Classification with Asymmetric Label Noise: Consistency and Maximal Denoising

C. Scott\textsuperscript{1}, G. Blanchard\textsuperscript{2}, G. Handy\textsuperscript{1}

\textsuperscript{1} U. Michigan
\textsuperscript{2} U. Potsdam
Contamination model

One contaminated class

Mutual contamination
STANDARD (GENERATIVE) SETTING FOR CLASSIFICATION

- $P_i \equiv P(X | Y = i)$: generating probability distributions for objects of class $i = 0, 1$ on space $\mathcal{X}$.
- Observed: samples
  $$S^i = (X^i_1, \ldots, X^i_{n_i}) \sim P_i$$
- **Goal**: estimate decision function $f : \mathcal{X} \rightarrow \{0, 1\}$
- Various performance error criteria: average classification error, min-max error, Neyman-Pearson error, ...
STANDARD CLASSIFICATION: GENERAL PRINCIPLES

- Approximate $P_i$ by corresponding empirical distribution $\hat{P}_i$
- For all error criteria, key quantities to estimate for classifiers $f$ are

$$R_i(f) := P_i [f(X) \neq i] \rightarrow \hat{R}_i(f) := \frac{1}{n_i} \sum_{j=1}^{n_i} 1\{f(X_j^i) \neq i\}$$

- agnostic/distribution-free philosophy:
  - don’t want a specific (parametric) model for $P_i$.
  - (first) theoretical goal is universal consistency
- basic strategy: uniform probabilistic control of $|R_i(f) - \hat{R}_i(f)|$ over function/set classes $C_k$
- use structural risk minimization to choose adapted class $C_k$
Assume the observed samples are drawn according to a contaminated distribution:

$$\begin{align*}
(X_1^0, \ldots, X_{n_0}^0) & \sim_{i.i.d.} \tilde{P}_0 = (1 - \kappa_0)P_0 + \kappa_0 P_1, \\
(X_1^1, \ldots, X_{n_1}^1) & \sim_{i.i.d.} \tilde{P}_1 = (1 - \kappa_1)P_1 + \kappa_1 P_0
\end{align*}$$

Goal: find a classification function $f$ that performs well for the true distributions.

Can only access/estimate

$$\tilde{R}_i(f) := \tilde{P}_i(f(X) \neq i)$$

via

$$\tilde{R}_i(f) := \frac{1}{n_i} \sum_{j=1}^{n_i} 1 \left\{ f(X_{j}^{(i)}) \neq i \right\}$$
EQUIVALENT MODEL: (ASYMMETRIC) RANDOM LABEL NOISE MODEL

Assume

$$(X_i, Y_i) \overset{i.i.d.}{\sim} P;$$

- true labels $Y_i$ unobserved, instead $\tilde{Y}_i$
- corrupted labels $P[\tilde{Y} = i | Y = j, X] = \zeta_{ij}$
- label corruption assumed not to depend on $X$
- label corruption not symmetric
Detect neutrons and gamma rays; need to classify between them
Training using gamma ray source (e.g. Na-22) and neutron source (e.g. Cf-252)
But: no pure neutron source – always mixed neutron/gamma ray
Additionally, background radiation (both particles)
Previous work on related topics include:

- Learning on positive and unlabeled data (LPUE) (Denis et al. 05, Liu et al. 03)
- Co-training (Blum and Mitchell 98)
- Label noise models and noise-tolerant PAC learning (Angluin and Laird 88, Kearns 93, Aslam and Deactur 96, Cesa-Bianchi et al. 97, Bshouty et al. 98, Kalai and Servedio 03, Stempfel and Ralaivola 09, Jabbari 10)

Generally one or several of the following is assumed:

- \(P_0, P_1\) have non-overlapping support (\(\iff\) deterministic target concept)
- symmetric label noise
- known noise proportions
- criterion is probability of error

