_0\Delta p_o}{p_{s1}} \\
&= (1-\pi_1)-\frac{1-\rho_1-\rho_0}{p_{s1}}\Delta p_o
\end{aligned}$$

Hence,

$$P(\hat{y}=0|s=1) = 1-P(\hat{y}=1|s=1) = \pi_1+\frac{1-\rho_1-\rho_0}{p_{s1}}\Delta p_o$$

Therefore,

$$\begin{aligned}
LB_{y=1} &= (1-\rho_1)\cdot P(\hat{y}=1|s=1)+\rho_0\cdot P(\hat{y}=0|s=1) \\
&= (1-\rho_1)\cdot\left((1-\pi_1)-\frac{1-\rho_1-\rho_0}{p_{s1}}\Delta p_o\right)+\rho_0\cdot\left(\pi_1+\frac{1-\rho_1-\rho_0}{p_{s1}}\Delta p_o\right) \\
&= LB^*_{y=1}-\frac{(1-\rho_1-\rho_0)^2}{p_{s1}}\Delta p_o
\end{aligned} \tag{6}$$

where $LB^*_{y=1}$ is the $LB_{y=1}$ value when $g(x)$ is ideal. We see in Eq. (6) that the overlapping support introduces a non-positive correction to $LB^*_{y=1}$ compared with the ideal condition.

Similarly, we have

$$UB_{y=0} = UB^*_{y=0}+\frac{(1-\rho_1-\rho_0)^2}{1-p_{s1}}\Delta p_o \tag{7}$$

II. Calculating $LB_{y=1}$ and $UB_{y=0}$ when g is unassuming

Define $\Delta g(x) := g(x)-g^*(x)$. When the support may overlap, we have

$$\begin{aligned}
LB_{y=1} &= E_{x\in\tilde{P}}[g(x)] \\
&= E_{x\in\tilde{P}}[g^*(x)]+E_{x\in\tilde{P}}[\Delta g(x)] \\
&= LB^*_{y=1}-\frac{(1-\rho_1-\rho_0)^2}{p_{s1}}\Delta p_o+E_{x\in\tilde{P}}[\Delta g(x)]
\end{aligned} \tag{8}$$

Similarly, we have

$$\begin{aligned}
UB_{y=0} &= E_{x\in\tilde{N}}[g(x)] \\
&= E_{x\in\tilde{N}}[g^*(x)]+E_{x\in\tilde{N}}[\Delta g(x)] \\
&= UB^*_{y=0}+\frac{(1-\rho_1-\rho_0)^2}{1-p_{s1}}\Delta p_o+E_{x\in\tilde{N}}[\Delta g(x)]
\end{aligned} \tag{9}$$

In summary, Eq. (8) (9) give the expressions for $LB_{y=1}$ and $UB_{y=0}$, respectively, when g is unassuming.

(2) $\hat{\rho}_i^{conf}$ under unassuming condition

Now let's calculate $\hat{\rho}_i^{conf}$, $i = 0, 1$. For simplicity, define

$$\begin{cases} PP = \{x \in P | s = 1\} \\ PN = \{x \in P | s = 0\} \\ NP = \{x \in N | s = 1\} \\ NN = \{x \in N | s = 0\} \\ \Delta_{PP_1} = \{x \in PP | g(x) < LB_{y=1}\} \\ \Delta_{NP_1} = \{x \in NP | g(x) \geq LB_{y=1}\} \\ \Delta_{PN_1} = \{x \in PN | g(x) < LB_{y=1}\} \\ \Delta_{NN_1} = \{x \in NN | g(x) \geq LB_{y=1}\} \end{cases} \tag{10}$$

For $\hat{\rho}_1^{conf}$, we have:

$$\hat{\rho}_1^{conf} = \frac{|\tilde{N}_{y=1}|}{|\tilde{P}_{y=1}| + |\tilde{N}_{y=1}|}$$

Here

$$\begin{aligned} \tilde{P}_{y=1} &= \{x \in \tilde{P} | g(x) \geq LB_{y=1}\} \\ &= \{x \in PP | g(x) \geq LB_{y=1}\} \cup \{x \in NP | g(x) \geq LB_{y=1}\} \\ &= (PP \setminus \Delta_{PP_1}) \cup \Delta_{NP_1} \end{aligned}$$

Similarly, we have

$$\tilde{N}_{y=1} = (PN \setminus \Delta_{PN_1}) \cup \Delta_{NN_1}$$

Therefore

$$\begin{aligned} \hat{\rho}_1^{conf} &= \frac{|PN| - |\Delta_{PN_1}| + |\Delta_{NN_1}|}{[(|PP| - |\Delta_{PP_1}|) + (|PN| - |\Delta_{PN_1}|)] + (|\Delta_{NN_1}| + |\Delta_{NP_1}|)} \\ &= \frac{|PN| - |\Delta_{PN_1}| + |\Delta_{NN_1}|}{|P| - |\Delta P_1| + |\Delta N_1|} \end{aligned} \tag{11}$$

where in the second equality we have used the definition of ΔP_1 and ΔN_1 in Eq. (5).

Using the definition of ρ_1, we have

$$\begin{aligned} \frac{|PN| - |\Delta_{PN_1}|}{|P| - |\Delta P_1|} &= \frac{|\{x \in PN | g(x) \geq LB_{y=1}\}|}{|\{x \in P | g(x) \geq LB_{y=1}\}|} \\ &= \frac{P(x \in PN, g(x) \geq LB_{y=1})}{P(x \in P, g(x) \geq LB_{y=1})} \\ &= \frac{P(x \in PN | x \in P, g(x) \geq LB_{y=1}) \cdot P(x \in P, g(x) \geq LB_{y=1})}{P(x \in P, g(x) \geq LB_{y=1})} \\ &= \frac{P(x \in PN | x \in P) \cdot P(x \in P, g(x) \geq LB_{y=1})}{P(x \in P, g(x) \geq LB_{y=1})} \\ &= \rho_1 \end{aligned}$$

Here we have used the property of CNP that $(s \perp\!\!\!\perp x)|y$, leading to $P(x \in PN|x \in P, g(x) \geq LB_{y=1}) = P(x \in PN|x \in P) = \rho_1$.

Similarly, we have

$$\frac{|\Delta_{NN_1}|}{|\Delta N_1|} = 1 - \rho_0$$

Combining with Eq. (11), we have

$$\hat{\rho}_1^{conf} = \rho_1 + \frac{1 - \rho_1 - \rho_0}{|P| - |\Delta P_1| + |\Delta N_1|}|\Delta N_1| \tag{12}$$

Similarly, we have

$$\hat{\rho}_0^{conf} = \rho_0 + \frac{1 - \rho_1 - \rho_0}{|N| - |\Delta N_0| + |\Delta P_0|}|\Delta P_0| \tag{13}$$

From the two equations above, we see that

$$\hat{\rho}_1^{conf} \geq \rho_1, \hat{\rho}_0^{conf} \geq \rho_0 \tag{14}$$

In other words, $\hat{\rho}_i^{conf}$ is an ***upper bound*** of ρ_i, $i = 0, 1$. The equality for $\hat{\rho}_1^{conf}$ holds if $|\Delta N_1| = 0$. The equality for $\hat{\rho}_0^{conf}$ holds if $|\Delta P_0| = 0$.

