A Estimation of ψ and η

Proof of Lemma 2. We first recall the following formula, derived in Parisi et al. [2014], for the vector $\boldsymbol{\mu}$ containing the mean values of the m classifiers,

$$\boldsymbol{\mu} = 2\boldsymbol{\delta} + b(2\boldsymbol{\pi} - 1) \tag{27}$$

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

$$\boldsymbol{\delta} = \frac{\boldsymbol{\psi} - \boldsymbol{\eta}}{2}. \tag{28}$$

Next, recall from Lemma 1 (also proven in Parisi et al. [2014]) that the off-diagonal elements of the covariance matrix R correspond to a rank-1 matrix $\mathbf{v}\mathbf{v}^T$ where,

$$\mathbf{v} = \sqrt{1 - b^2}(2\boldsymbol{\pi} - 1). \tag{29}$$

Inverting the relation between $\mathbf{v}$ and $\boldsymbol{\pi}$ in Eq. (29) gives

$$\boldsymbol{\pi} = \frac{1}{2}\left(\frac{\mathbf{v}}{\sqrt{1 - b^2}} + 1\right). \tag{30}$$

Plugging (30) into (27), we obtain the following expression for the vector $\boldsymbol{\delta}$, in terms of $\mathbf{v}$ and $\boldsymbol{\mu}$,

$$\boldsymbol{\delta} = \frac{1}{2}\left(\boldsymbol{\mu} - b\frac{\mathbf{v}}{\sqrt{1 - b^2}}\right). \tag{31}$$

Combining (28), (30) and (31) we obtain $\boldsymbol{\psi}(b)$ and $\boldsymbol{\eta}(b)$,

$$\boldsymbol{\psi} = \boldsymbol{\pi} + \boldsymbol{\delta} = \frac{1}{2}\left(1 + \boldsymbol{\mu} + \mathbf{v}\sqrt{\frac{1-b}{1+b}}\right),$$
$$\boldsymbol{\eta} = \boldsymbol{\pi} - \boldsymbol{\delta} = \frac{1}{2}\left(1 - \boldsymbol{\mu} + \mathbf{v}\sqrt{\frac{1+b}{1-b}}\right).$$

□

B Statistical Properties of ψ and η

Proof of Lemma 3. Eq. (5) provides an explicit expression for $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\eta}}$ as a function of the estimates $\hat{\mathbf{v}}$ and $\hat{\boldsymbol{\mu}}$. The empirical mean $\hat{\boldsymbol{\mu}}$ is clearly not only unbiased, but by the law of large numbers also a consistent estimate of $\boldsymbol{\mu}$, and its error indeed satisfies

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

The estimate $\hat{\mathbf{v}}$, computed by one of the methods described in Parisi et al. [2014] may be biased, but as proven there is still consistent, and assuming at least three classifiers are different than random (in particular, implying that the eigenvalue of the rank one matrix is non-zero), its error also decreases as $\mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right)$,

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

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

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,

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

Note, that the output space $\mathcal{Y}$ of the new classifiers is $\{0, 1\}$, with class probabilities equal to $1 - p$ and p respectively. Let us also denote by $\tilde{\eta}_i$ and $\tilde{\psi}_i$ the following probabilities,

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

Note that $\tilde{\eta}_i$ is not the specificity of classifier i, but rather its complement, $\tilde{\eta}_i = 1 - \eta_i$.

The mean of classifier $\tilde{f}_i$, denoted $\tilde{\mu}_i$, is given by

$$\tilde{\mu}_i = \mathbb{E}[\tilde{f}_i(X))] = \Pr(\tilde{f}_i(X) = 1) = p\tilde{\psi}_i + (1 - p)\tilde{\eta}_i \tag{32}$$

Next, let us calculate the (un-centered) covariance between two different classifiers $i \neq j$,

$$\begin{aligned}\mathbb{E}[\tilde{f}_i(X)\tilde{f}_j(X)] &= \Pr(\tilde{f}_i(X) = 1, \tilde{f}_j(X) = 1) \\ &= p\tilde{\psi}_i\tilde{\psi}_j + (1 - p)\tilde{\eta}_i\tilde{\eta}_j \quad (33)\end{aligned}$$