We do not assume the above here

Main assumption: label noise independent of \(X\) – no adversarial noise
Assume $P_0, P_1$ have densities $p_0, p_1$

Then $\tilde{P}_0, \tilde{P}_1$ have densities

\[
\begin{align*}
\tilde{p}_0 &= (1 - \kappa_0)p_0 + \kappa_0 p_1 \\
\tilde{p}_1 &= (1 - \kappa_1)p_1 + \kappa_1 p_0
\end{align*}
\]

Simple algebra:

\[
\frac{p_1(x)}{p_0(x)} \leq \lambda \iff \frac{\tilde{p}_1(x)}{\tilde{p}_0(x)} \leq \gamma,
\]

where

\[
\lambda(\gamma) = \frac{\kappa_1 + \gamma(1 - \kappa_1)}{1 - \kappa_0 - \gamma \kappa_0}
\]
Standard classifier trained on data with noisy labels $\tilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\tilde{P}$.

Consider criterion: misclassification probability

$$\mathcal{E}(f) = P[f(X) \neq Y]$$

Identical decisions only for symmetric label noise $\zeta_{01} = \zeta_{10}$
Standard classifier trained on data with noisy labels $\tilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\tilde{P}$.

Consider criterion: \textit{max error}

$$\varepsilon(f) = \max (R_0(f), R_1(f))$$

Identical decisions if $\kappa_0 = \kappa_1$, or $P_0 = P_1$
Label Noise under Different Error Criteria

- Standard classifier trained on data with noisy labels $\tilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\tilde{P}$.
- Consider criterion: balanced error

$$\mathcal{E}(f) = R_0(f) + R_1(f),$$

Then:

$$(1 - \mathcal{E}(f)) = (1 - \kappa_0 - \kappa_1)(1 - \tilde{\mathcal{E}}(f)).$$

- In this case, identical decisions: OK to train on contaminated data (implicit in Blum and Mitchell, 98)
LABEL NOISE UNDER DIFFERENT ERROR CRITERIA

- Standard classifier trained on data with noisy labels $\tilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\tilde{P}$.
- Consider criterion: balanced error

$$\mathcal{E}(f) = R_0(f) + R_1(f),$$

Then:

$$(1 - \mathcal{E}(f)) = (1 - \kappa_0 - \kappa_1)(1 - \tilde{\mathcal{E}}(f))$$

- In this case, identical decisions: OK to train on contaminated data (implicit in Blum and Mitchell, 98)

Overall: training a regular classifier on contaminated data leads to asymptotic bias and inconsistency except in very particular circumstances.
We can surely estimate $\tilde{R}_i(f)$ from its empirical counterpart

$$\tilde{R}_i(f) = \frac{1}{n_i} \sum_{j=1}^{n_i} 1 \{ f(X_j^i) \neq i \},$$

uniformly in $f$ in a limited complexity classifier class $C_K$.

Observe

$$\tilde{P}_0 = (1 - \kappa_0)P_0 + \kappa_0 P_1 \implies \tilde{R}_0(f) = (1 - \kappa_0)R_0(f) + \kappa_0 R_1(f)$$
$$\tilde{P}_1 = (1 - \kappa_1)P_1 + \kappa_1 P_0 \implies \tilde{R}_1(f) = (1 - \kappa_1)R_1(f) + \kappa_1 R_0(f)$$

implying

$$R_0(f) = \frac{(1 - \kappa_1)R_0(f) - \kappa_0 R_1(f)}{1 - (\kappa_0 + \kappa_1)},$$
$$R_1(f) = \frac{(1 - \kappa_0)R_1(f) - \kappa_1 R_0(f)}{1 - (\kappa_0 + \kappa_1)}.$$ 

Key point: estimation of contamination proportions $\kappa_0, \kappa_1$. 

- Key point: estimation of contamination proportions $\kappa_0, \kappa_1$. 

- Key point: estimation of contamination proportions $\kappa_0, \kappa_1$. 

