# Robust Bayesian Classification Using an Optimistic Score Ratio

**Viet Anh Nguyen** [1] **Nian Si** [1] **Jose Blanchet** [1]

## Abstract

We build a Bayesian contextual classification model using an optimistic score ratio for robust binary classification when there is limited information on the class-conditional, or contextual, distribution. The optimistic score searches for the distribution that is most plausible to explain the observed outcomes in the testing sample among all distributions belonging to the contextual ambiguity set which is prescribed using a limited structural constraint on the mean vector and the covariance matrix of the underlying contextual distribution. We show that the Bayesian classifier using the optimistic score ratio is conceptually attractive, delivers solid statistical guarantees and is computationally tractable. We showcase the power of the proposed optimistic score ratio classifier on both synthetic and empirical data.

## 1. Introduction

We consider a binary classification setting in which we are provided with training samples from two classes but there is little structure within the classes, e.g., data with heterogeneous distributions except for means and covariance. The ultimate goal is to correctly classify an unlabeled test sample of a given feature. This supervised learning task is the cornerstone of modern machine learning, and its diverse applications are flourishing in promoting healthcare (Naraei et al., 2016; Tomar & Agarwal, 2013), speeding up technological progresses (Rippl et al., 2016; Zhu et al., 2014), and improving societal values (Bhagat et al., 2011; Bodendorf & Kaiser, 2009). Confronting the unstructured nature of the problem, it is natural to exercise a Bayesian approach which employs subjective belief and available information, and then determine an optimal classifying decision that minimizes a certain loss function integrated under the posterior distribution. In this modelling framework, a consensus on the selection of the loss function can be easily reached. Unfortunately, the choice of a class prior and a class-conditional distribution (i.e., the likelihood given the class), two compulsory inputs to the Bayesian machinery to devise the posterior, is more difficult to be agreed upon due to conflicting beliefs among involving parties and limited available data.

Robust Bayesian statistics, which explicitly aims to assemble a posterior inference model with multiple priors and/or multiple class-conditional distributions, emerges as a promising remedy to this longstanding problem. Existing research in this field mainly focuses on robust divergences in the general Bayesian inference framework. Walker (2013) identifies the behaviour of Bayesian updating in the context of model misspecification to show that standard Bayesian updating method learns a model that minimizes the Kullback-Leibler (KL) divergence to the true data generating model. To achieve robustness in Bayesian inference, existing works often target robust divergences, including maximum mean discrepancy, Rényi's alpha-divergences, Hellinger-based divergences, and density power divergence (Chérief-Abdellatif & Alquier, 2019; Knoblauch et al., 2019; Bissiri et al., 2016; Jewson et al., 2018; Ghosh & Basu, 2016). Learning the learning rate in the general Bayesian inference framework is also gaining more recent attention (Holmes & Walker, 2017; Knoblauch, 2019). Besides, Miller & Dunson (2019) use approximate Bayesian computation to obtain a 'coarsened' posterior to achieve robustness and Grünwald (2012) proposes a safe Bayesian method.

Despite being an active research field, alleviating the impact of the model uncertainty in the class-conditional distribution (i.e., the likelihood conditional on the class) using ideas from distributional robustness is left largely unexplored even though this uncertainty arises naturally for numerous reasons. Even if we assume a proper parametric family, the plug-in estimator still carries statistical error from finite sampling and rarely matches the true distribution. The uncertainty is amplified when one relaxes to the nonparametric setting where no hardwired likelihood specification remains valid, and we are not aware of any guidance on a reasonable choice of a likelihood in this case. The situation deteriorates further when the training data violates the independent or identically distributed assumptions, or when the test distribution differs from the training distribution as in the setting

[1]Stanford University. Correspondence to: Viet Anh Nguyen <viet-anh.nguyen@stanford.edu>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Vienna, Austria, PMLR 119, 2020. 

of covariate shift (Gretton et al., 2009; Bickel et al., 2009; Moreno-Torres et al., 2012).

We endeavor in this paper to provide the precise mathematical model for binary classification with uncertain likelihood under the Bayesian decision analysis framework. Consider a binary classification setting where $Y \in \{0,1\}$ represents the random class label and $X \in \mathbb{R}^d$ represents the random features. With a new observation $x$ to be classified, we consider the problem of finding an optimal action $a \in \{0,1\}$,

$$a = \begin{cases} 0 & \text{if classify } x \text{ in class } 0, \\ 1 & \text{if classify } x \text{ in class } 1, \end{cases}$$

to minimize the probability of misclassification by solving the optimization problem

$$\min_{a\in\{0,1\}} \; a\mathbb{P}(Y=0|X=x) + (1-a)\mathbb{P}(Y=1|X=x),$$

where $\mathbb{P}(Y|X=x)$ denotes the posterior probability. If a class-proportion prior $\pi$ and the class-conditional (parametric) densities $f_0$ and $f_1$ are known, then this posterior probability can be calculated by using the Bayes' theorem (Schervish, 1995, Theorem 1.31). Unluckily, we rarely have access to the true conditional densities in real life.

To tackle this problem in the data-driven setting, for any class $c \in \{0,1\}$, the decision maker first forms, to the best of its belief and on the availability of data, a nominal class-conditional distribution $\widehat{\mathbb{P}}_c$. We assume now that the true class-conditional distribution belongs to an ambiguity set $\mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c)$, defined as a ball, prescribed via an appropriate measure of dissimilarity, of radius $\rho_c \geq 0$ centered at the nominal distribution $\widehat{\mathbb{P}}_c$ in the space of class-conditional probability measures. Besides, we allow to constrain the class-conditional distributions to lie in a subspace $\mathcal{P}$ of probability measures to facilitate the injection of optional parametric information, should the need arise.