1.4 PROOF OF LEMMA 4

Let's restate Theorem 4 below:

Theorem 4 *Given non-overlapping support condition,*
If $\forall x \in N, \Delta g(x) < LB_{y=1} - \rho_0$, *then* $\hat{\rho}_1^{conf} = \rho_1$.
If $\forall x \in P, \Delta g(x) > -(1 - \rho_1 - UB_{y=0})$, *then* $\hat{\rho}_0^{conf} = \rho_0$.

Theorem 4 directly follows from Eq. (12) and (13). Assuming non-overlapping support, we have $g^*(x) = P(s = 1|x) = (1 - \rho_1) \cdot \mathbb{1}[[y = 1]] + \rho_0 \cdot \mathbb{1}[[y = 0]]$. In other words, the contribution of overlapping support to $|\Delta N_1|$ and $|\Delta P_0|$ is 0. The only source of deviation comes from imperfect $g(x)$.

For the first half of the theorem, since $\forall x \in N, \Delta g(x) < LB_{y=1} - \rho_0$, we have $\forall x \in N, g(x) = \Delta g(x) + g^*(x) < (LB_{y=1} - \rho_0) + \rho_0 = LB_{y=1}$, then $|\Delta N_1| = |\{x \in N|g(x) \geq LB_{y=1}\}| = 0$, so we have $\hat{\rho}_1^{conf} = \rho_1$.

Similarly, for the second half of the theorem, since $\forall x \in P, \Delta g(x) > -(1 - \rho_1 - UB_{y=0})$, then $|\Delta P_0| = |\{x \in P|g(x) \leq UB_{y=0}\}| = 0$, so we have $\hat{\rho}_0^{conf} = \rho_0$.

1.5 PROOF OF LEMMA 5

Theorem 5 reads as follows:

Theorem 5 *If g range separates P and N and $\hat{\rho}_i = \rho_i$, $i = 0, 1$, then for any classifier f_θ and any bounded loss function $l(\hat{y}_i, y_i)$, we have*

$$R_{\tilde{l},\mathcal{D}_\rho}(f_\theta) = R_{l,\mathcal{D}}(f_\theta) \tag{15}$$

where $\tilde{l}(\hat{y}_i, s_i)$ is Rank Pruning's loss function given by

$$\tilde{l}(\hat{y}_i, s_i) = \frac{1}{1-\hat{\rho}_1} l(\hat{y}_i, s_i) \cdot \mathbb{1}[[x_i \in \tilde{P}_{conf}]] + \frac{1}{1-\hat{\rho}_0} l(\hat{y}_i, s_i) \cdot \mathbb{1}[[x_i \in \tilde{N}_{conf}]] \tag{16}$$

and $\tilde{P}_{conf}$ and $\tilde{N}_{conf}$ are given by

$$\tilde{P}_{conf} := \{x \in \tilde{P} \mid g(x) \geq k_1\}, \tilde{N}_{conf} := \{x \in \tilde{N} \mid g(x) \leq k_0\} \tag{17}$$

where k_1 is the $(\hat{\pi}_1|\tilde{P}|)^{th}$ smallest $g(x)$ for $x \in \tilde{P}$ and k_0 is the $(\hat{\pi}_0|\tilde{N}|)^{th}$ largest $g(x)$ for $x \in \tilde{N}$

Proof:

Since $\tilde{P}$ and $\tilde{N}$ are constructed from P and N with noise rates π_1 and π_0 using the class-conditional extension of the Classification Noise Process (Angluin & Laird, 1988), we have

$$\begin{cases} P = PP \cup PN \\ N = NP \cup NN \\ \tilde{P} = PP \cup NP \\ \tilde{N} = PN \cup NN \end{cases} \tag{18}$$

where

$$\begin{cases} PP = \{x \in P | s = 1\} \\ PN = \{x \in P | s = 0\} \\ NP = \{x \in N | s = 1\} \\ NN = \{x \in N | s = 0\} \end{cases} \tag{19}$$

satisfying

$$\begin{cases} PP \sim PN \sim P \\ NP \sim NN \sim N \\ \frac{|NP|}{|\tilde{P}|} = \pi_1, \frac{|PP|}{|\tilde{P}|} = 1 - \pi_1 \\ \frac{|PN|}{|\tilde{N}|} = \pi_0, \frac{|NN|}{|\tilde{N}|} = 1 - \pi_0 \\ \frac{|PN|}{|P|} = \rho_1, \frac{|PP|}{|P|} = 1 - \rho_1 \\ \frac{|NP|}{|N|} = \rho_0, \frac{|NN|}{|N|} = 1 - \rho_0 \end{cases} \tag{20}$$

Here the $\sim$ means obeying the same distribution.

Since g range separates P and N, there exists a real number z such that $\forall x_1 \in P$ and $\forall x_0 \in N$, we have $g(x_1) > z > g(x_0)$. Since $P = PP \cup PN$, $N = NP \cup NN$, we have

$$\begin{aligned} &\forall x \in PP, g(x) > z;\ \forall x \in PN, g(x) > z; \\ &\forall x \in NP, g(x) < z;\ \forall x \in NN, g(x) < z \end{aligned} \tag{21}$$

Since $\hat{\rho}_1 = \rho_1$ and $\hat{\rho}_0 = \rho_0$, we have

$$\begin{cases} \hat{\pi}_1 = \frac{\hat{\rho}_0}{p_{s1}} \frac{1-p_{s1}-\hat{\rho}_1}{1-\hat{\rho}_1-\hat{\rho}_0} = \frac{\rho_0}{p_{s1}} \frac{1-p_{s1}-\rho_1}{1-\rho_1-\rho_0} = \pi_1 \equiv \frac{\rho_0|N|}{|\tilde{P}|} \\ \hat{\pi}_0 = \frac{\hat{\rho}_1}{1-p_{s1}} \frac{p_{s1}-\hat{\rho}_0}{1-\hat{\rho}_1-\hat{\rho}_0} = \frac{\rho_1}{1-p_{s1}} \frac{p_{s1}-\rho_0}{1-\rho_1-\rho_0} = \pi_0 \equiv \frac{\rho_1|P|}{|\tilde{N}|} \end{cases} \tag{22}$$

