# Classification with Asymmetric Label Noise: Consistency and Maximal Denoising

C. Scott[1], **G. Blanchard**[2], G. Handy[1]

[1] U. Michigan
[2] U. Potsdam



# Plan

1. Contamination model
2. One contaminated class
3. 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

$$S^i = (X_1^i, \ldots, X_{n_i}^i) \overset{i.i.d}{\sim} P_i$$

- **Goal:** estimate decision function $f : \mathcal{X} \to \{0, 1\}$
- Various performance error criteria: average classification error, min-max error, Neyman-Pearson error, ...

# STANDARD CLASSIFICATION: GENERAL PRINCIPLES

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

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

- 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 $\left|R_i(f) - \widehat{R}_i(f)\right|$ over function/set classes $\mathcal{C}_k$
- use structural risk minimization to choose adapted class $\mathcal{C}_k$

# CONTAMINATION MODEL

- Assume the observed samples are drawn according to a contaminated distribution:

$$\begin{cases} (X_1^0, \ldots, X_{n_0}^0) \overset{i.i.d.}{\sim} \widetilde{P}_0 & = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ (X_1^1, \ldots, X_{n_1}^1) \overset{i.i.d.}{\sim} \widetilde{P}_1 & = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases}$$

- Goal: find a classification function $f$ that performs well for the true distributions.
- Can only access/ estimate

$$\widetilde{R}_i(f) := \widetilde{P}_i\left(f(X) \neq i\right)$$

via

$$\widehat{\widetilde{R}}_i(f) := \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{1}\left\{f(X_j^{(i)}) \neq i\right\}$$

# Equivalent Model: (Asymmetric) Random Label Noise Model

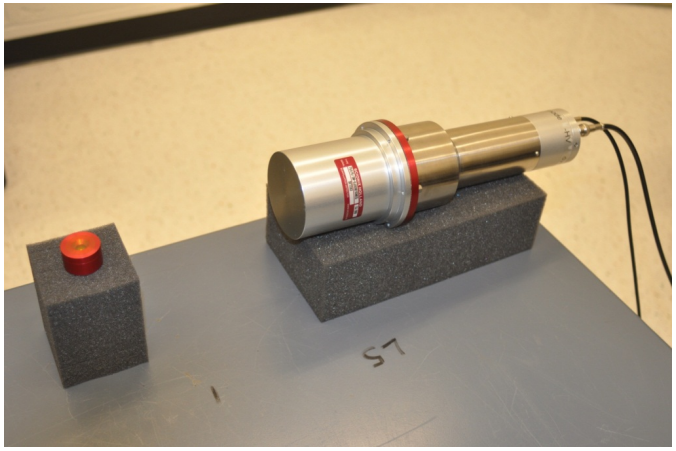
Assume

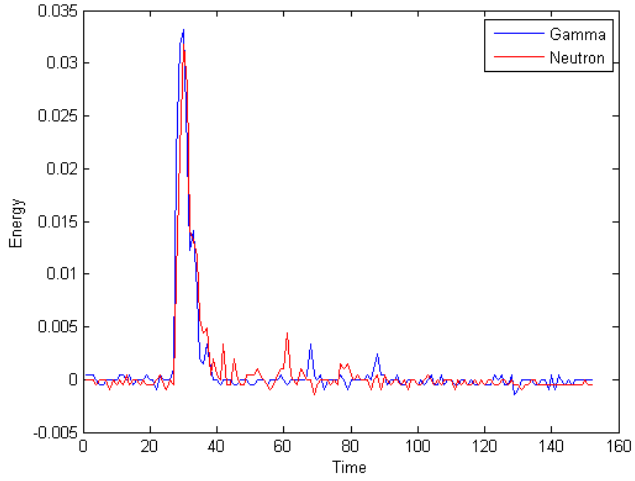
$$(X_i, Y_i) \overset{i.i.d.}{\sim} P;$$

- true labels $Y_i$ unobserved, instead $\widetilde{Y}_i$
- corrupted labels $P\left[\widetilde{Y} = i | Y = j, X\right] = \zeta_{ij}$
- label corruption assumed not to depend on $X$
- label corruption not symmetric

# Motivating Application

## Organic Scincillation detector






- 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)

# Related Work, Previous Assumptions

- 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 ( $\leftrightarrow$ deterministic target concept )
  - symmetric label noise
  - known noise proportions
  - criterion is probablity of error
- We do not assume the above here
- Main asumption: label noise independent of $X$ – no adversarial noise

# UNDERSTANDING LABEL NOISE

- Assume $P_0, P_1$ have densities $p_0, p_1$
- Then $\widetilde{P}_0, \widetilde{P}_1$ have densities

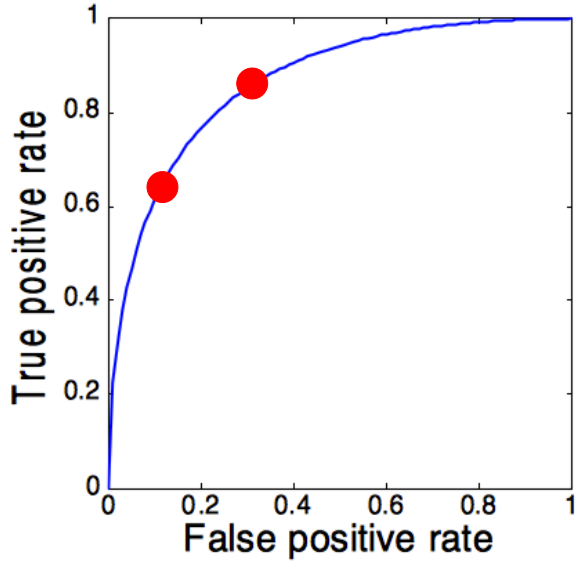
$$\begin{cases} \widetilde{p}_0 = (1-\kappa_0)p_0 + \kappa_0 p_1 \\ \widetilde{p}_1 = (1-\kappa_1)p_1 + \kappa_1 p_0 \end{cases}$$

Simple algebra:

$$\frac{p_1(x)}{p_0(x)} \lessgtr \lambda \iff \frac{\widetilde{p}_1(x)}{\widetilde{p}_0(x)} \lessgtr \gamma,$$

where

$$\lambda(\gamma) = \frac{\kappa_1 + \gamma(1-\kappa_1)}{1-\kappa_0-\gamma\kappa_0}$$



