Estimating the Accuracies of Multiple Classifiers
Without Labeled Data

Ariel Jaffe 1, Boaz Nadler 1, Yuval Kluger 2,3,∗

1 Weizmann Institute of Science, Department of Computer Science and Applied Mathematics, Rehovot, 76100 Israel. 2 Yale University School of Medicine, Department of Pathology, 333 Cedar St., New Haven, CT 06520, USA. 3 NYU Center for Health Informatics and Bioinformatics New York University Langone Medical Center 227 East 3030th Street, New York, NY 10016, USA. ∗ E-mail: yuval.kluger@yale.edu.

Abstract

In various situations one is given the predictions of multiple classifiers over a large unlabeled test data. This scenario raises the following questions: Without any labeled data and without any a-priori knowledge about the reliability of these different classifiers, is it possible to consistently and computationally efficiently estimate their accuracy? Furthermore, also without any labeled data can one construct a more accurate unsupervised ensemble classifier? In this paper we make the following contributions: Under the standard assumption that classifiers make independent errors, we provide (partial) positive answers to these questions. In the binary case, we present two different methods to estimate the class imbalance and classifiers specificities and sensitivities. This directly gives a novel unsupervised ensemble learner. In the multi-class case, we show how to estimate the class probabilities and the diagonal entries of the classifiers confusion matrices. We illustrate our algorithms with empirical results on both artificial and real data.

1 Introduction

Consider a classification problem from an instance space $\mathcal{X}$ to an output label set $\mathcal{Y} = \{1, \ldots, K\}$. In contrast to the classical supervised setting, in various contemporary applications, one has access only to the predictions or forecasts of multiple experts or classifiers over a large number of unlabeled instances. Moreover, the reliability of these experts is unknown, and at test time there is no labeled data to assess it. A notable example of this setup is in crowdsourcing, where an annotation task over many instances is distributed to many annotators whose reliability is a-priori unknown [10, 12, 13]. Other notable application domains including decision science, economics, medicine and more, see [7, 8, 11].

Given only the $m \times n$ matrix $Z$ (or a significant part of it) with $Z_{ij} = f_i(x_j)$ holding the predictions of the given $m$ classifiers over $n$ instances, and without any labeled data, two fundamental questions arise: (i) is it possible to consistently estimate the accuracies of the $m$ classifiers in a computationally efficient way, and (ii) is it possible to construct, again by some computationally efficient procedure, an unsupervised ensemble learner, more accurate than most if not all of the original $m$ classifiers.

The first question is important in cases where obtaining the predictions of these $m$ classifiers is by itself an expensive task, and after collecting a certain number of instances and their predictions, we wish to pick only a few of the most accurate ones, see [9]. The second question, also known as offline consensus, is of utmost importance in improving the quality of automatic decision making systems based on multiple sources of information.

Beyond the simplest approach of majority voting, perhaps the first to define and address these questions were Dawid and Skene [2]. With the increasing popularity of crowdsourcing and large scale expert opinion systems, the last years have seen a surge of interest in these problems. Yet, the most common methods to address questions (i) and (ii) above are based on the expectation maximization (EM) algorithm, already proposed in this context by Dawid and Skene, and whose only guarantee is convergence to a local maxima.

Two recent exceptions, proposing spectral (and thus computationally efficient) methods with strong consistency guarantees are [5] and [7]. In [5], the authors assume a spammer-hammer model, where each classifier is either perfectly correct or totally random and present a spectral method to detect which one is which. Parisi et. al. [7] presented a spectral approach to address questions (i) and (ii) above in the context of binary classification. Their approach, however, has several limitations. First, they did not actually estimate each classifier sensitivity and specificity, but only showed how to consistently rank them according to their balanced accuracies. Second, their unsupervised learner assumes that all classifiers have balanced accuracies close to 1/2 (random). Hence, their method may be suboptimal, for example, when few classifiers are significantly more accurate than all others.
In this paper we extend and generalize the results of \cite{7} in several directions and make the following contributions. We show that under standard independent assumptions on different classifier errors, it is possible, in the binary classification setting, to consistently estimate in a computationally efficient manner both the class imbalance and the individual sensitivities and specificities of the m classifiers. Given the value of the class imbalance, we present in Sec. 3 a simple and consistent spectral method to estimate each classifier sensitivity and specificity. Hence, the problem boils down to estimating a single scalar - the class imbalance, when unknown. In Sec. 4 we present two computationally efficient and consistent methods to do so. The first is based on the 3-dimensional covariance tensor of triplets of classifiers, and the other based on the full likelihood of all the data.

In Sec. 5 we consider the multi-class setting. Building upon the results for the binary case, here we present a method to estimate the probabilities of the different classes and the diagonal entries of each classifier confusion matrix. Unfortunately, our approach does not enable us to estimate all off-diagonal entries of the m confusion matrices, and this remains an interesting challenge for future research. Finally, in Sec. 6 we present simulation results on both real and artificial data.

2 Problem Setup

We start with the following binary classification problem setup, as also studied is several previous works \cite{7,8}. The multi-class case appears in Sec. 5. Let \( X \) be an instance space with an output space \( Y = \{-1, 1\} \). A labeled instance \((x, y) \in X \times Y\) is a realization of the random variable \((X, Y)\), which has an unknown probability density \(p(x, y)\), and \(X\) and \(Y\) marginals \(p_X(x)\) and \(p_Y(y)\), respectively. We further denote by \( b \) the class imbalance of \( Y \),

\[
b = \Pr(Y = 1) - \Pr(Y = -1) = p_Y(1) - p_Y(-1).
\]

Let \( \{f_i\}_{i=1}^m \) be a set of \( m \) binary classifiers operating on \( X \). As our classification problem is binary, the accuracy of the \( i \)-th classifier is fully characterized by its sensitivity \( \psi_i \) and specificity \( \eta_i \),

\[
\psi_i = \Pr(f_i(X) = 1|Y = 1), \quad \eta_i = \Pr(f_i(X) = -1|Y = -1).
\]

For future use, we denote by \( \pi_i \) its balanced accuracy,

\[
\pi_i = (\psi_i + \eta_i)/2.
\]

In this paper we consider the following totally unsupervised scenario. We are given only an \( m \times n \) matrix \( Z_{ij} = f_j(x_i), \ i = 1 \ldots m, \ j = 1 \ldots n, \) where \( f_j(x_i) \) is the label predicted at instance \( x_i \) by classifier \( f_j \). In particular, we assume no prior knowledge about the \( m \) classifiers, so their accuracies (their sensitivities \( \psi_i \) and specificities \( \eta_i \)) are all unknown. This may occur, for example, when each classifier was trained with its own possibly proprietary labeled data set, unavailable to us. For other examples, including crowdsourcing applications, combination of experts and more, see \cite{7,8}.

Given only this \( m \times n \) matrix of binary predictions, we consider the following two problems: (i) consistently and computationally efficiently estimate the sensitivity and specificity of each classifier, and (ii) construct a more accurate ensemble classifier. As discussed below, under certain assumptions, a solution to the first problem readily yields a solution to the second one.

