# Decontamination of Mutually Contaminated Models

**Gilles Blanchard**
Universität Potsdam

**Clayton Scott**
University of Michigan

## Abstract

A variety of machine learning problems are characterized by data sets that are drawn from multiple different convex combinations of a fixed set of base distributions. We call this a *mutual contamination* model. In such problems, it is often of interest to recover these base distributions, or otherwise discern their properties. This work focuses on the problem of classification with multiclass label noise, in a general setting where the noise proportions are unknown and the true class distributions are nonseparable and potentially quite complex. We develop a procedure for decontamination of the contaminated models from data, which then facilitates the design of a consistent discrimination rule. Our approach relies on a novel method for estimating the error when projecting one distribution onto a convex combination of others, where the projection is with respect to a statistical distance known as the separation distance. Under sufficient conditions on the amount of noise and purity of the base distributions, this projection procedure successfully recovers the underlying class distributions. Connections to novelty detection, topic modeling, and other learning problems are also discussed.

## 1 Introduction

This paper considers a general framework for multiclass classification with label noise. As we develop later, this framework encompasses or relates to several other machine learning problems, including novelty detection, crowdsourcing, topic modeling, and learning from partial labels. Each of these applications involves *mutual contamination models*, meaning that observed data are drawn from mixtures of the underlying probability distributions of interest, and therefore may be viewed as being contaminated.

We begin by stating the problem of multiclass classification with label noise in a general setting. There are $L$ classes, each governed by a class-conditional distribution $P_i$. The learner observes training random samples drawn from the contaminated distributions

$$\tilde{P}_i = \sum_{j=1}^{L} \pi_{ij} P_j, \qquad (1)$$

$i = 1, \ldots, L$, where $\pi_{ij} \geq 0$ and $\sum_j \pi_{ij} = 1$ for each $i$. Here $\pi_{ij}$ is the probability that an instance with observed label $i$ is actually a realization of $P_j$. We consider a quite general setting where the $\pi_{ij}$ are *unknown*, and the distributions $P_i$ are not amenable to parametric modeling. We want to stress that the issue of contamination of training data comes *in addition to*, and is *different from*, the usual source of uncertainty about the labels in "noisy classification", which traditionally simply means that the supports of the true $P_i$ can overlap (which we also allow here). In the contamination setting, training a conventional classifier directly on the contaminated training data will in general lead to a biased classifier at test time (Scott et al., 2013), and in particular be asymptotically inconsistent.

Our contribution is to establish general sufficient conditions on the true class-conditional distributions and label noise mixing proportions for recovery of the true distributions, which then enables the design of a consistent discrimination rule. The sufficient conditions essentially state that the examples drawn from $\tilde{P}_i$ come "mostly" from $P_i$, and that the distributions $\{P_i\}$ are "pure" with respect to each other. Both of these notions are made precise below.

At the heart of our approach is a novel technique for "decontamination" of the contaminated models. In particular, we show that under the sufficient conditions, if we project the contaminated distribution $\tilde{P}_i$

Appearing in Proceedings of the $17^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

onto the convex hull of the other contaminated distributions, the "residual" distribution is the true class-conditional distribution $P_i$. The projection is with respect to a statistical distance known as the separation distance, and a key part of our contribution is the development of a universally consistent estimator of the error of this projection. This estimator is valid regardless of whether the sufficient conditions for label noise recovery hold, and is therefore of independent interest.

## 1.1 Motivation

The label noise model (1) is relevant in a number of applications. For example, Scott et al. (2013) describe a problem in nuclear particle classification. Each class corresponds to a different type of nuclear particle, and label noise comes from the fact that it is impossible to eliminate different particle types from the background, so that every training sample for a given particle type is always contaminated with particles of other types.

Another potential application of the label noise model is crowdsourcing, where the labels of training data sets are provided by a mixture of expert and non-expert sources. One approach to crowdsourcing is to assume each annotator makes mistakes according to model (1), where the contamination probabilities are specific to each annotator. Then there is a label noise problem for each annotator. The "two-coin" model of Raykar et al. (2010) studies this setup in the two-class setting.

As a second formulation of crowdsourcing, consider unlabeled data representing some mixture of the true class-conditional distributions $P_i$, and suppose the data is labeled by crowdsourcing. One simple labelling model is as follows: with probability $1 - \alpha$, an example is labeled by an expert, in which case the correct label is always assigned. With probability $\alpha$, a random guesser assigns the label according to some predetermined distribution over class that is independent of the actual example. This fits the contamination model (1), and furthermore, as we argue in Sec. 5, satisfies our sufficient conditions on the noise.

A third relevant application is novelty detection. Suppose that for $i < L$ and $j \neq i$, we have $\pi_{ij} = 0$. In other words, the first $L - 1$ data samples are uncontaminated. The last data sample, on the other hand, is drawn from a mixture of all classes. We may think of this as a semi-supervised learning problem, where the first $L - 1$ random samples represent training data, while the last sample is an unlabeled testing sample. Then $P_L$ corresponds to a novel class that is not represented among the $L - 1$ training classes. We argue in Sec. 5 that this setting also satisfies our sufficient conditions on the noise, thus yielding a consistent discrimination rule for multiclass novelty detection.

Several previous works on classification with label noise also adopt (1) or an equivalent model, and we refer the reader to these works for additional applications of the label noise setting (Lawrence and Schölkopf, 2001; Bouveyron and Girard, 2009; Long and Servido, 2010; Manwani and Sastry, 2011; Stempfel and Ralaivola, 2009; Natarajan et al., 2013). Significantly, our works provides estimates of the proportions $\pi_{ij}$, which are assumed known in several of these earlier works.