To avoid any unnecessary measure theoretic complications, we position ourselves temporarily in the parametric setting and assume that we can generically write $\mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c) \cap \mathcal{P}$ parametrically as

$$\mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c) \cap \mathcal{P} = \{f_c(\,\cdot\,|\theta_c) : \theta_c \in \Theta_c\} \quad \forall c \in \{0,1\},$$

where $\Theta_c$ are non-empty (sub)sets on the finite-dimensional parameter space $\Theta$, and $\Theta_c$ satisfy the additional regularity condition that the density evaluated at point $x$ is strictly positive, i.e., $f_c(x|\theta_c) > 0$ for all $\theta_c \in \Theta_c$. Notice that the parametric subspace of probability distributions $\mathcal{P}$ is now explicitly described through the set of admissible parameters $\Theta$. If we denote the prior proportions by $\pi_0 = \pi(Y=0) > 0$ and $\pi_1 = \pi(Y=1) > 0$, then the ambiguity set over the posterior distributions induced by the class-conditional ambiguity sets $\mathbb{B}_0$ and $\mathbb{B}_1$ can be written as

$$\mathcal{B} = \left\{ \mathbb{P} : \begin{array}{c} \exists f_0 \in \mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0) \cap \mathcal{P}, f_1 \in \mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1) \cap \mathcal{P} : \\ \mathbb{P}(Y=c|X=x) = \dfrac{f_c(x)\pi_c}{\sum_{c'\in\{0,1\}} f_{c'}(x)\pi_{c'}} \; \forall c \end{array} \right\},$$

where the constraint in the set $\mathcal{B}$ links the class-conditional densities $f_c$ and the prior distribution of the class proportions $\pi$ to the posterior distribution. Facing with the uncertainty in the posterior distributions, it is reasonable to consider now the distributionally robust problem

$$\min_{a\in\{0,1\}} \sup_{\mathbb{P}\in\mathcal{B}} \; a\mathbb{P}(Y=0|X=x) + (1-a)\mathbb{P}(Y=1|X=x), \tag{1}$$

where the action $a$ is chosen so as to minimize the worst-case mis-classification probability over all posterior distribution $\mathbb{P} \in \mathcal{B}$. The next proposition asserts that the optimal action $a^\star$ belongs to the class of ratio decision rule.

**Proposition 1.1** (Optimal action)**.** *The optimal action that minimizes the worst-case mis-classification probability* (1) *has the form*

$$a^\star = \begin{cases} 1 & \textit{if } \dfrac{\sup_{f_1\in\mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1)\cap\mathcal{P}} f_1(x)}{\sup_{f_0\in\mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0)\cap\mathcal{P}} f_0(x)} \geq \tau(x), \\ 0 & \textit{otherwise,} \end{cases}$$

*for some threshold $\tau > 0$ that is dependent on $x$.*

Motivated by this insight from the parametric setting, we now promote the following classification decision rule

$$\mathcal{C}(x) = \begin{cases} 1 & \text{if } \mathcal{R}(x) \geq \tau(x), \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{R}(x)$ is the ratio defined as

$$\mathcal{R}(x) \triangleq \frac{\sup\limits_{\mathbb{Q}\in\mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1)\cap\mathcal{P}} \ell(x,\mathbb{Q})}{\sup\limits_{\mathbb{Q}\in\mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0)\cap\mathcal{P}} \ell(x,\mathbb{Q})},$$

and $\tau(x) > 0$ is a positive threshold which is potentially dependent on the observation $x$. The *score function* $\ell(x,\mathbb{Q})$ quantifies the plausibility of observing $x$ under the probability measure $\mathbb{Q}$, and the value $\mathcal{R}(x)$ quantifies how plausible an observation $x$ can be generated by *any* class-conditional probability distribution in $\mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1) \cap \mathcal{P}$ relatively to *any* distribution in $\mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0) \cap \mathcal{P}$. Because both the numerator and the denominator search for the distribution in the respective ambiguity set that maximizes the score of observing $x$, $\mathcal{R}(x)$ is thus termed the *ratio of optimistic scores*, and the classification decision $\mathcal{C}$ is hence called the *optimistic score ratio classifier*.

The classifying decision $\mathcal{C}(x)$ necessitates the solution of two *optimistic score evaluation problem*s of the form

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}})\cap\mathcal{P}} \quad \ell(x,\mathbb{Q}), \tag{2}$$

where the dependence of the input parameters on the label $c \in \{0,1\}$ has been omitted to avoid clutter. The performance of $\mathcal{C}$ depends critically on the specific choice of $\ell$ and $\mathbb{B}_\rho(\widehat{\mathbb{P}})$. Typically, $\ell$ is subjectively tailored to the choice of a parametric or a nonparametric view on the conditional distribution, as we shall see later on in this paper. The construction of $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ is principally governed by choice of the dissimilarity measure that specifies the $\rho$-neighborhood of the nominal distribution $\widehat{\mathbb{P}}$. Ideally, $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ should allow a coherent transition between the parametric and nonparametric setting via its interaction with the set $\mathcal{P}$. Furthermore, it should render problem (2) computationally tractable with meaningful optimal value, and at the same time provide the flexibility to balance between exerting statistical guarantees and modelling domain adaptation. These stringent criteria precludes the utilization of popular dissimilarity measures in the emerging literature. Indeed, the likelihood problem using the $f$-divergence (Ben-Tal et al., 2013; Namkoong & Duchi, 2016) delivers unreasonable estimate in the nonparametric setting (Nguyen et al., 2019a, Section 2), the Wasserstein distance (Mohajerin Esfahani & Kuhn, 2018; Kuhn et al., 2019; Blanchet et al., 2019; Gao & Kleywegt, 2016; Zhao & Guan, 2018) typically renders the Gaussian parametric likelihood problem non-convex, and the maximum mean discrepancy (Iyer et al., 2014; Staib & Jegelka, 2019) usually results in an infinite-dimensional optimization problem which is challenging to solve. This fact prompts us to explore an alternative construction of $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ that meets the criteria as mentioned above.