To tackle these problems, we make the following three assumptions: (i) The \( n \) instances \( x_i \) are i.i.d. realizations from the marginal \( p_X(x) \). (ii) The \( m \) classifiers are conditionally independent. That is, for every pair of classifiers \( f_i, f_j \) with \( i \neq j \) and for all labels \( a_i, a_j \in \{-1, 1\} \),

\[
\Pr(f_i = a_i, f_j = a_j|Y = y) = \Pr(f_i = a_i|Y = y) \Pr(f_j = a_j|Y = y).
\]

(iii) Most of the classifiers are better than random, in the sense that for more than half of all classifiers, \( \pi_i > 0.5 \). Note that (i)-(ii) are standard assumptions in both the supervised and unsupervised settings, see \cite{2,3,7,8}. Assumption (iii) or a variant thereof is needed, given an inherent \( \pm 1 \) sign ambiguity of the fully unsupervised problem.

3 Estimating the sensitivity and specificity in case of a known bias

For some classification problems, the value of the class imbalance \( b \) may be known a-priori. One example is in epidemiology, where the overall prevalence of a certain disease in the population is known, and the classification problem is to predict its presence (or future onset) in individuals, based on their observed features (such as blood results, height, weight, age, genetic profiles, etc).

Assuming \( b \) is known, we build upon the spectral approach in \cite{7}, and present a computationally efficient algorithm to consistently estimate the sensitivities and specificities of all \( m \) classifiers. To motivate our approach, it is instructive to study the limit of an infinite unlabeled set size, \( n \to \infty \). In this case, the mean values of the \( m \) classifiers, denoted \( \mu_i = \mathbb{E}[f_i(X)] \), and their \( m \times m \) population covariance matrix, \( R = \mathbb{E}[(f_i(x) - \mu_i)(f_j(x) - \mu_j)] \) are perfectly known.

\footnote{For simplicity of exposition, we assume the matrix is fully observed. While beyond the scope of this paper, our proposed methods and theory continue to hold if few entries are missing (at random), such that accurate estimates of various means, covariances and tensors, as detailed in Sections 3-4 are still possible.}
The following two lemmas show that $R$ and $\{\mu_i\}_{i=1}^m$, contain the information needed to extract the specificities and sensitivities of the $m$ classifiers. Lemma 1 appeared in [7], and implies that given the value of $b$ one may compute the balanced accuracy of each classifier. Lemma 2 is new and shows how to extract the sensitivity and specificity. Its proof appears in the appendix.

**Lemma 1.** The off-diagonal elements of the matrix $R$ are identical to those of a rank one matrix, $v v^T$ whose vector $v$, up to a $\pm 1$ sign ambiguity, is equal to

$$v = \sqrt{1 - b^2(2\mu - 1)},$$

where the vector $\pi = (\pi_1, \ldots, \pi_m)$ contains the balanced accuracies of the $n$ classifiers.

**Lemma 2.** Given the class imbalance $b$, the vector $\mu = (\mu_1, \ldots, \mu_m)$ containing the mean values of the $m$ classifiers, and the vector $v$ defined in (5), the vectors $\psi = (\psi_1, \ldots, \psi_m)$ and $\eta = (\eta_1, \ldots, \eta_m)$ with the specificities and sensitivities of the $m$ classifiers are given by

$$\psi = \frac{1}{2} \left(1 + \mu + \sqrt{\frac{1 - b}{1 + b}}\right), \quad \eta = \frac{1}{2} \left(1 - \mu + \sqrt{\frac{1 + b}{1 - b}}\right).$$

In practice, the mean values $\{\mu_i\}_{i=1}^m$, the population matrix $R$ and consequently the vector $v$ are unknown to us. Thus we estimate them from the given data, and plug into Eq. (6). Let us denote by $\hat{\mu}$ and $\hat{\psi}$ the sample mean and covariance matrix of all classifiers, whose entries are given by

$$\hat{\mu}_i = \frac{1}{n} \sum_{k=1}^n f_i(x_k), \quad \hat{\psi}_{ij} = \frac{1}{n - 1} \sum_{k=1}^n (f_i(x_k) - \hat{\mu}_i) (f_j(x_k) - \hat{\mu}_j).$$

Similarly, let $\hat{\psi}$ be an estimate of the vector $v$ computed from the matrix $\hat{R}$. Ref. [7] presents several methods to construct such an estimator, including the resolution of its inherent $\pm 1$ sign ambiguity, via assumption (iii). Inserting $\hat{\mu}$ and $\hat{\psi}$ into (6), gives the following estimates for $\psi$ and $\eta$,

$$\hat{\psi} = \frac{1}{2} \left(1 + \hat{\mu} + \sqrt{\frac{1 - b}{1 + b}}\right), \quad \hat{\eta} = \frac{1}{2} \left(1 - \hat{\mu} + \sqrt{\frac{1 + b}{1 - b}}\right).$$

The following lemma, proven in the appendix, presents some statistical properties of $\hat{\psi}$ and $\hat{\eta}$.

**Lemma 3.** Under assumptions (i)-(iii) of Section 3, $\hat{\psi}$ and $\hat{\eta}$ are consistent estimators of $\psi$ and $\eta$. Furthermore, as $n \to \infty$, their error decreases as $O_P \left(\frac{1}{\sqrt{n}}\right)$.

$$\hat{\psi}_i = \psi_i + O_P \left(\frac{1}{\sqrt{n}}\right), \quad \hat{\eta}_i = \eta_i + O_P \left(\frac{1}{\sqrt{n}}\right).$$

In summary, assuming the class imbalance $b$ is known, Eq. (6) gives a computationally efficient way to estimate the sensitivities and specificities of all classifiers. Lemma 3 ensures that this approach is also consistent. As far as we know, this is the first computationally efficient method, with strong convergence guarantees, to directly estimate these parameters. This resolves the long-standing problem originally posed by Dawid and Skene [2], whose previous solutions were mostly based on expectation maximization approaches to the full likelihood function, see [2,6]. In the next section we show that the assumption of explicit knowledge of $b$ can be removed, whereas in Section 5 we show that a similar approach can also (partly) handle the multiclass case.