Topic modeling is another problem that is closely related to multiclass label noise. In topic modeling, the base distributions $P_i$ correspond to topics, and each $\tilde{P}_i$ corresponds to a document, viewed as a mixture of topics. An important difference between topic modeling and classification with label noise is that in topic modeling, typically far more documents are observed than topics. Nonetheless, as we discuss in Sec. 5, our sufficient conditions on the purity of the base distributions, when specialized to the discrete setting of topic modeling, coincide with the "separability" condition that has been widely adopted in that area.

Finally, we note that multiclass label noise has strong similarities to the problem of learning from partial labels (Cour et al., 2011). In this problem, every training data point is labeled by a subset $S \subset \{1, \dots, L\}$ of possible labels, as opposed to a single label as in standard classification. The true label is one of the elements of the subset, but it is not known which one. Therefore, collecting all training examples sharing a common $S$ gives a contamination model $\tilde{P}_S = \sum_{j \in S} \pi_{S,j} P_j$, and there is a contamination model (with associated data) for every observed $S$. We conjecture that this problem can be converted to a multiclass label noise problem, satisfying our sufficient conditions on the noise, through appropriate resampling of the data. Some insight is given in Sec. 5, although a full development is deferred to future work.

## 1.2 Related Work

Our work extends the recent work of Scott et al. (2013), reviewed in the next section, which studies label noise for binary classification ($L = 2$). We find there to be some significant differences between the multiclass and binary cases. Indeed the decontamination procedure is considerably more complex, as are the sufficient conditions for recovery. Multiclass label noise has received little attention in the literature. Existing theoretical work on label noise (of which we are aware) focuses on the binary case; we refer the reader to Scott et al. (2013) for a recent review.

The aforementioned projection in distribution space is accomplished by means of a problem we call multi-sample mixture proportion estimation, which is apparently new. It generalizes a two-sample version developed by Blanchard et al. (2010). It is also similar to the problem of semi-supervised class proportion estimation; the latter problem essentially assumes that the projection error is zero, and is only concerned with estimating the mixing weights giving the projection (Hall, 1981; Titterington, 1983; Latinne et al., 2001; Du Plessis and Sugiyama, 2012). When the projection error is nonzero, as in our work, such methods are inconsistent.

### 1.3 Outline

In the next section we review the work of Scott et al. (2013) on label noise in the binary setting. Section 3 presents the aforementioned mixture proportion estimation problem and our universally consistent estimator. Section 4 introduces sufficient conditions under which this estimator successfully decontaminates the noisy distributions in the multiclass label noise setting. The final section connects our results more concretely to the other learning problems mentioned above. Proofs are contained in the supplemental file.

## 2 Label Noise in the Binary Case

Scott et al. (2013) study label noise in the binary case, $L = 2$. Their work hinges on the following result, which appears originally in Blanchard et al. (2010).

**Proposition 1.** *Given probability distributions $F_0, F_1$ on a measurable space $(\mathcal{X}, \mathcal{C})$, define*

$$\kappa^*(F_0|F_1) = \max\left\{\kappa \in [0,1] \middle| \exists \text{ a distribution } G \text{ s.t. } F_0 = (1-\kappa)G + \kappa F_1\right\}; \quad (2)$$

*If $F_0 \neq F_1$, then $\kappa^*(F_0|F_1) < 1$ and the above supremum is attained for a unique distribution $G$ (which we refer to as the* residue *of $F_0$ w.r.t. $F_1$). Furthermore, the following equivalent characterization holds:*

$$\kappa^*(F_0|F_1) = \inf_{C \in \mathcal{C}, F_1(C)>0} \frac{F_0(C)}{F_1(C)}. \quad (3)$$

The number $\kappa^*(F_0|F_1)$ can be understood at the maximum possible proportion of $F_1$ present in $F_0$. The result implies that the function $1 - \kappa^*(F_0|F_1)$ is a statistical distance, i.e., a functional that is nonnegative and equal to zero iff $F_0 = F_1$. This quantity has been called the *separation distance*, and has arisen previously in studies of Markov chain convergence (Aldous and Diaconis, 1987).

For the label noise/contamination model (1), the following two conditions are shown by Scott et al. (2013) to be sufficient for decontamination:

- $\pi_{12} + \pi_{21} < 1\,,$
- $\kappa^*(P_1|P_2) = \kappa^*(P_2|P_1) = 0.$

The former condition bounds the total amount of label noise, while the latter condition is referred to as *mutual irreducibility*, and says that it is not possible to write $P_1$ as a nontrivial mixture of $P_2$ and some other distribution, and *vice versa*.

In particular, by a simple algebraic manipulation of (1), it is shown that under the first condition

$$\tilde{P}_1 = (1-\kappa_1)P_1 + \kappa_1 \tilde{P}_2 \quad (4)$$

$$\tilde{P}_2 = (1-\kappa_2)P_2 + \kappa_2 \tilde{P}_1, \quad (5)$$

for unique $\kappa_1, \kappa_2 < 1$. It is then shown that if $P_1$ and $P_2$ are mutually irreducible, then $\kappa_1 = \kappa^*(\tilde{P}_1|\tilde{P}_2)$ and $\kappa_2 = \kappa^*(\tilde{P}_2|\tilde{P}_1)$ and therefore $P_1$ and $P_2$ are the respective residues of $\tilde{P}_1$ w.r.t. $\tilde{P}_2$ and $\tilde{P}_2$ w.r.t. $\tilde{P}_1$. This establishes decontamination in the population case.

In the sample case, the problem of estimating $\kappa^*$ is referred to as *mixture proportion estimation*. The universally consistent estimator $\widehat{\kappa}$ of Blanchard et al. (2010) for $\kappa^*$ is applied to estimate $\kappa_1$ and $\kappa_2$. Now all terms in (4) and (5) can be estimated except for $P_1$ in (4) and $P_2$ in (5). Therefore (4) and (5) can be solved for (estimates of) $P_1$ and $P_2$, which are then available for the design of a consistent discrimination rule.

