Classifier Design Given an Uncertainty Class of Feature Distributions via Regularized Maximum Likelihood and the Incorporation of Biological Pathway Knowledge in Steady-State Phenotype Classification

Mohammad Shahrokh Esfahani[a,*], Jason Knight[a], Amin Zollanvari[a], Byung-Jun Yoon[a], Edward R. Dougherty[a,b]

[a] *Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX USA*
[b] *The Computational Biology Division of the Translational Genomics Research Institute, Phoenix, AZ USA*

Abstract

Contemporary high-throughput technologies provide measurements of very large numbers of variables but often with very small sample sizes. This paper proposes an optimization-based paradigm for utilizing prior knowledge to design better performing classifiers when sample sizes are limited. We derive approximate expressions for the first and second moments of the true error rate of the proposed classifier under the assumption of two widely-used models for the uncertainty classes; ε-contamination and p-point classes. The applicability of the approximate expressions is discussed by defining the problem of finding optimal regularization parameters through minimizing the expected true error. Simulation results using the Zipf model show that the proposed paradigm yields improved classifiers that outperform traditional classifiers that use only training data. Our application of interest involves discrete gene regulatory networks possessing labeled steady-state distributions. Given prior operational knowledge of the process, our goal is to build a classifier that can accurately label future observations obtained in the steady state by utilizing both the available prior knowledge and the training data. We examine the proposed paradigm on networks containing NF-κB pathways, where it shows significant improvement in classifier performance over the classical data-only approach to classifier design. Companion website: `http://gsp.tamu.edu/Publications/supplementary/shahrokh12a`.

Keywords: Steady-state classifier, biological-pathway knowledge, uncertainty class, regularized

*Corresponding author
Email addresses: `m.shahrokh@tamu.edu` (Mohammad Shahrokh Esfahani), `jknight@tamu.edu` (Jason Knight), `amin_zoll@neo.tamu.edu` (Amin Zollanvari), `bjyoon@ece.tamu.edu` (Byung-Jun Yoon), `edward@ece.tamu.edu` (Edward R. Dougherty)

maximum-likelihood, prior knowledge

1. Introduction

In recent years, phenotypic classification based on genomic data has confronted the pattern recognition community with very small samples. There can be tens of thousands of potential features (gene expressions) but the sample sizes tend to be small, typically under 100 and often less than 50. This makes classification problematic. A promising approach to alleviate the problem is the use of prior knowledge. For example, the usual procedure for classifier design is to apply a classification rule to a set of features and sample data with the result being a designed classifier that will be applied to the population (all future observations). Prior knowledge can play a role in deciding upon the nature of the data and the original list of features. Knowledge may also be used in choosing a classification rule based on the nature of physical characteristics. The salient point from our perspective herein is that, once the features, sampling procedure, and classification rule are decided upon, from that point on the typical classification rule proceeds without operational knowledge concerning the features. In particular, no assumptions are made regarding the feature-label distribution (population) from which the sample data have been drawn. It is in this regard that the classification procedure is said to be "model-free." If knowledge concerning the feature-label distribution is available, then it can be used in classifier design.

A good bit of attention has been paid to the difficulty of error estimation in such circumstances. This has led to the incorporation of prior knowledge in error estimation, for instance, sample-size requirements based on an uncertainty class of feature-label distributions [1] and minimum-mean-square-error (MMSE) error estimation based on a prior distribution over an uncertainty class of feature-label distributions [2, 3].

Here the issue is incorporation of prior knowledge into the design of the classifier itself, not the estimation of its error. A number of recent studies have proposed various methods that can enhance classifiers by incorporating prior knowledge. For example, to improve classification performance, several studies have proposed to interpret the gene expression data at the level of functional modules (i.e., pathways), instead of at the level of individual genes, by utilizing available pathway knowledge [4, 5]. These pathway-based methods try to infer the activity level of a given pathway by analyzing the expression of its member genes, which is then used as a potential feature. These studies have

shown that such "pathway markers" are generally more reproducible compared to "gene markers" and that they lead to better classification performance. Another example is the network-based classification approach [6, 7], which has been gaining interest in recent years. These network-based methods try to identify "subnetwork markers" by overlaying the gene expression data on a large-scale PPI (protein-protein interaction) network, where each gene is mapped to the corresponding protein, and searching for differentially expressed subnetwork regions. It has been shown that these subnetwork markers often yield more accurate classification results and have better reproducibility compared to both gene and pathway markers. Considering that pathways are partial representations of the gene regulatory network and that the PPI network provides a skeleton of the biological network underlying cells, the aforementioned methods can be viewed as attempts to construct better classifiers by integrating partial network knowledge with measurement data. A Bayesian approach to using prior knowledge for classification has been taken by defining a prior distribution on an uncertainty class of feature-label distributions and deriving a classifier that is optimal with respect to the posterior distribution resulting from utilizing sample data in conjunction with the prior distribution [8, 9]. This approach has been applied to classify the mammalian cell cycle as normal or mutated [10].

Although recent advances in pathway-based and network-based classification have demonstrated the potential for utilizing prior knowledge to improve genomic classification, currently available methods mostly rely on heuristics. In this paper, we propose a general paradigm for classification that incorporates prior knowledge along with the data in the context of an optimization procedure. We address optimal discrete classification where prior knowledge is restricted to an uncertainty class of feature distributions absent a prior distribution on the uncertainty class, a problem that arises directly for biological classification using pathway information.

In our case, the application in mind is phenotype classification based on gene (or protein) expression measurements in the steady-state of a biological network. This "biomarker problem" is perhaps the most active area of research in genomics owing to the potential for disease diagnosis and prognosis. Rather than depend only on expression data, one can use classical genetic pathway information to provide prior knowledge and augment classifier design. The example laid out in this paper involves the following chain: {pathways} $\rightarrow$ {class of networks} $\rightarrow$ {class of steady-state distributions}. Prior knowledge in the form of a set of pathways constrains the possible behaviors of

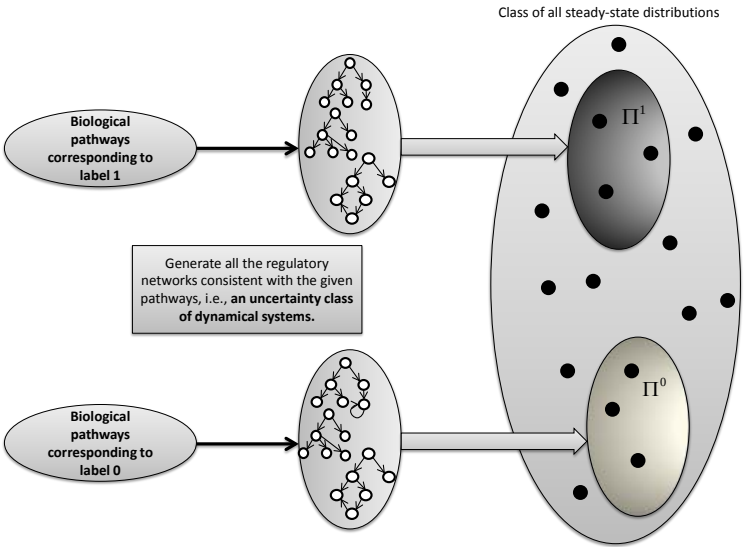


Figure 1: An illustrative example of the chain: {pathways} $\rightarrow$ {class of networks} $\rightarrow$ {class of steady-state distributions}. In this schematic view, an intermediate step is applied to construct a class of dynamical systems whose behaviors are consistent with the given pathways, for example, see the methods in [11] and [12]. Two uncertainty classes are shown by Π^0 and Π^1 for labels zero and one, respectively. These classes will be employed as the prior knowledge in the classifier design.

the dynamical system to an "uncertainty class" of networks consistent with the pathway information [11]. Each of these possesses a steady-state distribution, thereby yielding an uncertainty class of steady-state distributions. Figure 1 shows an illustrative view of this process chain. Detailed description of this figure is given in Section 6. Hence, rather than assume nothing is known about the feature-label distribution than what can be extracted from the data during classifier design, we can impose the constraint that the feature distribution belongs to the uncertainty class of steady-state distributions shown by a box in the middle of Figure 1. Put simply, a classifier is designed based on the uncertainty class of steady-state distributions, denoted by Π^0 and Π^1 in Figure 1, and the steady-state data.

We emphasize that while the particular application motivating our interest involves the generation of a steady-state uncertainty class from genetic pathway information, the theoretical content of this paper lies solely within classification theory – classifier design assuming an uncertainty class of feature distributions. In line with that focus, we provide analytic characterization of the first and second moments of the true error for two well-known uncertainty models, ε-contamination

and p-point uncertainty classes, under the assumption of stratified sampling. Characterization of these moments is basic to understanding the behavior of a classification rule and has a long history in pattern recognition, most commonly with stratified sampling [1],[13]-[29]. Recently, the issue of true-error moments has been addressed in the context of the joint distribution of the true and estimated error moments, in this case the most important moment being the second-order mixed moment between the true and estimated errors because this mixed moment is critical to characterizing the accuracy of the error estimate [1, 13, 18, 30].

The paper is organized in the following manner. In Section 2, we introduce our proposed paradigm. True error statistics for the stratified sampling case are derived in Section 3. Section 4 contains a brief discussion on the regularization parameter defined and used throughout the paper. Simulation results are shown in Sections 5 and 6 where we show the improvement of the designed classifier over the histogram rule in synthetic and biologically inspired cases, respectively. Finally, Section 7 contains concluding remarks.

We use the following notation and abbreviations. Boldface lower case letters denote column vectors. The cardinality of the set, Π is denoted by $|\Pi|$. $\pi(k)$ and $\boldsymbol{\pi}^T$ denote the $k-$th element and the transpose of the vector $\boldsymbol{\pi}$, respectively . $\Pr(A)$ denotes the probability of event A. The binomial distribution is shown by $bin(n,p)$. $bin(n,p) = x$ is used to denote the binomial random variable having value x. The trinomial distribution is shown by $trin(n,p_1,p_2)$. Thus,

$$\Pr(trin(n,p_1,p_2) = (x_1,x_2)) = \binom{n}{p_1, p_2, 1-p_1-p_2} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2}.$$

To show the comparison between two vectors, we use $\boldsymbol{\pi_1} \preceq \boldsymbol{\pi_2}$ meaning that the vector $\boldsymbol{\pi_1}$ is element-wise less than or equal to $\boldsymbol{\pi_2}$. The notation $\mathrm{E}_x(g(x))$ is used to denote taking expectation of $g(x)$ with respect to the subscript x. The indicator function for the event A is shown by I_A.

2. Regularized maximum-likelihood

In this section, we propose an optimization paradigm for classifier design that utilizes both an uncertainty class (from prior knowledge) and the available training data. Let $\pi_{ac}^y(k) = \Pr(X = k|Y = y)$ be the true conditional distribution of the feature $X = k \in \{1, ..., b\}$ given the class label $y \in \{0, 1\}$, and let $c_y = \Pr(Y = y)$ be the prior distribution of the class label. We can build a classifier by first finding label conditional probabilities $\hat{\pi}^y(k)$ that estimate the true probabilities

$\pi^y_{ac}(k)$ and then defining

$$\psi(k) = \mathrm{I}_{\{c_1\hat{\pi}^1(k) \geq c_0\hat{\pi}^0(k)\}} = \begin{cases} 1, & \text{if } c_1\hat{\pi}^1(k) \geq c_0\hat{\pi}^0(k) \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

This can be viewed as using the "plug-in rule" in the Bayes classifier $\psi(k) = \mathrm{I}_{\{c_1\pi^1_{ac}(k) \geq c_0\pi^0_{ac}(k)\}}$. In the absence of prior knowledge, the label-conditional distribution $\Pr(X = k|Y = y) = \pi^y_{ac}(k)$ is estimated solely based on the training data by solving the following maximum log-likelihood problem:

$$\min_{\boldsymbol{\pi}^{y\,T}\mathbf{e}=1, \mathbf{0} \preceq \boldsymbol{\pi}^y} \quad -\sum_{k=1}^{b} u^y_k \log \pi^y(k), \tag{2}$$

where u^y_k is the number of sample points at state k with label y and $\mathbf{e}$ is the all-one column vector. The solution to (2) is

$$\hat{\pi}^y_{\mathbf{data}}(k) = \frac{u^y_k}{n_y}, \tag{3}$$

where n_y is the number of sample points with label y.

We now assume we have *uncertainty classes*, $\Pi^y = \{\boldsymbol{\pi}^y_1, \boldsymbol{\pi}^y_2, ..., \boldsymbol{\pi}^y_{|\Pi^y|}\}, y = 0, 1$, e.g., see Figure 1, conveying the prior network knowledge of the label-y conditional distribution, $\boldsymbol{\pi}^y_{ac}$. We adapt (2) to form the following weighted-sum optimization problem for the class labels $y = 0, 1$, which includes a term contributed by the uncertainty class:

$$\min_{\boldsymbol{\pi}^{y\,T}\mathbf{e}=1, \mathbf{0} \preceq \boldsymbol{\pi}^y} \quad -(1-\lambda_y)\sum_{k=1}^{b} u^y_k \log \pi^y(k) + \lambda_y \ell(\boldsymbol{\pi}^y, \Pi^y) \ . \tag{4}$$

The *regularization parameter* $\lambda_y \in [0, 1]$ reflects the uncertainty of the labeled training data compared to the total amount of uncertainty in our prior knowledge and $\ell : \mathcal{S}_b \times \mathcal{S}_b^{|\Pi^y|} \to [0, \infty)$, where $\mathcal{S}_b$ is the *standard unit* $(b-1)-$*simplex* and $\mathcal{S}_b^{|\Pi^y|}$ is any uncertainty class containing $|\Pi^y|$ $b-$dimensional distributions, is a nonnegative function to measure the *dissimilarity* between a given $\boldsymbol{\pi}^y$ and the uncertainty class.