ONLY ONE CONTAMINATED DISTRIBUTION

\[ P_0 \sim P_1 \]

- Observed / Uncontaminated
- Observed / Contaminated
ONLY ONE CONTAMINATED DISTRIBUTION

- Observed / Uncontaminated
- Observed / Contaminated
- Unobserved / Uncontaminated

Diagram with points $P_0$, $\tilde{P}_0$, $P_1$ connected by dashed lines.
**ONLY ONE CONTAMINATED DISTRIBUTION**

- Observed / Uncontaminated
- Observed / Contaminated
- Unobserved / Uncontaminated
**ONLY ONE CONTAMINATED DISTRIBUTION**

- Observed / Uncontaminated
- Observed / Contaminated
- Unobserved / Uncontaminated, irreducible wrt. $P_1$

Diagram:

- $P_0$
- $P_1$
- $\tilde{P}_0$

Equation:

$\tilde{P}_0 \sim P_0 \sim P_1$
ONLY $\tilde{P}_0$ CONTAMINATED: IDENTIFIABLITY
[BLANCHARD, SCOTT, LEE 2010]

\[
\begin{align*}
(X_0^1, \ldots, X_{n_0}^1) \overset{i.i.d.}{\sim} P_0 &= (1 - \kappa_0) P_0 + \kappa_0 P_1 \\
(X_1^0, \ldots, X_{n_1}^0) \overset{i.i.d.}{\sim} P_1
\end{align*}
\]

- Define the “maximum proportion of source $H$ in $F$”

$$\kappa^*(F|H) = \max \left\{ \kappa \in [0, 1] \mid \exists \text{ a distribution } G \text{ s.t. } F = (1 - \kappa) G + \kappa H \right\};$$

- The following holds:

$$\kappa_0 = \kappa^*(\tilde{P}_0|P_1) \leftrightarrow \kappa^*(P_0|P_1) = 0 \quad (P_0 \text{ is irreducible wrt. } P_1)$$
**ONLY \( \tilde{P}_0 \) CONTAMINATED: ESTIMATION**

[Blanchard, Scott, Lee 2010]

- \( F, H \) distributions; Lebesgue decomposition:
  \[
  F = F_H + F_H^\perp,
  \]
  with \( F_H \ll H \) and \( (F_H^\perp, H) \) mutually singular;

\[
\kappa^*(F|H) = \operatorname{Ess.Inf.} \frac{dF_H}{dH} = \inf_{C:H(C)>0} \frac{F(C)}{H(C)}
\]

- Suggests the estimator
  \[
  \hat{\kappa}(\tilde{P}_0|\hat{P}_1) = \inf_{C \in C_k} \frac{\tilde{P}_0(C) + \varepsilon_k}{\hat{P}_1(C) - \varepsilon_k^+}
  \]

- \( \hat{\kappa}(\tilde{P}_0|\hat{P}_1) \geq \kappa^*(\tilde{P}_0|P_1) \) with high probability

- Appropriate choice of \( \varepsilon_k \) + take inf. over sequence of nested classes \( C_1 \subset C_2 \subset \ldots \) with universal approximation property yields universally consistent estimator
MUTUAL CONTAMINATION

\[ \tilde{P}_0 \quad P_1 \]

- Blue dot: Observed / Uncontaminated
- Red dot: Observed / Contaminated
Mutual Contamination

\[ \tilde{P}_0 \quad \tilde{P}_1 \]

Observed / Contaminated
Mutual Contamination

\[
\begin{align*}
\tilde{P}_0 &= (1 - \kappa_0)P_0 + \kappa_0 P_1, \\
\tilde{P}_1 &= (1 - \kappa_1)P_1 + \kappa_1 P_0
\end{align*}
\]