Multiclass label noise can be treated analogously, although the generalization is far from trivial.

## 3 Multi-Sample Mixture Proportion Estimation

Let $(\mathcal{X}, \mathcal{C})$ be a measurable space, and let $F_i$ be probability distributions on this space, $i = 0, 1, \ldots, M$. Let $S_M$ denote the $(M-1)$-dimensional simplex $\left\{\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M) \in \mathbb{R}^M \middle| \forall i\ \mu_i \geq 0 \text{ and } \sum_i \mu_i = 1\right\}$. For $\boldsymbol{\mu} \in S_M$, denote the probability distribution $F_{\boldsymbol{\mu}} := \sum_{i=1}^M \mu_i F_i$. We first define the maximum collective contaminating proportion of distributions $(F_i)_{1 \leq i \leq M}$ in $F_0$ by the following generalization of the binary case:

**Definition 1.** *Given probability distributions $F_i, i = 0, \ldots, M$, define*

$$\kappa^*(F_0|F_1, \ldots, F_M) = \max_{\boldsymbol{\mu} \in S_M} \kappa^*(F_0|F_{\boldsymbol{\mu}}). \quad (6)$$

If there is a unique $F_{\boldsymbol{\mu}}$ achieving the maximum in this definition, it may be thought of as the projection of $F_0$ onto the convex hull of $F_1, \ldots, F_M$ with respect to the separation distance. However, there exist simple examples for which $F_{\boldsymbol{\mu}}$ attaining $\kappa^*$ is not unique; for example, suppose $M = 2$. Let $F_0$ be uniform on $\{0, 1, 2\}$,

$F_1$ uniform on $\{0,1\}$, and $F_2$ uniform on $\{1,2\}$. Then $\kappa^* = \frac{2}{3}$ and any $\boldsymbol{\mu}$ is optimal. Later, when we return to the label noise problem, we give sufficient conditions, analogous to those in the binary case, for the unicity of $\boldsymbol{\mu}$, so that the projection is well defined.

Observe that the following equivalent characterizations for $\kappa^*$ hold:

$$\begin{aligned}
&\kappa^*(F_0|F_1,\ldots,F_M)\\
&\quad= \max_{\boldsymbol{\mu}\in S_M} \inf_{C\in\mathcal{C}:F_{\boldsymbol{\mu}}(C)>0} \frac{F_0(C)}{F_{\boldsymbol{\mu}}(C)} \qquad (7)\\
&\quad= \max\Big\{\kappa\in[0,1]\Big|\exists\boldsymbol{\mu}\in S_M \text{ and a distribution } G\\
&\qquad\quad \text{s.t. } F_0=(1-\kappa)G+\kappa F_{\boldsymbol{\mu}}\Big\} \qquad (8)\\
&\quad= \max\Big\{\sum_{i=1}^M \nu_i \,\Big|\, \nu_i\geq 0, \sum_{i=1}^M \nu_i\leq 1, \text{ and}\\
&\qquad\quad \exists \text{ a distribution } G \text{ s.t.}\\
&\qquad\quad F_0=\left(1-\sum_{i=1}^M\nu_i\right)G+\sum_{i=1}^M\nu_iF_i\Big\}. \qquad (9)
\end{aligned}$$

The equivalence of (6), (7) and (8) is a straightforward consequence of the equivalent definitions (2) and (3) in the 2-class case. The equivalence of (8) and (9) is also clear from the representation $\kappa=\sum_i\nu_i$ and $\boldsymbol{\mu}=(\nu_1,\ldots,\nu_M)/\kappa$. The fact that the maxima are well-defined is justified formally in the supplemental material. Finally, if the maximum is attained for a unique $G$ in (8)-(9), we refer to $G$ as the *residue* of $F_0$ w.r.t. $F_1,\ldots,F_M$.

Suppose that for $m=0,1,\ldots,M$, we observe

$$X_1^m,\ldots,X_{n_m}^m \overset{iid}{\sim} F_m$$

The random samples need not be independent of each other. Let $\widehat{F}_i$ denote the corresponding empirical distributions. By way of notation, set $\boldsymbol{n} := (n_0,n_1,\ldots,n_M)$ and write $\boldsymbol{n}\to\infty$ to indicate $\min_i\{n_i\}\to\infty$.

To define the estimator, let $\mathcal{C}_k\subset\mathcal{C}$, $k\geq 1$ be VC classes with VC dimensions $V_k<\infty$. For each $0\leq m\leq M, k\geq 1$ and $\delta>0$, define

$$\epsilon_m^k(\delta) := 3\sqrt{\frac{V_k\log(n+1)-\log\delta/2}{n_m}}.$$

By the VC inequality (Devroye et al., 1996), the following property holds:

**P1** For each value of $k$, $m$, and $\delta>0$, the following holds with probability at least $1-\delta$:

$$\sup_{C\in\mathcal{C}_k}\left|F_m(C)-\widehat{F}_m(C)\right|\leq\epsilon_m^k(\delta),$$

where the probability is with respect to the draw of $X_1^m,\ldots,X_{n_m}^m \overset{iid}{\sim} F_m$.

We also require that $(\mathcal{C}_k)_{k\geq 1}$ possess the following universal approximation property:

**P2** For any probability distribution $Q$, any $C^*\in\mathcal{C}$:

$$\liminf_{k\to\infty}\inf_{C\in\mathcal{C}_k} Q(C\Delta C^*)=0,$$

where $C_1\Delta C_2 := (C_1\backslash C_2)\cup(C_2\backslash C_1)$ is the symmetric difference.

Examples of such classes include histograms, decision trees, neural networks, and generalized linear classifiers (Devroye et al., 1996).