Therefore, $\hat{\pi}_1|\tilde{P}| = \pi_1|\tilde{P}| = \rho_0|N|$, $\hat{\pi}_0|\tilde{N}| = \pi_0|\tilde{N}| = \rho_1|P|$. Using $\tilde{P}_{conf}$ and $\tilde{N}_{conf}$'s definition in Eq. (17), and $g(x)$'s property in Eq. (21), we have

$$\tilde{P}_{conf} = PP \sim P, \tilde{N}_{conf} = NN \sim N \tag{23}$$

Hence P_{conf} and N_{conf} can be seen as a uniform downsampling of P and N, with a downsampling ratio of $(1-\rho_1)$ for P and $(1-\rho_0)$ for N. Then according to Eq. (16), the loss function $\tilde{l}(\hat{y}_i, s_i)$ essentially sees a fraction of $(1-\rho_1)$ examples in P and a fraction of $(1-\rho_0)$ examples in N, with a final reweighting to restore the class balance. Then for any classifier f_θ that maps $x \to \hat{y}$ and any bounded loss function $l(\hat{y}_i, y_i)$, we have

$$\begin{aligned} &R_{\tilde{l},\mathcal{D}_\rho}(f_\theta) \\ &= E_{(x,s)\sim\mathcal{D}_\rho}[\tilde{l}(f_\theta(x), s)] \\ &= \frac{1}{1-\hat{\rho}_1} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in \tilde{P}_{conf}]]\right] + \frac{1}{1-\hat{\rho}_0} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in \tilde{N}_{conf}]]\right] \\ &= \frac{1}{1-\rho_1} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in \tilde{P}_{conf}]]\right] + \frac{1}{1-\rho_0} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in \tilde{N}_{conf}]]\right] \\ &= \frac{1}{1-\rho_1} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in PP]]\right] + \frac{1}{1-\rho_0} \cdot E_{(x,s)\sim\mathcal{D}_\rho}\left[l(f_\theta(x), s) \cdot \mathbb{1}[[x \in NN]]\right] \\ &= \frac{1}{1-\rho_1} \cdot (1-\rho_1) \cdot E_{(x,y)\sim\mathcal{D}}\left[l(f_\theta(x), y) \cdot \mathbb{1}[[x \in P]]\right] + \\ &\quad \frac{1}{1-\rho_0} \cdot (1-\rho_0) \cdot E_{(x,y)\sim\mathcal{D}}\left[l(f_\theta(x), y) \cdot \mathbb{1}[[x \in N]]\right] \\ &= E_{(x,y)\sim\mathcal{D}}\left[l(f_\theta(x), y) \cdot \mathbb{1}[[x \in P]] + l(f_\theta(x), y) \cdot \mathbb{1}[[x \in N]]\right] \\ &= E_{(x,y)\sim\mathcal{D}}\left[l(f_\theta(x), y)\right] \\ &= R_{l,\mathcal{D}}(f_\theta) \end{aligned}$$

Therefore, we see that the expected risk for Rank Pruning with corrupted labels, is exactly the same as the expected risk for the true labels, for any bounded loss function l and classifier f_θ. The reweighting ensures that after pruning, the two sets still remain unbiased w.r.t. to the true dataset.

Since the ideal condition is more strict than the range separability condition, we immediately have that when g is ideal and $\hat{\rho}_i = \rho_i$, $i = 0, 1$, $R_{\tilde{l},\mathcal{D}_\rho}(f_\theta) = R_{l,\mathcal{D}}(f_\theta)$ for any f_θ and bounded loss function l.

2 ADDITIONAL FIGURES

Figure S1 shows the sum of absolute difference between theoretically estimated $\hat{\rho}_i^{thry}$ (Eq. (8) in main paper) and empirical $\hat{\rho}_i$: $|\hat{\rho}_1 - \hat{\rho}_1^{thry}| + |\hat{\rho}_0 - \hat{\rho}_0^{thry}|$. The deviation of the theoretical and empirical estimates reflects the assumption that we have infinite examples, whereas empirically, the number of examples is finite.

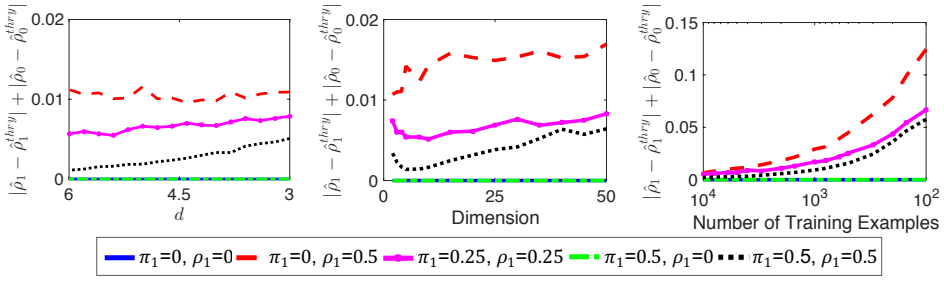

Figure S 1: Sum of absolute difference between theoretically estimated $\hat{\rho}_i^{thry}$ and empirical $\hat{\rho}_i$, $i = 0, 1$, with five different (π_1, ρ_1), for varying separability d, dimension, and number of training examples. Note that no figure exists for percent random noise because the theoretical estimates in Eq. (4) do not address added noise examples. The default parameters are: $d = 4$, 2 dimensional input, 0% random noise, and 5000 training examples with a fraction of $p_{y1} = 0.2$ examples as positive. The lines are an average of 200 trials.

Figure S2 shows the Rank Pruning's noise rate estimation of $\hat{\pi}_1$ for the MNIST dataset using a logistic regression classifier, for varying amount of ($\hat{\pi}_1$,$\hat{\rho}_1$), averaging over 10 digits.

Figure S3 shows the average image for each digit for the binary classification problem "1" or "not 1" in MNIST with logistic regression and high noise ($\rho_1 = 0.5, \pi_1 = 0.5$). The number on the bottom and on the right counts the total number of examples (images). From the figure we see that RP makes few mistakes, and when it does, the mistakes vary greatly in image from the typical digit.

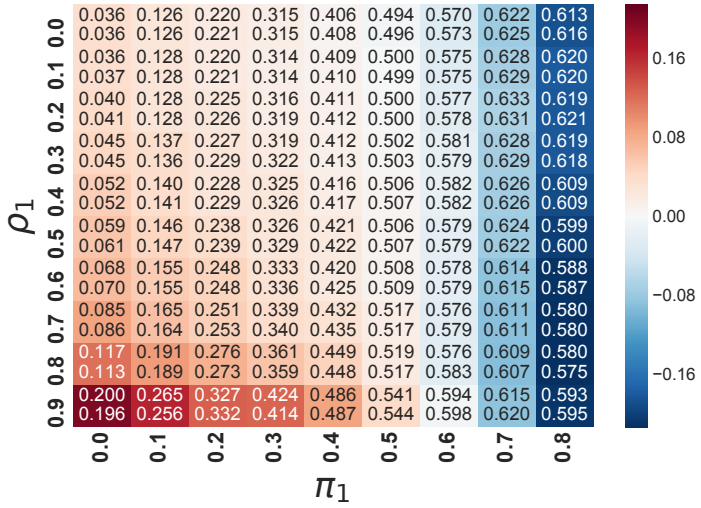

Figure S 2: Rank Pruning $\hat{\pi}_1$ estimation consistency, averaged over all digits in MNIST. Color depicts $\hat{\pi}_1$ - π_1 with $\hat{\rho}_1$ (upper) and theoretical $\hat{\pi}_1^{thry}$ (lower) in each block.