### 3.1 Unsupervised Ensemble Learning

We now consider the second problem discussed in Section 3, the construction of an unsupervised ensemble learner. To this end, note that under the stronger assumption that all classifiers make independent errors, the likelihood of a label $y$ at an instance $x$ with predicted labels $f_1(x), \ldots, f_m(x)$ is

$$L(f_1(x), \ldots, f_m(x)) | y) = \prod_{i=1}^m \Pr(f_i(x) | y).$$

In the above equation, each term $\Pr(f_i(x) | y)$ depends on the specificity and sensitivity $\psi_i$ and $\eta_i$ of the $i$-th classifier. While the likelihood is non-convex in both $\psi_i$, $\eta_i$ and $y$, if the former are known, there is an explicit closed-form solution for the maximum-likelihood value of the class label,

$$\hat{y}^{(ML)} = \text{sign} \left(\sum_i f_i(x) \ln \alpha_i + \ln \beta_i\right)$$

where

$$\alpha_i = \frac{\psi_i \eta_i}{(1 - \psi_i)(1 - \eta_i)}, \quad \beta_i = \frac{\psi_i (1 - \psi_i)}{\eta_i (1 - \eta_i)}.$$

In [2], the authors showed via a Taylor expansion near $\psi = \eta = 1/2$, that $\beta$ is approximately zero, and $\alpha_i \approx 1 + 4(2\psi_i - 1)$, and hence derived the following spectral meta-learner (SML),

$$\hat{y}^{(SML)} = \text{sign} \left(\sum_i f_i(x) \hat{\psi}_i\right).$$

Their motivation was that they only had estimates of the vector $v$, which according to Eq. (6) is proportional to $(2\mu - 1)$. Since we consistently estimate the individual specificities and sensitivities of the $m$ classifiers, we instead suggest to plug in their estimates directly into Eq. (12) and (13). Our approach, denoted $i$-SML, should provide a more accurate ensemble learner in cases where few classifiers are significantly better than random. We present such examples in Sec. 4.
4 Estimation of the class imbalance

We now consider the problem of estimating $\phi$ and $\eta$ when the class imbalance $b$ is unknown. Our proposed approach is to first estimate $b$, and plug this estimate into Eq. 6. We present two different methods for estimating the class imbalance. The first one uses the covariance matrix and the 3-dimensional covariance tensor of all $m$ classifiers. The second method exploits properties of the likelihood function. As detailed below, both methods are computationally efficient, but require stronger independence assumptions than Eq. 4 on classifier errors to prove their consistency.

4.1 Estimation of the class imbalance by the 3 dimensional joint covariance tensor

For the method derived in this subsection, we assume that the classifiers are conditionally independent in triplets. That is, for every $f_i, f_j, f_k$ with $i \neq j \neq k$ and for all labels $a_i, a_j, a_k \in \{-1, 1\}$,

$$\Pr(f_i = a_i, f_j = a_j, f_k = a_k | y) = \Pr(f_i = a_i | y)\Pr(f_j = a_j | y)\Pr(f_k = a_k | y).$$

Let $T = (T_{ijk})$ denote the 3-dimensional covariance tensor of the set of $m$ classifiers $\{f_i(X)\}_{i=1}^m$.

$$T_{ijk} = \mathbb{E}[(f_i(X) - \mu_i)(f_j(X) - \mu_j)(f_k(X) - \mu_k)].$$

The following lemma, proven in the appendix, characterizes the relation between the tensor $T$, the class imbalance $b$ and the balanced accuracies of the $m$ classifiers.

Lemma 4. Under assumption (14), the following holds for all $i \neq j \neq k$,

$$T_{ijk} = -2b(1 - b^2)/(2\pi_i - 1)(2\pi_j - 1)(2\pi_k - 1).$$

According to (16), the off diagonal elements of $T$ (with $i \neq j \neq k$) correspond to a rank one tensor,

$$T = w \circ w \circ w,$$

where $\circ$ denotes the outer product and the vector $w \in \mathbb{R}^m$ is equal to

$$w = (-2b(1 - b^2))^{\frac{1}{3}} \cdot (2\pi - 1).$$

Note that unlike the vector $v$ of the covariance matrix $R$, there is no sign ambiguity in the vector $w$.

From (15) and (13) it follows that the vectors $v$ of $R$ and $w$ of $T$ are both proportional to the vector $(2\pi - 1)$, where the proportionality factor depends on the class imbalance $b$. Hence,

$$w = \alpha(b)v,$$

where $\alpha(b) = (-2b)^{\frac{1}{3}}/(1 - b^2)^{\frac{1}{3}}$. Inverting this expression yields the following relation

$$b = \pm \sqrt[3]{\alpha(b)/(4 + \alpha(b))}$$

where the sign ambiguity can be resolved using assumption (iii). Eq. 20 thus shows, that in our setup, as $n \rightarrow \infty$, the $m \times m$ matrix $\{f_i(x_j)\}$ contains sufficient data to determine both the class imbalance and the sensitivity and specificity of all $m$ classifiers.

In practice, the tensor $T$ is unknown. However, we may estimate it by

$$\hat{T}_{ijk} = \frac{1}{n} \sum_{i=1}^n (f_i(x_i) - \bar{\mu}_i)(f_j(x_i) - \bar{\mu}_j)(f_k(x_i) - \bar{\mu}_k).$$

From the off diagonal elements of $\hat{T}$ we then estimate the vector $\hat{w}$. A simple and computationally efficient way to do so is described in the appendix. An estimate of $\alpha$ is possible by least squares,

$$\hat{\alpha} = \arg\min_\alpha \sum_{i=1}^m (\hat{w}_i - \alpha \hat{v}_i)^2.$$

Let us summarize the algorithm for estimating the class imbalance, using the results above.

Algorithm 1 Estimating class imbalance with the 3 dimension covariance tensor

1: Estimate covariance matrix $R$ by Eq. (7).
2: Estimate $v$ using the off diagonal values of $\hat{R}$ (see appendix).
3: Estimate the 3 dimensional tensor $T$ by Eq. 21.
4: Estimate $w$ using the off diagonal values of $\hat{T}$ (see appendix).
5: Estimate $\alpha$ via Eq. 22 and then estimate $b$ by Eq. 20.
6: Resolve the sign ambiguity in $b$ using assumption (iii).
4.2 Estimation of the class imbalance by a restricted-likelihood function

The algorithm described in Section 4.1 relied on estimates $R$ and $T$ of the covariance matrix and tensor of the $m$ classifiers. The latter, being three-dimensional, may require a large number of instances to be close to $T$. We now present a second method to estimate the class imbalance, which may require fewer unlabeled instances $n$. This method is based on the likelihood function of the data, but requires the following stronger assumption of joint conditional independence of all $m$ classifiers,

$$
\Pr \{f_1 = a_1, \ldots, f_m = a_m | y \} = \prod_{i=1}^{m} \Pr \{f_i = a_i | y \}.
$$

Our starting point is Eq. (3) which provides consistent estimates of $\psi$ and $\eta$ given the class imbalance $b$. In particular, any guess $\hat{b}$ of the class imbalance, yields corresponding guesses for the sensitivities and specificities of all $m$ classifiers, $\hat{\psi}(\hat{b})$ and $\hat{\eta}(\hat{b})$. Hence, our approach is to construct a suitable functional, that depends on both $\hat{b}$ and on the observed data, $\hat{G}_n(Z|\hat{b})$ whose maxima as a function of $\hat{b}$, as $n \to \infty$ is attained at the true class imbalance $b$.