# Label Noise under Different Error Criteria

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

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

- Identical decisions only for symmetric label noise $\zeta_{01} = \zeta_{10}$

# Label Noise under Different Error Criteria

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

$$\mathcal{E}(f) = \max\left(R_0(f), R_1(f)\right)$$

- 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 $\widetilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\widetilde{P}$.
- Consider criterion: balanced error

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

Then: $$(1 - \mathcal{E}(f)) = (1 - \kappa_0 - \kappa_1)(1 - \widetilde{\mathcal{E}}(f))$$

- 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 $\widetilde{Y}$ is consistent only when optimal decision is identical under $P$ and $\widetilde{P}$.
- Consider criterion: balanced error

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

Then: $$(1 - \mathcal{E}(f)) = (1 - \kappa_0 - \kappa_1)(1 - \widetilde{\mathcal{E}}(f))$$

- 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 $\widetilde{R}_i(f)$ from its empirical counterpart

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

uniformly in $f$ in a limited complexity classifier class $\mathcal{C}_K$

- Observe

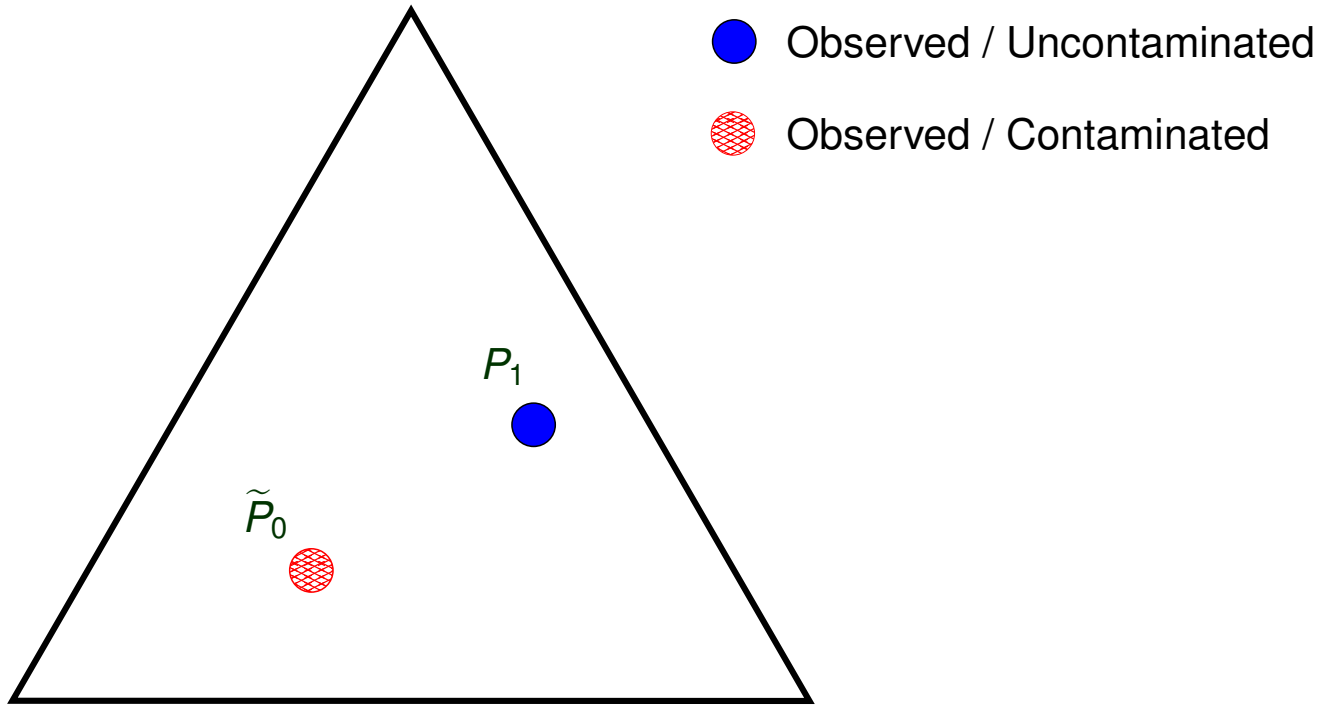
$$\widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1 \Longrightarrow \quad \widetilde{R}_0(f) = (1-\kappa_0)R_0(f) + \kappa_0 R_1(f)$$
$$\widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \Longrightarrow \quad \widetilde{R}_1(f) = (1-\kappa_1)R_1(f) + \kappa_1 R_0(f)$$