If the objective function in (4) is a convex function, then the optimization problem can be solved efficiently. Since the log-likelihood of the multinomial distribution is concave (i.e., the negative log-likelihood function for $\pi^y(k), k = 1, ..., b$, given the sample, is convex), it is sufficient to use a convex function for ℓ (i.e., the regularizer term) in (4) to make it a convex programming problem. We use

$$\ell(\boldsymbol{\pi}^y, \Pi^y) := \frac{1}{|\Pi^y|}\sum_{i=1}^{|\Pi^y|} D(\boldsymbol{\pi}^y_i || \boldsymbol{\pi}^y), \tag{5}$$

where $D(\boldsymbol{\pi}_i^y||\boldsymbol{\pi}^y) = \sum_{k=1}^{b} \pi_i^y(k) \log \frac{\pi_i^y(k)}{\pi^y(k)}$ is the *Kullback Leibler (information) distance* (KL-distance).

Lemma 1 (RML Classifier)**.** *Suppose that the dissimilarity function ℓ is defined as* (5)*. Then, the solution to the regularized maximum-likelihood (RML) problem in* (4) *is obtained bin-wise as*

$$\hat{\pi}^y_{\mathbf{RML}}(k) = \frac{(1-\lambda_y)u_k^y + \lambda_y \overline{\pi}^y(k)}{(1-\lambda_y)n_y + \lambda_y}; y \in \{0,1\}, \forall k = 1, ..., b, \tag{6}$$

where $\overline{\pi}^y(k)$ is the probability of the $k-$th bin obtained from the average of $\boldsymbol{\pi}_i^y, i = 1, 2, ..., |\Pi^y|$ in the corresponding uncertainty class $\Pi^y, y \in \{0,1\}$. The corresponding RML classifier can be found by plugging $\hat{\boldsymbol{\pi}}^0_{\mathbf{RML}}$ and $\hat{\boldsymbol{\pi}}^1_{\mathbf{RML}}$ in equation (1)*.*

Proof. Please refer to Appendix A. □

Consider the following two special cases:

1. Suppose the uncertainty in the information extracted from the training data is much less than that in the prior network knowledge. In the limiting case, $\lambda_y \to 0$ and

$$\lim_{\lambda_y \to 0} \hat{\pi}^y_{\mathbf{RML}}(k) = \frac{u_k^y}{n_y}, \forall k = 1, ..., b. \tag{7}$$

   This is consistent with our expectation: if there is infinite amount of training data (hence no uncertainty therein), the classifier can be perfectly estimated from the data.

2. Suppose we have very good prior network knowledge, so that the uncertainty in this knowledge is much smaller compared to that extracted from the data. In the limiting case, $\lambda_y \to 1$ and

$$\lim_{\lambda_y \to 1} \hat{\pi}^y_{\mathbf{RML}}(k) = \overline{\pi}^y(k), \forall k = 1, ..., b. \tag{8}$$

   If we have perfect knowledge of the steady-state distribution, then we do not need training data.

In this paper we consider two models having finite uncertainty classes:

2.0.1. ε-contamination uncertainty class

The ε-contamination class has been used for modeling uncertainty in a wide range of applications, including robust hypothesis testing [31], robust Wiener filtering (uncertainty about the spectral density) [32, 33], Bayesian robust optimal linear filter design [34], robust decision making

problems [35], and minimax robust quickest change detection (with the application in intrusion detection in computer networks and security systems) [36]. In [32]-[34], the ε-contamination class contains all the power spectral densities (PSD) in the vicinity of the nominal PSD. In [31] and [36], the ε-contamination contains all the probability densities in the vicinity of the nominal one.

Here, we use ε-contamination to model the uncertainty about the label-conditional probabilities. We define the ε-contamination class of multinomial distributions associated with each label as the class containing the distributions in the form of

$$\boldsymbol{\pi}^y = (1-\varepsilon_y)\boldsymbol{\pi}^y_{ac} + \varepsilon_y \boldsymbol{\pi}; y \in \{0,1\} \tag{9}$$

where $\varepsilon_y \in [0,1)$ is the degree of contamination and $\boldsymbol{\pi}$ is one of a finite number of randomly chosen densities from $\mathcal{S}_b$. Increasing ε_y corresponds to increasing the variance of prior knowledge about the true distribution. We assume a uniform distribution for the contamination part whose domain is the relative interior of the volume under the $(b-1)$-simplex. Since our application of interest is related to steady-state classifiers, we assume that in the simplex the corners and axes have measure zero.

2.0.2. p-point uncertainty class

The p-point uncertainty class has been used to model uncertainty in rate distortion problems, detection problems, robust Wiener filter design, and robust non-stationary signal estimation [33],[37]-[42]. In our application of interest, we often only know that the cell, in its steady state, spends a specific portion of time in a subset of states but know nothing about the details of these states individually. Hence, to model this prior knowledge, we can see the problem as a partitioning scenario: if we partition the state space, then the amount of time that the cell spends in each subset in the partition is known. Therefore, we can say that the label-conditional distributions belong to an uncertainty class of distributions satisfying the following constraints:

$$\sum_{k=1}^{b} \pi(k) \mathrm{I}_{\{k \in s^y_p\}} = \sum_{k=1}^{b} \pi^y_{ac}(k) \mathrm{I}_{\{k \in s^y_p\}}; p = 1, ..., m_y, \tag{10}$$

where $\boldsymbol{\pi}^y_{ac}$ is the actual steady-state distribution , $s^y_1, ..., s^y_{m_y}$ form a partition of the state space denoted by $\mathcal{P}^y$, and $\boldsymbol{\pi} \in \mathcal{S}_b$ is any density function.

We will use the following notation throughout the paper for the probability mass cumulated in

each partition:

$$\sum_{k=1}^{b} \pi_{ac}^{y}(k) \mathrm{I}_{\{k \in s_p^y\}} = \omega_p^y; p = 1, ..., m_y. \tag{11}$$

Moreover, we define the following mapping from state space to the partition:

$$P^y : \{1, ..., b\} \to \{1, ..., m_y\}; y = 0, 1. \tag{12}$$

In the extreme case, $m_y = 1$ means that we only know that the label-conditional probabilities for the bins sum up to 1, which corresponds to a minimal amount of prior knowledge. On the other hand, $m_y = b$, i.e. $|s_p^y| = 1$, for any $p \in \{1, ..., m_y\}, y \in \{0, 1\}$, means that we are certain about the label-conditional distributions, because we are given all bin probabilities – hence minimal variance in the uncertainty class (for more details refer to Section 1 of the supplementary materials on the companion website).

3. Moments for the true error

For a classifier ψ_n trained on the sample data S_n, the probability of error is defined as $\epsilon_{\mathbf{data}} = \Pr(\psi_n(X) \neq Y | S_n)$. The overall performance of the classification rule can be evaluated by the expected classification error, $\mathrm{E}(\epsilon_{\mathbf{data}}) = \mathrm{E}_{S_n}\left[\Pr(\psi_n(X) \neq Y | S_n)\right]$, over all samples of size n. When prior knowledge (denoted by "**uc**" for uncertainty class) is incorporated into classifier design, we rewrite the probability of error as

$$\epsilon_{\mathbf{data,uc}} = \Pr(\psi_{n,\Pi^0,\Pi^1}(X) \neq Y | S_n, \Pi^0, \Pi^1).$$

In this section we provide analytic representation of the first and second moments for the error in the ε-contamination and p-point uncertainty models under stratified sampling, in which sampling is performed from classes 0 and 1 in accordance with their prior probabilities. Since we incorporate prior knowledge, the moments are computed relative to all samples of size n and the uncertainty-class space. They take the form

$$\mathrm{E}(\epsilon_{\mathbf{RML}}) = \mathrm{E}(\epsilon_{\mathbf{data,uc}}) = \mathrm{E}_{\Pi^0,\Pi^1}\left[\mathrm{E}_{S_n}\left[\Pr(\psi_{n,\Pi^0,\Pi^1}(X) \neq Y | S_n)\right] | \Pi^0, \Pi^1\right], \tag{13}$$

$$\mathrm{E}(\epsilon_{\mathbf{RML}}^2) = \mathrm{E}(\epsilon_{\mathbf{data,uc}}^2) = \mathrm{E}_{\Pi^0,\Pi^1}\left[\mathrm{E}_{S_n}\left[\Pr(\psi_{n,\Pi^0,\Pi^1}(X) \neq Y | S_n)\right]^2 | \Pi^0, \Pi^1\right]. \tag{14}$$

We derive tight approximations for these moments for $\lambda_y \in (0, 1)$. The cases $\lambda_y \in \{0, 1\}$ can be handled with a slight modification to the proof.

Theorem 1 (First-Order Moment of the True Error: $\varepsilon-$Contamination Class)**.** *Suppose that the uncertainty classes,* Π^0 *and* Π^1*, come from* ε_0- *and* ε_1- *contamination classes, respectively. Then, the first-order moment of the true-error of the RML classifier defined in Lemma 1 is given by*

$$\begin{aligned} E(\epsilon_{\mathbf{RML}}) &= c_0 \sum_{k=1}^{b} \pi_{ac}^0(k) \left[\sum_{l_0=0}^{n_0} \sum_{j=0}^{n_1} \sum_{m=j}^{n_1} \Pr(bin(n_0, \pi_{ac}^0(k)) = l_0) \Pr(\underline{\zeta}_{k,l_0}^0 = j) \Pr(bin(n_1, \pi_{ac}^1(k)) = m) \right] \\ &+ c_1 \sum_{k=1}^{b} \pi_{ac}^1(k) \left[\sum_{l_1=0}^{n_1} \sum_{j=0}^{n_0} \sum_{m=j}^{n_0} \Pr(bin(n_1, \pi_{ac}^1(k)) = l_1) \Pr(\underline{\zeta}_{k,l_1}^1 = j) \Pr(bin(n_0, \pi_{ac}^0(k)) = m) \right]. \end{aligned} \tag{15}$$

where the random variables $\underline{\zeta}_{k,l_y}^y$*,* $k = 1, ..., b$*,* $\forall l_y = 0, .., n_y; y \in \{0, 1\}$*, approximately have the following probability mass function (pmf):*

$$\begin{cases} \Pr(\underline{\zeta}_{k,l_y}^y = 0) = \Phi(\frac{-\mu_{k,l_y}^y}{\sigma_{k,y}}) \\ \Pr(\underline{\zeta}_{k,l_y}^y = m) = \Phi(\frac{m-\mu_{k,l_y}^y}{\sigma_{k,y}}) - \Phi(\frac{m-1-\mu_{k,l_y}^y}{\sigma_{k,y}}); m = 1, .., n_y \\ \Pr(\underline{\zeta}_{k,l_y}^y = m) = 0; m \geq n_y + 1 \end{cases}, \tag{16}$$

$\Phi(.)$ *being the standard normal distribution. In equation* (16) *we have*

$$\begin{aligned} \mu_{k,l_y}^y &= \frac{g_y l_y + (1-\varepsilon_y)\alpha_y \pi_{ac}^y(k) - (1-\varepsilon_{\overline{y}})\alpha_{\overline{y}} \pi_{ac}^{\overline{y}}(k) + \frac{\varepsilon_y \alpha_y - \varepsilon_{\overline{y}} \alpha_{\overline{y}}}{b}}{g_{\overline{y}}} \\ \sigma_{k,y}^2 &= (\frac{\alpha_y^2 \varepsilon_y^2 (b-1)}{b^2 |\Pi^y| (b+1)} + \frac{\alpha_{\overline{y}}^2 \varepsilon_{\overline{y}}^2 (b-1)}{b^2 |\Pi^{\overline{y}}| (b+1)}) / g_{\overline{y}}^2; \forall k = 1, ..., b \end{aligned} \tag{17}$$

where $\overline{y}$ *denotes* $1 - y$ *and*

$$\begin{aligned} g_y &:= (1 - \lambda_y) n_y \left[n_{\overline{y}} (1 - \lambda_{\overline{y}}) + \lambda_{\overline{y}} \right] \\ \alpha_y &:= \frac{g_y \lambda_y}{1 - \lambda_y}. \end{aligned} \tag{18}$$