3 ADDITIONAL TABLES

Here we provide additional tables for the comparison of error, Precision-Recall AUC (AUC-PR, Davis & Goadrich (2006)), and F1 score for the algorithms *RP*, *Nat13*, *Elk08*, *Liu16* with ρ_1, ρ_0 given to all methods

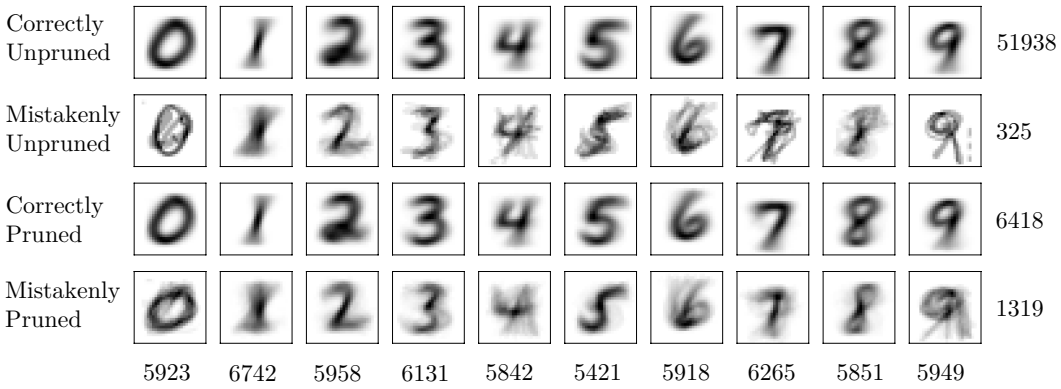

Figure S 3: Average image for each digit for the binary classification problem "1" or "not 1" in MNIST with logistic regression and significant mislabeling ($\rho_1 = 0.5, \pi_1 = 0.5$). The right and bottom numbers count the total number of example images averaged in the corresponding row or column.

for fair comparison. Additionally, we provide the performance of the ground truth classifier (*true*) trained with uncorrupted labels (X, y), as well as the complete Rank Pruning algorithm (RP_ρ) trained using the noise rates estimated by Rank Pruning. The top model scores are in bold with RP_ρ in red if its performance is better than non-RP models. The $\pi_1 = 0$ quadrant in each table represents the "PU learning" case of $\tilde{P}\tilde{N}$ learning.

Whenever $g(x) = P(\hat{s} = 1|x)$ is estimated for any algorithm, we use a 3-fold cross-validation to estimate the probability $g(x)$. For improved performance, a higher fold may be used.

For the logistic regression classifier, we use scikit-learn's LogisticRegression class (scikit learn (2016)) with default settings (L2 regularization with inverse strength $C = 1$).

For the convolutional neural networks (CNN), for MNIST we use the structure in Chollet (2016b) and for CIFAR-10, we use the structure in Chollet (2016a). A 10% holdout set is used to monitor the weighted validation loss (using the sample weight given by each algorithm) and ends training when there is no decrease for 10 epochs, with a maximum of 50 epochs for MNIST and 150 epochs for CIFAR-10.

The following list comprises the MNIST and CIFAR-10 experimental result tables for error, AUC-PR and F1 score metrics:

Table S1: Error for MNIST with logisitic regression as classifier.

Table S2: AUC-PR for MNIST with logisitic regression as classifier.

Table S3: Error for MNIST with CNN as classifier.

Table S4: AUC-PR for MNIST with CNN as classifier.

Table S5: F1 score for CIFAR-10 with logistic regression as classifier.

Table S6: Error for CIFAR-10 with logistic regression as classifier.

Table S7: AUC-PR for CIFAR-10 with logistic regression as classifier.

Table S8: Error for CIFAR-10 with CNN as classifier.

Table S9: AUC-PR for CIFAR-10 with CNN as classifier.

Due to its sensitivity to imperfect probability estimation, here *Liu16* always predicts all labels to be positive or negative, resulting in the same metric score for every digit/image in each scenario. Since $p_{y1} \simeq 0.1$, when

predicting all labels as positive, *Liu16* has an F1 score of 0.182, error of 0.90, and AUC-PR of 0.55; when predicting all labels as negative, *Liu16* has an F1 score of 0.0, error of 0.1, and AUC-PR of 0.55.

4 ADDITIONAL RELATED WORK

In this section we include tangentially related work which was unable to make it into the final manuscript.

4.1 ONE-CLASS CLASSIFICATION

One-class classification (Moya et al., 1993) is distinguished from binary classification by a training set containing examples from only one class, making it useful for outlier and novelty detection (Hempstalk et al., 2008). This can be framed as $\tilde{P}\tilde{N}$ learning when outliers take the form of mislabeled examples. The predominant approach, one-class SVM, fits a hyper-boundary around the training class (Platt et al., 1999), but often performs poorly due to boundary over-sensitivity (Manevitz & Yousef, 2002) and fails when the training class contains mislabeled examples.

4.2 $\tilde{P}\tilde{N}$ LEARNING FOR IMAGE RECOGNITION AND DEEP LEARNING

Variations of $\tilde{P}\tilde{N}$ learning have been used in the context of machine vision to improve robustness to mislabeling (Xiao et al., 2015). In a face recognition task with 90% of non-faces mislabeled as faces, a bagging model combined with consistency voting was used to remove images with poor voting consistency (Angelova et al., 2005). However, no theoretical justification was provided. In the context of deep learning, consistency of predictions for inputs with mislabeling enforces can be enforced by combining a typical cross-entropy loss with an auto-encoder loss (Reed et al., 2015). This method enforces label consistency by constraining the network to uncover the input examples given the output prediction, but is restricted in architecture and generality.

Table S 1: Comparison of **error** for one-vs-rest MNIST (averaged over all digits) using a **logistic regression** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if better (smaller) than non-RP models.