Proposition (Decoupled Representation)
Assume \( P_0 \neq P_1 \) and
\[(A) \quad \kappa_1 + \kappa_2 < 1; \]
then \( \tilde{P}_0 \neq \tilde{P}_1 \), and there exist unique \( 0 \leq \tilde{\kappa}_0, \tilde{\kappa}_1 < 1 \) such that
\[
\begin{align*}
\tilde{P}_0 &= (1 - \tilde{\kappa}_0)P_0 + \tilde{\kappa}_0 P_1, \\
\tilde{P}_1 &= (1 - \tilde{\kappa}_1)P_1 + \tilde{\kappa}_1 P_0.
\end{align*}
\]
with
\[
\tilde{\kappa}_0 = \frac{\kappa_0}{1 - \kappa_1} < 1; \quad \tilde{\kappa}_1 = \frac{\kappa_1}{1 - \kappa_0} < 1.
\]
The Two Representations

Decoupled representation

\[
\begin{align*}
\tilde{P}_0 &= (1 - \tilde{\kappa}_0)P_0 + \tilde{\kappa}_0\tilde{P}_1, \\
\tilde{P}_1 &= (1 - \tilde{\kappa}_1)P_1 + \tilde{\kappa}_1\tilde{P}_0.
\end{align*}
\]

Original representation

\[
\begin{align*}
\tilde{P}_0 &= (1 - \kappa_0)P_0 + \kappa_0P_1, \\
\tilde{P}_1 &= (1 - \kappa_1)P_1 + \kappa_1P_0.
\end{align*}
\]
IDENTIFIABILITY

Decoupled model:
\[
\begin{align*}
\tilde{P}_0 &= (1 - \tilde{\kappa}_0)P_0 + \tilde{\kappa}_0\tilde{P}_1, \\
\tilde{P}_1 &= (1 - \tilde{\kappa}_1)P_1 + \tilde{\kappa}_1\tilde{P}_0.
\end{align*}
\]

From the results on mixture proportion estimation: we can estimate \( \tilde{\kappa}_0 \) consistently if \( \kappa(P_0, \tilde{P}_1) = 0 \)

**Lemma**

*Under assumption (A): \( \kappa_0 + \kappa_1 < 1 \), it holds*

\[
(C) \quad \begin{cases}
\kappa(P_0|\tilde{P}_1) = 0 \\
\kappa(P_1|\tilde{P}_0) = 0
\end{cases} \iff \begin{cases}
\kappa(P_0|P_1) = 0 \\
\kappa(P_1|P_0) = 0
\end{cases} \quad (B)
\]

*(C): \( P_0 \) and \( P_1 \) are mutually irreducible*
Identifiability

\[ \tilde{P}_0 \quad \tilde{P}_1 \quad \text{observed / contaminated} \]
IDENTIFIABILITY

mutually irreducible

unobserved / uncontaminated, observed / contaminated

\[ P_0 \quad \tilde{P}_0 \quad \tilde{P}_1 \quad P_1 \]

Observed / Contaminated
Unobserved / Uncontaminated, mutually irreducible
MUTUAL IRREDUCIBILITY

- Top: mutually irreducible
- Middle: mutually irreducible
- Bottom: $P_1$ irreducible wrt $P_0$, but $P_0$ not irreducible wrt $P_0$. 
**Mutual Irreducibility**

Under joint distribution model

\[(X, Y) \sim \mathbb{P}_{XY}, \quad \eta(x) = \mathbb{P}_{XY}[Y = 1|X = x]\]

Then:

\[
\begin{align*}
\kappa(P_0|P_1) = 0 \quad &\iff\quad \text{Ess.Sup.}_x \eta(x) = 1, \\
\kappa(P_1|P_0) = 0 \quad &\iff\quad \text{Ess.Inf.}_x \eta(x) = 0,
\end{align*}
\]
CHARACTERIZING THE IRREDUCIBLE SOLUTION