Proof. Please refer to Appendix B. □

Theorem 2 (First-Order Moment of the True Error: $p-$Point Class)**.** *Let the uncertainty classes,* Π^0 *and* Π^1*, be modeled by the* $p-$*point model with partition probabilities* ω_p^0 *and* ω_p^1 *with* $p = 1, ..., m_y$ *for labels* 0 *and* 1*, respectively. Then, the first-order moment of the true-error of the RML classifier defined in Lemma 1 can be written as in equation* (15) *in which the random variables* $\underline{\zeta}_{k,l_y}^y$*,* $k = 1, ..., b$*, for any* $l_y = 0, .., n_y$*, approximately have the pmf as defined in equation* (16)*,*

whereas assuming the definitions in equation (18)*, we have*

$$\mu^{y}_{k,l_y} = \frac{g_y l_y + \alpha_y \frac{\omega^{y}_{P^y(k)}}{|s^{y}_{P^y(k)}|} - \alpha_{\overline{y}} \frac{\omega^{\overline{y}}_{P^{\overline{y}}(k)}}{|s^{\overline{y}}_{P^{\overline{y}}(k)}|}}{g_{\overline{y}}}$$

$$\sigma^2_{k,y} = \Big[\alpha^2_y(\omega^{y}_{P^y(k)})^2 \frac{(|s^{y}_{P^y(k)}|-1)}{|s^{y}_{P^y(k)}|^2(|s^{y}_{P^y(k)}|+1)|\Pi^y|} + \alpha^2_{\overline{y}}(\omega^{\overline{y}}_{P^{\overline{y}}(k)})^2 \frac{(|s^{\overline{y}}_{P^{\overline{y}}(k)}|-1)}{|s^{\overline{y}}_{P^{\overline{y}}(k)}|^2(|s^{\overline{y}}_{P^{\overline{y}}(k)}|+1)|\Pi^{\overline{y}}|}\Big]/g^2_{\overline{y}}, \tag{19}$$

where the mapping $P^y(.)$ *is defined in equation* (12)*.*

Proof. Please refer to Appendix B. □

Theorem 3 (Second-Order Moment of the True Error)**.** *The second-order moment of the true-error of the RML classifier defined in Lemma 1 can be decomposed as*

$$\begin{aligned} E(\epsilon^2_{\mathbf{RML}}) &= E_{\Pi^0,\Pi^1}\left[c^2_0 \sum_{k=1}^{b} (\pi^0_{ac}(k))^2 A^1 + c^2_1 \sum_{k=1}^{b} (\pi^1_{ac}(k))^2 A^0\right] \\ &+ E_{\Pi^0,\Pi^1}\left[c^2_0 \sum_{k_1\neq k_2}^{b} \pi^0_{ac}(k_1)\pi^0_{ac}(k_2) B^1 + c^2_1 \sum_{k_1\neq k_2}^{b} \pi^1_{ac}(k_1)\pi^1_{ac}(k_2) B^0\right] \\ &+ E_{\Pi^0,\Pi^1}\left[c_0 c_1 \sum_{k_1\neq k_2}^{b} \pi^0_{ac}(k_1)\pi^1_{ac}(k_2) C^1 + c_0 c_1 \sum_{k_1\neq k_2}^{b} \pi^1_{ac}(k_1)\pi^0_{ac}(k_2) C^0\right]. \end{aligned} \tag{20}$$

where $A^0 := E_{S_n}[I_{\{\psi(X=k)=0\}}]$ *and* $A^1 := E_{S_n}[I_{\{\psi(X=k)=1\}}]$ *can be found similarly as in Theorem 1.* B^0*,* B^1*,* C^0*, and* C^1 *are computed as follows:*

$$\begin{aligned} B^0 &:= \sum_{t^1_1,t^1_2}\Big[\sum_{(t^0_1,t^0_2)\succeq(\underline{\zeta}^1_{k_1,t^1_1},\underline{\zeta}^1_{k_2,t^1_2})} \Pr(u^0_{k_1}=t^0_1, u^0_{k_2}=t^0_2)\Pr(u^1_{k_1}=t^1_1, u^1_{k_2}=t^1_2)\Big] \\ B^1 &:= \sum_{t^0_1,t^0_2}\Big[\sum_{(t^1_1,t^1_2)\succeq(\underline{\zeta}^0_{k_1,t^0_1},\underline{\zeta}^0_{k_2,t^0_2})} \Pr(u^1_{k_1}=t^1_1, u^1_{k_2}=t^1_2)\Pr(u^0_{k_1}=t^0_1, u^0_{k_2}=t^0_2)\Big] \\ C^0 &:= \sum_{t^1_1,t^1_2}\Big[\sum_{t^0_1\geq\underline{\zeta}^1_{k_1,t^1_1},t^0_2\leq,\overline{\zeta}^1_{k_2,t^1_2}} \Pr(u^0_{k_1}=t^0_1, u^0_{k_2}=t^0_2)\Pr(u^1_{k_1}=t^1_1, u^1_{k_2}=t^1_0)\Big] \\ C^1 &:= \sum_{t^0_1,t^0_2}\Big[\sum_{t^1_1\geq\underline{\zeta}^0_{k_1,t^0_1},t^1_2\leq\overline{\zeta}^0_{k_2,t^0_2}} \Pr(u^1_{k_1}=t^1_1, u^1_{k_2}=t^1_2)\Pr(u^0_{k_1}=t^0_1, u^0_{k_2}=t^0_1)\Big]. \end{aligned} \tag{21}$$

Proof. Please refer to Appendix C. □

The joint distribution of $\underline{\zeta}^0_{k_1,t^0_1}$ and $\underline{\zeta}^0_{k_2,t^0_2}$ (similarly for $\underline{\zeta}^1_{k_1,t^1_1}$ and $\underline{\zeta}^1_{k_2,t^1_2}$) and the joint distribution of $\underline{\zeta}^0_{k_1,t^0_1}$ and $\overline{\zeta}^0_{k_1,t^0_2}$ (similarly for $\underline{\zeta}^1_{k_1,t^1_1}$ and $\overline{\zeta}^1_{k_2,t^1_2}$), which depend on the uncertainty classes, are given in Appendix D for $\varepsilon-$contamination and $p-$point classes.

4. The regularization parameter

The regularization parameter λ_y in (4) should be adjusted based on the relative uncertainty between the training data and the prior knowledge. We propose three approaches for tuning the regularization parameter.

4.1. Minimizing the expected true error

The optimal value of the regularization parameter, based on expected true error, can be found by solving the following optimization problem:

$$\boldsymbol{\lambda}^* = \arg \min_{\mathbf{0} \preceq \boldsymbol{\lambda} \preceq \mathbf{1}} \mathrm{E}(\epsilon_{\mathbf{RML}}), \tag{22}$$

where $\boldsymbol{\lambda} = [\lambda_0, \lambda_1]$, $\mathbf{1} = [1, 1]$, $\mathbf{0} = [0, 0]$ and $\mathrm{E}(\epsilon_{\mathbf{RML}})$ is given in equation (15). In (15), the only parameters affected by $\boldsymbol{\lambda}$ are $\Pr(\underline{\zeta}^y_{k,l_y} = j), y \in \{0, 1\}$, approximated in Theorems 1 and 2. (22) is a constrained non-linear programming problem whose global minimum is not guaranteed to be found by classic gradient-based methods.

4.2. SURE-tuning of regularization parameter

One way to evaluate the performance of the estimator in Lemma 1 is to use the mean-squared error (MSE) of the estimator. In the problem of multinomial distribution estimation, the MSE can be expanded as follows

$$\mathrm{MSE}^y = \mathrm{E}\Big[\sum_{k=1}^{b} \Big[\hat{\pi}^y_{\lambda_y}(k) - \pi^y_{ac}(k) \Big]^2 \Big], y = 0, 1, \tag{23}$$

where we drop the subscript RML and instead use the regularization parameter λ_y to show that the estimate depends on λ_y. One strategy to find the regularization parameter is to minimize MSE^y in (23) [43, 44]; however, MSE^y depends on the parameter for estimating $\boldsymbol{\pi}^y_{ac}$. We use an approach called SURE (Stein's Unbiased Risk Estimator) [45], proposed for the i.i.d. Gaussian model. Here, an unbiased estimate of the MSE of the designed estimator is found and then one can do optimization to find the required parameters of the estimator. For the sake of simplicity, in the following lemma we omit the superscript y.

Lemma 2. *Let the uncertainty class,* Π*, be given and fixed. Denoting the RML estimator of* $\boldsymbol{\pi_{ac}}$ *in Lemma 1 using* λ *as the regularization parameter by* $\hat{\boldsymbol{\pi}}_\lambda$*, an unbiased estimate of the MSE of the estimate in Lemma 1 is given by*

$$\hat{MSE} = \sum_{k=1}^{b} \Big[\hat{\pi}_\lambda^2(k) + \pi_{ac}^2(k) - 2\{ \frac{\delta_\lambda}{n-1} u_k^2 - u_k (\frac{\delta_\lambda}{n-1} - \frac{\theta_\lambda(k)}{n}) \} \Big] \tag{24}$$

where $\delta_\lambda = \frac{1-\lambda}{(1-\lambda)n+\lambda}$ *and* $\theta_\lambda(k) = \frac{\lambda \overline{\pi}(k)}{(1-\lambda)n+\lambda}$.

Proof. Please refer to Appendix E. □

Minimizing the SURE-estimate of the MSE with respect to the regularization parameter λ yields the following result for case of $n \geq 2$.

Corollary 1 (SURE-Optimal Regularization Parameter)**.** *The SURE-optimal regularization parameter of the estimator defined in Lemma 1 is given by*

$$\lambda^*_{SURE} = \begin{cases} \tilde{\lambda} & 0 \leq \tilde{\lambda} \leq 1 \\ I_{\sum_{k=1}^{b} \left[\overline{\pi}(k)^2 - 2\frac{u_k \overline{\pi}(k)}{n} \right] < \frac{2}{n-1} - \frac{n+1}{n^2(n-1)} \sum_{k=1}^{b} u_k^2} & \textit{otherwise} \end{cases} \tag{25}$$

in which we have $\tilde{\lambda} = \dfrac{n \left[1 - \sum_{k=1}^{b} (u_k/n)^2 \right]}{(n-1) \left[1 + \sum_{k=1}^{b} \overline{\pi}(k) \left[\overline{\pi}(k) - 2u_k/n \right] \right]}$.

Proof. The corollary results from equating the derivative of (24) (with respect to λ) to zero, while considering the boundary of the feasible region of the λ (the SURE estimate in equation (24) is continuous in $[0, 1]$). □

Fixing the uncertainty class, as $n \to \infty$, we obtain

$$\lim_{n \to \infty} \lambda^*_{SURE} = \frac{1 - ||\boldsymbol{\pi}_{ac}||_2^2}{1 - ||\boldsymbol{\pi}_{ac}||_2^2 + ||\boldsymbol{\pi}_{ac} - \overline{\boldsymbol{\pi}}||_2^2}, \tag{26}$$

in which $||\mathbf{x}||_2$ denotes the ℓ_2-norm of vector $\mathbf{x}$.

To illustrate the effects of different sample sizes and different amounts of uncertainty on λ^*_{SURE}, we have run a simulation assuming an $\varepsilon-$contamination uncertainty class and that the actual distribution follows a Zipf model with parameter $a = 1$ (a detailed description of the Zipf model will be provided in Section 5). We observe the behavior of $\overline{\lambda}_{SURE} = \mathrm{E}_\Pi \Big[\mathrm{E}_{S_n} [\lambda^*_{SURE} | \Pi] \Big]$ using Monte-Carlo expectation over 4000 training data sets (for each fixed sample size) and 500 uncertainty

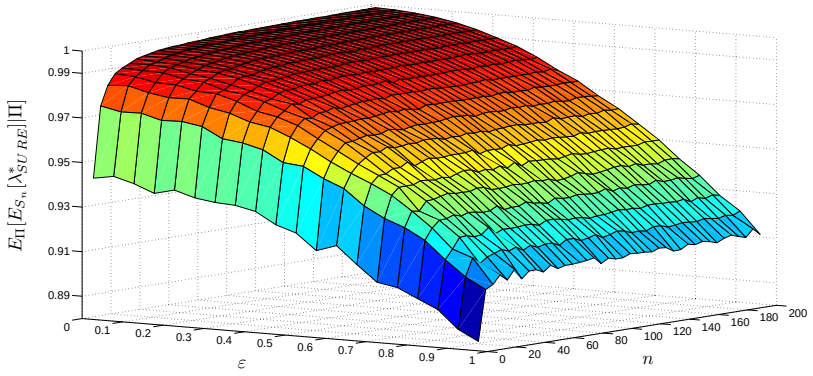


Figure 2: Illustrating the expected value of λ^*_{SURE} for different amount of uncertainty and sample sizes. The result is for $\varepsilon-$contamination classes. The uncertainty class size, $|\Pi|$ is set to 50.

classes. We consider different values for $\varepsilon \in [0, 1)$ and sample size n. Figure 2 shows the 3-D figure with n as the x-axis and ε as the y-axis. As $\varepsilon \to 1$ (uncertainty is increased), for a fixed sample size, $\overline{\lambda}_{SURE}$ decreases as in equation (26).

4.3. A heuristic approach

Although one can use a stochastic algorithm to solve (22) (which is not necessarily guaranteed to achieve the global minimum), or use the result in Corollary 1, we can take a heuristic approach for specifying λ_y. Suppose $|\Pi^0|$ and $|\Pi^1|$ are the sizes of the uncertainty classes for labels 0 and 1, respectively. Proceeing heuristically and denoting the ith distribution with label y as $\boldsymbol{\pi}^y_i$, we form a network-based estimate, $\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}} = \overline{\boldsymbol{\pi}}^y$, by averaging the $\boldsymbol{\pi}^y_i$, $i = 1, ..., |\Pi^y|$. A data-based estimate, $\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}$, is obtained from (3). Under this setting, we can estimate the relative uncertainty by

$$\lambda_y := \frac{\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}})}{\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}) + \mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}})}. \tag{27}$$

where

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}) &= \sum_{k=1}^{b} \mathrm{Var}(\hat{\pi}^y_{\mathbf{data}}(k)) \\ \mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}}) &= \sum_{k=1}^{b} \mathrm{Var}(\hat{\pi}^y_{\mathbf{uc}}(k)). \end{aligned} \tag{28}$$

In (28), the variance of the training data is independent of the uncertainty class model and can therefore be analytically computed by

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}) = \sum_{k=1}^{b} \frac{\pi^y_{ac}(k)(1 - \pi^y_{ac}(k))}{n_y}. \tag{29}$$

The variance of the uncertainty class depends on the underlying model of the uncertainty class. We obtain

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}}) = \frac{\varepsilon_y^2 b(b-1)}{(b+1)b^2}. \tag{30}$$

for a $\varepsilon-$contamination class and

$$\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}}) = \sum_{p=1}^{m_y} \frac{\omega_p^y |s_p^y|(|s_p^y|-1)}{(|s_p^y|+1)|s_p^y|^2}. \tag{31}$$

for a $p-$point uncertainty class (please refer to Section 1 of the supplementary materials on the companion website).