The contributions of this paper are summarized as follows.

- We introduce a novel ambiguity set based on a divergence defined on the space of mean vector and covariance matrix. We show that this divergence manifests numerous favorable properties and evaluating the optimistic score is equivalent to solving a non-convex optimization problem. We prove the asymptotic statistical guarantee of the divergence, which directs an optimal calibration the size of the ambiguity set.

- We show that, despite its inherent non-convexity and hence intractability, the optimistic score evaluation problem can be efficiently solved in both nonparametric and parametric Gaussian settings. We reveal that the optimistic score ratio classifier generalizes the Mahalanobis distance classifier and the linear/quadratic discriminant analysis.

Because evaluating the plausibility of an observation $x$ is a fundamental problem in statistics, the results of this paper have far-reaching implications beyond the scope of the classification task. These include Bayesian inference using synthetic likelihood (Wood, 2010; Price et al., 2018), approximate Bayesian computation (Csilléry et al., 2010; Toni et al., 2009), variational Bayes inference (Blei et al., 2017; Ong et al., 2018), and composite hypothesis testing using likelihood ratio (Cox, 1961; 2013). These connections will be explored in future research.

All proofs are relegated to the appendix.

**Notations.** We let $\mathcal{M}$ be the set of probability measures supported on $\mathbb{R}^d$ with finite second moment. The set of (symmetric) positive definite matrices is denoted by $\mathbb{S}_{++}^d$. For any $\mathbb{Q} \in \mathcal{M}$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{S}_{++}^d$, we use $\mathbb{Q} \sim (\mu, \Sigma)$ to express that $\mathbb{Q}$ has mean vector $\mu$ and covariance matrix $\Sigma$. The $d$-dimensional identity matrix is denoted by $I_d$. The space of Gaussian distributions is denoted by $\mathcal{N}$, and $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. The trace and determinant operator are denoted by $\mathrm{Tr}\left[A\right]$ and $\det(A)$, respectively.

## 2. Moment-based Divergence Ambiguity Set

We specifically study the construction of the ambiguity set using the following divergence on the space of moments.

**Definition 2.1** (Moment-based divergence)**.** *For any vectors $\mu_1$, $\mu_2 \in \mathbb{R}^d$ and matrices $\Sigma_1$, $\Sigma_2 \in \mathbb{S}_{++}^d$, the divergence from the tuple $(\mu_1, \Sigma_1)$ to the tuple $(\mu_2, \Sigma_2)$ amounts to*

$$\mathbb{D}\big((\mu_1,\Sigma_1) \parallel (\mu_2,\Sigma_2)\big) \triangleq (\mu_2-\mu_1)^\top \Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{Tr}\left[\Sigma_1\Sigma_2^{-1}\right] - \log\det(\Sigma_1\Sigma_2^{-1}) - d.$$

To avoid any confusion, it is worthy to note that contrary to the usual utilization of the term 'divergence' to specify a dissimilarity measure on the probability space, in this paper, the divergence is defined on the finite-dimensional space of mean vectors and covariance matrices.

It is straightforward to show that $\mathbb{D}$ is a divergence on $\mathbb{R}^d \times \mathbb{S}_{++}^d$ by noticing that $\mathbb{D}$ is a sum of the log-determinant divergence (Chebbi & Moakher, 2012) from $\Sigma_1$ to $\Sigma_2$ and a non-negative Mahalanobis distance between $\mu_1$ and $\mu_2$ weighed by $\Sigma_2$. As a consequence, $\mathbb{D}$ is non-negative, and perishes to 0 if and only if $\Sigma_1 = \Sigma_2$ and $\mu_1 = \mu_2$. With this property, $\mathbb{D}$ is an attractive candidate for the divergence on the joint space of mean vector and covariance matrix of $d$-dimensional random vectors. One can additionally verify that $\mathbb{D}$ is affine-invariant in the following sense. Let $\xi$ be a $d$-dimensional random vector and $\zeta$ be the affine-transformation of $\xi$, that is, $\zeta = A\xi + b$ for an invertible matrix $A$ and a vector $b$ of matching dimensions, then the value of the divergence $\mathbb{D}$ is preserved between the space of moments of $\xi$ and $\zeta$. In fact, if $\xi$ is a random vector with mean vector $\mu_j \in \mathbb{R}^d$ and covariance matrix $\Sigma_j \in \mathbb{S}_{++}^d$, then $\zeta$ has mean $A\mu_j + b$ and covariance matrix $A\Sigma_j A^\top$

for $j \in \{1,2\}$, and we have

$$\begin{aligned}&\mathbb{D}\big((\mu_1,\Sigma_1) \parallel (\mu_2,\Sigma_2)\big) \\ &= \mathbb{D}\big((A\mu_1+b, A\Sigma_1 A^\top) \parallel (A\mu_2+b, A\Sigma_2 A^\top)\big).\end{aligned} \tag{3}$$

A direct consequence is that $\mathbb{D}$ is also scale-invariant. Furthermore, the divergence $\mathbb{D}$ is closely related to the KL divergence[1], or the relative entropy, between two non-degenerate Gaussian distributions as

$$\mathbb{D}\big((\mu_1,\Sigma_1)\|(\mu_2,\Sigma_2)\big) = 2\,\mathrm{KL}\left(\mathcal{N}(\mu_1,\Sigma_1)\|\mathcal{N}(\mu_2,\Sigma_2)\right).$$

However, we emphasize that $\mathbb{D}$ is not symmetric, and in general $\mathbb{D}\big((\mu_1,\Sigma_1) \parallel (\mu_2,\Sigma_2)\big) \neq \mathbb{D}\big((\mu_2,\Sigma_2) \parallel (\mu_1,\Sigma_1)\big)$. Hence, $\mathbb{D}$ is not a distance on $\mathbb{R}^d \times \mathbb{S}_{++}^d$.