implying

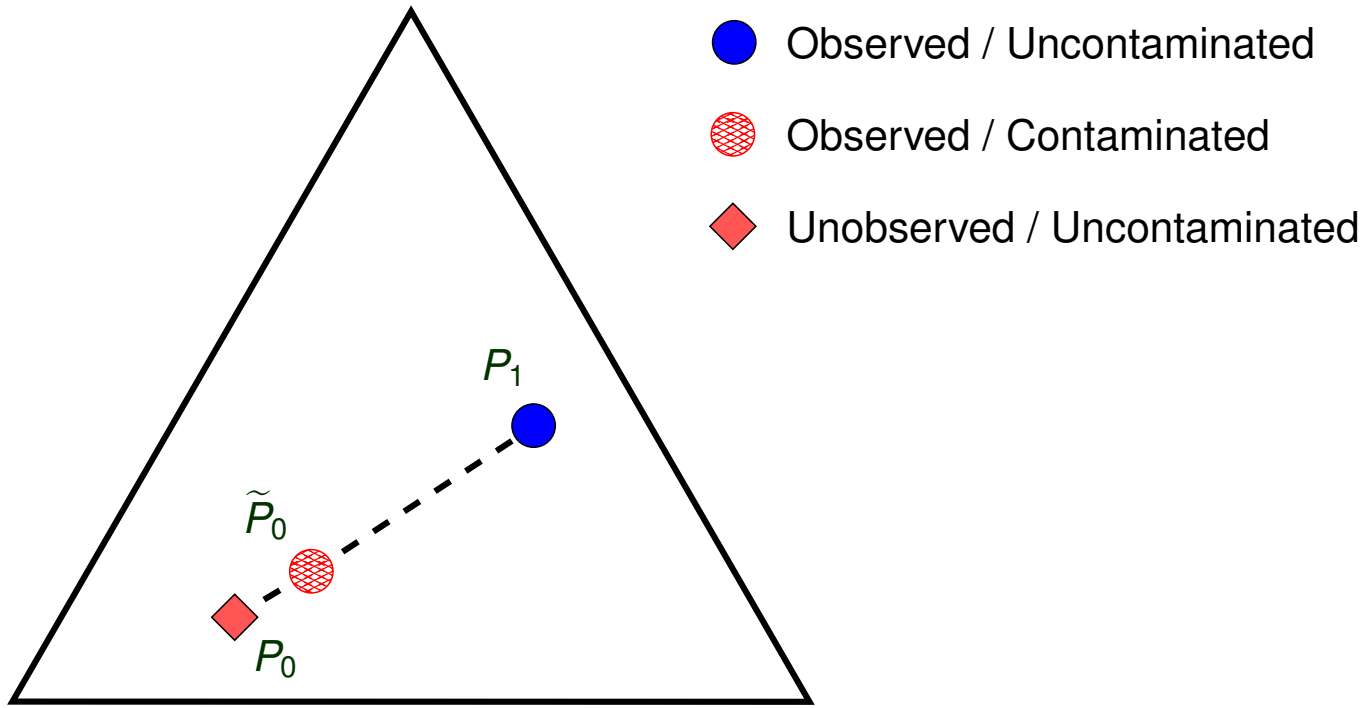
$$R_0(f) = \frac{(1-\kappa_1)R_0(f) - \kappa_0 R_1(f)}{1-(\kappa_0+\kappa_1)},$$
$$R_1(f) = \frac{(1-\kappa_0)R_1(f) - \kappa_1 R_0(f)}{1-(\kappa_0+\kappa_1)}$$