	π_1 = 0			π_1 = 0.25				π_1 = 0.5				π_1 = 0.75			
MODEL,ρ_1 =	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**
TRUE	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
RP$_\rho$	**0.023**	**0.025**	**0.031**	**0.024**	**0.025**	**0.027**	**0.038**	0.040	0.037	0.039	0.049	0.140	0.128	0.133	0.151
RP	**0.022**	**0.025**	**0.031**	**0.021**	**0.024**	**0.027**	**0.035**	**0.023**	**0.027**	**0.031**	**0.043**	**0.028**	**0.036**	**0.045**	0.069
NAT13	0.025	0.030	0.038	0.025	0.029	0.034	0.042	0.030	0.033	0.038	0.047	0.035	0.039	0.046	**0.067**
ELK08	0.025	0.030	0.038	0.026	0.028	0.032	0.042	0.030	0.031	0.035	0.051	0.092	0.093	0.123	0.189
LIU16	0.187	0.098	0.100	0.100	0.738	0.738	0.419	0.100	0.820	0.821	0.821	0.098	0.760	0.741	0.820

Table S 2: Comparison of **AUC-PR** for one-vs-rest MNIST (averaged over all digits) using a **logistic regression** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if greater than non-RP models.

	π_1 = 0			π_1 = 0.25				π_1 = 0.5				π_1 = 0.75			
MODEL,ρ_1 =	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**	**0.00**	**0.25**	**0.50**	**0.75**
TRUE	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935	0.935
RP$_\rho$	0.921	**0.913**	**0.882**	**0.928**	**0.920**	**0.906**	**0.853**	**0.903**	**0.902**	**0.879**	**0.803**	0.851	0.835	**0.788**	0.640
RP	**0.922**	**0.913**	**0.882**	**0.930**	**0.921**	**0.906**	**0.858**	**0.922**	**0.903**	**0.883**	**0.811**	**0.893**	**0.841**	**0.799**	0.621
NAT13	**0.922**	0.908	0.878	0.918	0.909	0.890	0.839	0.899	0.892	0.862	0.794	0.863	0.837	0.784	**0.645**
ELK08	0.921	0.903	0.864	0.917	0.908	0.884	0.821	0.898	0.892	0.861	0.763	0.852	0.837	0.772	0.579
LIU16	0.498	0.549	0.550	0.550	0.500	0.550	0.505	0.550	0.550	0.550	0.549	0.503	0.512	0.550	0.550

Table S 3: Comparison of **error** for one-vs-rest MNIST (averaged over all digits) using a **CNN** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if better (smaller) than non-RP models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5				
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_1 = 0				ρ_1 = 0.5
IMAGE	**TRUE**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**
0	0.0013	**0.0018**	**0.0023**	0.0045	0.0047	0.9020	**0.0017**	**0.0016**	0.0034	0.0036	0.9020	**0.0017**	**0.0016**	0.0031	0.0026	0.0029	**0.0021**	**0.0022**	0.0116	0.0069	0.9020
1	0.0015	**0.0022**	**0.0020**	0.0025	0.0034	0.8865	**0.0019**	**0.0019**	0.0035	0.0030	0.8865	0.0023	0.0020	0.0018	**0.0016**	0.0023	**0.0025**	**0.0025**	0.0036	0.0027	0.8865
2	0.0027	**0.0054**	**0.0049**	0.0057	0.0062	0.8968	**0.0032**	**0.0035**	0.0045	0.0051	0.8968	0.0030	0.0029	0.0031	0.0029	**0.0024**	**0.0059**	**0.0050**	0.0066	0.0083	0.8968
3	0.0020	**0.0032**	**0.0032**	0.0055	0.0038	0.8990	**0.0029**	**0.0029**	0.0043	0.0043	0.8990	**0.0021**	0.0027	**0.0023**	**0.0023**	0.0032	**0.0038**	**0.0042**	0.0084	0.0057	0.8990
4	0.0012	**0.0037**	0.0040	**0.0038**	0.0044	0.9018	**0.0029**	**0.0025**	0.0055	0.0069	0.9018	0.0026	0.0020	**0.0019**	0.0021	0.0030	**0.0044**	**0.0035**	0.0086	0.0077	0.9018
5	0.0019	**0.0032**	**0.0035**	0.0039	0.0038	0.9108	**0.0027**	**0.0031**	0.0062	0.0060	0.9108	**0.0021**	0.0024	0.0024	0.0028	**0.0023**	**0.0061**	**0.0056**	0.0066	0.0074	0.9108
6	0.0021	**0.0027**	**0.0028**	0.0053	0.0035	0.9042	**0.0028**	**0.0025**	0.0042	0.0036	0.9042	0.0029	0.0029	**0.0022**	0.0024	0.0028	**0.0032**	**0.0035**	0.0098	0.0075	0.9042
7	0.0026	**0.0039**	**0.0041**	0.0066	0.0103	0.8972	**0.0050**	**0.0052**	0.0058	0.0058	0.8972	0.0049	0.0040	**0.0030**	0.0037	0.0035	**0.0054**	**0.0064**	0.0113	0.0085	0.8972
8	0.0022	**0.0047**	**0.0043**	0.0106	0.0063	0.9026	**0.0034**	**0.0036**	0.0062	0.0091	0.9026	0.0036	**0.0030**	0.0035	0.0041	0.0032	**0.0044**	**0.0048**	0.0234	0.0077	0.9026
9	0.0036	0.0067	**0.0052**	0.0056	0.0124	0.8991	**0.0048**	**0.0051**	0.0065	0.0064	0.8991	0.0048	0.0050	0.0051	**0.0043**	0.0059	**0.0081**	0.0114	0.0131	**0.0112**	0.8991
AVG	0.0021	**0.0038**	**0.0036**	0.0054	0.0059	0.9000	**0.0031**	**0.0032**	0.0050	0.0054	0.9000	0.0030	**0.0028**	**0.0028**	0.0029	0.0032	**0.0046**	**0.0049**	0.0103	0.0074	0.9000