To this end, let $f(x) = (f_1(x), \ldots, f_m(x))$ denote the vector of labels predicted by the $m$ classifiers at a single instance $x$. We define the following approximate log-likelihood, assuming class imbalance $\hat{b}$

$$
\hat{g}_n(f(x)|\hat{b}) = \log \Pr \left\{ f(x) | \hat{\psi}(\hat{b}), \hat{\eta}(\hat{b}), \hat{b} \right\}
$$

where $\hat{\psi}$ and $\hat{\eta}$ are given by Eq. (3), and an expression for the above probability is given in Eq. (54) in the appendix. Our functional $\hat{G}_n(Z|\hat{b})$ is the average of $\hat{g}_n(f(x)|\hat{b})$ over all instances $x_j$,

$$
\hat{G}_n(Z|\hat{b}) = \frac{1}{n} \sum_{j=1}^{n} \hat{g}_n(f(x_j)|\hat{b}).
$$

As the estimates of $\psi, \eta$ become singular (in fact not defined) when $b = \pm 1$, in what follows we assume there is an apriori-known $\delta > 0$, such that the true class imbalance $b \in [-1 + \delta, 1 - \delta]$. The estimate of the class imbalance is then defined as

$$
\hat{b}_n = \arg \max_{\hat{b} \in [-1 + \delta, 1 - \delta]} \hat{G}_n(Z|\hat{b}).
$$

To justify this method, it is again constructive to consider the limit $n \to \infty$. First, at any instance $x$, and any $\hat{b} \in [-1 + \delta, 1 - \delta]$, the convergence of $\hat{\psi}$ and $\hat{\eta}$ to $\psi$ and $\eta$, respectively, implies that

$$
\lim_{n \to \infty} \hat{g}_n(f(x)|\hat{b}) = g(f(x)|\hat{b}) \equiv \log \Pr \left\{ f(x) | \hat{\psi}(\hat{b}), \hat{\eta}(\hat{b}), \hat{b} \right\}.
$$

Next, since the $n$ instances $x_j$ are i.i.d., by the law of large numbers, combined with the delta method

$$
\lim_{n \to \infty} \hat{G}_n(Z|\hat{b}) = G(\hat{b}) \equiv E(x,y) \left\{ g(f(X)|\hat{b}) \right\}.
$$

The following theorem, proven in the appendix, shows that the maxima of $G(\hat{b})$ is obtained at the true class imbalance $\hat{b} = b$, and that $b_n \to b$ in probability. Since $b_n$ is the maximizer of an approximate likelihood, its convergence to $b$ is not a direct consequence of the consistency of ML estimators. Instead, what is needed is uniform convergence in probability of $G_n(\hat{b})$ to $G(b)$, as discussed in [6].

**Theorem 1.** Assume Eq. (23) holds. Further assume that $b \in [-1 + \delta, 1 - \delta]$ for an a-priori known $\delta > 0$, and that classifiers sensitivities and specificities satisfy $\epsilon < \psi, \eta < 1 - \epsilon$ for an a-priori known $\epsilon > 0$. Then,

$$
\hat{b} = \arg \max_{b \in [-1 + \delta, 1 - \delta]} E(x,y) \left\{ g(f(X)|\hat{b}) \right\}
$$

and as $n \to \infty$ the estimate $\hat{b}_n$ of Eq. (26) converges to $b$ in probability.

Algorithm 2 summarizes the method for estimating $b$ by the restricted-likelihood functional.

**Algorithm 2** Estimating the class imbalance using the log likelihood functional

1: Estimate the covariance matrix $\hat{R}$, the mean values $\{\hat{\mu}_i\}_{i=1}^m$ and the vector $\hat{\psi}$.
2: for $\hat{b} \in (-1 + \delta, 1 - \delta)$ do
3: Estimate the values $\hat{\psi}(\hat{b})$ and $\hat{\eta}(\hat{b})$ using Eq. (5).
4: Calculate $G_n(Z|\hat{b})$ by Eqs. (24) and (25).
5: end for
6: Estimate $b$ by Eq. (26).

This algorithm scans possible values of $\hat{b}$, where each evaluation of $G$ requires $O(mn)$ operations. Since $\hat{g}_n$ and consequently $\hat{G}_n$ are smooth functions of $\hat{b}$ on $(-1 + \delta, 1 - \delta)$, the finite grid of values of $\hat{b}$ can be of size polynomial in $n$ and the method is computationally efficient.
5 The multi-class case

We now consider the multi-class case, with $K$ distinct classes. Here we are given the predictions of $m$ multi-class classifiers, $f_i: X \rightarrow \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$. Instead of the class imbalance $b$, we now have a vector of $K$ a-priori class probabilities $p_k = \Pr(Y = k)$. Similarly, instead of specificity and sensitivity, now each classifier is characterized by a $K \times K$ confusion matrix $\psi^i$:

$$
\psi^i_{kk'} = \Pr(f_i(x) = k' | Y = k) \quad k, k' \in \mathcal{Y}.
$$

Given only an $m \times n$ matrix of predictions, with elements $f_i(x_j) \in \{1 \ldots K\}$, we consider the following problem, analogous to the one described in Section 2: consistently and computationally efficiently estimate the confusion matrices $\psi^i$ of all classifiers and the class probabilities $p_k$.

As we show below, by a simple reduction to the binary case, it is possible to consistently estimate the class probabilities $p_k$, as well as the diagonals of the confusion matrices, namely the various probabilities $\Pr(f_i(X) = k | Y = k)$, though unfortunately not all their entries.

Similarly to the binary case, we make an assumption regarding the mutual independence of errors made by different classifiers. Whether it is necessary for the independence to be in couples, triplets or for the full set of classifiers depends on the method used for solving the reduced binary problems.

Next, we build upon the methods developed in Sections 3 and 4 for binary problems. Consider a split of the group $\mathcal{Y} = \{1 \ldots K\}$ into two non-empty disjoint subsets, $\mathcal{Y} = A \cup (\mathcal{Y} \setminus A)$, where $A \subset \mathcal{Y}$ is a non-trivial subset of $\mathcal{Y}$, with $0 < |A| < K$. Next, define the binary classifiers $\{f^A_i\}_{i=1}^m$:

$$
\begin{align*}
    f^A_i(X) = \begin{cases} 
        1 & f_i(X) \in A \\
        -1 & f_i(X) \not\in A 
    \end{cases}
\end{align*}
$$

Using one of the algorithms described in Section 4, we estimate the a-prior probability of group $A$:

$$
p^A = \Pr(Y \in A) = \sum_{k \in A} p_k
$$

and the sensitivity of each classifier $f^A_i$ by Eq. (30):

$$
\psi^A_i = \Pr(f_i(X) \in A | Y \in A) = \frac{1}{p^A} \sum_{k,k' \in A} p_k \psi^i_{kk'},
$$

By considering all 1-vs-all splits, with $A = \{k\}$, and $k = 1, \ldots, K$, we may thus consistently and computationally efficiently estimate all class probabilities $p_k$, and all diagonal entries $\psi^i_{kk}$.