5. Numerical experiments

In this section, we evaluate the performance of the classifiers designed using the proposed optimization paradigm. Let $\epsilon_{\mathbf{RML}}$ denote the error of the RML classifier designed via (4) using the estimated probabilities given in Lemma 1. Let $\epsilon_{\mathbf{hist}}$ denote the error of the traditional histogram rule obtained by designing the classifier as in (1) using the data-based estimate $\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}$ given in (3). The exact expression for $\mathrm{E}(\epsilon_{\mathbf{hist}})$ is given in [18].

We use both the approximation in (15) as well as Monte Carlo simulations for assessing $\mathrm{E}(\epsilon_{\mathbf{RML}})$. In the Monte-Carlo estimation, based on the given assumption for the structure of the uncertainty classes, we generate T pairs of uncertainty classes denoted by $(\Pi_l^0, \Pi_l^1), l = 1, ..., T$. Then for each pair, based on the given model for the true distributions $\boldsymbol{\pi}^y_{ac}, y = 0, 1$, we generate M sample sets with size n denoted by $S_n^{l,m}, m = 1, ..., M$. For each sample $S_n^{l,m}$, we estimate the conditional probabilities using Lemma 1. The estimates $\hat{\pi}^y_{\mathbf{RML}}(k)$ are then used to construct the classifier, as defined in (1). The error of the classifier designed using $S_n^{l,m}$ (i.e., mth sample set generated for the lth pair) is then computed analytically using the actual distribution $\boldsymbol{\pi}^y_{ac}$ which was used to generate the sample. We denote this error by $\epsilon^{l,m}_{\mathbf{RML}}$. The first- and the second-order moments of the true error are approximated by

$$\mathrm{E}(\epsilon_{\mathbf{RML}}) \approx \frac{1}{MT} \sum_{l=1}^{T} \sum_{m=1}^{M} \epsilon^{l,m}_{\mathbf{RML}}, \tag{32}$$

$$\mathrm{E}(\epsilon^2_{\mathbf{RML}}) \approx \frac{1}{MT} \sum_{l=1}^{T} \sum_{m=1}^{M} (\epsilon^{l,m}_{\mathbf{RML}})^2 \tag{33}$$

via Monte Carlo simulation. We estimate the variances, $\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}})$ and $\mathrm{Var}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}})$ in (27) as

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\pi}}^y_{\mathbf{data}}) = \sum_{k=1}^{b} \frac{\frac{u_k^y}{n_y}(1-\frac{u_k^y}{n_y})}{n_y}, \tag{34}$$

$$\hat{\mathrm{Var}}(\hat{\boldsymbol{\pi}}^y_{\mathbf{uc}}) = \frac{1}{|\Pi^y|-1} \sum_{k=1}^{b} \sum_{i=1}^{|\Pi^y|} (\pi_i^y(k) - \hat{\pi}^y_{\mathbf{uc}}(k))^2. \tag{35}$$

5.1. Performance assessment using a Zipf model

We first assume that the true label-conditional distributions (i.e., $\boldsymbol{\pi}^y_{ac}, y = 0, 1$) follow a Zipf model,

$$\pi^0_{ac}(k) = \frac{\xi}{k^a}, \quad \pi^1_{ac}(k) = \pi^0_{ac}(b-k+1), \tag{36}$$

where ξ is a normalizing constant. The Zipf distribution, originally introduced by G.K. Zipf to model the frequency of words in common text [46], is a well-known power-law discrete distribution, encountered in many applications. In particular, it has been used as a model to study the moments of error estimators for discrete classifiers [18]. As $a \to 0$, both conditional distributions ($y \in \{0, 1\}$) tend to become uniform. Hence the classification problem becomes more difficult, resulting in a larger Bayes error. We assume $c_y = 0.5; y \in \{0, 1\}$ are known. We consider $b = 8$ (which corresponds to the number of states in a three-gene Boolean network when modeling genomic regulatory networks [47]). We evaluate the proposed framework under two different scenarios. First, we examine the accuracy of our approximate expressions by comparing them with the Monte-Carlo simulation while one has access to the exact regularization parameters defined by applying (29)-(31). The motivation is to test the accuracy of our approximation when the regularization parameters are found off-line, independent of the given sample data. In the second scenario, we assume one has to estimate the regularization parameters based on the given data and uncertainty classes using equations (34)-(35). Depending on the underlying assumption for the uncertainty classes, for each size n and each set of model parameters (e.g., ε_0, ε_1, or partitions in the $p-$point class), we generate $T = 1000$ different pairs of uncertainty classes, $(\Pi^0_l, \Pi^1_l), l = 1, ..., 1000$, for which we generate $M = 2,000$ samples, $S_n^{l,m}, l = 1, ..., 1000; m = 1, ..., 2000$, for estimating the first- and the second-order moments of the true error, $\mathrm{E}(\epsilon_{\mathbf{RLM}})$ and $\mathrm{E}(\epsilon^2_{\mathbf{RLM}})$. For the approximate second-order moments, where there are double-integrals, we use the adaptive Simpson algorithm for

approximating the integral values. Some results for the various experiments are shown in Figure 3 for, ε-contamination, $b = 8$, and uncertainty class size $|\Pi^y| = 250$ for $y = 0, 1$ (more results, including those for p-point uncertainty, are shown in Section 2 of the supplementary materials on the companion website). In the figure, the Bayes error corresponding to the optimal classifier is denoted as ϵ_{Bayes}.

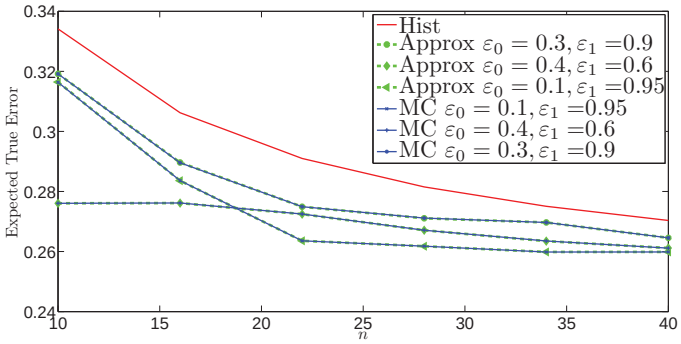
We use the algorithm proposed in [48] for generating the contaminating distribution generated uniformly under a unit-simplex. Figure 3(a) shows the results for the first scenario. Three cases are considered for the pair: $(\varepsilon_0, \varepsilon_1)$: $(0.3, 0.9)$, $(0.4, 0.6)$, and $(0.1, 0.95)$. The expected true error of the proposed scheme is smaller than that of the histogram rule in all cases. Moreover, the results from the Monte-Carlo simulations are very close to those obtained from our approximations in (15), shown by "Approx" in the legends of plots. The expected true error for the case $(0.4, 0.6)$ is significantly smaller than the others for small sample sizes. This is due to the reliable prior knowledge compared to other cases, for small samples. However, when the sample size increases, $(0.1, 0.95)$ outperforms $(0.4, 0.6)$. Owing to a small contamination degree ε_0 in $(0.1, 0.95)$, the proposed RML framework provides a good estimate of $\boldsymbol{\pi}^0_{ac}$ for any sample size. Furthermore, by increasing the sample size, we achieve a better estimate of $\boldsymbol{\pi}^1_{ac}$, making the designed classifier perform close to the optimal classifier. Therefore, it outperforms $(0.4, 0.6)$, which has less accurate estimates of the two conditional distributions for these sample sizes.

Figure 3(b) shows the results for the data-dependent regularization parameter, where one can see that our approximation and the Monte-Carlo simulations are slightly different. This happens only for small sample sizes, owing to having a poor estimate of $\lambda_y; y = 0, 1$, defined in (29)-(30). Figures 3(c) and 3(d) correspond to Figures 3(a) and 3(b) for the second-order moment.

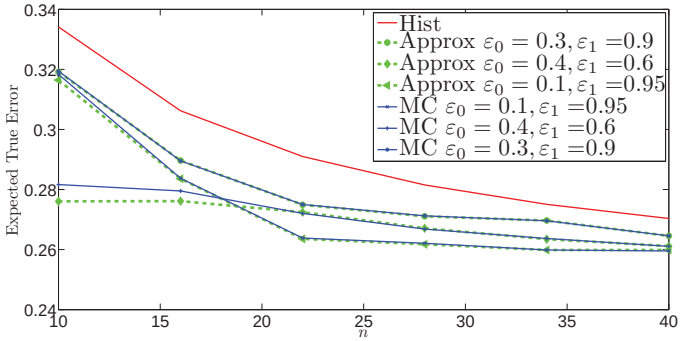
6. Performance assessment using networks containing NF-κB pathways

While the theoretical development of the paper pertains to uncertainty classes of distributions for classification, as stated at the outset, our original motivation for the theory comes from our desire to apply prior pathway knowledge in biological network steady-state classification.

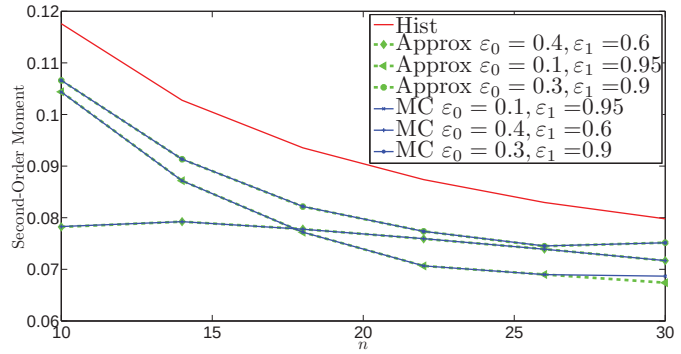
In this section, we use prior pathway knowledge and an associated cellular context in order to improve the performance of a classifier which discriminates between biologically relevant states of a biological system. More specifically, a biological system can be modeled by a discrete, dynamical



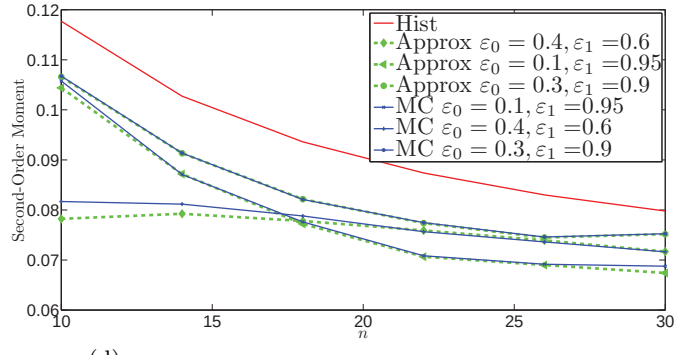
(a) Expected true error with exact λ_0 and λ_1



(b) Expected true error with data-dependent λ_0 and λ_1



(c) Second-order moment with exact λ_0 and λ_1



(d) Second-order moment with data-dependent λ_0 and λ_1

Figure 3: The first- and second-moments of the true error of the RML classifier and Histogram rule with $\varepsilon-$contamination uncertainty classes with size, $|\Pi^0| = |\Pi^1| = 250$. Steady-state distributions with $b = 2^3$ states are considered. In (a) and (c) the regularization parameters, are exact as in (29)- (30). In (b) and (d), they are estimated as in (34)-(35).

system that is subject to external stimuli and behaves according to interactions amongst its constitutive components. These interactions between components are often referred to as pathways and are time invariant in most biological processes. It is instead the varying cellular context that activates or deactivates pathways in order for a cell to respond to the demands of life. For many classification problems of interest and this example here, these pathways will be identical in each class and it is the cellular context of available nutrients, signaling proteins, or other agents that are of interest. However, the general method can be used with differing pathways if the goal is to discriminate against such things as the presence of mutations, separate organisms, or cancer. In all of these examples, we would expect the two classes to have different pathways through differing genetics.

To set up the classification example, we use a single set of pathways describing our biological system of interest, and choose two different cellular contexts which describe the biological phenomena we are interested in classifying. Then for each (context, pathways) tuple we generate an intermediate class of dynamical systems that have behavior described by the the biological pathways under this context. These classes represent all possible dynamical systems that can behave according to the constraints of the pathways and cellular context. Each dynamical system in these two classes possesses a unique steady-state distribution, and we can therefore obtain two classes of steady state distributions from our two tuples of (context, pathways).

6.1. The NF-κB system

Nuclear factor-κB (NF-κB) is a family of transcription factors that control the expression of over 100 genes. Its primary role is in the immune system as a central regulator of inflammation. This makes it important in cancer research as inflammation contributes to the reduction of apoptosis and increased angiogenesis in the tumor microenvironment [49].

Biologically the NF-κB transcription factor can be activated through several parallel signaling pathways. In this paper we use a model containing three stimulating external inputs which are shaded in Figure 4. When a bacterial infection occurs, the lipopolysaccharide (LPS) molecule present in the cell wall of the bacteria binds to TLR4 receptors in immune cell membranes and initiates a strong NF-κB response [50]. Tumor necrosis factor α (TNFα) is a cytokine produced primarily by macrophages to induce an endogenous inflammatory response by binding to the TNFR receptor. And finally, NF-κB responses can be initiated through the 'alternative pathway' with

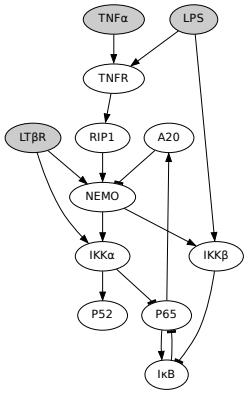


Figure 4: The interactions between members of this model are shown using directed edges where an edge from species A to species B indicates that species A regulates species B. Pointed edges represent promoting influences while tee edges represent down regulating influences. LPS, TNFα, and LTβR are shaded indicating their role as external stimuli to the cell. These three inputs provide the cellular context for the model as described in [12].