Table S 4: Comparison of **AUC-PR** for one-vs-rest MNIST (averaged over all digits) using a **CNN** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if greater than non-RP models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5				
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_1 = 0				ρ_1 = 0.5
IMAGE	**TRUE**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**	**RP$_\rho$**	**RP**	**NAT13**	**ELK08**	**LIU16**
0	0.9998	**0.9992**	**0.9990**	0.9986	0.9982	0.5490	**0.9996**	**0.9996**	0.9986	0.9979	0.5490	**0.9989**	**0.9995**	0.9976	0.9979	0.9956	**0.9984**	**0.9982**	0.9963	0.9928	0.5490
1	0.9999	**0.9995**	**0.9995**	0.9976	0.9974	0.5568	**0.9996**	0.9993	**0.9995**	**0.9995**	0.5568	**0.9995**	**0.9998**	0.9982	0.9972	0.9965	**0.9995**	**0.9994**	0.9978	0.9985	0.5568
2	0.9994	**0.9971**	**0.9969**	0.9917	0.9942	0.5516	**0.9980**	**0.9977**	0.9971	0.9945	0.5516	**0.9988**	**0.9992**	0.9958	0.9934	0.9940	**0.9938**	**0.9947**	0.9847	0.9873	0.5516
3	0.9996	**0.9986**	**0.9987**	0.9983	0.9984	0.5505	**0.9991**	**0.9989**	0.9982	0.9980	0.5505	**0.9993**	**0.9994**	0.9991	0.9971	0.9974	**0.9969**	**0.9959**	0.9951	**0.9959**	0.5505
4	0.9997	0.9982	**0.9989**	0.9939	0.9988	0.0891	**0.9992**	**0.9991**	0.9976	0.9965	0.5491	**0.9994**	**0.9996**	0.9985	0.9978	0.9986	**0.9983**	**0.9977**	0.9961	0.9919	0.5491
5	0.9993	**0.9982**	**0.9976**	0.9969	0.9956	0.5446	**0.9986**	**0.9987**	0.9983	0.9979	0.5446	**0.9984**	**0.9982**	0.9971	0.9963	0.9929	**0.9958**	**0.9965**	0.9946	0.9934	0.5446
6	0.9987	**0.9976**	**0.9970**	0.9928	0.9931	0.5479	**0.9974**	**0.9980**	0.9956	0.9959	0.5479	**0.9968**	**0.9983**	0.9933	0.9950	0.9905	**0.9964**	0.9957	0.9942	**0.9961**	0.5479
7	0.9989	**0.9973**	**0.9972**	0.9965	0.9944	0.0721	0.9968	0.9973	0.9966	**0.9979**	0.5514	0.9969	**0.9983**	0.9961	0.9958	0.9974	**0.9933**	**0.9937**	0.9896	0.9886	0.5514
8	0.9996	**0.9974**	**0.9964**	**0.9964**	0.9946	0.5487	**0.9981**	**0.9981**	0.9973	0.9971	0.5487	**0.9983**	0.9988	0.9984	0.9976	**0.9989**	**0.9976**	**0.9975**	0.9873	0.9893	0.5487
9	0.9979	**0.9931**	**0.9951**	0.9901	0.9922	0.5504	**0.9935**	**0.9951**	0.9933	0.9920	0.5504	**0.9961**	**0.9951**	0.9924	0.9922	0.9912	**0.9877**	**0.9876**	0.9819	0.9828	0.5504
AVG	0.9993	**0.9976**	**0.9976**	0.9953	0.9957	0.4561	**0.9980**	**0.9982**	0.9972	0.9967	0.5500	**0.9983**	**0.9986**	0.9966	0.9960	0.9953	**0.9958**	**0.9957**	0.9918	0.9917	0.5500

Table S 5: Comparison of **F1 score** for one-vs-rest CIFAR-10 (averaged over all images) using a **logistic regression** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if greater than non-RP models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5									
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_0 = 0					ρ_1 = 0.5				
IMAGE	TRUE	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16
PLANE	0.272	**0.311**	**0.252**	0.217	0.220	0.182	**0.329**	**0.275**	0.222	0.224	0.182	**0.330**	**0.265**	0.231	0.259	0.0	**0.266**	**0.188**	0.183	0.187	0.182
AUTO	0.374	**0.389**	**0.355**	0.318	0.320	0.182	**0.388**	**0.368**	0.321	0.328	0.182	**0.372**	**0.355**	0.308	0.341	0.0	**0.307**	0.287	0.287	**0.297**	0.182
BIRD	0.136	**0.241**	0.167	0.143	0.136	**0.182**	**0.248**	**0.185**	0.137	0.137	0.182	**0.258**	**0.147**	0.100	0.126	0.0	**0.206**	0.153	0.132	0.150	**0.182**
CAT	0.122	**0.246**	0.170	0.141	0.150	**0.182**	**0.232**	0.163	0.112	0.127	**0.182**	**0.241**	**0.125**	0.068	0.103	0.0	**0.209**	0.148	0.119	0.157	**0.182**
DEER	0.166	**0.250**	**0.184**	0.153	0.164	0.182	**0.259**	0.175	0.146	0.163	**0.182**	**0.259**	**0.177**	0.126	0.164	0.0	**0.222**	0.162	0.132	0.164	**0.182**
DOG	0.139	**0.245**	0.174	0.146	0.148	**0.182**	**0.262**	0.171	0.115	0.126	**0.182**	**0.254**	**0.152**	0.075	0.120	0.0	**0.203**	0.151	0.128	0.137	**0.182**
FROG	0.317	**0.322**	**0.315**	0.289	0.281	0.182	**0.350**	**0.319**	0.283	0.299	0.182	**0.346**	**0.305**	0.239	0.279	0.0	**0.308**	0.252	0.244	**0.269**	0.182
HORSE	0.300	**0.300**	**0.299**	0.283	0.263	0.182	**0.334**	**0.313**	0.272	0.281	0.182	**0.322**	**0.310**	0.260	0.292	0.0	**0.275**	**0.258**	0.240	0.245	0.182
SHIP	0.322	**0.343**	**0.322**	0.297	0.272	0.182	**0.385**	**0.319**	0.287	0.289	0.182	**0.350**	**0.303**	0.250	0.293	0.0	**0.304**	**0.248**	0.230	0.237	0.182
TRUCK	0.330	**0.359**	**0.323**	0.273	0.261	0.182	**0.369**	**0.327**	0.293	0.290	0.182	**0.343**	**0.302**	0.278	0.299	0.0	**0.313**	0.246	0.252	**0.262**	0.182
AVG	0.248	**0.301**	**0.256**	0.226	0.221	0.182	**0.316**	**0.262**	0.219	0.226	0.182	**0.308**	**0.244**	0.194	0.228	0.000	**0.261**	0.209	0.195	**0.210**	0.182