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

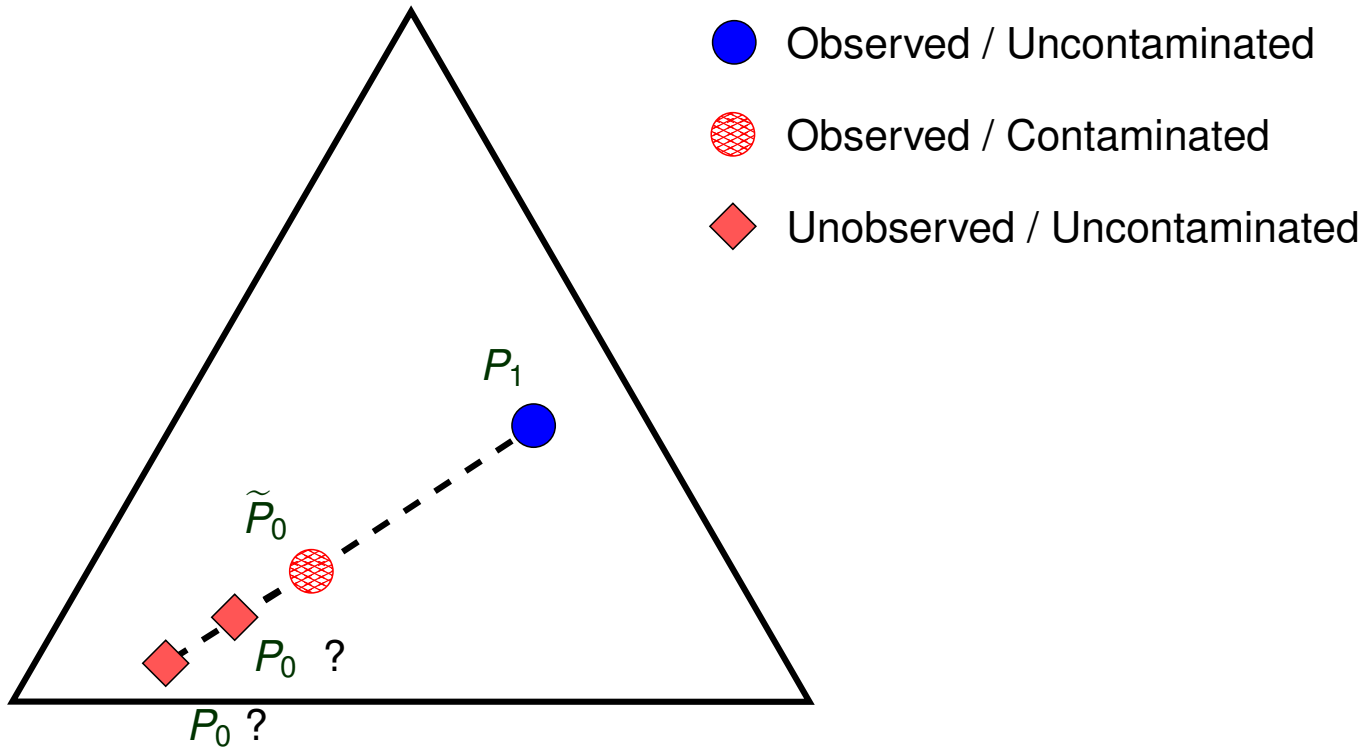
# Only One Contaminated Distribution



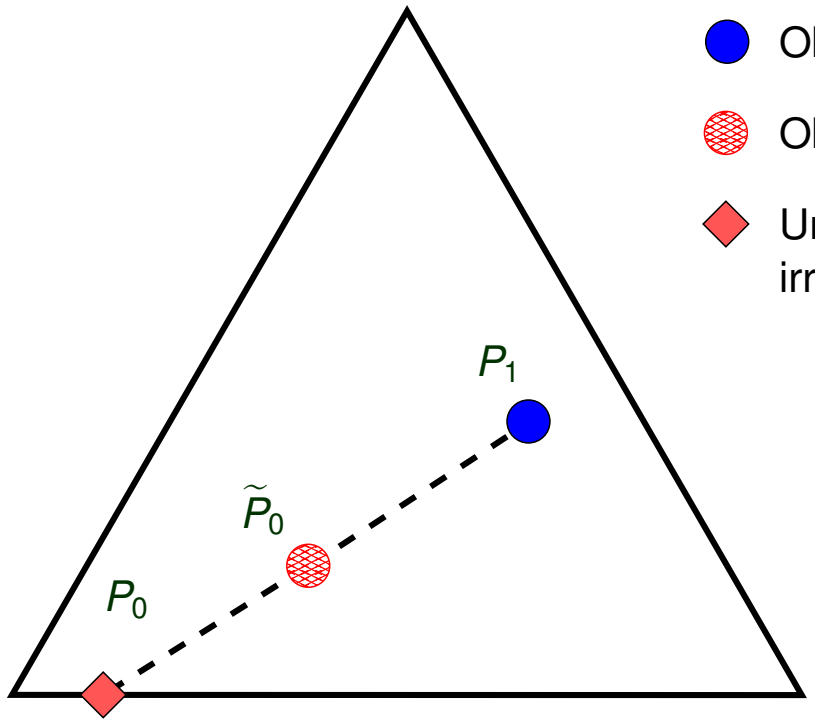
# Only One Contaminated Distribution



# Only One Contaminated Distribution



# Only One Contaminated Distribution




Observed / Uncontaminated

Observed / Contaminated

Unobserved / Uncontaminated, irreducible wrt. $P_1$

# ONLY $\widetilde{P}_0$ CONTAMINATED: IDENTIFIABLITY

[BLANCHARD, SCOTT, LEE 2010]

$$\begin{cases} (X_1^0, \ldots, X_{n_0}^0) \overset{i.i.d.}{\sim} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1 \\ (X_1^1, \ldots, X_{n_1}^1) \overset{i.i.d.}{\sim} P_1 \end{cases}$$

- Define the "maximum proportion of source $H$ in $F$"

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

- The following holds:

$$\kappa_0 = \kappa^*(\widetilde{P}_0|P_1) \Leftrightarrow \kappa^*(P_0|P_1) = 0 \qquad (P_0 \text{ is irreducible wrt.} P_1)$$

# ONLY $\widetilde{P}_0$ 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;

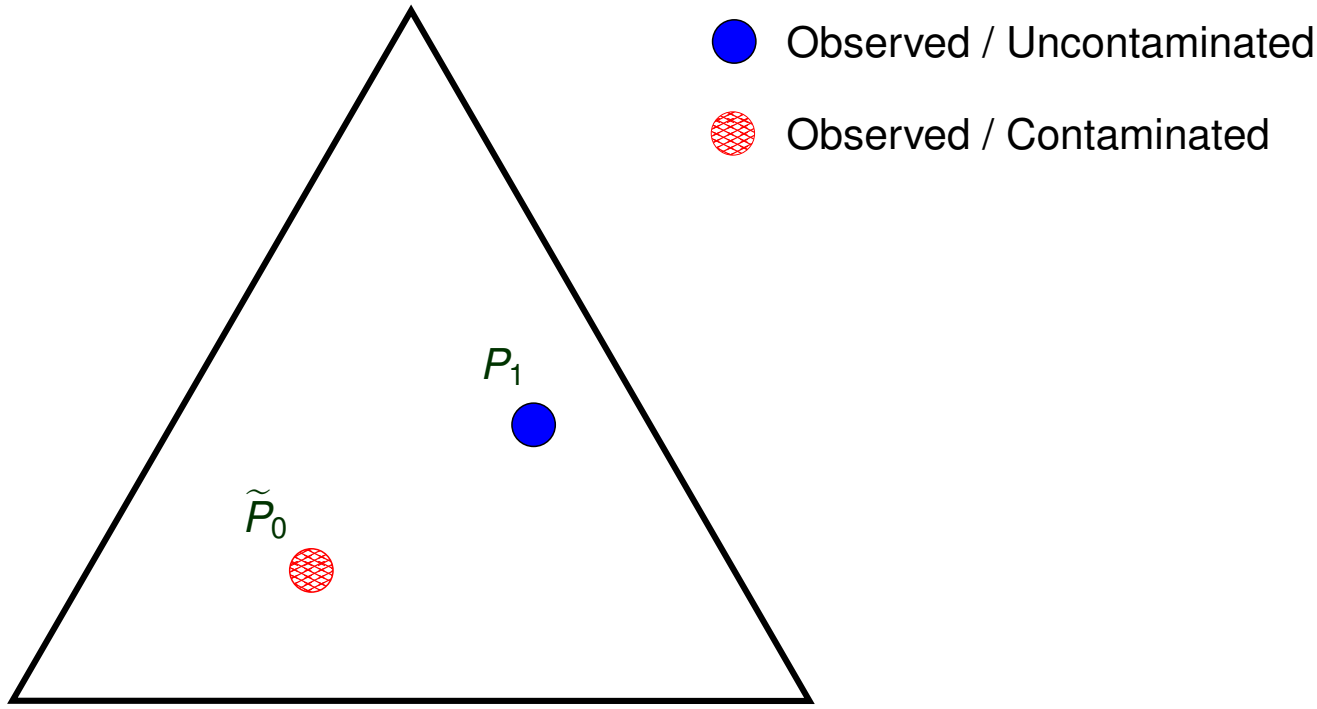
$$\kappa^*(F|H) = \text{Ess.Inf.} \frac{dF_H}{dH} = \inf_{C:H(C)>0} \frac{F(C)}{H(C)}$$