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

$$\begin{aligned}\mathbb{E}[\tilde{f}_i(X)\tilde{f}_j(X)\tilde{f}_k(X)] &= \Pr(\tilde{f}_i(X) = \tilde{f}_j(X) = \tilde{f}_k(X) = 1) \\ &= p\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k + (1 - p)\tilde{\eta}_i\tilde{\eta}_j\tilde{\eta}_k \quad (34)\end{aligned}$$

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,

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

where

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

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

$$\tilde{T}_{ijk} = \mathbb{E}\left[\tilde{f}_i(X)\tilde{f}_j(X)\tilde{f}_k(X)\right] \\ - \tilde{\mu}_i \mathbb{E}\left[\tilde{f}_j(X)\tilde{f}_k(X)\right] - \tilde{\mu}_j \mathbb{E}\left[\tilde{f}_i(X)\tilde{f}_k(X)\right] \\ - \tilde{\mu}_k \mathbb{E}\left[\tilde{f}_i(X)\tilde{f}_j(X)\right] + 2\tilde{\mu}_i\tilde{\mu}_j\tilde{\mu}_k \quad (35)$$

Plugging (32),(33) and (34) into (35) we get,

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

Opening the brackets and collecting similar terms yields

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

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

$$\tilde{T}_{ijk} = p(1-p)(1-2p)(\tilde{\psi}_i\tilde{\psi}_j\tilde{\psi}_k - \tilde{\eta}_i\tilde{\psi}_j\tilde{\psi}_k - \tilde{\eta}_j\tilde{\psi}_k\tilde{\psi}_i - \\ \tilde{\eta}_k\tilde{\psi}_i\tilde{\psi}_j + \tilde{\eta}_i\tilde{\eta}_j\tilde{\psi}_k + \tilde{\eta}_j\tilde{\eta}_k\tilde{\psi}_i + \tilde{\eta}_k\tilde{\eta}_i\tilde{\psi}_j - \tilde{\eta}_i\tilde{\eta}_k\tilde{\eta}_j) \quad (36)$$

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

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

□

Proof of Lemma 5. To prove that $\hat{b}_n$ is consistent with an asymptotic error $\mathcal{O}_P(1/\sqrt{n})$, we first recall that according to Parisi et al. [2014], it follows that

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

By its definition, each entry of $\hat{T}_{ijk}$ also incurs an error of $O_P(1/\sqrt{n})$. Hence, by the delta method, the estimate $\hat{\alpha}$ of Eq. (19), being a least squares minimizer, also satisfies

$$\hat{\alpha} = \alpha + \mathcal{O}_P(1/\sqrt{n}).$$

Since $\hat{b}_n$ is found by the smooth relation of Eq. (17), again by the delta method, $\hat{b}_n = b + \mathcal{O}_P(1/\sqrt{n})$. Finally, the fact that the corresponding estimates $\hat{\psi}_i$ and $\hat{\eta}_i$ also have errors $\mathcal{O}_P(1/\sqrt{n})$ follows by standard application of the delta method to Eq. (5), where all quantities $\hat{\boldsymbol{\mu}}, \hat{\mathbf{v}}$ and $\hat{b}$ have errors $\mathcal{O}_P(1/\sqrt{n})$. □

Dependence of estimated parameters on number of classifiers and their accuracies. Beyond the fact that $\hat{\alpha}$ and consequently $\hat{b}_n, \hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}$ are all $\mathcal{O}(1/\sqrt{n})$ consistent, it is of interest to study the dependence of these estimates on the number of classifiers and their accuracies. To this end, we first prove the following simple result.