Table S 6: Comparison of **error** for one-vs-rest CIFAR-10 (averaged over all images) using a **logistic regression** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if better (smaller) than non-RP models. Here the logistic regression classifier severely underfits CIFAR, resulting in Rank Pruning pruning out some correctly labeled examples that "confuse" the classifier, hence in this scenario, RP and RP_ρ generally have slightly smaller precision, much higher recall, and hence larger F1 scores than other models and even the ground truth classifier (Table S5). Due to the class inbalance ($p_{y1} = 0.1$) and their larger recall, RP and RP_ρ here have larger error than the other models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5									
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_0 = 0					ρ_1 = 0.5				
IMAGE	TRUE	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16
PLANE	0.107	0.287	0.133	0.123	**0.122**	0.900	0.177	0.128	**0.119**	0.123	0.900	0.248	0.124	0.110	0.118	**0.100**	0.202	0.147	**0.142**	0.160	0.900
AUTO	0.099	0.184	0.120	**0.110**	**0.110**	0.900	0.132	0.114	**0.105**	0.109	0.900	0.189	0.110	0.105	0.110	**0.100**	0.159	0.129	**0.125**	0.139	0.900
BIRD	0.117	0.354	0.148	0.133	**0.131**	0.900	0.217	0.135	**0.120**	0.125	0.900	0.277	0.135	0.115	0.123	**0.100**	0.226	0.147	**0.139**	0.158	0.900
CAT	0.114	0.351	0.138	**0.129**	**0.129**	0.900	0.208	0.139	**0.122**	0.125	0.900	0.303	0.132	0.114	0.122	**0.100**	0.225	0.151	**0.141**	0.158	0.900
DEER	0.112	0.336	0.143	**0.128**	0.130	0.900	0.194	0.135	**0.120**	0.122	0.900	0.271	0.133	0.118	0.126	**0.100**	0.209	0.150	**0.147**	0.161	0.900
DOG	0.119	0.370	0.150	**0.136**	0.138	0.900	0.205	0.142	**0.129**	0.132	0.900	0.288	0.135	0.120	0.128	**0.100**	0.229	0.154	**0.147**	0.168	0.900
FROG	0.107	0.228	0.128	**0.117**	**0.117**	0.900	0.155	0.124	**0.113**	0.115	0.900	0.228	0.118	0.110	0.116	**0.100**	0.167	0.137	**0.130**	0.142	0.900
HORSE	0.104	0.251	0.127	**0.114**	0.116	0.900	0.153	0.123	**0.110**	0.112	0.900	0.224	0.116	0.108	0.113	**0.100**	0.178	0.134	**0.129**	0.144	0.900
SHIP	0.112	0.239	0.134	**0.121**	0.126	0.900	0.160	0.131	**0.119**	0.123	0.900	0.236	0.122	0.113	0.120	**0.100**	0.193	0.145	**0.139**	0.159	0.900
TRUCK	0.106	0.210	0.130	**0.121**	0.122	0.900	0.145	0.125	**0.113**	0.117	0.900	0.213	0.121	0.108	0.117	**0.100**	0.165	0.142	**0.134**	0.150	0.900
AVG	0.110	0.281	0.135	**0.123**	0.124	0.900	0.175	0.130	**0.117**	0.120	0.900	0.248	0.125	0.112	0.119	**0.100**	0.195	0.144	**0.137**	0.154	0.900

Table S 7: Comparison of **AUC-PR** for one-vs-rest CIFAR-10 (averaged over all images) using a **logistic regression** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if greater than non-RP models. Since $p_{y1} = 0.1$, here *Liu16* always predicts all labels as positive or negative, resulting in a constant AUC-PR of 0.550.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5									
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_0 = 0					ρ_1 = 0.5				
IMAGE	TRUE	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16
PLANE	0.288	0.225	0.224	0.225	0.207	**0.550**	0.261	0.235	0.225	0.217	**0.550**	0.285	0.251	0.245	0.248	**0.550**	0.196	0.171	0.171	0.159	**0.550**
AUTO	0.384	0.350	0.317	0.312	0.316	**0.550**	0.342	0.335	0.331	0.331	**0.550**	0.328	0.348	0.334	0.333	**0.550**	0.256	0.257	0.259	0.261	**0.550**
BIRD	0.198	0.160	0.169	0.166	0.161	**0.550**	0.188	0.185	0.179	0.177	**0.550**	0.186	0.173	0.174	0.175	**0.550**	0.150	0.154	0.150	0.147	**0.550**
CAT	0.188	0.164	0.175	0.174	0.175	**0.550**	0.163	0.169	0.168	0.170	**0.550**	0.148	0.156	0.154	0.152	**0.550**	0.145	0.143	0.140	0.145	**0.550**
DEER	0.215	0.161	0.177	0.180	0.183	**0.550**	0.194	0.180	0.180	0.182	**0.550**	0.174	0.175	0.176	0.175	**0.550**	0.151	0.152	0.146	0.151	**0.550**
DOG	0.188	0.162	0.161	0.165	0.155	**0.550**	0.175	0.160	0.161	0.158	**0.550**	0.173	0.169	0.162	0.164	**0.550**	0.145	0.142	0.139	0.133	**0.550**
FROG	0.318	0.246	0.264	0.262	0.258	**0.550**	0.292	0.277	0.272	0.273	**0.550**	0.276	0.274	0.277	0.277	**0.550**	0.239	0.212	0.206	0.212	**0.550**
HORSE	0.319	0.242	0.267	0.269	0.260	**0.550**	0.283	0.264	0.264	0.263	**0.550**	0.288	0.282	0.279	0.278	**0.550**	0.223	0.218	0.208	0.207	**0.550**
SHIP	0.317	0.257	0.267	0.271	0.248	**0.550**	0.296	0.266	0.267	0.259	**0.550**	0.279	0.268	0.259	0.262	**0.550**	0.220	0.212	0.207	0.191	**0.550**
TRUCK	0.329	0.288	0.261	0.271	0.263	**0.550**	0.298	0.275	0.286	0.284	**0.550**	0.289	0.272	0.276	0.277	**0.550**	0.241	0.213	0.208	0.204	**0.550**
AVG	0.274	0.226	0.228	0.229	0.223	**0.550**	0.249	0.235	0.233	0.231	**0.550**	0.243	0.237	0.234	0.234	**0.550**	0.197	0.187	0.183	0.181	**0.550**

Table S 8: Comparison of **error** for one-vs-rest CIFAR-10 (averaged over all images) using a **CNN** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if better (smaller) than non-RP models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5									
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_1 = 0					ρ_1 = 0.5				
IMAGE	TRUE	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16
PLANE	0.044	**0.054**	**0.057**	0.059	0.063	0.900	**0.050**	**0.051**	0.054	0.057	0.900	**0.048**	**0.045**	0.049	0.048	0.100	**0.063**	**0.061**	0.074	0.065	0.900
AUTO	0.021	**0.040**	**0.037**	0.041	0.043	0.100	**0.032**	**0.034**	0.040	0.039	0.900	0.028	**0.026**	**0.026**	**0.026**	0.100	**0.047**	**0.049**	0.062	0.070	0.900
BIRD	0.055	0.083	**0.078**	0.080	0.082	0.900	**0.074**	**0.074**	0.077	0.078	0.900	0.072	**0.066**	0.072	0.070	0.100	0.124	**0.084**	0.089	0.093	0.900
CAT	0.077	0.108	**0.091**	0.092	0.095	0.100	0.111	0.090	**0.086**	0.089	0.900	0.113	**0.084**	0.086	0.088	0.100	0.117	0.098	**0.094**	0.100	0.900
DEER	0.049	0.081	**0.078**	**0.078**	0.079	0.900	0.080	**0.069**	0.075	0.070	0.900	0.076	0.062	**0.061**	0.062	0.100	0.106	**0.086**	0.091	0.093	0.900
DOG	0.062	**0.075**	**0.071**	0.079	0.080	0.100	0.071	0.069	0.070	**0.067**	0.900	0.069	0.061	**0.057**	0.076	0.100	0.103	**0.081**	0.084	0.086	0.900
FROG	0.038	0.050	**0.048**	**0.048**	0.054	0.100	**0.047**	**0.052**	0.056	0.062	0.900	0.045	**0.040**	0.042	0.043	0.100	**0.058**	**0.062**	0.066	0.071	0.900
HORSE	0.035	**0.050**	**0.052**	0.057	0.054	0.900	**0.048**	**0.051**	0.052	0.057	0.900	0.045	**0.040**	0.042	0.046	0.100	**0.065**	**0.063**	0.066	0.075	0.900
SHIP	0.028	**0.042**	**0.042**	0.046	**0.042**	0.900	**0.037**	**0.036**	0.042	0.047	0.900	0.035	0.033	**0.031**	0.033	0.100	**0.051**	**0.049**	0.064	0.058	0.900
TRUCK	0.027	**0.044**	**0.046**	0.054	0.056	0.900	**0.034**	**0.032**	0.038	0.043	0.900	**0.034**	**0.031**	0.034	0.034	0.100	**0.060**	0.066	0.067	**0.065**	0.900
AVG	0.043	**0.063**	**0.060**	0.064	0.065	0.580	**0.059**	**0.056**	0.059	0.061	0.900	0.056	**0.049**	0.050	0.053	0.100	0.080	**0.070**	0.076	0.077	0.900