- Suggests the estimator

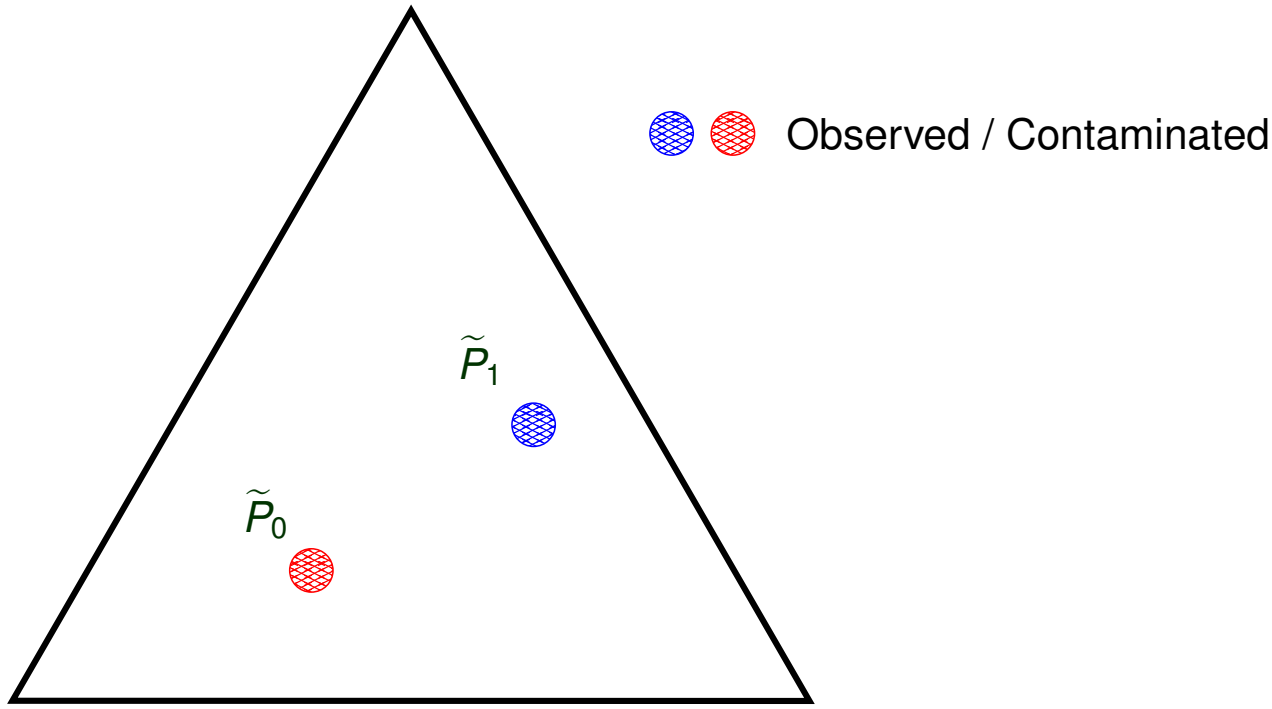
$$\widehat{\kappa}(\widehat{\widetilde{P}}_0|\widehat{P}_1) = \inf_{C \in \mathcal{C}_k} \frac{\widehat{\widetilde{P}}_0(C) + \varepsilon_k}{\left(\widehat{P}_1(C) - \varepsilon_k\right)_+}$$

- $\widehat{\kappa}(\widehat{\widetilde{P}}_0|\widehat{P}_1) \geq \kappa^*(\widetilde{P}_0|P_1)$ with high probability
- Appropriate choice of $\varepsilon_k$ + take inf. over sequence of nested classes $\mathcal{C}_1 \subset \mathcal{C}_2 \subset \ldots$ with universal approximation property yields universally consistent estimator

# Mutual Contamination



# Mutual Contamination



# Mutual Contamination

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases}$$

## Proposition (Decoupled Representation)

*Assume $P_0 \neq P_1$ and*

$$(\mathbf{A}) \qquad \kappa_1 + \kappa_2 < 1\,;$$

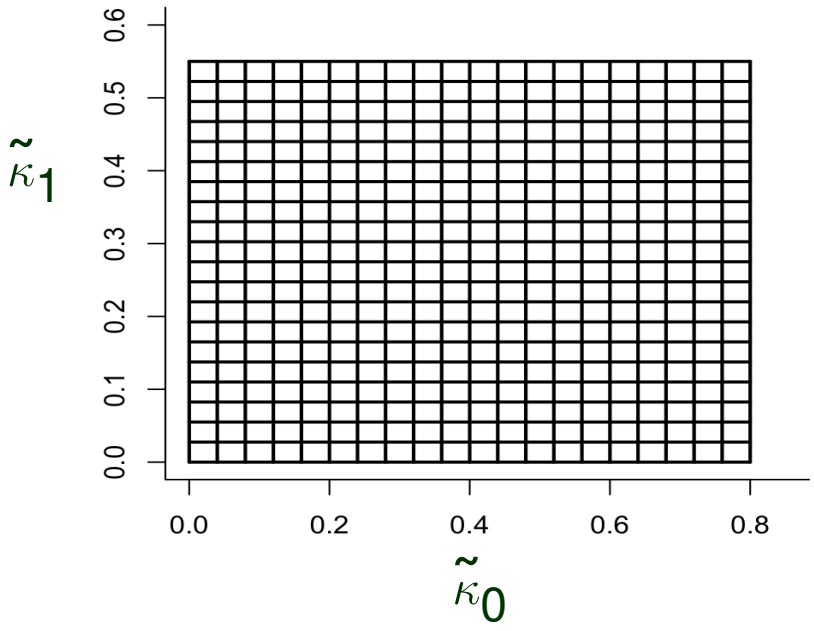
*then $\widetilde{P}_0 \neq \widetilde{P}_1$, and there exist unique $0 \leq \widetilde{\kappa}_0, \widetilde{\kappa}_1 < 1$ such that*

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0 \widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1 \widetilde{P}_0. \end{cases}$$