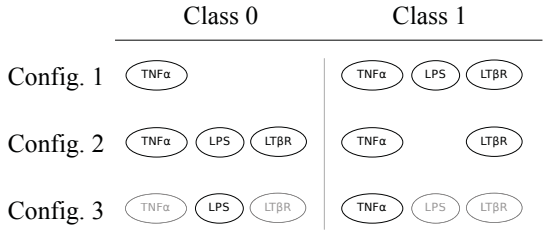


Figure 5: The three classification problems (configurations) considered in this paper are defined by a pair of biologically interesting cellular contexts. For each configuration we attempt to classify samples as coming from class 0 or class 1 given measurements of the 9 downstream signaling proteins. The presence of an input indicates activation, absence indicates inactivation, and a shaded input indicates the input may either be active or inactive.

the lymphotoxin β receptor (LTβR). Once activated, each of these inputs initiates a downstream signaling cascade activating the NF-κB system. As there is no feedback from the system back onto these three external signaling molecules, their state is constant once chosen and helps determine the behavior of the other nine genes.

6.2. NF-κB classification

In a biological system, we are often unable to directly measure or quantify the cellular context which controls the behavior of some cells of interest. We consider such a scenario as a classification problem. Given two possible cellular contexts and some data samples of the 9 proteins whose behaviors are constrained by the context, determine which context the samples were taken from.

In Figure 5 we graphically depict the two contexts (or classes) in three such classification problems (or configurations). The presence of an input indicates activation, absence indicates inactivation, and a shaded input indicates the input may either be active or inactive.

Qualitatively the three configurations in Figure 5 can be described in the following manner: configuration 1 considers an endogenous macrophage induced inflammatory insult in class 0 versus inflammation as a result of bacteria and the response of immune cells in class 1 [50]. Configuration 2 considers an inflammatory insult resulting from bacteria and immune cells in class 0 versus an endogenous inflammatory insult arising from many types of immune cells signaling in class 1. Configuration 3 compares inflammation resulting from a bacterial infection (either in the early stage with no immune cells present or late stage after immune cells have arrived) in class 0 versus an inflammatory injury with immune cells present (possibly resulting from a bacterial infection in class 1).

In these three configurations we measure the ability for the classifier to distinguish the underlying context for an inflammatory response. The classification problem is of significant medical and translational science import.

6.3. Modeling the NF-κB system

Previously, we have used pathways collected from the literature to develop and validate a discrete-time, finite-state Markov chain model of the NF-κB system [12]. This method was then generalized in [51] to generate a parameterized class of Markov chains from the pathway knowledge instead of a single Markov chain.

The pathways which define the NF-κB model (which can be seen in [52]) constrain the possible behaviors and interactions of the nine genes. As these pathways are incomplete and sometimes conflicting, the evolution of the Markov chain in some states is often uncertain. We model these uncertainties as independent Bernoulli random variables in the state transition graph with unknown parameters. We then consider the collection of these parameters in the vector $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, where $\theta_i \in [0, 1]$, in order to parameterize the uncertainty class of system behavior.

In the NF-κB model, there are only three uncertainties that arise from the pathways. These determine the parameterization of the uncertainty class via the vector $\theta \in [0, 1]^3$. Choosing θ gives a single well-defined Markov chain from the uncertainty class. For a small example see the companion website (Section 3 of the supplementary materials) and for more details we refer to [12]

and [51]. For the true network, we choose a network from [12]. It is at the center of the parameter space, $\theta_{ac} = (0.5, 0.5, 0.5)$. From the standpoint of classification this network is unknown; it is chosen here to generate samples. *A priori* we only know that the true network exists inside our uncertainty class.

6.4. Results

To utilize this modeling technique with the proposed RML framework we define two uncertainty classes of models for each configuration by fixing the inputs according to Figure 5. Since the RML framework requires finite uncertainty classes, we discretize the continuous $[0, 1]^3$ space as explained in the companion website. Then, adding a perturbation probability $p = 10^{-3}$ in our simulations to each network, we obtain a class of ergodic irreducible Markov chains and, accordingly, a class of steady-state distributions [47]. The perturbation probability for the true model is set to $p = 10^{-5}$. We generate data from the true network in each class. These two data sets along with the two uncertainty classes allow us to compare the RML classification framework with the classical histogram rule. Figure 6 shows the results for the histogram-rule and proposed method for

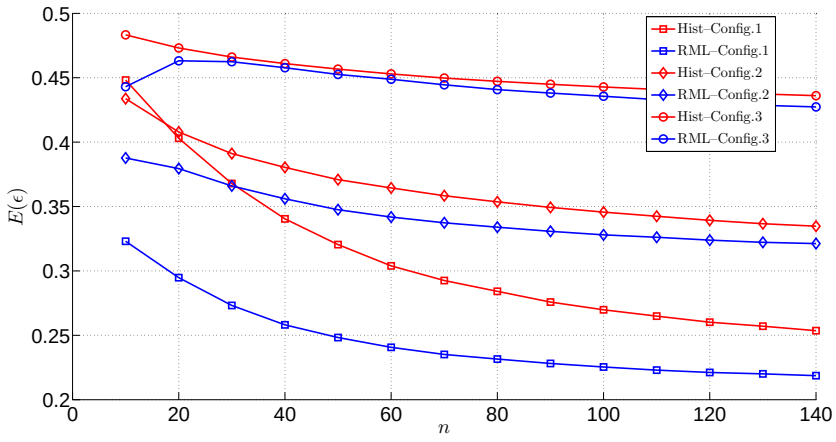


Figure 6: Performance comparison between the Histogram-rule and the RML framework. The x axis shows the number of samples n, with $n = n_0 + n_1, n_0 = n_1$. We have $\epsilon_{Bayes} = 0.193$, $\epsilon_{Bayes} = 0.299$, and $\epsilon_{Bayes} = 0.371$ for Configurations 1, 2, and 3, respectively.

different configurations. In configuration 3, the error of the classifier briefly increases as a function of the sample size at the beginning. The regularization parameter is set according to Corollary 1, denoted by λ_{SURE}. Both the histogram and RML classifiers converge to the Bayes errors as $n \to \infty$.

In all cases, the RML approach outperforms the histogram-rule, illustrating the benefit of prior knowledge, if available.

6.4.1. Comparison to MAP

Designing the RML classifier begins with the assumption of having finite uncertainty classes of feature distributions, in the absence of a prior distribution governing these classes, i.e., no prioritization of any uncertainty class member in favor of the others. Nonetheless, one would still solve the maximum *a posteriori* (MAP) to find the most likely multinomial distribution existing in the uncertainty class and build the "plug-in rule" classifier according to equation (1). Hence, using the log-likelihood function in equation (2), we define the MAP distribution as

$$\hat{\boldsymbol{\pi}}^y_{\mathbf{MAP}} := \arg\max_{\boldsymbol{\pi}^y \in \Pi^y} \sum_{k=1}^{b} u_k^y \log \pi^y(k). \tag{37}$$

Thereafter, we define the MAP classifier by plugging the estimates $\hat{\boldsymbol{\pi}}^y_{\mathbf{MAP}}$ in equation (1). In Figure 7, we compare performance of the RML given in Lemma 1 with that of MAP given in equation (37) by plotting the difference between the corresponding expected true errors, i.e., $\mathrm{E}_{S_n}[\epsilon_{\mathbf{MAP}} - \epsilon_{\mathbf{RML}}]$ as a function of sample size for the three configurations considered in Figure 6. Figure 7 illustrates that for configurations 1 and 2 the RML classifier performs always better than

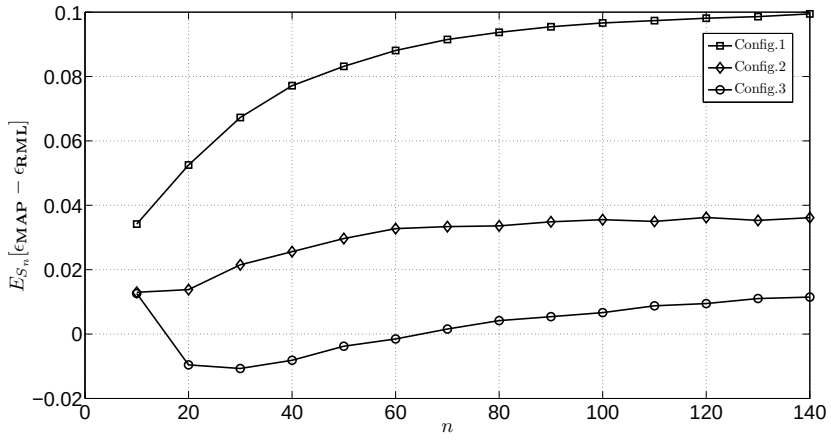


Figure 7: Performance comparison between the RML and MAP classifier defined in Lemma 1 and the one designed using estimates in equation (37), respectively. The x axis shows the number of samples n.

the MAP. For category 3, the MAP classifier performs better than the RML in some range, but then, the RML classifier outperforms the MAP after increasing the sample size.

7. Conclusion

We have proposed a novel classifier design paradigm that allows us to design enhanced classifiers by incorporating available prior knowledge of the process generating the observation data. As shown in our simulations, such knowledge can significantly improve the performance of the designed classifier, especially, when the sample size is small. Having laid the theoretical groundwork for enhancing steady-state classifier design via the use of prior process knowledge, our plan is to apply the methodology to developing better biomedical classifiers in the presence of partial knowledge of the underlying genetic regulatory network. More generally, given the ubiquity of large feature sets and relatively small sample sizes now common in many disciplines, including medicine, material science, environmental science, and transportation, there will no doubt be an increasing number of methods proposed for using prior knowledge in classifier design. We believe it is important to provide analytic performance characterization of the classifiers on standard models, as we have done in this paper, so that their behavior can be understood.

Appendix A. Proof of Lemma 1

Plugging (5) in (4), we obtain

$$\begin{aligned}
\hat{\boldsymbol{\pi}}^y_{RML} &= \arg\min -(1-\lambda_y)\sum_{k=1}^{b} u_k^y \log \pi^y(k) + \tfrac{\lambda_y}{|\Pi^y|}\sum_{i=1}^{|\Pi^y|}\sum_{k=1}^{b}\pi_i^y(k)\log\tfrac{\pi_i^y(k)}{\pi^y(k)} \\
&= \arg\min -\Big[(1-\lambda_y)\sum_{k=1}^{b} u_k^y \log \pi^y(k) + \lambda_y \sum_{k=1}^{b}\log\pi^y(k)\{\tfrac{1}{|\Pi^y|}\sum_{i=1}^{|\Pi^y|}\pi_i^y(k)\}\Big] \qquad \text{(A.1)} \\
&= \arg\min -\Big[\sum_{k=1}^{b}[(1-\lambda_y)u_k^y + \lambda_y\overline{\pi}^y(k)]\log\pi^y(k)\Big]
\end{aligned}$$

The solution to this problem can be obtained using a *Lagrangian multiplier* similar to (2), which leads to the label conditional probabilities in (6).

Appendix B. Proof of Theorems 1 and 2

In this appendix, we prove Theorems 1 and 2 for $y = 0$. The case $y = 1$ can be handled similarly. Let the inner expectation in (13), $\mathrm{E}_{S_n}\Big[\Pr(\psi_{n,\Pi^0,\Pi^1}(X) \neq Y|S_n)\Big]$, be denoted by EXP_1. Then

$$\begin{aligned}\mathrm{EXP}_1 &= \mathrm{E}_{S_n}\left[\sum_k \Pr(X=k,Y=0)\mathrm{I}_{\{\psi_{n,\Pi^0.\Pi^1}=1\}} + \Pr(X=k,Y=1)I_{\{\psi_{n,\Pi^0.\Pi^1}=0\}}\right] \\ &= c_0\sum_k\left[\pi^0(k)\Pr(\hat{c_1}\tfrac{(1-\lambda_1)u_k^1+\lambda_1\overline{\pi}^1(k)}{(1-\lambda_1)n_1+\lambda_1} \geq \hat{c_0}\tfrac{(1-\lambda_0)u_k^0+\lambda_0\overline{\pi}^0(k)}{(1-\lambda_0)n_0+\lambda_0})\right] \\ &+c_1\sum_k\left[\pi^1(k)\Pr(\hat{c_0}\tfrac{(1-\lambda_0)u_k^0+\lambda_0\overline{\pi}^0(k)}{(1-\lambda_0)n_0+\lambda_0} > \hat{c_1}\tfrac{(1-\lambda_1)u_k^1+\lambda_1\overline{\pi}^1(k)}{(1-\lambda_1)n_1+\lambda_1})\right],\end{aligned} \tag{B.1}$$

in which we apply $\hat{c_y} = \frac{n_y}{n}; y = 0, 1$. We denote the average distribution by $\overline{\boldsymbol{\pi}}_y; y = 0, 1$ which can be computed by $\overline{\boldsymbol{\pi}}_y = (1-\varepsilon_y)\boldsymbol{\pi}_{ac}^y + \varepsilon_y\overline{\boldsymbol{\pi}}$, where $\overline{\boldsymbol{\pi}}$ is the average of contaminating distributions. Now, for $y = 0, 1$, define

$$\begin{aligned} g_y &:= (1-\lambda_y)n_y(n_{1-y}(1-\lambda_{1-y})+\lambda_{1-y}) \\ \alpha_y &:= \tfrac{g_y\lambda_y}{1-\lambda_y} \\ p_y(k) &:= \alpha_y\overline{\pi}_y(k).\end{aligned} \tag{B.2}$$