For any vector $\widehat{\mu} \in \mathbb{R}^d$, invertible matrix $\widehat{\Sigma} \in \mathbb{S}_{++}^d$ and radius $\rho \in \mathbb{R}_+$, we define the uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ over the mean vector and covariance matrix space as

$$\begin{aligned}&\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma}) \triangleq \\ &\quad \{(\mu,\Sigma) \in \mathbb{R}^d \times \mathbb{S}_{++}^d : \mathbb{D}\big((\widehat{\mu},\widehat{\Sigma}) \parallel (\mu,\Sigma)\big) \leq \rho\}.\end{aligned} \tag{4}$$

By definition, $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ includes all tuples $(\mu,\Sigma)$ which is of a divergence not bigger than $\rho$ from the tuple $(\widehat{\mu},\widehat{\Sigma})$. Because $\mathbb{D}$ is not symmetric, it is important to note that $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is defined with the tuple $(\widehat{\mu},\widehat{\Sigma})$ being the first argument of the divergence $\mathbb{D}$, and this uncertainty set can be written in a more expressive form as

$$\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma}) = \left\{ \begin{array}{l} (\mu,\Sigma) \in \mathbb{R}^d \times \mathbb{S}_{++}^d : \\ (\mu-\widehat{\mu})^\top \Sigma^{-1}(\mu-\widehat{\mu}) + \mathrm{Tr}\left[\widehat{\Sigma}\Sigma^{-1}\right] + \log\det\Sigma \leq \overline{\rho} \end{array} \right\}$$

for a scalar $\overline{\rho} \triangleq \rho + d + \log\det\widehat{\Sigma}$. Moreover, one can assert that $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$ is non-convex due to the log-determinant term, and this non-convexity cannot be eliminated using the reparametrization to the space of inverse covariance matrices (or equivalently called the precision matrices).

Equipped with $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, the ambiguity set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ is systematically constructed as follows. If the nominal distribution $\widehat{\mathbb{P}}$ admits a nominal mean vector $\widehat{\mu}$ and a nominal nondegenerate covariance matrix $\widehat{\Sigma}$, then $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ is a ball that contains all probability measures whose mean vector and covariance matrix are contained in $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, that is,

$$\mathbb{B}_\rho(\widehat{\mathbb{P}}) \triangleq \{\mathbb{Q} \in \mathcal{M} : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})\}. \tag{5}$$

The set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$, by construction, differentiates only through the information about the first two moments: if a distribution $\mathbb{Q}$ belongs to $\mathbb{B}_\rho(\widehat{\mathbb{P}})$, then *any* distribution $\mathbb{Q}'$ with the same mean vector and covariance matrix with $\mathbb{Q}$ also belongs to $\mathbb{B}_\rho(\widehat{\mathbb{P}})$. Further, $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ embraces all types of probability distributions, including discrete, continuous and even mixed continuous/discrete distributions.

We now delineate a principled approach to solve the optimistic score evaluation problem (2) for a generic score function $\ell : \mathbb{R}^d \times \mathcal{M} \to \mathbb{R}$. We denote by $\mathcal{M}(\mu,\Sigma)$ the Chebyshev ambiguity set that contains all probability measures with *fixed* mean vector $\mu \in \mathbb{R}^d$ and *fixed* covariance matrix $\Sigma \in \mathbb{S}_{++}^d$, that is,

$$\mathcal{M}(\mu,\Sigma) \triangleq \{\mathbb{Q} \in \mathcal{M} : \mathbb{Q} \sim (\mu,\Sigma)\}.$$

The moment-based divergence ambiguity set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ then admits an equivalent representation

$$\mathbb{B}_\rho(\widehat{\mathbb{P}}) = \bigcup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \mathcal{M}(\mu,\Sigma),$$

which is an infinite union of Chebyshev ambiguity sets, where the union operator is taken over all tuples of mean vector-covariance matrix belonging to $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$. Leveraging on this representation, problem (2) can now be decomposed as a two-layer optimization problem

$$\sup_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}})} \ell(x,\mathbb{Q}) = \sup_{(\mu,\Sigma)\in\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})} \sup_{\mathbb{Q}\in\mathcal{M}(\mu,\Sigma)\cap\mathcal{P}} \ell(x,\mathbb{Q}). \tag{6}$$

The inner subproblem of (6) is a distributionally robust optimization problem with a Chebyshev second moment ambiguity set, hence there is a strong potential to exploit existent results from the literature, see Delage & Ye (2010) and Wiesemann et al. (2014), to reformulate this inner problem into a finite dimensional convex optimization problem. Unfortunately, the outer subproblem of (6) is a robust optimization problem over a non-convex uncertainty set $\mathcal{U}_\rho(\widehat{\mu},\widehat{\Sigma})$, thus the two-layer decomposition problem (6) remains computationally intractable in general. As a direct consequence, solving the optimistic score evaluation problem requires an intricate adaptation of non-convex optimization techniques applied on a case-by-case basis. Two exemplary settings in which problem (6) can be efficiently solved will be depicted subsequently in Sections 3 and 4.

We complete this section by providing the asymptotic statistical guarantees of the divergence $\mathbb{D}$, which serves as a potential guideline for the construction of the ambiguity set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ and the tuning of the radius parameter $\rho$.