While beyond the scope of this paper, we remark that these and other splits provide additional linear equations on the entries of the confusion matrices $\psi^i$. For example, assume for simplicity that all $p_k = 1/K$, then subsets $A = \{k, k'\}$ provide the values of $\psi^i_{kk} + \psi^i_{kk'}$. However, even for $K = 3$ classes, the resulting equations are under-constrained. An interesting question for further research is whether all entries of $\psi^i$ may be consistently estimated in a computationally efficient manner.

6 Results for artificial and real data

6.1 Artificial Data

First, we present simulated results with artificial binary data. In the following we considered an ensemble of $m = 10$ classifiers, whose sensitivities and specificities were chosen uniformly at random from the interval $[0.4, 0.8]$. Thus, assumption (iii) on the values of the balanced accuracies $\pi$ holds. The vector of true labels $y$ was randomly generated according to the class imbalance $b$, and the data matrix $Z$ was randomly generated according to $y$, $\psi$, and $\eta$.

Figs. 1A and 1B present the accuracy (mean and standard deviation) of the estimates $\hat{b}$ of the class imbalance, achieved by the two different algorithms presented in Sections 3.1 and 3.2, as a function of the number of unlabeled instances $n$, for several values of $b$ ($b = 0, 0.3, 0.6$). As expected, and for both methods, the accuracy increases as the number of instances increase. Further simulations based on artificial data appear in the appendix.

6.2 Real data results

Here we present the results of our algorithms on various binary and multi-class problems using the MNIST data. To create an ensemble of classifiers we used $m = 10$ classification methods implemented in the software package Weka [4]. (see supplementary for further details). The unlabeled test data of size $n \approx 5 \cdot 10^4$ was composed of roughly 40K instances from the original training set augmented by the 10K instances of the original test instances.

To mimic a potentially realistic setting where different classifiers have significantly different accuracies, we used a different number of training instances ranging from 50 to 2000 (from the remaining 20K training data) for each classifier. The simulation was repeated 10 times, each time with a different randomly chosen training set.
(a) Estimating $b$ via the 3-D tensor $T$. (b) Estimating $b$ via the likelihood $G$

Fig. 1: The mean and variance of two estimators of the class imbalance $b$, vs. the number of instances $n$, for several values of $b$.

**Unsupervised Ensemble Learning.** To define a binary problem, we divided the MNIST data set into two classes as follows: 0 – 4 vs. 5 – 9. We compared 4 ensemble learners: Voting, SML, i-SML and oracle ML. Fig. 2a shows the balanced accuracy of the three top methods (voting had a much lower balanced accuracy of 0.78). As seen from the figure, i-SML improves upon the SML by approximately 0.5%. Fig. 2b displays the error rate $1 - \pi_{i-SML}$ vs. $1 - \pi_{SML}$. As all points are below the diagonal, the improvement over SML was consistent in all 10 simulation runs. Further simulations based on real data appear in the appendix.

(a) The balanced accuracy for 3 ensemble methods: SML, i-SML and optimal ML.

(b) The value of $1 - \pi_{i-SML}$ as a function of $1 - \pi_{SML}$.

Fig. 2: Comparing 3 ensemble learning algorithms. The number of classifiers is $m = 10$. 
References

[1] C.A. Andersson and R. Bro. The n-way toolbox for matlab. *Journal of Chemometrics*, 52:1–4, 2000.

[2] A. P. Dawid and A. M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C*, 28:20–28, 1979.

[3] T.G. Dietterich. Ensemble methods in machine learning. In *Lecture Notes in Computer Science*, volume 1857, pages 1–15. Springer, Berlin, 2000.

[4] M. Hall, E. Frank, G. Holmes, G. Pfahringer, P. Reutemann, and I.H Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[5] D.R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *IEEE Alerton Conference on Communication, Control and Computing*, pages 284–291, 2011.

[6] W. K. Newey. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59:1161–1167, 1991.

[7] F. Parisi, F. Strino, B. Nadler, and Y. Kluger. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences*, 111:1253–1258, 2014.

[8] V.C. Raykar, Y. Shipeng, L.H. Zhao, G.H. Valdez, C. Florin, L. Bogoni, and Moy L. Learning from crowds. *J. Machine Learning Research*, 11:1297–1322, 2010.

[9] L. Rokach. Collective-agreement-based pruning of ensembles. *Computational Statistics and Data Analysis*, 53:1015–1026, 2009.

[10] A. Sheshadri and M. Lease. Square: A benchmark for research on computing crowd concensus. In *AAAI conference on human computation and crowdsourcing*, 2013.

[11] R. Snow, B. O’Connor, D. Jurafsky, and A.Y. Ng. Cheap and fast but is it good? In *Conference on Empirical Methods in Natural Language Processing*, 2008.

[12] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, 2010.

[13] J. Whitehill, P. Ruvolo, T. Wu, J Bergsma, and J.R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, 2009.
A Estimation of $\psi$ and $\eta$

Proof of Lemma 2. We first recall the following expression, derived in [7], for the vector $\mu$ containing the mean values of the $m$ classifiers,

$$\mu = 2\delta + b(2\pi - 1)$$  \hspace{1cm} (32)

where $\delta = (\delta_1, \ldots, \delta_m)$ denotes the vector containing half of the difference between $\psi$ and $\eta$,

$$\delta = \frac{\psi - \eta}{2}$$  \hspace{1cm} (33)

Next, recall from Lemma 1 (also proven in [7]) that the off-diagonal elements of the covariance matrix $R$ correspond to a rank-1 matrix $vv^T$ where,

$$v = \sqrt{1 - \frac{b^2}{4}(2\pi - 1)}.$$  \hspace{1cm} (34)

Inverting the relation between $v$ and $\pi$ in Eq. (34) gives

$$\pi = \frac{1}{2} \left( \frac{v}{\sqrt{1 - \frac{b^2}{4}}} + 1 \right).$$  \hspace{1cm} (35)

Plugging (35) into (32), we obtain the following expression for the vector $\delta$, in terms of $v$ and $\mu$,

$$\delta = \frac{1}{2} \left( \mu - b \frac{v}{\sqrt{1 - \frac{b^2}{4}}} \right).$$  \hspace{1cm} (36)

Combining (33), (35) and (36) we obtain $\psi(b)$ and $\eta(b)$,

$$\psi = \pi + \delta = \frac{1}{2} \left( 1 + \mu + \sqrt{\frac{1 - b}{1 + b}} \right), \quad \eta = \pi - \delta = \frac{1}{2} \left( 1 - \mu + \sqrt{\frac{1 + b}{1 - b}} \right).$$  \hspace{1cm} (37)