Lemma 6. *Let $\hat{\alpha}$ be the estimate of α in Eq. (19). Then asymptotically as $n \to \infty$, its estimation error is given by*

$$\hat{\alpha} - \alpha = \frac{\langle \hat{T} - T, \mathbf{v}^{\otimes 3}\rangle}{\langle \mathbf{v}^{\otimes 3}, \mathbf{v}^{\otimes 3}\rangle} - \alpha\frac{\langle \hat{\mathbf{v}}^{\otimes 3} - \mathbf{v}^{\otimes 3}, \mathbf{v}^{\otimes 3}\rangle}{\langle \mathbf{v}^{\otimes 3}, \mathbf{v}^{\otimes 3}\rangle} + O_P\left(\frac{1}{n}\right) \quad (37)$$

where $\mathbf{v}^{\otimes 3} = \mathbf{v} \otimes \mathbf{v} \otimes \mathbf{v}$, and for any two tensors T, S, $\langle T, S\rangle = \sum_{i<j<k} T_{ijk}S_{ijk}$.

Proof. The minimizer of Eq. (19) is given by

$$\hat{\alpha} = \frac{\langle \hat{T}, \hat{\mathbf{v}}^{\otimes 3}\rangle}{\langle \hat{\mathbf{v}}^{\otimes 3}, \hat{\mathbf{v}}^{\otimes 3}\rangle}$$

According to Parisi et al. [2014], as $n \to \infty$, the estimate $\hat{\mathbf{v}}$ is $O(1/\sqrt{n})$ consistent, namely $\hat{\mathbf{v}} = \mathbf{v} + \delta\mathbf{v}$, where $\delta\mathbf{v} = O_P(1/\sqrt{n})$. Writing $\hat{T} = T + (\hat{T} - T)$ where the latter is also $O_P(1/\sqrt{n})$ and inserting these into the expression for $\hat{\alpha}$ above gives that

$$\hat{\alpha} = \frac{\langle T, \mathbf{v}^{\otimes 3}\rangle + \langle \hat{T} - T, \mathbf{v}^{\otimes 3}\rangle + \langle T, \hat{\mathbf{v}}^{\otimes 3} - \mathbf{v}^{\otimes 3}\rangle + O_P(1/n)}{\langle \mathbf{v}^{\otimes 3}, \mathbf{v}^{\otimes 3}\rangle + 2\langle \mathbf{v}^{\otimes 3}, \hat{\mathbf{v}}^{\otimes 3} - \mathbf{v}^{\otimes 3}\rangle + O_P(1/n)}$$

Next, recall that $T = \alpha\mathbf{v}^{\otimes 3}$. Now, keeping only the leading order error terms yields Eq. (37). □

According to Eq. (37), the estimation error depends on the statistical properties of the deviations $\hat{\mathbf{v}} - \mathbf{v}$ and $\hat{T} - T$ and their correlations. While these are quite complicated, we may gain insight by looking at some particular instances. Assume for simplicity that all classifiers have comparable accuracies. Then, $\langle \mathbf{v}^{\otimes 3}, \mathbf{v}^{\otimes 3}\rangle \propto m(m-1)(m-2)/6 \cdot (2\pi - 1)^6$. Hence, the estimation error in $\hat{\alpha}$ should decrease with the number of classifiers. Moreover, for a balanced problem with $b = 0$ and hence $\alpha = 0$, to leading order, the errors in $\hat{\alpha}$ and consequently also in $\hat{b}_n$ should not depend on the errors in estimating the eigenvector $\mathbf{v}$. Figure 4 shows this empirically. The x-axis is the number of classifiers, the y-axis is the mean absolute deviation $\mathbb{E}[|\hat{b}_n - b|]$ (MAE), both on a log scale. We considered