Equation (B.1) can be written as

$$\begin{aligned}\mathrm{EXP}_1 &= \sum_{k=1}^{b}\Big[\Pr(X=k|Y=0)c_0\Pr(g_1u_k^1+p_1(k) \geq g_0u_k^0+p_0(k))\} \\ &+\Pr(X=k|Y=1)c_1\Pr(g_0u_k^0+p_0(k) > g_1u_k^1+p_1(k))\Big],\end{aligned} \tag{B.3}$$

$$\begin{aligned}\mathrm{EXP}_1 &= \sum_{k=1}^{b}\pi_{ac}^0(k)c_0\Big[\textstyle\sum_{l_0=0}^{n_0}\{\sum_{m=\zeta_{k,l_0}^0}^{n_1}(\pi_{ac}^1(k))^m(1-\pi_{ac}^1(k))^{n_1-m}\binom{n_1}{m}\} \\ &(\pi_{ac}^0(k)^{l_0}(1-\pi_{ac}^0(k))^{n_0-l_0}\binom{n_0}{l_0})\Big] \\ &+\textstyle\sum_{k=1}^{b}\pi_{ac}^1(k)c_1\Big[\sum_{l_1=0}^{n_1}\{\sum_{m=\zeta_{k,l_1}^1}^{n_0}(\pi_{ac}^0(k))^m(1-\pi_{ac}^0(k))^{n_0-m}\binom{n_0}{m}\} \\ &\times(\pi_{ac}^1(k)^{l_1}(1-\pi_{ac}^1(k))^{n_1-l_1}\binom{n_1}{l_1})\Big],\end{aligned} \tag{B.4}$$

where

$$\begin{aligned}\underline{\zeta}_{k,l_0}^0 &= \max\{0, \left\lfloor\tfrac{g_0l_0+p_0(k)-p_1(k)}{g_1}\right\rfloor+1\}, \\ \underline{\zeta}_{k,l_1}^1 &= \max\{0, \left\lfloor\tfrac{g_1l_1+p_1(k)-p_0(k)}{g_0}\right\rfloor+1\}.\end{aligned} \tag{B.5}$$

In (B.4), we have two random variables $\underline{\zeta}_{k,l_0}^0$ and $\underline{\zeta}_{k,l_1}^1$ depending on the uncertainty classes Π^0 and Π^1, respectively. We present the distributions of these random variables for the uncertainty class models described in Section 2 in the following subsections:

Appendix B.0.2. $\varepsilon-$contamination class

We first show that the contaminating part $\pi(k)$ in (9) has a Beta distribution $B(1, b-1)$, where b is the number of states. Suppose that the contaminating distributions come from a uniform distribution on a $(b-1)$- simplex. Thus, as $\Delta x \to 0$,

$$\Pr(x - \Delta x/2 < \pi(k) < x + \Delta x/2) \quad = \Delta x \frac{\text{Vol}(\mathcal{S}_{b-2}^{1-x})}{\text{Vol}(\mathcal{S}_{b-1})} = \Delta x \frac{\frac{(1-x)^{b-2}}{(b-2)!}}{\frac{1}{(b-1)!}} = \Delta x (b-1)(1-x)^{b-2} \quad \text{(B.6)}$$

where Vol(.) denotes volume under the specified argument and $\mathcal{S}_{b-1}$ and $\mathcal{S}_{b-2}^{1-x}$ are the unit $(b-1)-$simplex and $(b-2)-$simplex with corners on $1-x$, respectively. (B.6) can be written as a density function according to

$$f_{\pi(k)}(x) = (b-1)(1-x)^{b-2}, x \in (0,1) \quad \text{(B.7)}$$

which is a Beta distribution with parameters 1 and $b-1$ whose mean and variance are $\frac{1}{b}$ and $\frac{b-1}{b^2(b+1)}$, respectively. Using the *Edgeworth expansion* to approximate the cumulative density function of $\overline{\pi}(k)$, [53], we obtain

$$\Pr(\overline{\pi}(k) < x) = \Phi(z) + R_{|\Pi^0|} \quad \text{(B.8)}$$

where $z = \sqrt{|\Pi^0|}\frac{x-\frac{1}{b}}{\sqrt{\frac{b-1}{b^2(b+1)}}}$, and we have

$$R_{|\Pi^0|} := \lim_{w\to\infty} \frac{\sum_{v=1}^{w} r_v(\Pi^0)}{\exp(c|\Pi^0|)}, c > 0. \quad \text{(B.9)}$$

In (B.9), according to the Edgeworth expansion, we have

$$r_v(|\Pi^0|) = O(|\Pi^0|^{\frac{v}{2}-1}). \quad \text{(B.10)}$$

Considering (B.9) and (B.10), one can conclude that $R_{|\Pi^0|} \to 0$ for large enough uncertainty classes. Therefore, for large uncertainty classes, we will approximately have $\frac{\overline{\pi}(k)-\frac{1}{b}}{\sqrt{\frac{b-1}{b^2(b+1)}}} \sim N(0, \frac{1}{|\Pi^y|})$. Hence, considering the last line of equation (B.2), we get the following result:

$$\begin{aligned} p_0(k) &\sim N(\alpha_0 \left[(1-\varepsilon_0)\pi_{ac}^0(k) + \tfrac{\varepsilon_0}{b}\right], \alpha_0^2\varepsilon_0^2\tfrac{(b-1)}{b^2(b+1)|\Pi^0|}) \\ p_1(k) &\sim N(\alpha_1 \left[(1-\varepsilon_1)\pi_{ac}^1(k) + \tfrac{\varepsilon_1}{b}\right], \alpha_1^2\varepsilon_1^2\tfrac{(b-1)}{b^2(b+1)|\Pi^1|}). \end{aligned} \quad \text{(B.11)}$$

Thus, since $p_0(k)$ and $p_1(k)$ are independent random variables, we get

$$\frac{g_0 l_0 + p_0(k) - p_1(k)}{g_1} \sim N(\mu_{k,l_0}^0, \sigma_0^2),$$

where μ_{k,l_0}^0, σ_0^2 are defined in (17). It is now straightforward to find the distribution of $\underline{\zeta}_{k,l_0}^0$ (and similarly $\underline{\zeta}_{k,l_1}^1$) using equation (B.5).

Appendix B.0.3. p−point class

From the mapping defined in 12, we know that state k belongs to $s^0_{P^0(k)}$ and $s^1_{P^1(k)}$ under labels zero and one, respectively. Considering class Π^0, similar to (B.6), one can show that

$$p_{\pi(k)}(x) = \frac{|s^0_{P^0(k)}| - 1}{\omega^0_{P^0(k)}}(1 - \frac{x}{\omega^0_{P^0(k)}})^{|s^0_{P^0(k)}|-2}, x \in (0, \omega^0_{P^0(k)}). \tag{B.12}$$

which is equivalent to the random variable $\omega^0_{P^0(k)}Y$ with $Y \sim Beta(1, |s^0_{P^0(k)}| - 1)$. Therefore, similar to (B.11), we obtain

$$\begin{aligned} p_0(k) &\sim N(\alpha_0 \tfrac{\omega^0_1}{|s^0_1|}, \alpha_0^2(\omega^0_1)^2 \tfrac{(|s^0_1|-1)}{|s^0_1|^2(|s^0_1|+1)|\Pi^0|}) \\ p_1(k) &\sim N(\alpha_1 \tfrac{\omega^1_1}{|s^1_1|}, \alpha_1^2(\omega^1_1)^2 \tfrac{(|s^1_1|-1)}{|s^1_1|^2(|s^1_1|+1)|\Pi^1|}), \end{aligned}$$

from which we obtain $\frac{g_0 l_0 + p_0(k) - p_1(k)}{g_1} \sim N(\mu^0_{k,l_0}, \sigma_0^2)$, whereas

$$\begin{aligned} \mu^0_{k,l_0} &= \frac{g_0 l_0 + \alpha_0 \frac{\omega^0_{P^0(k)}}{|s^0_{P^0(k)}|} - \alpha_1 \frac{\omega^1_{P^1(k)}}{|s^1_{P^1(k)}|}}{g_1} \\ {\sigma_0}^2 &= \Big[\alpha_0^2(\omega^0_{P^0(k)})^2 \tfrac{(|s^0_{P^0(k)}|-1)}{|s^0_{P^0(k)}|^2(|s^0_{P^0(k)}|+1)|\Pi^0|} + \alpha_1^2(\omega^1_{P^1(k)})^2 \tfrac{(|s^1_{P^1(k)}|-1)}{|s^1_{P^1(k)}|^2(|s^1_{P^1(k)}|+1)|\Pi^1|}\Big]/g_1^2. \end{aligned} \tag{B.13}$$

Now, one can find the distribution of $\underline{\zeta}^0_{k,l_0}$ according to (B.5). The distribution of $\underline{\zeta}^1_{k,l_1}$ can be found similarly. Afterwards, we obtain equation (15).

Appendix C. Proof of Theorem 3

The second-order moment of the true error of the RML classifier can be written as

$$\mathrm{E}(\epsilon^2_{\mathbf{RML}}) = \mathrm{E}_{\Pi^0,\Pi^1}\left[\mathrm{E}_{S_n}\left[\Pr(\psi_{n,\Pi^0,\Pi^1}(X) \neq Y|S_n)\right]^2 |\Pi^0, \Pi^1\right]. \tag{C.1}$$

For simplicity, we drop the subscript of ψ_{n,Π^0,Π^1}, noting that the classifier depends S_n and Π^0, Π^1. The proof has two parts shown in two appendices. First, we take the expectation with respect to the training data, S_n. Later, we will see that the dependency of the second-order moment on the uncertainty classes manifests itself in the indices of the double-summations (found from combinatorial parts). In the next section, then we find the distribution of those indices, knowing that the randomness comes from the uncertainty classes. Let us start the proof by expanding

equation (C.1):

$$\begin{aligned}
\mathrm{E}(\epsilon^2_{\mathbf{RML}}) &= \mathrm{E}_{\Pi^0,\Pi^1}\Big[c_0^2\sum_k(\pi^0_{ac}(k))^2\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k)=1\}}]}_{A^1}+c_1^2\sum_k(\pi^1_{ac}(k))^2\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k)=0\}}]}_{A^0}\\
&+c_0^2\sum_{k_1\neq k_2}\pi^0_{ac}(k_1)\pi^0_{ac}(k_2)\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k_1)=1\}}\mathrm{I}_{\{\psi(X=k_2)=1\}}]}_{B^1}\\
&+c_1^2\sum_{k_1\neq k_2}\pi^1_{ac}(k_1)\pi^1_{ac}(k_2)\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k_1)=0\}}\mathrm{I}_{\{\psi(X=k_2)=0\}}]}_{B^0}\\
&+c_0c_1\sum_{k_1\neq k_2}\pi^0_{ac}(k_1)\pi^1_{ac}(k_2)\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k_1)=1\}}\mathrm{I}_{\{\psi(X=k_2)=0\}}]}_{C^1}\\
&+c_0c_1\sum_{k_1\neq k_2}\pi^1_{ac}(k_1)\pi^0_{ac}(k_2)\underbrace{\mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k_1)=0\}}\mathrm{I}_{\{\psi(X=k_2)=1\}}]}_{C^0}\Big].
\end{aligned}\tag{C.2}$$

In (C.2), parts A_0 and A_1 can be found similarly as in Appendix 1. In the following, whenever we sum over $t_1^y, t_2^y; y\in\{0,1\}$ we implicitly consider $t_1^y, t_2^y\geq 0$ and $t_1^y+t_2^y\leq n_y$. Furthermore, for any pair of $(t_1^y,t_2^y)\succeq\mathbf{0}$ with $t_1^y+t_2^y\leq n_y$, we have

$$\Pr(u^y_{k_1}=t_1^y,u^y_{k_2}=t_2^y)=\Pr(trin(n_y,\pi^y_{ac}(k_1),\pi^y_{ac}(k_2))=(t_1^y,t_2^y)).$$

Hence, for the B^1, we may write

$$\begin{aligned}
B^1 &= \mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=k_1)=1\}}\mathrm{I}_{\{\psi(X=k_2)=1\}}]=\Pr(\psi(X=k_1)=1,\psi(X=k_2)=1)\\
&=\Pr(g_1u^1_{k_1}+p_1(k_1)\geq g_0u^0_i+p_0(i),g_1u^1_{k_2}+p_1(k_2)\geq g_0u^0_{k_2}+p_0(k_2))\\
&=\sum_{t_1^0,t_2^0}\Pr(g_1u^1_{k_1}+p_1(k_1)\geq g_0t_1^0+p_0(k_1),g_1u^1_{k_2}+p_1(k_2)\geq g_0t_2^0+p_0(k_2))\\
&\times\Pr(u^0_{k_1}=t_1^0,u^0_{k_2}=t_2^0)\\
&=\sum_{t_1^0,t_2^0}\Pr(u^1_{k_1}\geq\underline{\zeta}^0_{k_1,t_1^0},u^1_{k_2}\geq\underline{\zeta}^0_{k_2,t_2^0})\Pr(u^0_{k_1}=t_1^0,u^0_{k_2}=t_2^0)\\
&=\sum_{t_1^0,t_2^0}\left[\sum_{(t_1^1,t_2^1)\succeq(\underline{\zeta}^0_{k_1,t_1^0},\underline{\zeta}^0_{k_2,t_2^0})}\Pr(u^1_{k_1}=t_1^1,u^1_{k_2}=t_2^1)\Pr(u^0_{k_1}=t_1^0,u^0_{k_2}=t_2^0)\right]
\end{aligned}\tag{C.3}$$

Similarly, we can get

$$B^0=\sum_{t_1^1,t_2^1}\left[\sum_{(t_1^0,t_2^0)\succeq(\underline{\zeta}^1_{k_1,t_1^1},\underline{\zeta}^1_{k_2,t_2^1})}\Pr(u^0_{k_1}=t_1^0,u^0_{k_2}=t_2^0)\Pr(u^1_{k_1}=t_1^1,u^1_{k_2}=t_2^1)\right].\tag{C.4}$$