**Theorem 2.2** (Asymptotic guarantee of $\mathbb{D}$)**.** *Suppose that a $d$-dimensional random vector $\xi$ has mean vector $m \in \mathbb{R}^d$, covariance matrix $S \in \mathbb{S}_{++}^d$ and admits finite fourth moment under a probability measure $\mathbb{P}$. Let $\widehat{\xi}_t \in \mathbb{R}^d$, $t = 1,\ldots,n$ be independent and identically distributed samples*

[1] If $\mathbb{Q}_1$ is absolutely continuous with respect to $\mathbb{Q}_2$, then the Kullback-Leibler divergence from $\mathbb{Q}_1$ to $\mathbb{Q}_2$ amounts to $\mathrm{KL}(\mathbb{Q}_1 \parallel \mathbb{Q}_2) \triangleq \mathbb{E}_{\mathbb{Q}_1}[\log \mathrm{d}\mathbb{Q}_1/\mathrm{d}\mathbb{Q}_2]$, where $\mathrm{d}\mathbb{Q}_1/\mathrm{d}\mathbb{Q}_2$ is the Radon-Nikodym derivative of $\mathbb{Q}_1$ with respect to $\mathbb{Q}_2$.

*of $\xi$ from $\mathbb{P}$. Denote by $\widehat{\mu}_n \in \mathbb{R}^d$ and $\widehat{\Sigma}_n \in \mathbb{S}_+^d$ the sample mean vector and sample covariance matrix defined as*

$$\widehat{\mu}_n = \frac{1}{n}\sum_{t=1}^{n}\widehat{\xi}_t, \quad \widehat{\Sigma}_n = \frac{1}{n}\sum_{t=1}^{n}(\widehat{\xi}_t - \widehat{\mu}_n)(\widehat{\xi}_t - \widehat{\mu}_n)^\top. \quad (7)$$

*Let $\eta = S^{-\frac{1}{2}}(\xi - m)$ be the isotropic transformation of the random vector $\xi$, let $H$ be a $d$-dimensional Gaussian random vector with mean vector 0 and covariance matrix $I_d$, and let $Z$ be a $d$-by-$d$ random symmetric matrix with the upper triangle component $Z_{jk}$ $(j \leq k)$ following a Gaussian distribution with mean 0 and the covariance coefficient between $Z_{jk}$ and $Z_{j'k'}$ is*

$$\mathrm{cov}(Z_{jk}, Z_{j'k'}) = \mathbb{E}_{\mathbb{P}}[\eta_j\eta_k\eta_{j'}\eta_{k'}] - \mathbb{E}_{\mathbb{P}}[\eta_j\eta_k]\,\mathbb{E}_{\mathbb{P}}[\eta_{j'}\eta_{k'}].$$

*Furthermore, $H$ and $Z$ are jointly Gaussian distributed with the covariance between $H_i$ and $Z_{jk}$ as*

$$\mathrm{cov}(H_i, Z_{jk}) = \mathbb{E}_{\mathbb{P}}[\eta_i\eta_j\eta_k].$$

*As $n \uparrow \infty$, we have*

$$n \times \mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big) \longrightarrow H^\top H + \frac{1}{2}\,\mathrm{Tr}\left[Z^2\right] \quad \textit{in distribution.}$$

We were not able to locate Theorem 2.2 in the existing literature. Interestingly, Theorem 2.2 also sheds light upon the asymptotic behavior of the KL divergence from an empirical Gaussian distribution $\mathcal{N}(\widehat{\mu}_n, \widehat{\Sigma}_n)$ to the data-generating Gaussian distribution $\mathcal{N}(m, S)$.

**Corollary 2.3** (Asymptotic guarantee of $\mathbb{D}$ – Gaussian distributions)**.** *Suppose that $\widehat{\xi}_t \in \mathbb{R}^d$, $t = 1, \ldots, n$ are independent and identically distributed samples of $\xi$ from $\mathbb{P} = \mathcal{N}(m, S)$ for some $m \in \mathbb{R}^d$ and $S \in \mathbb{S}_{++}^d$. Let $\widehat{\mu}_n \in \mathbb{R}^d$ and $\widehat{\Sigma}_n \in \mathbb{S}_+^d$ be the sample mean vector and covariance matrix defined as in (7). As $n \uparrow \infty$, we have*

$$n \times \mathrm{KL}\left(\mathcal{N}(\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel \mathcal{N}(m, S)\right) \longrightarrow \frac{1}{2}\chi^2\left(d(d+3)/2\right) \quad \textit{in distribution,}$$

*where $\chi^2\left(d(d+3)/2\right)$ is a chi-square distribution with $d(d+3)/2$ degrees of freedom.*

If we use independent and identically distributed (i.i.d.) samples to estimate the nominal mean vector and covariance matrix of $\widehat{\mathbb{P}}$, then the radius $\rho$ should be asymptotically scaled at the rate $n^{-1}$ as the sample size $n$ increases. Indeed, Theorem 2.2 and Corollary 2.3 suggest that $n^{-1}$ is the optimal asymptotic rate which ensures that the true but unknown mean vector and covariance matrix of the data-generating distribution fall into the set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ with high probability. While the limiting distribution under the Gaussian setting is a typical chi-square distribution, the general limiting distribution $H^\top H + \mathrm{Tr}\left[Z^2\right]/2$ in Theorem 2.2 does not have any analytical form. This limiting distribution can be numerically approximated, for example, via Monte Carlo simulations. If the i.i.d. assumption of the training samples is violated or if we expect a covariate shift at test time, then the radius $\rho$ reflects the modeler's belief regarding the moment mismatch measured using the divergence $\mathbb{D}$ and in this case, the radius $\rho$ should be considered as an exogenous input to the problem.