B Statistical Properties of $\psi$ and $\eta$

Proof of Lemma 3. Eq. (8) provides an explicit expression for $\hat{\psi}$ and $\hat{\eta}$ as a function of the estimates $\hat{v}$ and $\hat{\mu}$. The empirical mean $\hat{\mu}$ is clearly not only unbiased, but by the law of large numbers also a consistent estimate of $\mu$, and its error indeed satisfies

$$\hat{\mu} = \mu + \mathcal{O}_P\left( \frac{1}{\sqrt{n}} \right).$$

The estimate $\hat{v}$, computed by one of the methods described in [7] may be biased, but as proven there is still consistent, and assuming at least two classifiers are different than random (namely, the eigenvalue of the rank one matrix is non-zero), its error also decreases as $\mathcal{O}_P\left( \frac{1}{\sqrt{n}} \right)$,

$$\hat{v} = v + \mathcal{O}_P\left( \frac{1}{\sqrt{n}} \right).$$  \hspace{1cm} (38)

Given the exact value of the class imbalance $b$, since the dependency of $\hat{\psi}$ and $\hat{\eta}$ on $\hat{v}$ and $\hat{\mu}$ is linear, it follows that both are also consistent and that their estimation error is $\mathcal{O}_P\left( \frac{1}{\sqrt{n}} \right).$ \hspace{1cm} \square

C The joint covariance tensor $T$

Proof of Lemma 4. To simplify the proof, we first introduce the following linear transformation to the original classifiers,

$$\tilde{f}_i(x) = \frac{f_i(x) + 1}{2}.$$  \hspace{1cm} (39)

Note, that the output space $Y$ of the new classifiers is $\{0, 1\}$. In addition, let us also denote by $p$ the a-priori probability $\Pr(Y = 1)$,

$$p = \frac{1 + b}{2}.$$  \hspace{1cm} (40)

Let us also denote by $\tilde{\eta}_i$ and $\tilde{\psi}_i$ the following probabilities,

$$\tilde{\eta}_i = \Pr(\tilde{f}_i(x) = 1|Y = 0), \quad \tilde{\psi}_i = \Pr(\tilde{f}_i(x) = 1|Y = 1).$$  \hspace{1cm} (41)

Note that the meaning of $\tilde{\eta}_i$ is different from the meaning of the specificity $\eta_i$ of the original classifiers. The relation between the two probabilities is given by $\tilde{\eta}_i = 1 - \eta_i$.

In terms of the new variables, the mean of the new classifier, denoted by $\tilde{\mu}_i$, is given by

$$\tilde{\mu}_i = \mathbb{E}[\tilde{f}_i(X)] = \Pr(\tilde{f}_i(X) = 1) = p\tilde{\psi}_i + (1 - p)\tilde{\eta}_i.$$  \hspace{1cm} (42)
Next, let us calculate the (un-centered) covariance between two different classifiers $i \neq j$,

$$E[\tilde{f}_i(X)\tilde{f}_j(X)] = Pr(\tilde{f}_i(X) = 1, \tilde{f}_j(X) = 1) = p\tilde{\psi}_i\tilde{\psi}_j + (1-p)\tilde{\eta}_i\tilde{\eta}_j$$ \hspace{1cm} (43)

Last, the joint covariance between 3 different classifiers $i \neq j \neq k$ is given by

$$E[\tilde{f}_i(X)\tilde{f}_j(X)\tilde{f}_k(X)] = Pr(\tilde{f}_i(X) = 1, \tilde{f}_j(X) = 1, \tilde{f}_k(X) = 1) = p\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k + (1-p)\tilde{\eta}_i\tilde{\eta}_j\tilde{\eta}_k$$ \hspace{1cm} (44)

The first step in calculating the joint covariance tensor of the original classifiers is to note that $f_i = 2\tilde{f}_i - 1$ and $\mu_i = 2\tilde{\mu}_i - 1$. Hence,

$$T_{ijk} = E[(f_i(X) - \mu_i)(f_j(X) - \mu_j)(f_k(X) - \mu_k)] = 8\tilde{T}_{ijk}$$

where

$$\tilde{T}_{ijk} = E[(\tilde{f}_i(X) - \tilde{\mu}_i)(\tilde{f}_j(X) - \tilde{\mu}_j)(\tilde{f}_k(X) - \tilde{\mu}_k)].$$

Upon opening the brackets, the later can be equivalently written as

$$\tilde{T}_{ijk} = E\left[\tilde{f}_i(X)\tilde{f}_j(X)\tilde{f}_k(X)\right] - \tilde{\mu}_i E\left[\tilde{f}_j(X)\tilde{f}_k(X)\right] - \tilde{\mu}_j E\left[\tilde{f}_i(X)\tilde{f}_k(X)\right] - \tilde{\mu}_k E\left[\tilde{f}_i(X)\tilde{f}_j(X)\right] + 2\tilde{\mu}_i\tilde{\mu}_j\tilde{\mu}_k$$

Plugging (42), (43) and (44) into (45) we get,

$$\tilde{T}_{ijk} = p\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k + (1-p)\tilde{\eta}_i\tilde{\eta}_j\tilde{\eta}_k - \left(p\tilde{\psi}_i + (1-p)\tilde{\eta}_i\right)\left(p\tilde{\psi}_j + (1-p)\tilde{\eta}_j\right)\left(p\tilde{\psi}_k + (1-p)\tilde{\eta}_k\right)$$

$$- \left(p\tilde{\psi}_j + (1-p)\tilde{\eta}_j\right)\left(p\tilde{\psi}_k + (1-p)\tilde{\eta}_k\right)\left(p\tilde{\psi}_i + (1-p)\tilde{\eta}_i\right)$$

$$+ 2\left(p\tilde{\psi}_i + (1-p)\tilde{\eta}_i\right)\left(p\tilde{\psi}_j + (1-p)\tilde{\eta}_j\right)\left(p\tilde{\psi}_k + (1-p)\tilde{\eta}_k\right)$$

(46)

Opening the brackets and collecting similar terms yields

$$\tilde{T}_{ijk} = \left(p - 3p^2 + 2p^3\right)\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k + \left(2p^2(1-p) - p(1-p)^2\right)\left(\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k + \tilde{\eta}_i\tilde{\psi}_j\tilde{\psi}_k + \tilde{\eta}_j\tilde{\psi}_i\tilde{\psi}_k + \tilde{\eta}_k\tilde{\psi}_i\tilde{\psi}_j\right)$$

$$+ \left(2p(1-p)^2 - p(1-p)^3\right)\left(\tilde{\eta}_i\tilde{\psi}_j\tilde{\psi}_k + \tilde{\eta}_j\tilde{\psi}_i\tilde{\psi}_k + \tilde{\eta}_k\tilde{\psi}_i\tilde{\psi}_j\right)$$

$$+ \left(1 - p - 3(1-p)^2 + 2(1-p)^3\right)\tilde{\eta}_i\tilde{\eta}_j\tilde{\eta}_k.$$ \hspace{1cm} (47)

Note that all polynomials in $p$ in the above expression are equal to $p^2(1-p)(1-2p)$. Hence,

$$\tilde{T}_{ijk} = p(1-p)(1-2p)(\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k - \tilde{\eta}_i\tilde{\psi}_j\tilde{\psi}_k - \tilde{\eta}_j\tilde{\psi}_i\tilde{\psi}_k - \tilde{\eta}_k\tilde{\psi}_i\tilde{\psi}_j + \tilde{\eta}_i\tilde{\eta}_j\tilde{\psi}_k + \tilde{\eta}_j\tilde{\eta}_k\tilde{\psi}_i + \tilde{\eta}_k\tilde{\eta}_i\tilde{\psi}_j - \tilde{\eta}_i\tilde{\eta}_j\tilde{\eta}_k)$$