Next, we can obtain C^1

$$\begin{aligned}
C^1 &= \mathrm{E}_{S_n}[\mathrm{I}_{\{\psi(X=i)=1\}}\mathrm{I}_{\{\psi(X=j)=0\}}] = \Pr(\psi(X=k_1)=1,\psi(X=k_2)=0)\\
&= \Pr(g_1u^1_{k_1}+p_1(k_1)\geq g_0u^0_{k_1}+p_0(k_1), g_1u^1_{k_2}+p_1(k_2)<g_0u^0_{k_2}+p_0(j))\\
&= \sum_{t^0_1,t^0_2}\Pr(g_1u^1_{k_1}+p_1(k_1)\geq g_0t^0_1+p_0(k_1), g_1u^1_{k_2}+p_1(k_2)<g_0t^0_2+p_0(k_2))\\
&\times \Pr(u^0_{k_1}=t^0_1,u^0_{k_2}=t^0_2) \qquad\qquad (C.5)\\
&= \sum_{t^0_1,t^0_2}\Pr(\underline{\zeta}^0_{k_1,t^0_1}\leq u^1_{k_1},u^1_{k_2}\leq\overline{\zeta}^0_{k_2,t^0_2})\Pr(u^0_{k_1}=t^0_1,u^0_{k_2}=t^0_2)\\
&= \sum_{t^0_1,t^0_2}\left[\sum_{t^1_1\geq\underline{\zeta}^0_{k_1,t^0_1},t^1_2\leq\overline{\zeta}^0_{k_2,t^0_2}}\Pr(u^1_{k_1}=t^1_1,u^1_{k_2}=t^1_2)\Pr(u^0_{k_1}=t^0_1,u^0_{k_2}=t^1_1)\right],
\end{aligned}$$

Similarly, we obtain

$$C^0=\sum_{t^1_1,t^1_2}\left[\sum_{t^0_1\geq\underline{\zeta}^1_{k_1,t^1_1},t^0_2\leq,\overline{\zeta}^1_{k_2,t^1_2}}\Pr(u^0_{k_1}=t^0_1,u^0_{k_2}=t^0_2)\Pr(u^1_{k_1}=t^1_1,u^1_{k_2}=t^1_0)\right]. \qquad (C.6)$$

In (C.5)-(C.6), we have

$$\begin{aligned}
\overline{\zeta}^0_{k,l_0} &= \min\{\left\lceil\tfrac{g_0l_0+p_0(k)-p_1(k)}{g_1}\right\rceil-1,n_1\},\\
\overline{\zeta}^1_{k,l_1} &= \min\{\left\lceil\tfrac{g_1l_1+p_1(k)-p_0(k)}{g_0}\right\rceil-1,n_0\}.
\end{aligned} \qquad (C.7)$$

In order to take the last expectation in (C.2) with respect to the uncertainty classes, we need to find the joint distribution of $\underline{\zeta}^0_{k_1,t^0_1}$ and $\underline{\zeta}^0_{k_2,t^0_2}$ (similarly for $\underline{\zeta}^1_{k_1,t^1_1}$ and $\underline{\zeta}^1_{k_2,t^1_2}$), and the joint distribution between $\underline{\zeta}^0_{k_1,t^0_1}$ and $\overline{\zeta}^0_{k_2,t^0_2}$ (similarly for $\underline{\zeta}^1_{k_1,t^1_1}$ and $\overline{\zeta}^1_{k_2,t^1_2}$). These distributions are found in Appendix D.

Appendix D. Joint distributions

To find the joint distribution of $(\underline{\zeta}^0_{k_1,t^0_1},\underline{\zeta}^0_{k_2,t^0_2})$, we need to approximate the joint distribution of $(p_0(k_1),p_0(k_2))$ defined in equation (B.2). We do this by a (zero-order) Edgeworth expansion. Thus, similar to the single variate case in (B.11), for the multivariate case we have $(p_0(k_1),p_0(k_2))\sim\mathcal{N}(\boldsymbol{\mu}^0_{k_1,k_2},\boldsymbol{\Sigma}^0_{k_1,k_2})$, whereas we find the parameters for different uncertainty classes in the following subsections.

Appendix D.1. $\varepsilon-$contamination class

From the definition of the joint probability distribution, for $x_1, x_2 > 0, x_1 + x_2 \leq 1$, we have

$$\begin{aligned}\Pr(\pi(k_1) = x_1, \pi(k_2) = x_2) &= \lim_{\Delta x_1 \to 0 \Delta x_2 \to 0} \frac{\Pr(|\pi(k_1) - x_1| < \frac{\Delta x_1}{2}, |\pi(k_2) - x_2| < \frac{\Delta x_2}{2})}{\Delta x_1 \Delta x_2} \\ &= \lim_{\Delta x_1 \to 0 \Delta x_2 \to 0} \frac{\frac{\Delta x_1 \Delta x_2 \mathrm{Vol}(\mathcal{S}_{b-3}^{1-x_1-x_2})}{\mathrm{Vol}(\mathcal{S}_{b-1})}}{\Delta x_1 \Delta x_2} \\ &= (b-1)(b-2)(1-x_1-x_2)^{b-3}.\end{aligned} \tag{D.1}$$

Since we are going to use the zero-order Edgeworth expansion, we only need to find the mean vector and the covariance matrix of these random variables. The variances are already found in the previous section of the Appendix. Therefore, we only find the covariance between these variables. Specifically,

$$\begin{aligned}\mathrm{Cov}[\pi(k_1), \pi(k_2)] &= E[\pi(k_1)\pi(k_2)] - E[\pi(k_1)]E[\pi(k_2)] \\ &= \int_0^1 \int_0^{1-x_1} x_1 x_2 (b-1)(b-2)(1-x_1-x_2)^{b-3} dx_2 dx_1 - \tfrac{1}{b^2} \\ &= \tfrac{-1}{b^2(b+1)},\end{aligned} \tag{D.2}$$

where in (D.2) we used integration by parts. Hence, considering our definitions in (B.2) for $p_0(k_1)$ and $p_0(k_2)$, we obtain the following for the normal distribution statistics:

$$\boldsymbol{\mu}_{k_1,k_2}^0 = \begin{bmatrix} \alpha_0(\frac{\varepsilon_0}{b} + (1-\varepsilon_0)\pi_{ac}^0(k_1)) \\ \alpha_0(\frac{\varepsilon_0}{b} + (1-\varepsilon_0)\pi_{ac}^0(k_2)) \end{bmatrix}, \tag{D.3}$$

$$\boldsymbol{\Sigma}_{k_1,k_2}^0 = \begin{bmatrix} \alpha_0^2\varepsilon_0^2 \frac{b-1}{b^2(b+1)|\Pi^0|} & -\alpha_0^2\varepsilon_0^2 \frac{1}{b^2(b+1)|\Pi^0|} \\ -\alpha_0^2\varepsilon_0^2 \frac{1}{b^2(b+1)|\Pi^0|} & \alpha_0^2\varepsilon_0^2 \frac{b-1}{b^2(b+1)|\Pi^0|} \end{bmatrix}. \tag{D.4}$$

Similarly, we can write for the joint distribution of $(p_1(k_1), p_1(k_2))$.

Appendix D.2. $p-$point class

Since we have partitions in this model, we need to know whether two states belong to the same partition or not. First, suppose that $P^0(k_1) \neq P^0(k_2)$. Then,

$$\Pr(\pi(k_1) = x_1, \pi(k_2) = x_2) = \Pr(\pi(k_1) = x_1) \Pr(\pi(k_2) = x_2), \tag{D.5}$$

from which we get

$$\boldsymbol{\mu}_{k_1,k_2}^0 = \begin{bmatrix} \frac{\omega_{P^0(k_1)}^0 \alpha_0}{|s_{P^0(k_1)}^0|} \\ \frac{\omega_{P^0(k_2)}^0 \alpha_0}{|s_{P^0(k_2)}^0|} \end{bmatrix}, \tag{D.6}$$

$$\boldsymbol{\Sigma}^0_{k_1,k_2} = \begin{bmatrix} \alpha_0^2(\omega^0_{P^0(k_1)})^2 \frac{|s^0_{P^0(k_2)}|-1}{|s^0_{P^0(k_1)}|^2(|s^0_{P^0(k_1)}|+1)|\Pi^0|} & 0 \\ 0 & \alpha_0^2(\omega^0_{P^0(k_2)})^2 \frac{|s^0_{P^0(k_2)}|-1}{|s^0_{P^0(k_2)}|^2(|s^0_{P^0(k_2)}|+1)|\Pi^0|} \end{bmatrix}. \quad \text{(D.7)}$$

Now, suppose $P^0(k_1) = P^0(k_2) = m_{k_1k_2}$. Then

$$\mathrm{Cov}[\pi(k_1), \pi(k_2)] = \frac{-(\omega^0_{m_{k_1k_2}})^2}{|s^0_{m_{k_1k_2}}|^2(|s^0_{m_{k_1k_2}}|+1)}, \quad \text{(D.8)}$$

and we have

$$\boldsymbol{\mu}^0_{k_1,k_2} = \begin{bmatrix} \frac{\omega^0_{m_{k_1k_2}}\alpha_0}{|s^0_{m_{k_1k_2}}|} \\ \frac{\omega^0_{m_{k_1k_2}}\alpha_0}{|s^0_{m_{k_1k_2}}|} \end{bmatrix}, \quad \text{(D.9)}$$

$$\boldsymbol{\Sigma}^0_{k_1,k_2} = \begin{bmatrix} \alpha_0^2(\omega^0_{m_{k_1k_2}})^2 \frac{|s^0_{m_{k_1k_2}}|-1}{|s^0_{m_{k_1k_2}}|^2(|s^0_{m_{k_1k_2}}|+1)|\Pi^0|} & -\alpha_0^2(\omega^0_{m_{k_1k_2}})^2 \frac{1}{|s^0_{m_{k_1k_2}}|^2(|s^0_{m_{k_1k_2}}|+1)|\Pi^0|} \\ -\alpha_0^2(\omega^0_{m_{k_1k_2}})^2 \frac{1}{|s^0_{m_{k_1k_2}}|^2(|s^0_{m_{k_1k_2}}|+1)|\Pi^0|} & \alpha_0^2(\omega^0_{m_{k_1k_2}})^2 \frac{|s^0_{m_{k_1k_2}}|-1}{|s^0_{m_{k_1k_2}}|^2(|s^0_{m_{k_1k_2}}|+1)|\Pi^0|} \end{bmatrix}. \quad \text{(D.10)}$$

In the following, $\Pr(p_0(k_1) = \alpha, p_0(k_2) = \beta)$ and $\Pr(p_1(k_1) = \alpha, p_1(k_2) = \beta)$ will be denoted by $F^0_{k_1,k_2}(\alpha,\beta)$ and $F^1_{k_1,k_2}(\alpha,\beta)$, respectively. Now, we start by computing the pmf of $(\zeta^0_{k_1,t^0_1}, \zeta^0_{k_2,t^0_2})$. After quite some computation we obtain

$$\Pr(\underline{\zeta}^0_{k_1,t^0_1} = m_1, \underline{\zeta}^0_{k_2,t^0_2} = m_2) = \begin{cases} \mathrm{Int}^0(\theta^0_{k_1,L}, \theta^0_{k_1,U}; \theta^0_{k_2,L}, \theta^0_{k_2,U}); m_1, m_2 \neq 0 \\ \mathrm{Int}^0(-\infty, -g_0t^0_1; \theta^0_{k_2,L}, \theta^0_{k_2,U}); m_1 = 0, m_2 \neq 0 \\ \mathrm{Int}^0(\theta^0_{k_1,L}, \theta^0_{k_1,U}; -\infty, -g_0t^0_2); m_2 = 0, m_1 \neq 0 \\ \mathrm{Int}^0(-\infty, -g_0t^0_1; -\infty, -g_0t^0_2); m_1 = m_2 = 0 \end{cases} \quad \text{(D.11)}$$

$$\Pr(\underline{\zeta}^1_{k_1,t^1_1} = m_1, \underline{\zeta}^1_{k_2,t^1_2} = m_2) = \begin{cases} \mathrm{Int}^1(\theta^1_{k_1,L}, \theta^1_{k_1,U}; \theta^1_{k_2,L}, \theta^1_{k_2,U}); m_1, m_2 \neq 0 \\ \mathrm{Int}^1(-\infty, -g_1t^1_1; \theta^1_{k_2,L}, \theta^1_{k_2,U}); m_1 = 0, m_2 \neq 0 \\ \mathrm{Int}^1(\theta_{k_1,L}, \theta^1_{k_1,U}; -\infty, -g_1t^1_2); m_2 = 0, m_1 \neq 0 \\ \mathrm{Int}^1(-\infty, -g_1t^1_1; -\infty, -g_1t^1_2); m_1 = m_2 = 0. \end{cases} \quad \text{(D.12)}$$

Furthermore, we have

$$\Pr(\overline{\zeta}^0_{k_1,t^0_1}=m_1,\underline{\zeta}^0_{k_2,t^0_2}=m_2)=\begin{cases}\mathrm{Int}^0(\overline{\theta}^0_{k_1,L},\overline{\theta}^0_{k_1,U};\overline{\theta}^0_{k_2,L},\overline{\theta}^0_{k_2,U});m_1\neq n_1,m_2\neq 0\\ \mathrm{Int}^0(\theta^0_{k_1,L},\overline{\theta}^0_{k_1,U};-\infty,-g_0t^0_2);m_2=0,m_1\neq n_1\\ \mathrm{Int}^0(-\infty,-g_0t^0_1;g_1(n_1-1)-g_0s,\infty);m_1=n_1,m_2\neq 0\\ \mathrm{Int}^0(g_1(n_1-1)-g_0t^0_1,\infty;-\infty,-g_0t^0_2);m_1=n_1,m_2=0\end{cases} \tag{D.13}$$

$$\Pr(\overline{\zeta}^1_{k_1,t^1_1}=m_1,\underline{\zeta}^1_{k_2,t^1_2}=m_2)=\begin{cases}\mathrm{Int}^1(\overline{\theta}^1_{k_1,L},\overline{\theta}^1_{k_1,U};\overline{\theta}^1_{k_2,L},\overline{\theta}^1_{k_2,U});m_1\neq n_1,m_2\neq 0\\ \mathrm{Int}^1(\overline{\theta}_{k_1,L},\overline{\theta}^1_{k_1,U};-\infty,-g_1s);m_2=0,m_1\neq n_1\\ \mathrm{Int}^1(-\infty,-g_1t^1_1;g_0(n_0-1)-g_1t^1_2,\infty);m_1=n_1,m_2\neq 0\\ \mathrm{Int}^1(g_0(n_0-1)-g_1t^1_1,\infty;-\infty,-g_1t^1_2);m_1=n_0,m_2=0\end{cases} \tag{D.14}$$