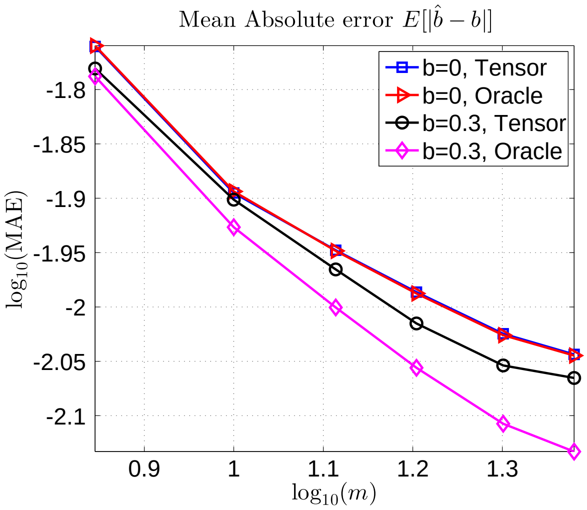


Fig. 4: Mean absolute error for the tensor based method, $\mathbb{E}[|\hat b_n - b|]$ vs. number of classifiers m, on log-log scale.

two values $b = 0$ and $b = 0.3$, and for each value of b we plotted two curves, one corresponding to the estimate $\hat b$ computed from $\hat\alpha$ based on $\hat{\mathbf{v}}$, and the second, an "oracle" one, where $\hat\alpha$ is estimated using the true $\mathbf{v}$. Indeed, for $b = 0$ both curves nearly coincide, in accordance to Eq. (37). In this simulation, all classifiers had a balanced accuracy in the range $[0.69, 0.71]$, and $n = 10,000$. These results suggest that it is potentially profitable to estimate the eigenvector $\mathbf{v}$ and the scalar α *jointly* from both the covariance matrix $\hat R$ and the tensor $\hat T$, and not separately as done in the present paper. This, as well as a more detailed study of the estimation errors are issues beyond the scope of the current work.

D The Restricted Likelihood Function

Proof of Theorem 1. By definition, the function $\hat g_n(\mathbf{f}(x)|\tilde b)$ in Eq. (22) is the log-likelihood of the observed vector $\mathbf{f}(x)$ of predicted labels at an instance x, assuming the class imbalance is $\tilde b$ and using the estimates $\hat{\boldsymbol\psi}$ and $\hat{\boldsymbol\eta}$ for the sensitivities and specificities of the m classifiers.

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

$$\begin{aligned}\Pr(\mathbf{f}|\tilde b) = \Pr(y=1|\tilde b)\Pr(\mathbf{f}|\tilde b, y=1)+\\ \Pr(y=-1|\tilde b)\Pr(\mathbf{f}|\tilde b, y=-1) =\\ \left(\tfrac{1+\tilde b}{2}\right)\prod_{i=1}^{m}\hat\psi_i^{\frac{1+f_i(x)}{2}}(1-\hat\psi_i)^{\frac{1-f_i(x)}{2}}+\\ \left(\tfrac{1-\tilde b}{2}\right)\prod_{i=1}^{m}\hat\eta_i^{\frac{1-f_i(x)}{2}}(1-\hat\eta_i)^{\frac{1+f_i(x)}{2}}\end{aligned} \quad (38)$$

We first prove Eq. (26), that upon using the exact log-likelihood function $g(\mathbf{f}|\tilde b)$, its mean is maximized at the true value b. To this end, we write the expectation explicitly,

$$\begin{aligned}\mathbb{E}[g(\mathbf{f}|\tilde b)] &= \sum_{\mathbf{f}\in\{-1,1\}^m}\Pr(\mathbf{f}|b)g(\mathbf{f}|\tilde b)\\ &= \sum_{\mathbf{f}\in\{-1,1\}^m}\Pr(\mathbf{f}|b)\log\Pr(\mathbf{f}|\tilde b)\end{aligned} \quad (39)$$

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

To prove Eq. (26), let us first present the following auxiliary lemma, which can be easily proved using Lagrange multipliers.