Given $\boldsymbol{\delta} := (\delta_0,\ldots,\delta_M)$, a vector with positive components, we define the estimator

$$\begin{aligned}
&\widehat{\kappa}(\widehat{F}_0|\widehat{F}_1,\ldots,\widehat{F}_M;\boldsymbol{\delta})\\
&\quad=\max_{\boldsymbol{\mu}\in S_M}\inf_k\inf_{C\in\mathcal{C}_k}\frac{\widehat{F}_0(C)+\epsilon_0^k(c\delta_0k^{-2})}{\left(\widehat{F}_{\boldsymbol{\mu}}(C)-\sum_i\mu_i\epsilon_i^k(c\delta_ik^{-2})\right)_+}, \qquad (10)
\end{aligned}$$

where $c := 6/\pi^2$, and the ratio is defined as $+\infty$ if the denominator is zero. Also, denote $\widehat{\boldsymbol{\mu}}$ an arbitrary point where the above maximum is attained; and finally $\widehat{\boldsymbol{\nu}}=\widehat{\kappa}\widehat{\boldsymbol{\mu}}$ (where the explicit dependence on $\widehat{F}_0$, etc., has been omitted to simplify notation). $\widehat{\boldsymbol{\nu}}$ is an estimate of the contaminating proportions $(\nu_1,\ldots,\nu_M)$ achieving the maximum in (9).

**Proposition 2.** *Let $(\mathcal{C}_k)_{k\geq 1}$ be a sequence of classes of sets with finite VC dimension and having the universal approximation property* **P2**.

- *(a) It holds that $\widehat{\kappa}\geq\kappa^*$ with probability at least $1-\sum_{i=0}^M\delta_i$.*
- *(b) If for all $i$, $\delta_i(\boldsymbol{n})=1/n_i$, then $\widehat{\kappa}$ converges in probability to $\kappa^*$ as $\boldsymbol{n}\to\infty$, for any family of generating distributions $(F_i, i=0,\ldots,M)$.*
- *(c) Let $\mathcal{B}_* :=$ Arg $\text{Max}_{\boldsymbol{\mu}\in S_M}\kappa^*(F_0|F_{\boldsymbol{\mu}})$, i.e., $\mathcal{B}_*$ is the set of all mixture weights $\boldsymbol{\mu}$ attaining the max. in (6). Then under the assumptions of point (b), $d(\widehat{\boldsymbol{\mu}},\mathcal{B}_*)$ converges to zero in probability as $\boldsymbol{n}\to\infty$ (where $d$ is any continuous distance function on the simplex). A similar result holds for the convergence in probability of $\widehat{\boldsymbol{\nu}}$ to the set of weight vectors attaining $\kappa^*$ in (9).*

The definition of the estimator (10) is an intuitive generalization of the consistent estimator $\widehat{\kappa}(\widehat{F}_0|\widehat{F}_1)$ of

Blanchard et al. (2010), with an added maximum operation over $\boldsymbol{\mu} \in S_M$. The proof of the above result, however, does not follow trivially from the pointwise consistency of $\widehat{\kappa}(\widehat{F}_0|\widehat{F}_{\boldsymbol{\mu}})$ to $\kappa^*(F_0|F_{\boldsymbol{\mu}})$ for all fixed $\boldsymbol{\mu} \in S_M$, and requires a careful compactness argument. Furthermore, point (c) entails that the weights defining the residue are estimated consistently if they are unique (i.e., when $\mathcal{B}_*$ contains a single element), which they are in the label noise setting under the sufficient conditions given below.

**Practical feasibility.** The focus of this work is on identifiability and existence of a consistent estimator. While a suitable practical implementation is left for future work, we observe that there exist a priori reasonable strategies to compute $\widehat{\kappa}$. Because the $\epsilon^k(\ldots)$ terms become eventually larger than 1 for large $k$, the infimum over $k$ can be limited to $k \leq k_{max}(n)$. The inner infimum loop can then be computed exactly if the classes $\mathcal{C}_k$ are finite (for instance, this is the case for the pieces of dyadic regular partitions of order $k$). The outer maximum loop can in turn be solved by gradient ascent (in alternating steps with the infimum solving step), which will converge to the global maximum since the inverse of the objective function is a linear function of $\boldsymbol{\mu}$.

Henceforth $\boldsymbol{\delta}$ will be omitted from the notation for $\widehat{\kappa}$, and $\delta_i$ taken to equal $1/n_i$.

# 4 Multiclass Label Noise

The model in (1) may be concisely expressed as $\tilde{\boldsymbol{P}} = \Pi \boldsymbol{P}$, where

$$\boldsymbol{P} = \begin{bmatrix} P_1 \\ \vdots \\ P_L \end{bmatrix}, \quad \tilde{\boldsymbol{P}} = \begin{bmatrix} \tilde{P}_1 \\ \vdots \\ \tilde{P}_L \end{bmatrix},$$

and $\Pi = [\pi_{ij}]$ is an $L \times L$ matrix with nonnegative entries and rows summing to one. We begin with identifiability assumptions that enable decontamination.

## 4.1 Assumptions

We start with a generalization of mutual irreducibility:

**Lemma 1.** *The following conditions on the family of distributions $P_1, \ldots, P_L$ are equivalent:*

- *It is not possible to write*

$$\sum_{i \in I} \epsilon_i P_i = \alpha \left( \sum_{i \notin I} \epsilon_i P_i \right) + (1-\alpha) H,$$

*where $I \subset \{1, \ldots, L\}$ such that $1 \leq |I| < L$, $\epsilon_i$ are such that $\epsilon_i \geq 0$ and $\sum_{i \in I} \epsilon_i = \sum_{i \notin I} \epsilon_i = 1$, $\alpha \in (0,1]$, and $H$ is a distribution.*

- *If $\sum_{i=1}^L \gamma_i P_i$ is a distribution, then $\gamma_i \geq 0 \ \forall i$.*

**Definition 2.** *We say the distributions $\{P_i\}_{1 \leq i \leq L}$ are* jointly irreducible *iff the conditions in Lemma 1 hold.*

Joint irreducibility says that every convex combination of some portion of the $P_i$s is irreducible w.r.t. every convex combination of the remaining $P_i$s. It generalizes the notion of mutual irreducibility from the two-class case (see Sec. 2). One case where it holds is when the support of each $P_i$ contains some region with positive probability that does not intersect the supports of the other $P_j, j \neq i$. If each $P_i$ is a class-conditional distribution, this means that every class has some exemplars that could not possibly arise from another class. This assumption is not unreasonable in many, and perhaps most, applications of interest. Joint irreducibility can still hold even when all $P_i$ have the same support, as in the case of Gaussian densities with a common variance (Scott et al., 2013).