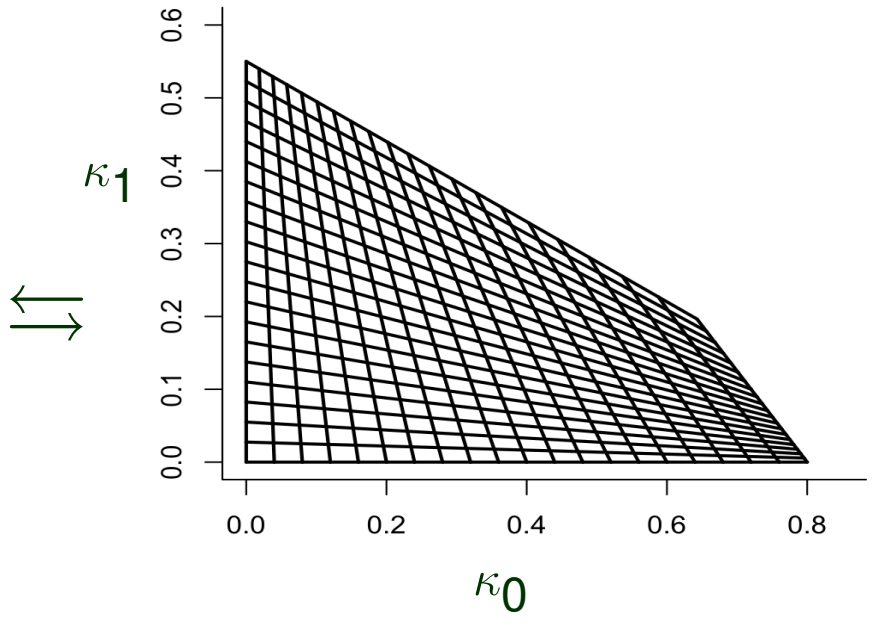
*with*

$$\widetilde{\kappa}_0 = \frac{\kappa_0}{1-\kappa_1} < 1; \qquad \widetilde{\kappa}_1 = \frac{\kappa_1}{1-\kappa_0} < 1.$$

# The Two Representations




Decoupled representation




Original representation

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0\widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1\widetilde{P}_0. \end{cases}$$

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases}$$

# IDENTIFIABILITY

Decoupled model:

$$\begin{cases} \widetilde{P}_0 = (1 - \widetilde{\kappa}_0) P_0 + \widetilde{\kappa}_0 \widetilde{P}_1, \\ \widetilde{P}_1 = (1 - \widetilde{\kappa}_1) P_1 + \widetilde{\kappa}_1 \widetilde{P}_0. \end{cases}$$