Lemma 7. *Consider the following function of k unknown variables $\{c_i\}_{i=1}^k$,*

$$h(\{c_i\}_{i=1}^k|\{a_i\}_{i=1}^k) = \sum_{i=1}^{k} a_i\log(c_i). \quad (40)$$

where $\{a_i\}_{i=1}^k$ are k non-negative constants. 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 use this lemma with $k = 2^m$ and the following set of 2^m constants $a_{\mathbf{f}}(b) = \Pr(\mathbf{f}|b)$, over all possible m-dimensional vectors $\mathbf{f}\in\{-1,1\}^m$, and the 2^m variables $c_{\mathbf{f}} = \Pr(\mathbf{f}|\tilde b)$. The expectation of g is now equal to

$$G(\tilde b) = \mathbb{E}[g(\mathbf{f}|\tilde b)] = \sum_{i=1}^{2^m} a_i\log(c_i) \quad (41)$$

By Eq. (40), over all possible choices of c_i, the expectation attains its maxima at $c_i = a_i$ for all i. Since at $\tilde b = b$, the corresponding probabilities $\Pr(\mathbf{f}|\tilde b = b) = a_{\mathbf{f}}$, Eq. (26) follows.

Next, we wish to prove that $\hat b_n \to b$ in probability. To this end, we follow the approach outlined in Newey [1991], and prove the following uniform convergence in probability of $\hat G_n$ to G,

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

This equation, coupled with the equicontinuity of G implies the convergence in probability of the maximizer of $\hat{G}_n$ (namely $\hat{b}_n$) to that of G, which by Eq. (26) is b.

As proved in [Newey, 1991, Theorem 2.1], this uniform convergence in probability is satisfied if and only if there is pointwise convergence of $\hat{G}_n(\tilde{b})$ to $G(\tilde{b})$, and $\hat{G}_n(\tilde{b})$ is stochastic equicontinuous. Fortunately, a sufficient condition for the latter property is that $\hat{G}_n(\tilde{b})$ is continuously differentiable and its derivative bounded, see Newey [1991] Corollary 2.2 and discussion after it.

In our case, since $\hat{G}_n(\tilde{b}) = 1/n \sum_i \hat{g}_n(\mathbf{f}(x_i)|\tilde{b})$, it suffices to prove that for any vector $\mathbf{f}$, the function $\hat{g}_n(\mathbf{f}|\tilde{b})$ is continuously differentiable with a bounded derivative. First note that by their definition, Eq. (5), the functions $\hat{\psi}_i(\tilde{b})$ and $\hat{\eta}_i(\tilde{b})$ are continuously differentiable with bounded derivative for all $\tilde{b} \in [-1+\delta, 1-\delta]$. Next, under the assumptions of the theorem, that ψ_i and η_i are ϵ bounded from 0 and from 1, and hence also their estimates can be restricted to $\epsilon < \hat{\psi}_i, \hat{\eta}_i < 1-\epsilon$, the term inside the logarithm in Eq. (22) is bounded away from zero. Hence, by its definition $\hat{g}_n$ satisfies the required condition. □

E Ambiguity in the Multi-Class Case

Proof of Theorem 2. For simplicity, let us assume that all K class probabilities are equal, $p_i = \frac{1}{K}$ for $i = 1, \ldots, K$. Let f_i be the set of original classifiers with confusion matrices $\{\psi^i\}_{i=1}^m$. We shall now construct another set of classifiers with different confusion matrices that nonetheless lead to the *same* values $\mu^i_{\mathcal{A}}$ and $R_{\mathcal{A}}$ for all subsets $\mathcal{A}$.

To this end, assume that all entries of the first confusion matrix ψ^1 are strictly positive and strictly smaller than one. Consider a second set of confusion matrices $\{\tilde{\psi}^i\}_{i=1}^m$ identical to the first, except for the following six changes in ψ^1: For three fixed indices $j \neq k \neq l$, let

$$\begin{aligned} \tilde{\psi}^1_{jk} &= \psi^1_{jk} + \Delta & \tilde{\psi}^1_{kj} &= \psi^1_{kj} - \Delta \\ \tilde{\psi}^1_{lj} &= \psi^1_{lj} + \Delta & \tilde{\psi}^1_{jl} &= \psi^1_{jl} - \Delta \\ \tilde{\psi}^1_{kl} &= \psi^1_{kl} + \Delta & \tilde{\psi}^1_{lk} &= \psi^1_{lk} - \Delta \end{aligned}$$

where Δ is sufficiently small so that all entries of $\tilde{\psi}^1$ are in $[0, 1]$.