We will also make assumptions on the contamination weight matrix $\Pi$. Let $\boldsymbol{\pi}_i$ be the transpose of the $i$-th row of $\Pi$, which is a discrete probability distribution on $\{1, \ldots, L\}$. Let $\boldsymbol{e}_i$ denote the length $L$ vector with 1 in the $i$th position and zeros elsewhere.

**Lemma 2.** *The following conditions on $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_L$ are equivalent:*

*(a) For each $\ell$, the residue of $\boldsymbol{\pi}_\ell$ with respect to $\{\boldsymbol{\pi}_j, j \neq \ell\}$ is $\boldsymbol{e}_\ell$.*

*(b) For every $\ell$ there exists a decomposition $\boldsymbol{\pi}_\ell = \kappa_\ell \boldsymbol{e}_\ell + (1-\kappa_\ell)\boldsymbol{\pi}'_\ell$ where $\kappa_\ell > 0$ and $\boldsymbol{\pi}'_\ell$ is a convex combination of $\boldsymbol{\pi}_j$ for $j \neq \ell$.*

*(c) $\Pi$ is invertible and $\Pi^{-1}$ is a matrix with strictly positive diagonal entries and nonpositive off-diagonal entries.*

**Definition 3.** *We say that $\Pi$ is* recoverable *iff the conditions in Lemma 2 hold.*

This assumption ensures that the amount of contamination is not too high. Some intuition is given by condition (b). Fig. 1 depicts the case $L = 3$. In panel (i), condition (b) is satisfied, while in panel (ii), condition (b) is not satisfied. Clearly, in panel (i), the diagonal entries $\pi_{\ell\ell}$ are much larger than they are in panel (ii), and consequently the off-diagonal entries of $\Pi$ will be smaller. Note that for $L = 2$ classes, from condition (c), recoverability is equivalent to $\pi_{12} + \pi_{21} < 1$, the condition developed by Scott et al. (2013).

We exhibit a setting where $\Pi$ is guaranteed to be recoverable. Assume $\boldsymbol{c} := \sum_{i=1}^M c_i \boldsymbol{e}_i$ is a contaminating "background noise" which is common to all observed classes, albeit in possibly different proportions, i.e.,

$\boldsymbol{\pi}_i = \kappa_i \boldsymbol{c} + (1-\kappa_i)\boldsymbol{e}_i$, with $\kappa_i \in [0,1)$. Geometrically, this means we shift by various amounts the vertices $\boldsymbol{e}_i$ of the simplex towards a common point $\boldsymbol{c}$, or equivalently that each $\boldsymbol{\pi}_\ell$ belongs to the segment $(\boldsymbol{c}\boldsymbol{e}_\ell]$. Panel (iii) illustrates this situation.

Then we have

$$\sum_i \frac{c_i}{1-\kappa_i}\boldsymbol{\pi}_i = \left(1 + \sum_i \frac{c_i\kappa_i}{1-\kappa_i}\right)\boldsymbol{c},$$

so that for instance for $\ell = 1$ and $\kappa_1 > 0$,

$$\begin{aligned}\frac{c_1}{1-\kappa_1}\boldsymbol{\pi}_1 &+ \sum_{i\geq 2}\frac{c_i}{1-\kappa_i}\boldsymbol{\pi}_i \\ &= \left(1+\sum_i \frac{c_i\kappa_i}{1-\kappa_i}\right)\left(\frac{1}{\kappa_1}\boldsymbol{\pi}_1 - \frac{1-\kappa_1}{\kappa_1}\boldsymbol{e}_1\right),\end{aligned}$$

and finally

$$\begin{aligned}\left(1+\sum_i \frac{c_i\kappa_i}{1-\kappa_i}\right)(1-\kappa_1)\boldsymbol{e}_1 + \kappa_1\sum_{i\geq 2}\frac{c_i}{1-\kappa_i}\boldsymbol{\pi}_i \\ = \boldsymbol{\pi}_1\left(1+\sum_{i\geq 2}\frac{c_i\kappa_i}{1-\kappa_i}\right),\end{aligned}$$

which is also valid when $\kappa_1 = 0$. Since all of the above coefficients are positive, this implies condition (b) (or (c) ) of Lemma 2 after normalization.

In Sec. 5 we consider various applications where the recoverability assumption is satisfied.

### 4.2 Decontamination

The following result shows that the recoverability and joint irreducibility conditions ensure decontamination in the population (infinite sample) case. This result is applied in the next subsection, in conjunction with the estimator of Section 3, to establish a consistent discrimination rule.