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

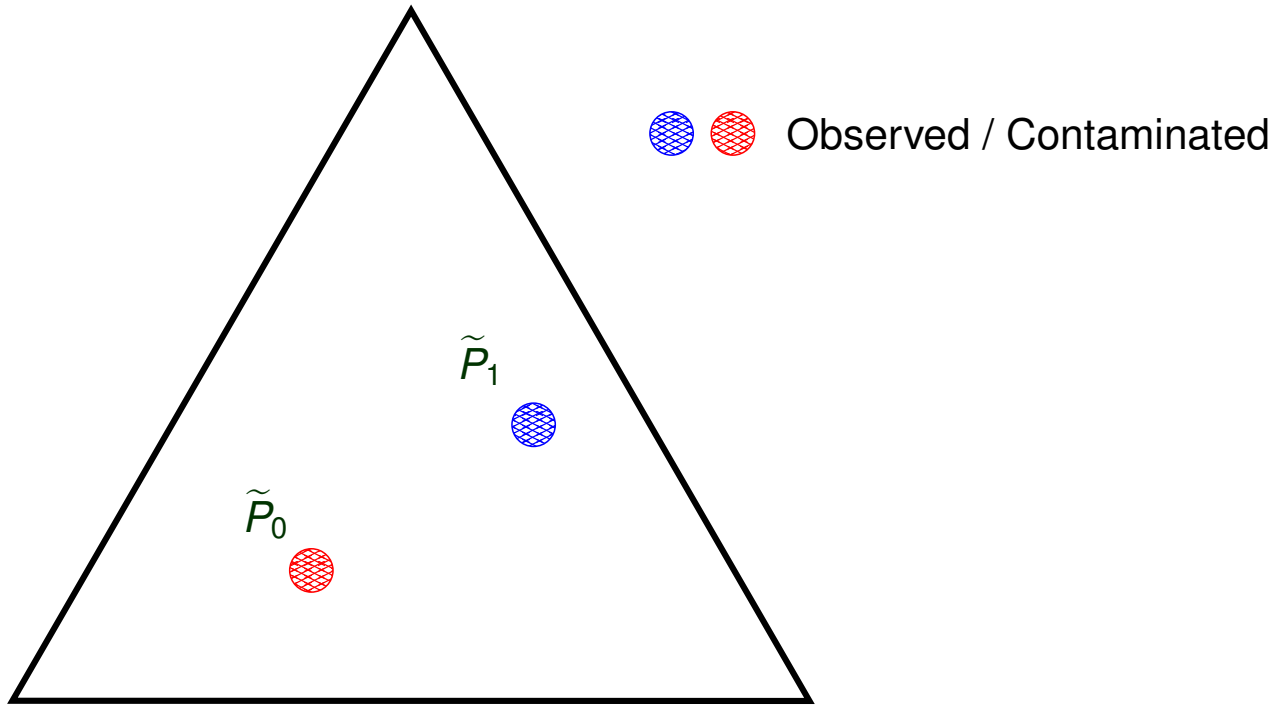
## Lemma

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

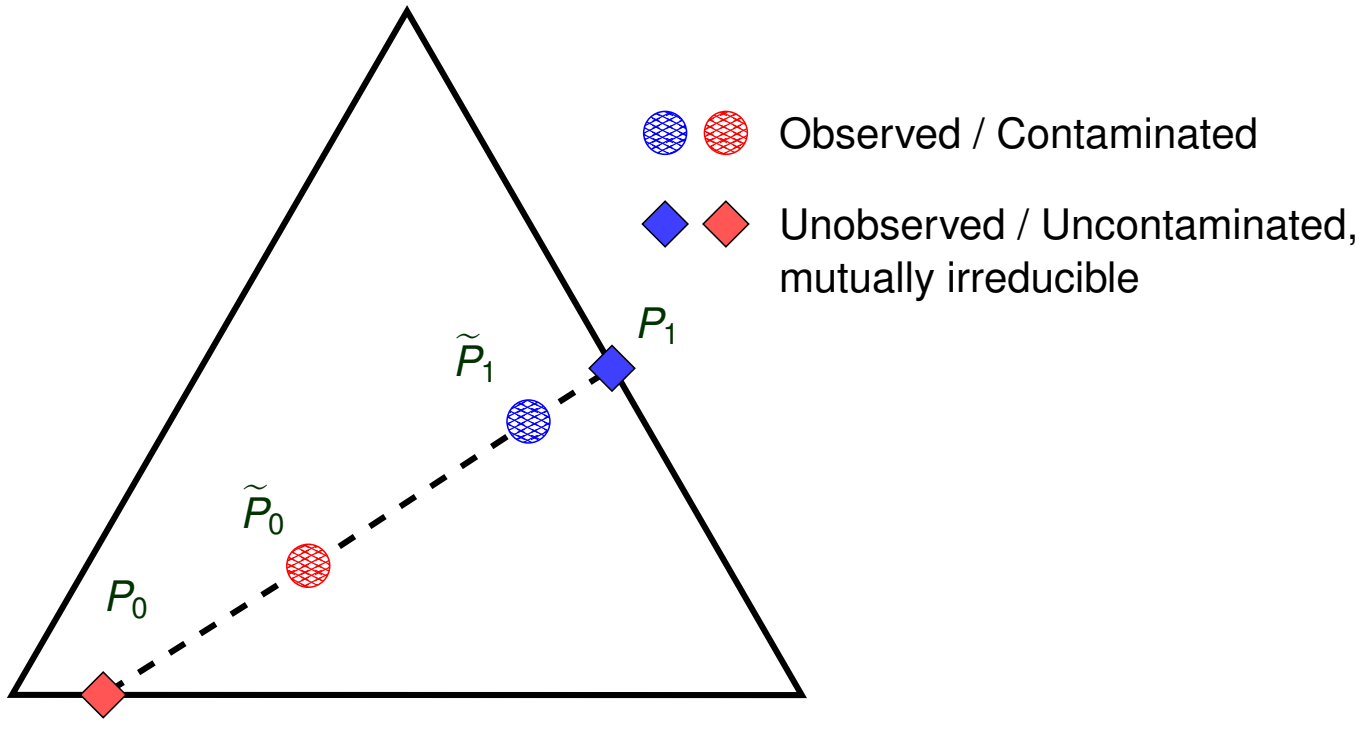
$$(\mathbf{B}) \left\{ \begin{array}{l} \kappa(P_0 | \widetilde{P}_1) = 0 \\ \kappa(P_1 | \widetilde{P}_0) = 0 \end{array} \right\} \iff \left\{ \begin{array}{l} \kappa(P_0 | P_1) = 0 \\ \kappa(P_1 | P_0) = 0 \end{array} \right\} (\mathbf{C})$$

**(C)**: $P_0$ and $P_1$ are mutually irreducible

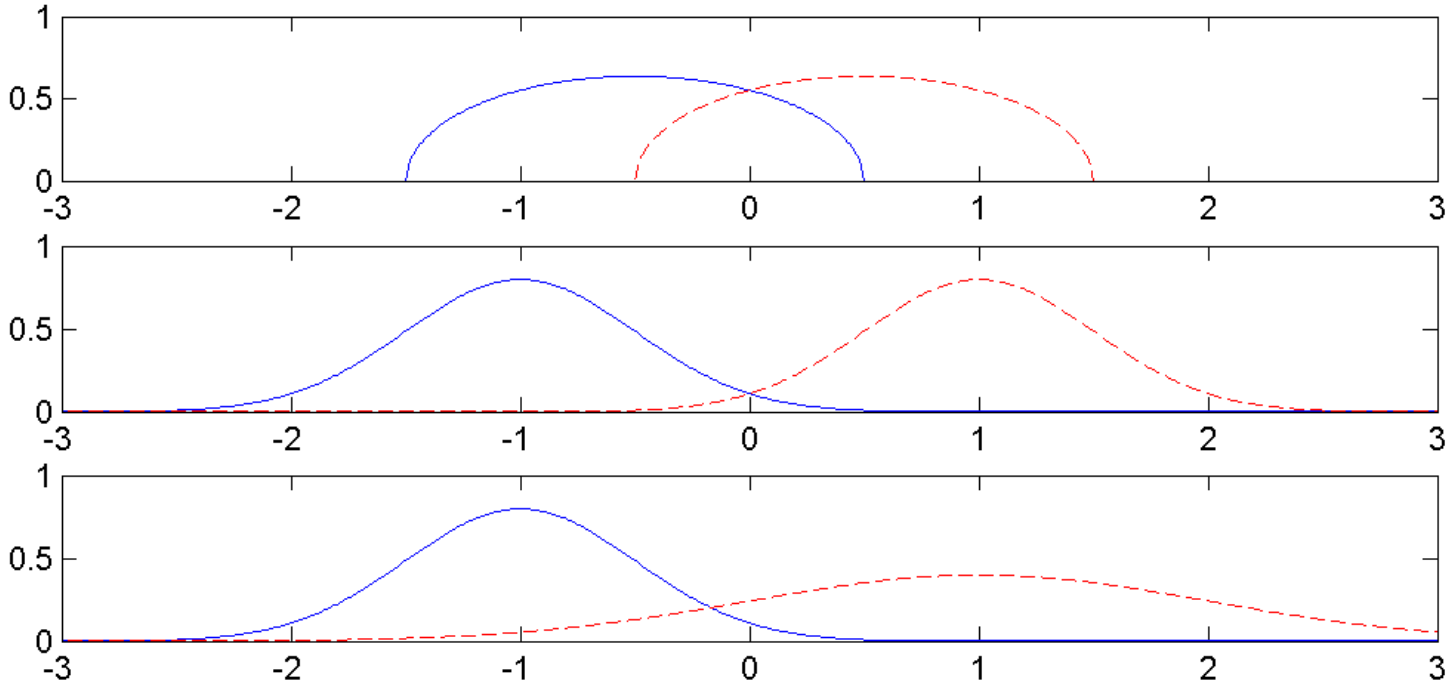
# IDENTIFIABILITY



# IDENTIFIABILITY



# 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

$$(X, Y) \sim \mathbb{P}_{XY}, \qquad \eta(x) = P_{XY}\left[Y = 1 | X = x\right]$$

Then:

$$\left.\begin{array}{l} \kappa(P_0|P_1) = 0 \\ \kappa(P_1|P_0) = 0 \end{array}\right\} \Leftrightarrow \begin{cases} \mathrm{Ess.Sup.}_x \eta(x) = 1, \\ \mathrm{Ess.Inf.}_x \eta(x) = 0, \end{cases}$$

# CHARACTERIZING THE IRREDUCIBLE SOLUTION

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

$$\begin{cases} \widetilde{P}_0 = (1-\kappa_0)P_0 + \kappa_0 P_1, \\ \widetilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 P_0 \end{cases} \tag{1}$$