Table S 9: Comparison of **AUC-PR** for one-vs-rest CIFAR-10 (averaged over all images) using a **CNN** classifier. Except for RP_ρ, ρ_1, ρ_0 are given to all methods. Top model scores are in bold with RP_ρ in red if greater than non-RP models.

		π_1 = 0					π_1 = 0.25					π_1 = 0.5									
		ρ_1 = 0.5					ρ_1 = 0.25					ρ_1 = 0					ρ_1 = 0.5				
IMAGE	TRUE	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16	RP_ρ	RP	NAT13	ELK08	LIU16
PLANE	0.856	0.779	0.780	**0.784**	0.756	0.550	**0.808**	**0.797**	0.770	0.742	0.550	**0.813**	**0.824**	0.792	0.794	0.550	**0.710**	**0.722**	0.662	0.682	0.550
AUTO	0.954	0.874	**0.889**	0.878	0.833	0.550	**0.905**	**0.900**	0.871	0.866	0.550	**0.931**	**0.927**	0.924	0.910	0.550	**0.824**	**0.814**	0.756	0.702	0.550
BIRD	0.761	0.559	0.566	**0.569**	0.568	0.550	**0.619**	**0.618**	0.584	0.597	0.550	**0.623**	**0.679**	0.613	0.619	0.115	0.465	0.492	0.436	0.434	**0.550**
CAT	0.601	0.387	0.447	**0.463**	0.433	0.550	0.423	0.454	0.487	0.480	**0.550**	0.483	**0.512**	0.493	0.473	0.050	0.373	0.375	0.382	0.371	**0.550**
DEER	0.820	**0.620**	0.600	**0.615**	0.573	0.550	0.646	**0.660**	0.610	0.657	0.550	0.658	**0.707**	0.700	0.703	0.550	0.434	0.487	0.414	0.435	**0.550**
DOG	0.758	**0.629**	**0.662**	0.617	0.573	0.550	**0.673**	**0.667**	0.658	0.660	0.550	0.705	0.722	**0.741**	0.705	0.550	0.541	0.545	0.496	0.519	**0.550**
FROG	0.891	0.812	**0.815**	0.812	0.776	0.550	**0.821**	**0.827**	0.808	0.749	0.550	**0.841**	**0.851**	0.828	0.831	0.550	**0.753**	**0.710**	0.691	0.620	0.550
HORSE	0.897	**0.810**	**0.817**	0.799	0.779	0.550	**0.824**	**0.809**	0.801	0.772	0.550	**0.826**	**0.844**	0.818	0.819	0.550	**0.736**	**0.699**	**0.699**	0.600	0.550
SHIP	0.922	**0.870**	0.862	**0.864**	0.853	0.550	**0.889**	**0.885**	0.843	0.848	0.550	0.889	**0.897**	0.891	0.887	0.550	**0.800**	**0.808**	0.767	0.741	0.550
TRUCK	0.929	**0.845**	**0.848**	0.824	0.787	0.550	**0.887**	**0.894**	0.873	0.853	0.550	**0.904**	**0.902**	0.898	0.883	0.550	**0.740**	**0.709**	0.695	0.690	0.550
AVG	0.839	0.719	**0.729**	0.722	0.693	0.550	**0.750**	**0.751**	0.730	0.722	0.550	0.767	**0.787**	0.770	0.762	0.457	**0.637**	**0.636**	0.600	0.579	0.550

References

Angelova, Anelia, Abu-Mostafam, Yaser, and Perona, Pietro. Pruning training sets for learning of object categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pp. 494–501. IEEE, 2005.

Angluin, Dana and Laird, Philip. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Chollet, Francois. *Keras CIFAR CNN*, 2016a. URL http://bit.ly/2mVKR3d.

Chollet, Francois. *Keras MNIST CNN*, 2016b. URL http://bit.ly/2nKiqJv.

Davis, Jesse and Goadrich, Mark. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 233–240, New York, NY, USA, 2006. ACM. ISBN 1-59593-383-2. doi: 10.1145/1143844.1143874. URL http://doi.acm.org/10.1145/1143844.1143874.

Hempstalk, Kathryn, Frank, Eibe, and Witten, Ian H. One-class classification by combining density and class probability estimation. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, ECML PKDD '08, pp. 505–519, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87478-2. doi: 10.1007/978-3-540-87479-9_51. URL http://dx.doi.org/10.1007/978-3-540-87479-9_51.

Manevitz, Larry M. and Yousef, Malik. One-class svms for document classification. *Journal of Machine Learning Research*, 2:139–154, March 2002. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944790.944808.

Moya, M. M., Koch, M. W., and Hostetler, L. D. One-class classifier networks for target recognition applications. *NASA STI/Recon Technical Report N*, 93, 1993.

Platt, John, Schlkopf, Bernhard, Shawe-Taylor, John, Smola, Alex J., and Williamson, Robert C. Estimating the support of a high-dimensional distribution. Technical report, Microsoft Research, November 1999. URL https://www.microsoft.com/en-us/research/publication/estimating-the-support-of-a-high-dimensional-distribution/.

Reed, Scott E., Lee, Honglak, Anguelov, Dragomir, Szegedy, Christian, Erhan, Dumitru, and Rabinovich, Andrew. Training deep neural networks on noisy labels with bootstrapping. In *ICLR 2015*, 2015. URL http://arxiv.org/abs/1412.6596.

scikit learn. *LogisticRegression Class at scikit-learn*, 2016. URL http://bit.ly/2o3y6r5.

Xiao, Tong, Xia, Tian, Yang, Yi, Huang, Chang, and Wang, Xiaogang. Learning from massive noisy labeled data for image classification. In *CVPR*, 2015.