**Proposition 3.** *If* $\Pi$ *is recoverable and* $P_1, \ldots, P_L$ *are jointly irreducible, then for each* $\ell$, $P_\ell$ *is the residue of* $\tilde{P}_\ell$ *w.r.t.* $\{\tilde{P}_j, j \neq \ell\}$. *Furthermore, in the representation*

$$\tilde{P}_\ell = (1-\kappa_\ell)P_\ell + \sum_{j\neq\ell}\nu_{\ell j}\tilde{P}_j, \tag{11}$$

*where* $\kappa_\ell = \kappa^*(\tilde{P}_\ell \,|\, \{\tilde{P}_j, j\neq\ell\})$, $\kappa_\ell < 1$ *and the* $\nu_{\ell j}$ *are unique.*

The proof of this result shows that under the joint irreducibility assumption, decontamination of the $\tilde{P}_\ell$ is equivalent to decontamination of the discrete distributions $\boldsymbol{\pi}_\ell$, which may be viewed as contaminated versions of the $\boldsymbol{e}_i$. In other words, the same weights $\kappa_\ell$ and $\nu_{\ell j}$ uniquely give the solutions of both approximation problems. The desired solution of the discrete problem is guaranteed by the recoverability assumption on $\Pi$ (so that $\boldsymbol{e}_\ell$ is the residue of $\boldsymbol{\pi}_\ell$ w.r.t. $\{\boldsymbol{\pi}_j\}_{j\neq\ell}$), and this ensures, by the equivalence of the decontamination problems, that $P_\ell$ is the residue of $\tilde{P}_\ell$ w.r.t. $\{\tilde{P}_j, j\neq\ell\}$.

This equivalence leads to a second insight: By a construction in the proof of Lemma 2, $\Pi^{-1}$ can be expressed explicitly in terms of the optimal weights $\kappa_\ell$ and $\nu_{\ell j}$. Therefore, under the assumptions of Proposition 3, $\Pi^{-1}$ can be consistently estimated (i.e., recovered) via the estimator in Section 3.

### 4.3 A Consistent Discrimination Rule

The decontamination result gives a way to accurately estimate the true class-conditional error probabilities, which can then be converted into a consistent discrimination rule. These details now follow. For any classifier, that is, any measurable function $f : \mathcal{X} \to \{1, \ldots, L\}$, denote $R_i(f) := P_i(f(X) \neq i)$, where $X$ follows $P_i$. As a performance measure we adopt the *minmax* criterion, which seeks to minimize

$$R(f) := \max_{1\leq i\leq L} R_i(f).$$

The optimal performance is the *minmax error*, $R^* := \inf_f R(f)$, where the inf is over all classifiers. Note that other performance measures could also be analyzed; we focus on the minmax criterion for concreteness.

The crux of classifier design is accurate error estimation. Denote $\tilde{R}_{j\ell}(f) := \tilde{P}_j(f(X)\neq\ell)$. To motivate an estimator for $R_\ell(f)$, the expression in (11) implies

$$R_\ell(f) = \frac{\tilde{R}_{\ell\ell}(f) - \sum_{j\neq\ell}\nu_{\ell j}\tilde{R}_{j\ell}(f)}{1-\kappa_\ell}.$$

If $X_1^j, \ldots, X_{n_j}^j \overset{iid}{\sim} \tilde{P}_j$ are the observed data from class $j$, then

$$\widehat{\tilde{R}}_{j\ell}(f) := \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{1}_{\{f(X_i^j)\neq\ell\}}$$

estimates $\tilde{R}_{j\ell}(f)$. Now let $(\widehat{\nu}_{\ell j})_{j\neq\ell}$ be any vector achieving the maximum in the definition of $\widehat{\kappa}_\ell := \widehat{\kappa}(\widehat{\tilde{P}}_\ell | \{\widehat{\tilde{P}}_j, j\neq\ell\})$, so that $\widehat{\kappa}_\ell = \sum_{j\neq\ell}\widehat{\nu}_{\ell j}$. By Proposition 2, this vector converges to the unique weights $\nu_{\ell j}$ in Proposition 3, motivating the following estimator:

$$\widehat{R}_\ell(f) := \frac{\widehat{\tilde{R}}_{\ell\ell}(f) - \sum_{j\neq\ell}\widehat{\nu}_{\ell j}\widehat{\tilde{R}}_{j\ell}(f)}{1-\widehat{\kappa}_\ell}.$$

It is well known in statistical learning theory that consistency of a learning algorithm follows from uniform

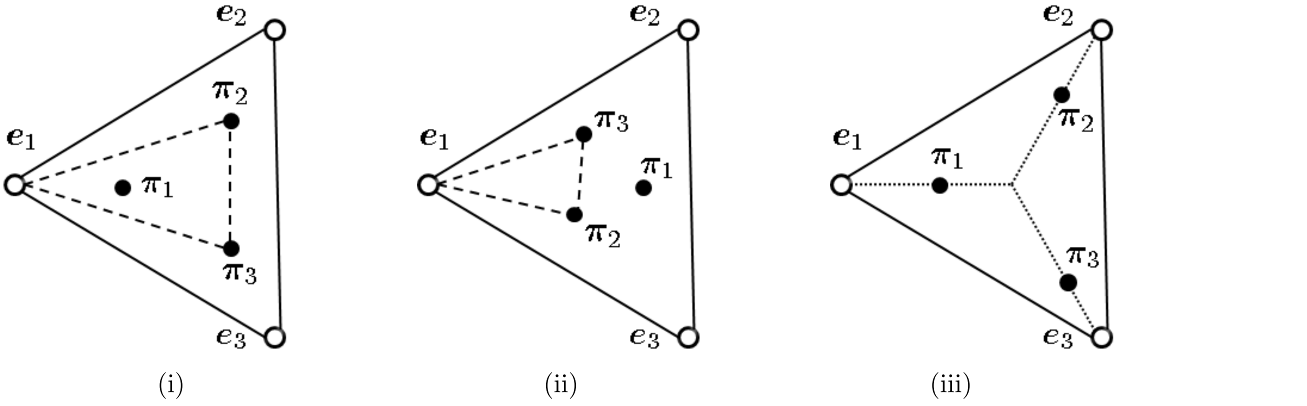

Figure 1: Illustration of the recoverability condition when $L = 3$. Panel (i): Low noise, $\Pi$ recoverable. Each $\boldsymbol{\pi}_\ell$ can be written as a convex combination of $\boldsymbol{e}_\ell$ and the other two $\boldsymbol{\pi}_j$ (with a positive weight on $\boldsymbol{e}_\ell$), depicted here for $\ell = 1$. Panel (ii): High noise, $\Pi$ not recoverable. Panel (iii): The setting of "common background noise."