Finally, replacing $\tilde{\psi}_i = \psi_i, \tilde{\eta}_i = 1 - \eta_i$ and $p = \frac{1+k}{2}$, yields

$$T_{ijk} = -2b(1-b^2)(\psi_i + \eta_i - 1)(\psi_j + \eta_j - 1)(\psi_k + \eta_k - 1)$$

$$= -2b(1-b^2)(2\pi_1 - 1)(2\pi_2 - 1)(2\pi_3 - 1).$$

$\square$

D Estimating the vector $\mathbf{w}$

The algorithm for estimating $\mathbf{w}$ consists of the following steps:

1. Estimate the vector $|\mathbf{w}| = (|w_1|, \ldots, |w_m|)$.
2. Estimate the sign of each component in the vector.

Steps (1-2) produce an initial possibly not very accurate estimate for $\mathbf{w}$. The purpose of the next two steps is to refine this initial estimate.
3. Estimate the diagonal elements of $\mathbf{T}$ that would correspond to a rank one tensor.
4. Perform tensor decomposition to obtain a refined estimate of $\mathbf{w}$.

We now present each step in more detail.

1. We build upon a method presented in [2] for the estimation of the vector $\mathbf{v}$. Under the following change of variables $|w_i| = e^{t_i}$, the absolute value of the off-diagonal elements of $\mathbf{T}$ are equal to

$$\log|T_{ijk}| = t_i + t_j + t_k.$$ \hspace{1cm} (48)

Of course, the values $T_{ijk}$ are unknown, and are estimated by $\tilde{T}_{ijk}$. We thus estimate the vector $\mathbf{t} = (t_1, \ldots, t_m)$ by least squares,

$$\mathbf{t} = \text{argmin}_{\mathbf{t}} \sum_{i<j<k} \left(\log|\tilde{T}_{ijk}| - t_i - t_j - t_k\right)^2,$$ \hspace{1cm} (49)

and then estimate the absolute value of the components of $\mathbf{w}$ by $|\tilde{w}_i| = e^{t_i}$.
2. From Eq. (56) it follows that the sign of the elements of $\mathbf{w}$ are related to the sign of the corresponding elements in $\mathbf{v}$ and the sign of the class imbalance $b$ via

$$\text{sign}(w_i) = -\text{sign}(v_i) \text{sign}(b).$$

(50)

Thus, if $b < 0$, then $\text{sign}(w_i) = \text{sign}(v_i)$ for $i = 1, \ldots, m$, and if $b > 0$ $\text{sign}(w_i) = -\text{sign}(v_i)$. This ambiguity can be easily resolved by the off diagonal elements in $T$:

$$\text{sign}(b) = \arg\min_{a = \{-1,1\}} \sum_{i < j < k} (\hat{T}_{ijk} - a \text{sign}(\hat{v}_i \hat{v}_j \hat{v}_k))|\hat{w}_i \hat{w}_j \hat{w}_k|^2$$

(51)

At this point, we obtained the following initial estimate for $\mathbf{w}$,

$$\hat{w}_i = -\text{sign}(b) \text{sign}(v_i)|\hat{w}_i|$$

(52)

However, this estimate is based on the log of the off-diagonal elements of $T$. The purpose of the next steps is to refine this initial estimate of $\mathbf{w}$,

3. Estimate the rank 1 diagonal elements of $\hat{T}$ by $w_i$,

$$\hat{T}_{iii} = \hat{T}_{i} = \hat{w}_i \hat{w}_i$$

(53)

4. Use one of the least square based methods for tensor decomposition in order to obtain an estimate of $\mathbf{w}$.

Note that since the tensor is a rank-one 3-way tensor, this is a relatively simple tensor decomposition problem, solved for example via PARAFAC [1].

### E The Approximate Log Likelihood Function

In this section we prove Theorem [1]. To this end, note that the function $\hat{g}_n(\mathbf{f}(x)|\hat{b})$ defined in Eq. (24) is the approximate log-likelihood of the observed vector $\mathbf{f}(x)$ of predicted labels at an instance $x$, assuming the class imbalance is $b$ and using the estimates $\hat{\psi}$ and $\hat{\eta}$ for the sensitivities and specificities of the $m$ classifiers.

Under the assumption that all classifiers make independent errors, the expression for $\Pr(\mathbf{f}(x)|\hat{\psi}, \hat{\eta}, b)$ is given by

$$\Pr(\mathbf{f} | \hat{b}) = \Pr(y = 1 | \hat{b}) Pr(\mathbf{f} | \hat{b}, y = 1) + \Pr(y = -1 | \hat{b}) Pr(\mathbf{f} | \hat{b}, y = -1)$$

$$= \left(1 - \frac{1}{2} \right) \prod_{i=1}^{m} \frac{\hat{\psi}_i^{1 - \hat{f}_i} \hat{f}_i^{\hat{f}_i - 1}}{(1 - \hat{\psi}_i) \hat{\psi}_i} + \left(1 - \frac{1}{2} \right) \prod_{i=1}^{m} \frac{\hat{\eta}_i^{1 - \hat{f}_i} \hat{f}_i^{\hat{f}_i - 1}}{(1 - \hat{\eta}_i) \hat{\eta}_i}$$

(54)

We first prove Eq. (29), that upon using the exact log-likelihood function $g(\mathbf{f} | \hat{b})$, its mean is maximized at the true value $b$.

To this end, we write the expectation explicitly,

$$E[g(\mathbf{f} | \hat{b})] = \sum_{\mathbf{r} \in \{-1,1\}^m} \Pr(\mathbf{f} | \hat{b}) g(\mathbf{f} | \hat{b}) = \sum_{\mathbf{r} \in \{-1,1\}^m} \Pr(\mathbf{f} | \hat{b}) \log \Pr(\mathbf{f} | \hat{b})$$

(55)

Note the difference between the assumed class imbalance $\hat{b}$, which appears inside the logarithm, and its true value $b$, over which we take the expectation.

To prove equation (29), let us first present the following auxiliary lemma, which can be easily proved by using Lagrange multipliers.

**Lemma 5.** Let us define the following function $h(\{a_i\}_{i=1}^{k}, \{c_i\}_{i=1}^{k})$,

$$h(\{a_i\}_{i=1}^{k}, \{c_i\}_{i=1}^{k}) = \sum_{i=1}^{k} a_i \log(c_i).$$

(56)

where $\{a_i\}_{i=1}^{k}$ are $k$ non-negative constants, and $\{c_i\}_{i=1}^{k}$ are $k$ unknown variables. Under the constraints that $\sum_{i=1}^{k} c_i = 1$, and $c_i \geq 0$, the function $h$ has a global maxima at $c_i = a_i$ for all $i$.