For illustrative purpose, we fix dimension $d = 20$ and consider the random vector $\xi = C\zeta + m$, where entries of $\zeta$ are mutually independent and the $i$-th entry follows a normalized chi-square distribution, i.e., $\zeta_i \sim (\chi^2(1) - 1)/\sqrt{2}$. Then the covariance matrix of $\xi$ is $S = CC^\top$. Notice that by the identity (3), $\mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big)$ is invariant of the choice of $C$ and $m$. We generate 10,000 datasets, each contains $n$ i.i.d. samples of $\xi$ and calculate for each dataset the empirical values of $n \times \mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big)$. We plot in Figure 1 the empirical distribution of $n \times \mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big)$ using 10,000 datasets versus the limiting distribution of $H^\top H + \mathrm{Tr}\left[Z^2\right]/2$ for different values of $n$. One can observe that for a small sample size ($n < 100$), there is a perceivable difference between the finite sample distribution and the limiting distribution, but as $n$ becomes larger ($n > 100$), this mismatch is significantly reduced.

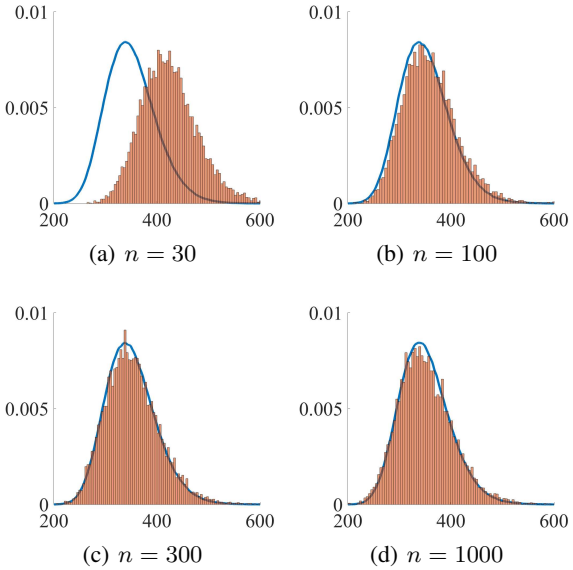



*Figure 1.* Empirical distribution of $n \times \mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big)$ collected from 10,000 datasets (orange histogram) versus the limiting distribution $H^\top H + \frac{1}{2}\,\mathrm{Tr}\left[Z^2\right]$ obtained by Monte Carlo simulations (blue curve) for different sample sizes $n$.

## 3. Optimistic Nonparametric Score

We consider in this section the nonparametric setting in which no prior assumption on the class-conditional distribution is imposed. A major difficulty in this nonparametric setting is the elicitation of a reasonable score function $\ell$ that can coherently encapsulate the plausibility of observing $x$ over the whole spectrum of admissible $\mathbb{Q}$, including continuous, discrete and mixed continuous/discrete distributions, while at the same time being amenable for optimization purposes. Taking this fact into consideration, we thus posit to choose the score function of the form

$$\ell(x, \mathbb{Q}) \equiv \mathbb{Q}(\{x\}),$$

which is the probability value of the singleton, measurable set $\{x\}$ under the measure $\mathbb{Q}$. If $\mathbb{Q}$ is a continuous distribution, then apparently $\mathbb{Q}(\{x\})$ is zero, hence this score function is admittedly not perfect. Nevertheless, it serves as a sensible proxy in the nonparametric setting and delivers competitive performance in machine learning tasks (Nguyen et al., 2019b). It is reasonable to set $\mathcal{P} \equiv \mathcal{M}$ in the nonparametric setting, and with this choice of $\ell$, the optimistic nonparametric score evaluation problem becomes

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}})} \mathbb{Q}(\{x\}),$$

which is inherently challenging because it is an infinite-dimensional optimization problem. The next theorem asserts that solving the nonparametric optimistic likelihood optimization problem is equivalent to solving a univariate convex optimization problem.

**Theorem 3.1** (Optimistic nonparametric probability)**.** *Suppose that $\widehat{\mathbb{P}} \sim (\widehat{\mu}, \widehat{\Sigma})$ for some $\widehat{\mu} \in \mathbb{R}^d$ and $\widehat{\Sigma} \in \mathbb{S}_{++}^d$. For any $\rho \in \mathbb{R}_+$, we have*

$$\begin{aligned} &\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}})} \mathbb{Q}(\{x\}) \\ &= \max_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} [1 + (\mu - x)^\top \Sigma^{-1} (\mu - x)]^{-1} \qquad (8a) \\ &= [1 + (\mu^\star - x)^\top (\Sigma^\star)^{-1} (\mu^\star - x)]^{-1}, \qquad (8b) \end{aligned}$$

*where $(\mu^\star, \Sigma^\star) \in \mathbb{R}^d \times \mathbb{S}_{++}^d$ satisfies*

$$\begin{aligned} \mu^\star &= \frac{x + \gamma^\star \widehat{\mu}}{1 + \gamma^\star}, \\ \Sigma^\star &= \widehat{\Sigma} + \frac{1}{(1 + \gamma^\star)} (x - \widehat{\mu})(x - \widehat{\mu})^\top, \end{aligned} \qquad (8c)$$

*and $\gamma^\star \in \mathbb{R}_+$ solves the univariate convex optimization problem*

$$\min_{\gamma \geq 0} \ \gamma \rho - \gamma \log \Big( 1 + \frac{(x - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (x - \widehat{\mu})}{1 + \gamma} \Big). \qquad (8d)$$

Because the feasible set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ is not weakly compact, the existence of an optimal measure that solves the optimistic likelihood problem on the left-hand side of (8a) is not trivial. However, equation (8a) asserts that this optimal measure exists, and it can be constructed by solving a non-convex optimization problem over the mean vector-covariance matrix tuple $(\mu, \Sigma)$. Notice that (8a) is a non-convex optimization problem because $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is a non-convex set. Surprisingly, one can show that the optimizer of (8a) can be found semi-analytically: the maximizer $(\mu^\star, \Sigma^\star)$ depends only on a single scalar $\gamma^\star$ through (8c), where $\gamma^\star$ solves the univariate optimization problem (8d). Because problem (8d) is convex, $\gamma^\star$ can be efficiently found using a bisection algorithm or using a Newton-Raphson method, and we expose in Appendix E the first- and second-order derivative of the objective function of (8d).