Note that the new matrix $\tilde{\psi}^1$ is a valid confusion matrix, since for any column $r \in \{1, \ldots, K\}$

$$\sum_{i=1}^K \tilde{\psi}^1_{ir} = 1.$$

Let $\tilde{f}_1$ be the classifier corresponding to the modified matrix $\tilde{\psi}^1$. Next, note that the first order statistics of $\tilde{f}_1$ and of f_1 are unchanged. Indeed, by definition

$$\Pr(\tilde{f}_1(X) = r) = \frac{1}{K} \sum_{i=1}^K \tilde{\psi}^1_{ri}$$

If $r \notin \{j, k, l\}$, then $\tilde{\psi}^1_{ri} = \psi^1_{ri}$ and thus

$$\Pr(\tilde{f}_1(X) = r) = \Pr(f_1(X) = r) \tag{42}$$

If $r \in \{j, k, l\}$, then by construction, in the r-th row of $\tilde{\psi}^1$ there are precisely two modified entries, one increased by Δ and the other reduced by Δ, so overall the above equation still holds. Eq. (42) directly implies that $\tilde{\mu}^1_{\mathcal{A}} = \mu^1_{\mathcal{A}}$ for all subsets $\mathcal{A}$.

Next, let us show that the covariance matrices $R_{\mathcal{A}}$ also remain unchanged. Recall that the entries of $R_{\mathcal{A}}$ are determined by the values $\psi^1_{\mathcal{A}} \ldots \psi^m_{\mathcal{A}}$ and $\eta^1_{\mathcal{A}} \ldots \eta^m_{\mathcal{A}}$. Hence, it suffices to show that for all subsets $\mathcal{A}$

$$\tilde{\psi}^1_{\mathcal{A}} = \psi^1_{\mathcal{A}} \quad \text{and} \quad \tilde{\eta}^1_{\mathcal{A}} = \eta^1_{\mathcal{A}} \tag{43}$$

To this end, recall that by definition

$$\tilde{\psi}^1_{\mathcal{A}} = \frac{1}{K} \sum_{i,i' \in \mathcal{A}} \tilde{\psi}^1_{ii'} \quad \text{and} \quad \tilde{\eta}^1_{\mathcal{A}} = \frac{1}{K} \sum_{i,i' \notin \mathcal{A}} \tilde{\psi}^1_{ii'}$$

First consider the case $|\mathcal{A} \cap \{j, k, l\}| = 0$. Here, all relevant entries in the sum for $\tilde{\psi}^1_{\mathcal{A}}$ are unchanged. In contrast, the sum for $\tilde{\eta}^1_{\mathcal{A}}$ includes all six modified entries. Both sums remain unchanged, and so Eq. (43) holds.

The proof for the other cases, where $\mathcal{A} \cap \{j, k, l\} \neq \emptyset$ follows similar arguments.