In equations (D.11)-(D.14) we use the following definitions (the notation $\int\!\!\!\!\!\!\int .$ is used to denote $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}.$)

$$\begin{aligned}\mathrm{Int}^0(b^{k_1}_L,b^{k_1}_U;b^{k_2}_L,b^{k_2}_U)&:=\int\!\!\!\!\!\!\int\Pr\Big[\begin{pmatrix}\alpha+b^{k_1}_L\\ \beta+b^{k_2}_L\end{pmatrix}\preceq\begin{pmatrix}p_0(k_1)\\ p_0(k_2)\end{pmatrix}\preceq\begin{pmatrix}\alpha+b^{k_1}_U\\ \beta+b^{k_2}_U\end{pmatrix}\Big]F^1_{k_1,k_2}(\alpha,\beta)d\alpha d\beta\\ \mathrm{Int}^0(b^{k_1}_L,b^{k_1}_U;b^{k_2}_L,b^{k_2}_U)&:=\int\!\!\!\!\!\!\int\Pr\Big[\begin{pmatrix}\alpha+b^{k_1}_L\\ \beta+b^{k_2}_L\end{pmatrix}\preceq\begin{pmatrix}p_1(k_1)\\ p_1(k_2)\end{pmatrix}\preceq\begin{pmatrix}\alpha+b^{k_2}_U\\ \beta+b^{k_2}_U\end{pmatrix}\Big]F^0_{k_1,k_2}(\alpha,\beta)d\alpha d\beta\end{aligned} \tag{D.15}$$

Table D.1 shows the parameters used in equations (D.11)- (D.14).

$\theta^0_{k_1,L}=g_1(m_1-1)-g_0t^0_1$	$\theta^0_{k_1,U}=g_1m_1-g_0t^0_1$	$\theta^0_{k_2,L}=g_1(m_2-1)-g_0t^0_2$	$\theta^0_{k_2,U}=g_1m_2-g_0t^0_2$
$\theta^1_{k_1,L}=g_0(m_1-1)-g_1t^1_1$	$\theta^1_{k_1,U}=g_0m_1-g_1t^1_1$	$\theta^1_{k_2,L}=g_0(m_2-1)-g_1t^1_2$	$\theta^1_{k_2,U}=g_0m_2-g_1t^1_2$
$\overline{\theta}^0_{k_1,L}=g_1m_1-g_0t^0_1$	$\overline{\theta}^0_{k_1,U}=g_1(m_1+1)-g_0t^0_1$	$\overline{\theta}^0_{k_2,L}=g_1(m_2-1)-g_0t^0_2$	$\overline{\theta}^0_{k_2,U}=g_1m_2-g_0t^0_2$
$\overline{\theta}^1_{k_1,L}=g_0m_1-g_1t^1_1$	$\overline{\theta}^1_{k_1,U}=g_0(m_1+1)-g_1t^1_1$	$\overline{\theta}^1_{k_2,L}=g_0(m_2-1)-g_1t^1_2$	$\overline{\theta}^1_{k_2,U}=g_0m_2-g_1t^1_2$

Table D.1: Defined parameters.

Appendix E. Proof of Lemma 2

Although we took a standard approach to find the unbiased estimator of the MSE, in this part, for simplicity, we only show that E($\hat{\mathrm{MSE}}$) = MSE (it is sufficient for the proof), where MSE can

be expanded as follows

$$\text{MSE} = \text{E}(\sum_{k=1}^{b}\Big[\hat{\pi}_\lambda(k) - \pi_{ac}(k)\Big]^2) = \sum_{k=1}^{b}\text{E}\left[\hat{\pi}_\lambda^2(k) + \pi_{ac}^2(k) - 2\hat{\pi}_\lambda(k)\pi_{ac}(k)\right]$$

The first and the second terms in the right summation do not need any manipulation. Therefore, in the remainder of the proof, we focus on the last term in the right summation. Using the definitions of δ_λ and $\theta_\lambda(k)$, and the fact that $\text{E}(u_k) = n\pi_{ac}(k)$, we have

$$\text{E}\Big[\sum_{k=1}^{b}\hat{\pi}_\lambda(k)\pi_{ac}(k)\Big] = \sum_{k=1}^{b}(\delta_\lambda n\pi_{ac}(k) + \theta_\lambda(k))\pi_{ac}(k)$$

Now, we return to the $\hat{\text{MSE}}$ in Lemma 2 and take the expectation of the last term in the summation (the term multiplied by 2). We obtain

$$\text{E}\left[\sum_{k=1}^{b}\frac{\delta_\lambda}{n-1}u_k^2 - \sum_{k=1}^{b}u_k(\frac{\delta_\lambda}{n-1} - \frac{\theta_\lambda(k)}{n})\right] = \sum_{k=1}^{b}\frac{\delta_\lambda}{n-1}[n(n-1)\pi_{ac}^2(k) + n\pi_{ac}(k)] - n\pi_{ac}(k)(\frac{\delta_\lambda}{n-1} - \frac{\theta_\lambda(k)}{n}). \tag{E.1}$$

in which we used the terms for the first and second-moments of the multinomial distribution. Some simplification completes the proof.

References

[1] A. Zollanvari, U. Braga-Neto, E. Dougherty, Analytic study of performance of error estimators for linear discriminant analysis, Signal Processing, IEEE Transactions on 59 (2011) 4238–4255.

[2] L. Dalton, E. Dougherty, Bayesian minimum mean-square error estimation for classification error part i: Definition and the bayesian mmse error estimator for discrete classification, Signal Processing, IEEE Transactions on 59 (2011) 115–129.

[3] L. Dalton, E. Dougherty, Bayesian minimum mean-square error estimation for classification error part ii: Linear classification of gaussian models, Signal Processing, IEEE Transactions on 59 (2011) 130–144.

[4] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. Topol, et al., Towards precise classification of cancers based on robust gene functional expression profiles, BMC bioinformatics 6 (2005) 58.

[5] J. Tomfohr, J. Lu, T. Kepler, Pathway level analysis of gene expression using singular value decomposition, BMC bioinformatics 6 (2005) 225.

[6] H. Chuang, E. Lee, Y. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Molecular systems biology 3 (2007).

[7] S. Junjie, Y. Byung-Jun, E. Dougherty, Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network, BMC Bioinformatics 11 (2010) S8.

[8] L. Dalton, E. Dougherty, Optimal classifiers with minimum expected error within a bayesian framework–part i: Discrete and gaussian models, Pattern Recognition (2012).

[9] L. Dalton, E. Dougherty, Optimal classifiers with minimum expected error within a bayesian framework–part ii: Properties and performance analysis, Pattern Recognition (2012).

[10] L. A. Dalton, E. Dougherty, Optimal bayesian classification and its application to gene regulatory networks, Genomic Signal Processing and Statistics (GENSIPS) (2012).

[11] R. Layek, A. Datta, E. Dougherty, From biological pathways to regulatory networks, Mol. BioSyst. 7 (2011) 843–851.

[12] J. Knight, A. Datta, E. Dougherty, Generating stochastic gene regulatory networks consistent with pathway information and steady-state behavior, Biomedical Engineering, IEEE Transactions on 59 (2012) 1701–1710.

[13] A. Zollanvari, U. Braga-Neto, E. Dougherty, Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis, Information Theory, IEEE Transactions on 56 (2010) 784–804.

[14] S. John, The distribution of wald's classification statistic when the dispersion matrix is known, Sankhyā: The Indian Journal of Statistics (1959) 371–376.

[15] S. John, On some classification problems: I, Sankhyā: The Indian Journal of Statistics (1960) 301–308.

[16] S. John, On some classification statistics, Sankhyā: The Indian Journal of Statistics (1960) 309–316.

[17] R. Sitgreaves, Some results on the distribution of the w-classification statistic, 1961.

[18] U. Braga-Neto, E. Dougherty, Exact performance of error estimators for discrete classifiers, Pattern Recognition 38 (2005) 1799–1814.

[19] A. BOWKER, R. Sitgreaves, An Asymptotic Expansion for The Distribution Function of The Classification Statistic W, Technical Report, DTIC Document, 1959.

[20] S. John, Errors in discrimination, The Annals of Mathematical Statistics (1961) 1125–1144.

[21] M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function, The Annals of Mathematical Statistics 34 (1963) 1286–1301.

[22] M. Okamoto, Correction notes: Correction to" an asymptotic expansion for the distribution of the linear discriminant function", The Annals of Mathematical Statistics 39 (1968) 1358–1359.

[23] G. McLachlan, Asymptotic results for discriminant analysis when the initial samples are misclassified, Technometrics 14 (1972) 415–422.

[24] T. Anderson, An asymptotic expansion of the distribution of the studentized classification statistic w1, The Annals of Statistics (1973) 964–972.

[25] J. Sayre, The distributions of the actual error rates in linear discriminant analysis, Journal of the American Statistical Association 75 (1980) 201–205.

[26] C. Lawoko, G. McLachlan, Asymptotic error rates of the w and z statistics when the training observations are dependent, Pattern recognition 19 (1986) 467–471.

[27] V. Berikov, An approach to the evaluation of the performance of a discrete classifier, Pattern recognition letters 23 (2002) 227–233.

[28] V. Berikov, A priori estimates of recognition quality for discrete features, Pattern Recognition and Image Analysis 12 (2002) 235–242.

[29] V. Berikov, A. Litvinenko, The influence of prior knowledge on the expected performance of a classifier, Pattern recognition letters 24 (2003) 2537–2548.

[30] A. Zollanvari, U. Braga-Neto, E. Dougherty, Exact representation of the second-order moments for resubstitution and leave-one-out error estimation for linear discriminant analysis in the univariate heteroskedastic gaussian model, Pattern Recognition 45 (2012) 908–917.

[31] P. Huber, A robust version of the probability ratio test, The Annals of Mathematical Statistics 36 (1965) 1753–1758.

[32] K. Vastola, H. Poor, An analysis of the effects of spectral uncertainty on wiener filtering, Automatica 19 (1983) 289–293.

[33] S. Kassam, H. Poor, Robust techniques for signal processing: A survey, Proceedings of the IEEE 73 (1985) 433–481.

[34] A. Grigoryan, E. Dougherty, Bayesian robust optimal linear filters, Signal processing 81 (2001) 2503–2521.

[35] J. Martín, C. Pérez, P. Müller, Bayesian robustness for decision making problems: Applications in medical contexts, International journal of approximate reasoning 50 (2009) 315–323.

[36] J. Unnikrishnan, V. Veeravalli, S. Meyn, Minimax robust quickest change detection, Information Theory, IEEE Transactions on 57 (2011) 1604–1614.

[37] K. Vastola, H. Poor, On the p–point uncertainty class (corresp.), Information Theory, IEEE Transactions on 30 (1984) 374–376.

[38] D. Sakrison, The rate of a class of random processes, Information Theory, IEEE Transactions on 16 (1970) 10–16.

[39] A. El-Sawy, V. VandeLinde, Robust detection of known signals, Information Theory, IEEE Transactions on 23 (1977) 722–727.

[40] A. El-Sawy, V. VandeLinde, Robust sequential detection of signals in noise, Information Theory, IEEE Transactions on 25 (1979) 346–353.

[41] L. Cimini, S. Kassam, Robust and Quantized Wiener Filters for p-Point Spectral Classes., Technical Report, DTIC Document, 1980.

[42] G. Matz, F. Hlawatsch, Minimax robust nonstationary signal estimation based on a p-point uncertainty model, Journal of the Franklin Institute 337 (2000) 403–419.

[43] P. Brown, P. Rundell, Kernel estimates for categorical data, Technometrics (1985) 293–299.

[44] Y. Eldar, Generalized sure for exponential families: Applications to regularization, Signal Processing, IEEE Transactions on 57 (2009) 471–481.

[45] C. Stein, Estimation of the mean of a multivariate normal distribution, The annals of Statistics (1981) 1135–1151.

[46] G. Zipf, K.(1968). the psycho-biology of language: An introduction to dynamic philology, 1935.

[47] I. Shmulevich, E. Dougherty, S. Kim, W. Zhang, Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks, Bioinformatics 18 (2002) 261–274.

[48] N. Smith, R. Tromble, Sampling uniformly from the unit simplex, Johns Hopkins University, Tech. Rep (2004).

[49] M. Karin, Nf-κb as a critical link between inflammation and cancer, Cold Spring Harbor perspectives in biology 1 (2009).

[50] P. Delves, I. Roitt, Roitt's essential immunology, Wiley-Blackwell, 2006.

[51] J. Knight, E. Dougherty, Attractor estimation and model refinement for stochastic regulatory network models, Genomic Signal Processing and Statistics (GENSIPS) (2011) 54–55.

[52] J. Knight, A. Datta, E. Dougherty, A stochastic nf-κb model consistent with pathway information, In Press, Transaction on Biomedical Engineering, IEEE (2012).

[53] P. Hall, The bootstrap and Edgeworth expansion, Springer Verlag, 1997.