A nonparametric classifier $\mathcal{C}_{\text{nonparam}}$ can be formed by utilizing the optimistic nonparametric score ratio

$$\mathcal{R}_{\text{nonparam}}(x) \triangleq \frac{\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1)} \mathbb{Q}(\{x\})}{\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0)} \mathbb{Q}(\{x\})}, \qquad (9)$$

where each nominal class-conditional distribution $\widehat{\mathbb{P}}_c$ has mean vector $\widehat{\mu}_c \in \mathbb{R}^d$ and covariance matrix $\widehat{\Sigma}_c \in \mathbb{S}_{++}^d$, and each ambiguity set $\mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c)$ is defined as in (5). The results of Theorem 3.1 can be used to compute the numerator and denominator of (9), thus the classification decision $\mathcal{C}_{\text{nonparam}}(x)$ can be efficiently evaluated. In particular, by substituting the expression (8a) into (9), we also find

$$\begin{aligned} &\mathcal{R}_{\text{nonparam}}(x) \\ &= \frac{\max_{(\mu,\Sigma) \in \mathcal{U}_{\rho_1}(\widehat{\mu}_1, \widehat{\Sigma}_1)} [1 + (\mu - x)^\top \Sigma^{-1} (\mu - x)]^{-1}}{\max_{(\mu,\Sigma) \in \mathcal{U}_{\rho_0}(\widehat{\mu}_0, \widehat{\Sigma}_0)} [1 + (\mu - x)^\top \Sigma^{-1} (\mu - x)]^{-1}} \\ &= \frac{1 + \min_{(\mu,\Sigma) \in \mathcal{U}_{\rho_0}(\widehat{\mu}_0, \widehat{\Sigma}_0)} (\mu - x)^\top \Sigma^{-1} (\mu - x)}{1 + \min_{(\mu,\Sigma) \in \mathcal{U}_{\rho_1}(\widehat{\mu}_1, \widehat{\Sigma}_1)} (\mu - x)^\top \Sigma^{-1} (\mu - x)}. \end{aligned}$$

Suppose that $\rho_0 = \rho_1 = 0$ and $\tau(x) = 1$, then the nonparametric classifier assigns $\mathcal{C}_{\text{nonparam}}(x) = 1$ whenever

$$(\widehat{\mu}_1 - x)^\top \widehat{\Sigma}_1^{-1} (\widehat{\mu}_1 - x) \leq (\widehat{\mu}_0 - x)^\top \widehat{\Sigma}_0^{-1} (\widehat{\mu}_0 - x),$$

and $\mathcal{C}_{\text{nonparam}}(x) = 0$ otherwise. In this case, the classifier coincides with the class-specific Mahalanobis distance classifier (MDC) where $\widehat{\Sigma}_0$ and $\widehat{\Sigma}_1$ denote the intra-class nominal covariance matrices. If in addition the nominal covariance matrices are homogeneous, that is, $\widehat{\Sigma}_0 = \widehat{\Sigma}_1$, then this classifier coincides with the Linear Discriminant Analysis (LDA) (Murphy, 2012, Section 4.2.2). The Bayesian version of LDA can be equivalently obtained from $\mathcal{C}_{\text{nonparam}}$

by setting a proper value of $\tau(x)$. This important observation reveals an intimate link between our proposed classifier $\mathcal{C}_{\text{nonparam}}$ using the optimistic nonparametric score ratio and the popular classifiers MDC and LDA. On the one hand, $\mathcal{C}_{\text{nonparam}}$ can now be regarded as a generalization of MDC and LDA, which takes into account the statistical imprecision of the estimated moments and/or the potential shift in the moment statistics in test data versus training data distributions. On the other hand, both MDC and LDA now admit a nonparametric, generative interpretation in which the class-conditional distribution is chosen in the set of all distributions with the same first- and second-moments as the nominal class-conditional measure $\widehat{\mathbb{P}}_c$. This novel interpretation goes beyond the classical Gaussian model, and it potentially explains the versatile performance of MDC and LDA when the conditional distribution are not normally distributed as empirically observed in (Lee et al., 2018).

## 4. Optimistic Gaussian Score

We now consider the optimistic score evaluation problem under a parametric setting. For simplicity, we assume that the true class-conditional distributions of the feature belong to the family of Gaussian distributions. Thus, a natural choice of the score value $\ell(x, \mathbb{Q})$ in this case is the Gaussian likelihood of an observation $x$ when $\mathbb{Q}$ is a Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$, that is,

$$\ell(x, \mathbb{Q}) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right).$$

It is also suitable to set $\mathcal{P}$ in problem (2) to the (sub)space of Gaussian distributions $\mathcal{N}$ and consider the following optimistic Gaussian score evaluation problem

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}) \cap \mathcal{N}} \ell(x, \mathbb{Q}). \tag{10}$$

One can verify that the maximizer of problem (10) coincides with the maximizer of

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}) \cap \mathcal{N}} \mathcal{L}(x, \mathbb{Q}),$$

where $\mathcal{L}$ is the translated Gaussian *log*-likelihood defined as

$$\begin{aligned}\mathcal{L}(x, \mathbb{Q}) &= 2\left(\log\left(\ell(x, \mathbb{Q})\right) + \frac{d}{2}\log(2\pi)\right) \\ &= -(\mu - x)^\top \Sigma^{-1}(\mu - x) - \log\det\Sigma.\end{aligned}$$

Theorem 4.1 is a counterpart to the optimistic nonparametric likelihood presented in Theorem 3.1.