control of error estimates over a class of classifiers $\mathcal{F}$, whose complexity may grow with sample size. Thus we introduce a sequence $(\mathcal{F}_k)_{k\geq 1}$ of sets of classifiers with VC dimension $V_k < \infty$. Since we are in a multiclass setting, we adopt the following generalization of VC dimension to multiclass, namely, define the VC dimension of class $\mathcal{F}$ to be the maximum (conventional) VC dimension of the family of sets $\{x : f(x) \neq \ell\}_{f\in\mathcal{F}}$, over $\ell = 1, \ldots, L$. This particular multiclass generalization of VC dimension allows our analysis to proceed using only the standard VC inequality for binary classifiers (or equivalently, for sets, as in Section 3). Indeed, the VC inequality (Devroye et al., 1996) gives uniform convergence of $\widehat{\tilde{R}}_{j\ell}(f)$ to $\tilde{R}_{j\ell}(f)$ over $f$ in $\mathcal{F}_k$, and together with consistency of $\widehat{\kappa}_\ell$ and $(\widehat{\nu}_{\ell j})$ implies the following result. Similar to Section 3, set $\boldsymbol{n} := (n_1, \ldots, n_L)$ and write $\boldsymbol{n} \to \infty$ to indicate $\min\{n_\ell\} \to \infty$.

**Proposition 4.** *Let $k(\boldsymbol{n})$ take positive integer values and be such that as $\boldsymbol{n} \to \infty$,*

$$\frac{V_{k(\boldsymbol{n})} \log n_\ell}{n_\ell} \to 0,$$

$1 \leq \ell \leq L$. *Then under the assumptions of Proposition 3,* $\sup_{f\in\mathcal{F}_{k(\boldsymbol{n})}} |R_\ell(f) - \widehat{R}_\ell(f)| \overset{i.p.}{\to} 0$ *as* $\boldsymbol{n} \to \infty$.

This result allows us to analyze the estimation error of a learning algorithm based on $\widehat{R}_\ell$. To control the approximation error, we choose $(\mathcal{F}_k)$ such that

**P3** For any $P_1, \ldots, P_L$, $\lim_{k\to\infty} \inf_{f\in\mathcal{F}_k} R(f) = R^*$.

This condition is analogous to the condition for our family of sets $(\mathcal{C}_k)$ in Section 3.

Let us now define a discrimination rule based on the above error estimates. Define $\widehat{R}(f) := \max_\ell \widehat{R}_\ell(f)$. Let $\tau_k$ be any sequence of positive numbers tending to zero. Let $\widehat{f}_k$ denote any classifier

$$\widehat{f}_k \in \Big\{ f \in \mathcal{F}_k \, : \, \widehat{R}(f) \leq \inf_{f\in\mathcal{F}_k} \widehat{R}(f) + \tau_k \Big\}.$$

The introduction of $\tau_k$ lets us avoid assuming the existence of an empirical risk minimizer. Finally, define the discrimination rule $\widehat{f} := \widehat{f}_{k(\boldsymbol{n})}$.

**Theorem 1.** *Let $k(\boldsymbol{n})$ take positive integer values and be such that as $\boldsymbol{n} \to \infty$, $k(\boldsymbol{n}) \to \infty$ and*

$$\frac{V_{k(\boldsymbol{n})} \log n_\ell}{n_\ell} \to 0, \tag{12}$$

$1 \leq \ell \leq L$. *Then under* **P3** *and the assumptions of Proposition 3,* $R(\widehat{f}) \overset{i.p.}{\to} R^*$ *as* $\boldsymbol{n} \to \infty$.

Thus, if the noise is recoverable and the distributions jointly irreducible, consistent classification is possible in the multiclass label noise setting.

## 5 Discussion

We now develop connections between the above theoretical framework for multiclass label noise and various applications mentioned in the introduction.

Consider the second crowdsourcing formulation, and suppose that unlabeled data are drawn according to $\sum_{j=1}^{L} \theta_j P_j$. Further suppose that with probability $1 - \alpha$, the label is assigned correctly, while with probability $\alpha$, the label is assigned according to some fixed distribution, say the uniform distribution $[\frac{1}{L}, \ldots, \frac{1}{L}]^T$. Let $Y$ denote the true label of an example, and $\tilde{Y}$ the crowdsourced label. By Bayes' rule,

$$\begin{aligned} \pi_{ij} &= Pr(Y = j \,|\, \tilde{Y} = i) \\ &\propto Pr(\tilde{Y} = i \,|\, Y = j) Pr(Y = j) \\ &= ((1-\alpha)\mathbf{1}_{\{i=j\}} + \frac{\alpha}{L})\theta_j. \end{aligned}$$

In vector form, this means $\boldsymbol{\pi}_i \propto (1-\alpha)\theta_i \boldsymbol{e}_i + \frac{\alpha}{L}\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_L]^T$. This clearly satisfies the common background noise model described previously, and therefore our noise condition is satisfied.

Next, consider the semi-supervised novelty detection problem described in the introduction. In this case

$$\Pi = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ \pi_{L1} & \pi_{L2} & \cdots & \pi_{LL} \end{bmatrix}$$

because the first $L-1$ data samples are uncontaminated, while the last sample is an unlabeled testing sample where $P_L$ governs the novelty class. It is easy to see that $\Pi$ is recoverable if and only if $\pi_{LL} > 0$, using condition (c) of Lemma 2. Therefore, as long as there are at least some novel exemplars in the unlabeled data, the noise condition is satisfied, and under joint irreducibility we have a consistent discrimination rule for multiclass, semi-supervised novelty detection. We also note that the above is not the only approach to novelty detection using mixture proportion estimation (Sanderson and Scott, 2014).