We shall use this lemma with $h = 2^m$ and the following set of $2^m$ constants $a_{\mathbf{r}}(\hat{b}) = \Pr(\mathbf{f} | \hat{b})$, over all possible $m$-dimensional vectors $\mathbf{f} \in \{-1,1\}^m$, and the $2^m$ variables $c_{\mathbf{r}} = \Pr(\mathbf{f} | \hat{b})$. The expectation of $g$ is now equal to

$$G(\hat{b}) = E[g(\mathbf{f} | \hat{b})] = \sum_{i=1}^{2^m} a_{\mathbf{r}} \log(c_{\mathbf{r}})$$

(57)

By Eq. (56), over all possible choices of $c_{\mathbf{r}}$, the expectation achieves its maxima at $c_{\mathbf{r}} = a_{\mathbf{r}}$ for all $i$. Since at $\hat{b} = b$, the corresponding probabilities $\Pr(\mathbf{f} | \hat{b} = b) = a_{\mathbf{r}}$, Eq. (55) follows.

Next, we wish to prove that $\hat{b}_n \rightarrow b$ in probability. To this end, we follow the approach outlined in [2], and prove the following uniform convergence in probability of $G_n$ to $G$,

$$\sup_{b \in [-1+ \delta, 1- \delta]} |G_n(b) - G(b)| = o_P(1)$$

11
This equation, coupled with the equicontinuity of $G$ implies the convergence in probability of the maximizer of $G_n$ (namely $b_n$) to that of $G$, which by Eq. (24) is $b$.

As proved in [6] Theorem 2.1, this uniform convergence in probability occurs if and only if there is pointwise convergence of $G_n(b)$ to $G(b)$, and $G_n(b)$ is stochastic equicontinuous. Fortunately, a sufficient condition for the latter property is that $G_n(b)$ is continuously differentiable and its derivative bounded, see [6] Corollary 2.2 and discussion after it.

In our case, since $G_n(b) = 1/n \sum \hat{g}_n(f(x_i))$, it suffices to prove that for any vector $f$, the function $\hat{g}_n(f(b))$ is continuously differentiable and has bounded derivative. First note that by their definition, Eq. (5), the functions $\psi_i(b)$ and $\eta_i(b)$ are continuously differentiable with bounded derivative for all $b \in [-1+\delta, 1-\delta]$. Next, under the assumptions of the theorem, that $\psi_i$ and $\eta_i$ are $\epsilon$ bounded from 0 and from 1, and hence also their estimates can be restricted to $\epsilon < \psi_i, \eta_i < 1 - \epsilon$, the term inside the logarithm in Eq. (24) is bounded away from zero. Hence, by its definition $\hat{g}_n$ satisfies the required condition.

### F Ensemble of Machine Learning Classifiers

The following table presents the 10 different classifiers used in our experiments, along with the number of labeled instances used to train each classifier, and the resulting mean specificity and sensitivity on the 50K test set, averaged over 10 different random training sets. Note that since some of these classifiers were trained with very limited labeled data, their accuracy on the test data is worse than random!

| classifier                        | Weka library          | Number of Instances | mean $\psi, \eta$ |
|----------------------------------|-----------------------|---------------------|-------------------|
| KNN - K nearest neighbours with $K=1$ | lazy.IBk              | 40                  | 0.4, 0.75        |
| KNN - K nearest neighbours with $K=2$ | lazy.IBk              | 40                  | 0.45, 0.73       |
| J48 - Decision tree              | trees.J48             | 40                  | 0.34, 0.41       |
| PART - Partial decision trees classifier | rules.PART          | 80                  | 0.27, 0.45       |
| LMT - Logistic model trees       | trees.LMT             | 80                  | 0.44, 0.53       |
| Random forest - with $n = 10$ trees | trees.RandomForest    | 80                  | 0.37, 0.46       |
| Logistic Regression              | functions.SimpleLogistic | 1200               | 0.52, 0.6        |
| Regularized Logistic regression  | functions.Logistic    | 1200               | 0.51, 0.56       |
| Sequential Minimal Optimization  | bayes.SMO             | 2000               | 0.81, 0.92       |
| NaiveBayes                       | bayes.NaiveBayes      | 2000               | 0.63, 0.85       |

### G Additional Simulation Results

#### G.1 Artificial Data

Fig. 3a and 3b show the accuracy of class imbalance estimation as a function of the number of classifiers $m$, for various values of $b$, with a fixed number of instances $n = 10^4$. As can be seen from the figure, both methods improve as the number of classifiers increase, though the improvement is larger for the likelihood method.

![Fig. 3: Two methods for estimating the class imbalance $b$. The number of instances is $n = 10^4$. $\psi_i$ and $\eta_i$ are randomly distributed in $[0.4, 0.8]$. The figures present the mean value and standard deviation of the estimation process, as a function of the number of classifiers.](image)

**Unsupervised Ensemble Learning** Using the same setup as in the main text, we compare four different ensemble learning algorithms: majority voting, SML [7], our improved method denoted i-SML, and an oracle method which uses Eq. (11) with the exact values of $\psi_i$ and $\eta_i$. 


Fig. 4: Comparing 3 ensemble learning algorithms. The number of classifiers is $m = 10$, $\psi_i$ and $\eta_i$ are uniformly distributed in $[0.4, 0.8]$.

(a) The balanced accuracy for 3 ensemble methods: SML, new-SML and optimal ML.
(b) The error rate $1 - \pi_{i\text{-SML}}$ vs. $1 - \pi_{\text{SML}}$.

Fig. 5: Comparison of i-SML to voting in a multi-class problem.

(a) The mean squared error $E\|\hat{p} - p\|^2$ of the estimated class probabilities.
(b) The mean square error between the diagonal values $\psi_{ii}$ and the estimated ones.

As can be seen in Figs. 5a and 5b, the estimates based on the i-SML approach are more accurate than the estimates based on majority voting.

G.2 Multi class results

The original 10-class problem in the MNIST data set was reduced to a 5 class problem by merging consecutive digits (0 with 1, 2 with 3, etc.). The 60K training set was split into a 40K test set (added to the original 10k test set), and a 20K training set. Once again, in order to mimic a potentially realistic setting where different classifiers have significantly different accuracies, we used a different number of training instances ranging from 50 to 2000 for each classifier. The simulation was repeated 10 times, each time with a different randomly chosen training set. We estimated the a-priori probability vector $p = (p_1, \ldots, p_5)$ of class probabilities, and the diagonal values of the set of confusion matrices $\{\psi_{ii}^{\text{est}}\}_{i=1}^{m}$. We compared the estimate produced via the algorithm described in Sec. 5 to the case where the estimation is done via the prediction results of majority voting. The comparison is done by the mean squared error between the actual and estimated vectors.

As can be seen in Figs. 5a and 5b, the estimates based on the i-SML approach are more accurate than the estimates based on majority voting.