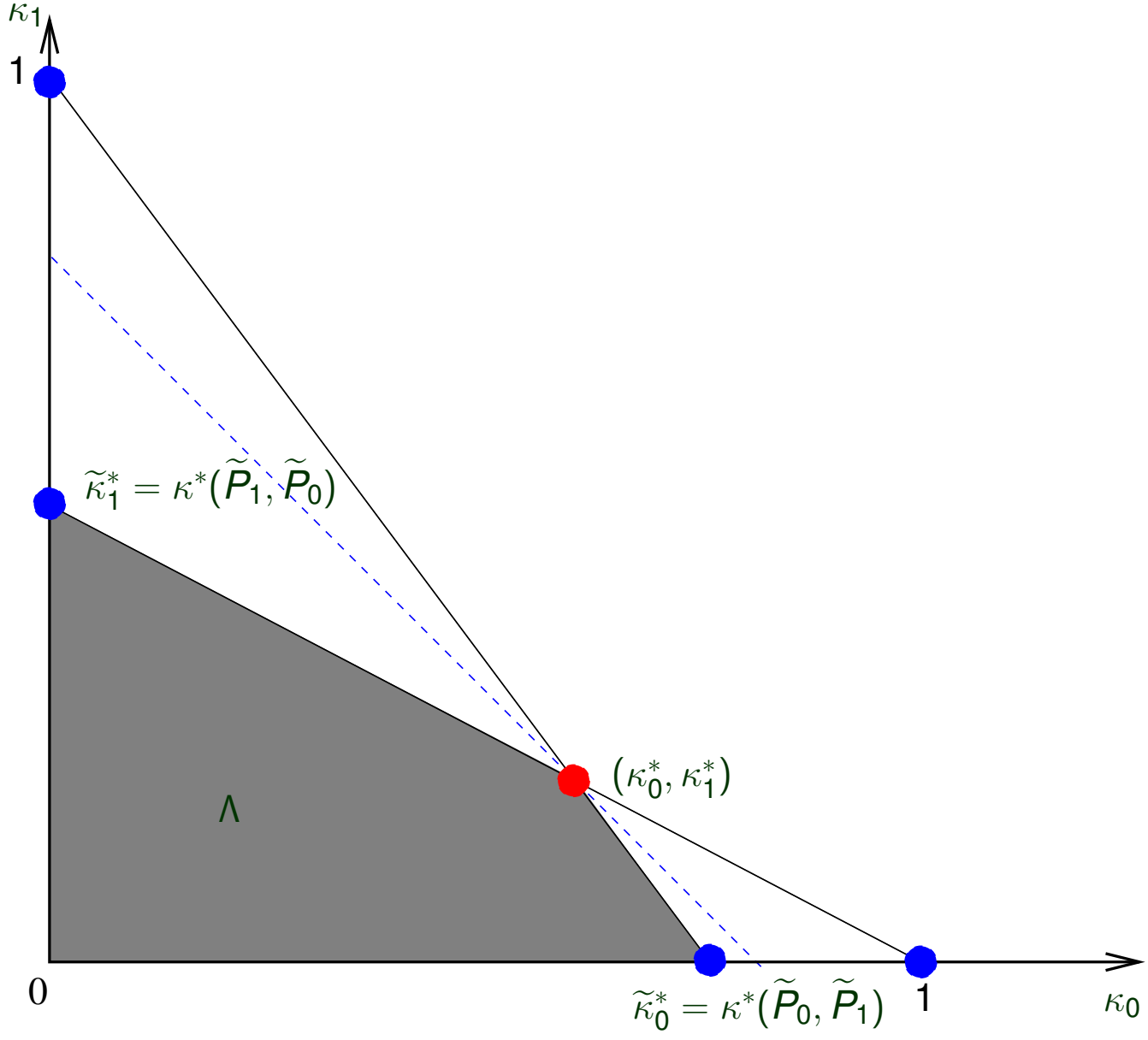
## Proposition

*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

# GEOMETRY OF SOLUTIONS



# Consistent Estimation of Contamination Proportions

Decoupled representation:

$$\begin{cases} \widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0\widetilde{P}_1, \\ \widetilde{P}_1 = (1-\widetilde{\kappa}_1)P_1 + \widetilde{\kappa}_1\widetilde{P}_0. \end{cases}$$

- $(P_0, P_1)$ mutually irreducible $\Rightarrow$ $P_0$ irreducible wrt $\widetilde{P}_1$, and $P_1$ irreducible wrt. $\widetilde{P}_1$
- leverage case of only one contaminated distribution (twice):

$$\widehat{\widetilde{\kappa}}_0 = \widehat{\kappa}(\widehat{\widetilde{P}}_0|\widehat{\widetilde{P}}_1); \qquad \widehat{\widetilde{\kappa}}_1 = \widehat{\kappa}(\widehat{\widetilde{P}}_1|\widehat{\widetilde{P}}_0)$$

- Then

$$\widehat{\kappa}_0 = \frac{\widehat{\widetilde{\kappa}}_0(1-\widetilde{\kappa}_1)}{1-\widetilde{\kappa}_0\widetilde{\kappa}_1}; \qquad \widehat{\kappa}_1 = \frac{\widehat{\widetilde{\kappa}}_1(1-\widetilde{\kappa}_0)}{1-\widetilde{\kappa}_0\widetilde{\kappa}_1}$$

  are universally consistent estimators of $\kappa_0, \kappa_1$ under **(A)**, **(C)**.

# Consistent Estimation of Risk

- Construction of estimator for type II error:

$$\widetilde{P}_0 = (1-\widetilde{\kappa}_0)P_0 + \widetilde{\kappa}_0\widetilde{P}_1 \Rightarrow R_0(f) = \frac{\widetilde{R}_0(f) - \widetilde{\kappa}_0(1-\widetilde{R}_1(f))}{1-\widetilde{\kappa}_0}$$

$$\rightarrow \widehat{R}_0(f) = \frac{\widehat{\widetilde{R}}_0(f) - \widehat{\widetilde{\kappa}}_0(1-\widehat{\widetilde{R}}_1(f))}{1-\widehat{\widetilde{\kappa}}_0}$$

- 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 $\rightarrow$ more challenging

# THANK YOU