Our work has an interesting connection to the problem of topic modeling, itself linked to nonnegative matrix factorization. In that problem, one also observes random samples from several contaminated distributions ("documents") $\tilde{P}_i$, but in the equation $\tilde{\boldsymbol{P}} = \Pi \boldsymbol{P}$, the matrix $\Pi$ is no longer square but now has far more rows than columns. The distributions $P_i$ are "topics," and the proportion $\pi_{ij}$ reflects the prevalence of topic $j$ in document $i$. Existing work on topic modeling typically represents the topics and documents as discrete distributions on a finite vocabulary. One interesting connection to our work is that in this discrete setting, our joint irreducibility condition is equivalent to a condition that has been previously shown to be sufficient for identifiability of the topics. In particular, in the discrete setting, this assumption states that for each topic, there exists at least one word occuring with a positive probability in the given topic, and with a probability of zero in the other topics (Donoho and Stodden, 2004; Arora et al., 2012).

Finally, we return to the problem of learning from partial labels (Cour et al., 2011). For the sake of argument we consider a simple form of the problem. Suppose there are $L = 3$ classes, and that each observed data point is labeled $A = \{1,2\}, B = \{1,3\}$, or $C = \{2,3\}$. The true label of each example is one of the labels in the associated subset. Grouping together observations according to the three "partial labels" ($A$, $B$, or $C$), these three data samples are described by the contamination models $\tilde{P}_S = \sum_{j \in S} \pi_{S,j} P_j$, where $S \in \{A, B, C\}$. Each data sample arises from a convex combination of two of the three base distributions, as depicted in Fig. 2. Since these contaminated models lie on the boundary of the probability simplex, they do not satisfy the recoverability assumption on the noise. However, by resampling the data, we may obtain random samples from distributions that do satisfy the noise assumption. For simplicity, let's assume $\pi_{S,j} = 1/2$ for each $j \in S$, i.e., each contaminated model is an equal mixture of the two associated base distributions. For every pair of contaminated models, corresponding to subsets $S$ and $T$, form a new random sample by resampling from the two observed samples, such that the new random samples are realizations of $\frac{1}{2}\tilde{P}_S + \frac{1}{2}\tilde{P}_T$. If $\Pi$ and $\ddot{\Pi}$ denote the contamination proportions before and after resampling, then

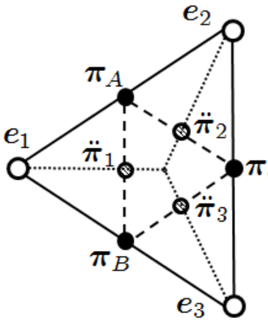

Figure 2: Learning from partial labels reduces to multiclass label noise in some cases.

$$\Pi = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad \ddot{\Pi} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix}.$$

The situation is depicted in Fig. 2. The rows $(\ddot{\boldsymbol{\pi}}_i)^T$ of $\ddot{\Pi}$ satisfy $\ddot{\boldsymbol{\pi}}_i = \frac{1}{4}\boldsymbol{e}_i + \frac{3}{4}\boldsymbol{c}$, where $\boldsymbol{c} = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]^T$, which is in the form of the common background noise model, implying our noise condition. Our framework can now be applied to the resampled data, assuming joint irreducibility, to give a consistent discrimination rule for learning from partial labels. If $\Pi$ is perturbed slightly, our noise assumption still holds, and we further conjecture that this kind of resampling strategy may be applied when $L > 3$ and when the pattern of observed subsets is more general.

## Acknowledgements

G. Blanchard was supported in part by the European Community's 7th Framework Programme under the E.U. grant agreement 247022 (MASH Project). C. Scott was supported in part by NSF Grants 0953135, 1047871, and 1217880.

## References

D. Aldous and P. Diaconis. Strong uniform times and finite random walks. *Adv. Appl. Math.*, 8(1):69–97, 1987.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *FOCS*. 2012.

G. Blanchard, G. Lee, and C. Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Journal of Pattern Recognition*, 42:2649–2658, 2009.

T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *J. Machine Learning Research*, 12:1501–1536, 2011.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, 1996.

D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

M. C. Du Plessis and M. Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In J. Langford and J. Pineau, editors, *Proc. 29th Int. Conf. on Machine Learning*, pages 823–830, 2012.

P. Hall. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 43(2):147–156, 1981.

P. Latinne, M. Saerens, and C. Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities may significantly improve classification accuracy: Evidence from a multi-class problem in remote sensing. In C. Sammut and A. H. Hoffmann, editors, *Proc. 18th Int. Conf. on Machine Learning*, pages 298–305, 2001.

N. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. *Proceedings of the International Conference in Machine Learning*, 2001.

P. Long and R. Servido. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78:287–304, 2010.

N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. Technical Report arXiv:1109.5231, 2011.

N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*, 2013.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *The Journal of Machine Learning Research*, 99:1297–1322, 2010.

T. Sanderson and C. Scott. Class proportion estimation with application to multiclass anomaly rejection. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2014.

C. Scott, G. Blanchard, and G. Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In *Proc. Conf. on Learning Theory*, JMLR W&CP, volume 30, pages 489–511. 2013.

G. Stempfel and L. Ralaivola. Learning SVMs from sloppily labeled data. In *Proc. 19th Int. Conf. on Artificial Neural Networks: Part I*, pages 884–893, 2009.

D. M. Titterington. Minimum distance nonparametric estimation of mixture proportions. *Journal of the Royal Statistical Society*, 45(1):37–46, 1983.