For given observed contaminated \( \tilde{P}_0 \neq \tilde{P}_1 \), let \( \Delta \) be the convex set of quadruples \((\kappa_0, \kappa_1, P_0, P_1)\) satisfying (A) and solution of:

\[
\begin{aligned}
\tilde{P}_0 &= (1 - \kappa_0)P_0 + \kappa_0 P_1, \\
\tilde{P}_1 &= (1 - \kappa_1)P_1 + \kappa_1 P_0
\end{aligned}
\]  

Proposition

The solution \((\kappa_0^*, \kappa_1^*, P_0^*, P_1^*)\) is characterized as either of:

- the unique quadruple for which \((P_0, P_1)\) are mutually irreducible;
- the unique nontrivial \((\kappa_0 \neq 0, \kappa_1 \neq 0)\) extremal point of \(\Lambda\);
- the unique maximizer of \((\kappa_0 + \kappa_1)\) over \(\Lambda\);
- the unique maximizer over \(\Lambda\) of \(\|P_0 - P_1\|_{TV}\);
- the unique minimizer over \(\Lambda\) of optimal balanced error for classifying \(P_0\) vs. \(P_1\).

Interpretation: maximal denoising / source separation
\[ \tilde{\kappa}^*_0 = \kappa^*(\tilde{P}_0, \tilde{P}_1) \]

\[ \tilde{\kappa}^*_1 = \kappa^*(\tilde{P}_1, \tilde{P}_0) \]
CONSISTENT ESTIMATION OF CONTAMINATION PROPORTIONS

Decoupled representation:

\[
\begin{align*}
\tilde{P}_0 &= (1 - \tilde{\kappa}_0)P_0 + \tilde{\kappa}_0 \tilde{P}_1, \\
\tilde{P}_1 &= (1 - \tilde{\kappa}_1)P_1 + \tilde{\kappa}_1 \tilde{P}_0.
\end{align*}
\]

- \((P_0, P_1)\) mutually irreducible \(\Rightarrow\) \(P_0\) irreducible wrt \(\tilde{P}_1\), and \(P_1\) irreducible wrt. \(\tilde{P}_1\)

- leverage case of only one contaminated distribution (twice):

\[
\begin{align*}
\tilde{\kappa}_0 &= \tilde{\kappa}(\tilde{P}_0|\tilde{P}_1); & \tilde{\kappa}_1 &= \tilde{\kappa}(\tilde{P}_1|\tilde{P}_0)
\end{align*}
\]

- Then

\[
\begin{align*}
\hat{\kappa}_0 &= \frac{\tilde{\kappa}_0(1 - \tilde{\kappa}_1)}{1 - \tilde{\kappa}_0 \tilde{\kappa}_1}; & \hat{\kappa}_1 &= \frac{\tilde{\kappa}_1(1 - \tilde{\kappa}_0)}{1 - \tilde{\kappa}_0 \tilde{\kappa}_1}
\end{align*}
\]

are universally consistent estimators of \(\kappa_0, \kappa_1\) under (A), (C).
Construction of estimator for type II error:

\[ \tilde{P}_0 = (1 - \tilde{\kappa}_0)P_0 + \tilde{\kappa}_0 \tilde{P}_1 \Rightarrow R_0(f) = \frac{\tilde{R}_0(f) - \tilde{\kappa}_0(1 - \tilde{R}_1(f))}{1 - \tilde{\kappa}_0} \]

\[ \Rightarrow \hat{R}_0(f) = \frac{\tilde{R}_0(f) - \hat{\kappa}_0(1 - \hat{R}_1(f))}{1 - \hat{\kappa}_0} \]

- Convergence uniform over e.g. VC-Classes of classifiers \( f \)
- Can apply SRM principle to choose appropriate model
- Can construct universally consistent estimators for various error measures
CONCLUSION

Contributions:
- nonparametric/distribution-free point of view of the (asymmetric) contamination problem
- existence and unicity of irreducible solution; characteristic properties
- consistent estimation of contamination weights under irreducibility
- consistent estimation of optimal decisions under different error criteria

Further work:
- rates
- multiclass case → more challenging
THANK YOU