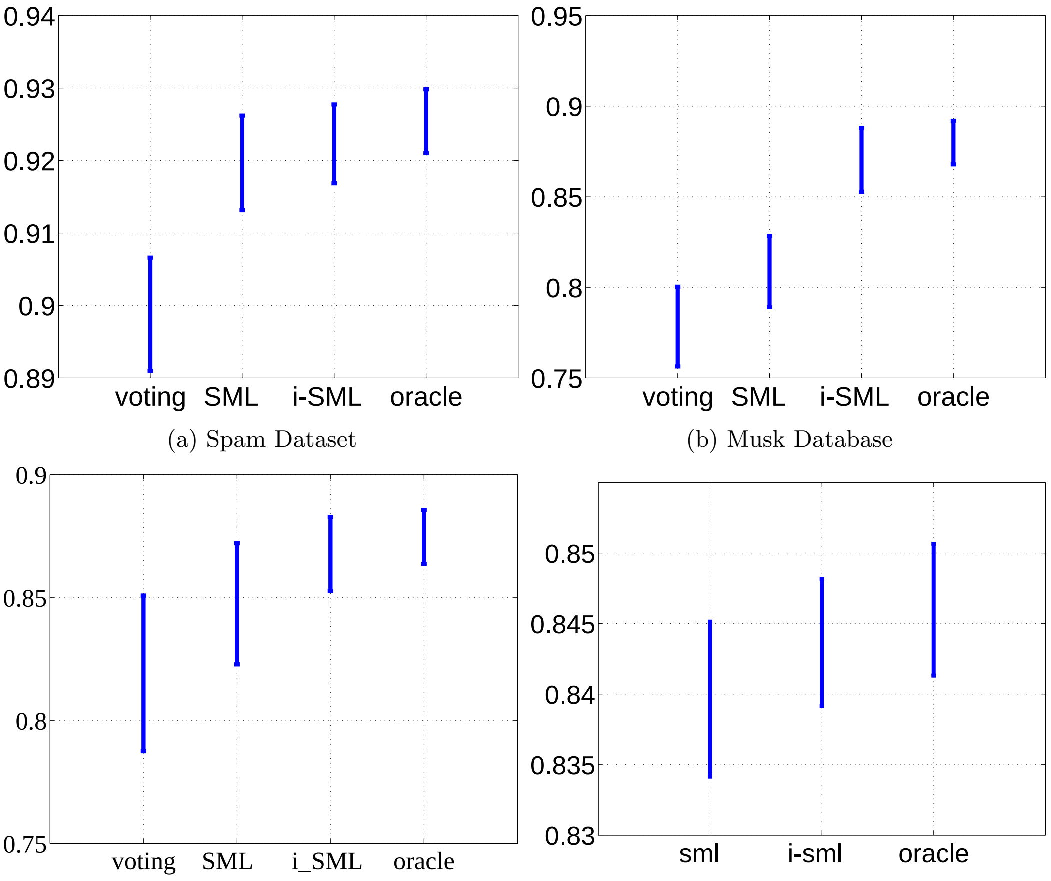
To conclude, both $\{\psi^i\}_{i=1}^m$ and $\{\tilde{\psi}^i\}_{i=1}^m$ have the same values $\mu^i_{\mathcal{A}}$ and covariance matrices $R_{\mathcal{A}}$. □

F Ensemble of Machine Learning Classifiers

Table 1 presents the 10 different classifiers used in our experiments. For each dataset, each classifier was trained with 200 different (randomly chosen) instances.

G Real Datasets

We tested our methods on a total of five datasets, 4 from the UCI repository and the MNIST digits data. A short description of each of the datasets is given in Table 2. A comparison of the performance of various ensemble learners on these datasets appears in Fig. 5.



(a) Spam Dataset

(b) Musk Database

(c) Miniboo Dataset

(d) 'MNIST' Dataset.

Fig. 5: The balanced accuracies of 4 unsupervised ensemble learning algorithms, all with $m = 10$ classifiers. In panel 5d we do not show the accuracy of majority voting which was significantly lower than all others.

classifier	**Weka library**
IBk - K nearest neighbours, $K = 1$	lazy.IBk
KStar - Instance based classifier	lazy.KStar
J48 - Decision tree	trees.J48
PART - Partial decision trees classifier	rules.PART
LMT - Logistic model trees	trees.LMT
Random forest - with $n = 10$ trees	trees.RandomForest
Logistic Regression	functions.SimpleLogistic
Decision Stump - One level decision tree	trees.DecisionStump
Sequential Minimal Optimization	functions.SMO
NaiveBayes	bayes.NaiveBayes

Table 1: 10 classification methods implemented in the software package Weka.

dataset	**Task**	**instances**	**attributes**
Magic	classifying gamma rays from background noise	19000	11
Spam	classifying spam from regular mail	4600	57
Musk	classifying different types of molecules to be ’musk’ or ’non musk’	6600	88
Miniboo	distinguish electron neutrinos (signal) from muon neutrinos (background)’	130000	50
Mnist	To define a binary problem, we divided the MNIST data set into two classes as follows: $0-4$ vs. $5-9$	40000	28^2

Table 2: Properties of datasets from the UCI repository