**Theorem 4.1** (Optimistic Gaussian log-likelihood)**.** *Suppose that $\widehat{\mathbb{P}} \sim \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ for some $\widehat{\mu} \in \mathbb{R}^d$ and $\widehat{\Sigma} \in \mathbb{S}_{++}^d$. For any $\rho \in \mathbb{R}_+$, we have*

$$\begin{aligned}&\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}) \cap \mathcal{N}} \mathcal{L}(x, \mathbb{Q}) \\ &= \max_{(\mu,\Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} -(\mu - x)^\top \Sigma^{-1}(\mu - x) - \log\det\Sigma \qquad (11a) \\ &= -(\mu^\star - x)^\top (\Sigma^\star)^{-1}(\mu^\star - x) - \log\det\Sigma^\star, \qquad (11b)\end{aligned}$$

*where $(\mu^\star, \Sigma^\star) \in \mathbb{R}^d \times \mathbb{S}_{++}^d$ satisfies*

$$\begin{aligned}\mu^\star &= \frac{x + \gamma^\star \widehat{\mu}}{1 + \gamma^\star}, \\ \Sigma^\star &= \frac{\gamma^\star}{1+\gamma^\star}\widehat{\Sigma} + \frac{\gamma^\star}{(1+\gamma^\star)^2}(x - \widehat{\mu})(x - \widehat{\mu})^\top,\end{aligned} \tag{11c}$$

*and $\gamma^\star \in \mathbb{R}_+$ solves the univariate convex optimization problem*

$$\begin{aligned}\min_{\gamma \geq 0} \Big\{ &\gamma\rho + d(\gamma+1)\log\left(1 + \frac{1}{\gamma}\right) \\ &- (1+\gamma)\log\left(1 + \frac{(x-\widehat{\mu})^\top \widehat{\Sigma}^{-1}(x-\widehat{\mu})}{(1+\gamma)}\right)\Big\}.\end{aligned} \tag{11d}$$

Notice that we impose the condition $\widehat{\mathbb{P}} \sim \mathcal{N}(\widehat{\mu}, \widehat{\Sigma})$ in Theorem 4.1 to conform with the belief that the true data generating distribution is Gaussian. This condition, in fact, can be removed without affecting the result presented in Theorem 4.1. Indeed, for any radius $\rho \geq 0$, the ambiguity set $\mathbb{B}_\rho(\widehat{\mathbb{P}})$ by definition contains a Gaussian distribution with the same mean vector and covariance matrix with the nominal distribution $\widehat{\mathbb{P}}$, and thus the feasible set of (10) is always non-empty and the value of the optimistic Gaussian log-likelihood is always finite. In Appendix E, we provide the first- and second-order derivatives of the objective function of (11d), which can be exploited to derive efficient algorithm to solve the convex program (11d).

Returning to the construction of the classifier, one can now construct the classifier $\mathcal{C}_{\mathcal{N}}(x)$ using the optimistic Gaussian score ratio $\mathcal{R}_{\mathcal{N}}(x)$ expressed by

$$\begin{aligned}\mathcal{R}_{\mathcal{N}}(x) &\triangleq \frac{\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1) \cap \mathcal{N}} \ell(x, \mathbb{Q})}{\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0) \cap \mathcal{N}} \ell(x, \mathbb{Q})} \\ &= \frac{\exp\left(\frac{1}{2}\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_1}(\widehat{\mathbb{P}}_1) \cap \mathcal{N}} \mathcal{L}(x, \mathbb{Q})\right)}{\exp\left(\frac{1}{2}\sup_{\mathbb{Q} \in \mathbb{B}_{\rho_0}(\widehat{\mathbb{P}}_0) \cap \mathcal{N}} \mathcal{L}(x, \mathbb{Q})\right)},\end{aligned} \tag{12}$$

where each nominal distribution $\widehat{\mathbb{P}}_c$ is a Gaussian distribution with mean vector $\widehat{\mu}_c \in \mathbb{R}^d$ and covariance matrix $\widehat{\Sigma}_c \in \mathbb{S}_{++}^d$, and each ambiguity set $\mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c)$ is defined as

in (5). Theorem 4.1 can be readily applied to evaluate the value $\mathcal{R}_{\mathcal{N}}(x)$, and classify $x$ using $\mathcal{C}_{\mathcal{N}}(x)$. Furthermore, suppose that $\rho_0 = \rho_1 = 0$ and $\tau(x) = 1$, then the resulting classifier recovers the Quadratic Discriminant Analysis (Murphy, 2012, Section 4.2.1). The Bayesian version of the QDA can be equivalently obtained from $\mathcal{C}_{\mathcal{N}}$ by setting a proper value for $\tau(x)$.

It is imperative to elaborate on the improvement of Theorem 4.1 compared to the result reported in Nguyen et al. (2019a, Section 3). While both results are related to the evaluation of the optimistic Gaussian log-likelihood, Nguyen et al. (2019a, Theorem 3.2) restricts the mean vector to its nominal value and optimizes only over the covariance matrix. On the other hand, Theorem 4.1 of this paper optimizes over both the mean vector and the covariance matrix, thus provides full flexibility to choose the optimal values of *all* sufficient statistics of the family of Gaussian distributions. From a technical standpoint, the non-convexity is overcome in Nguyen et al. (2019a, Theorem 3.2) through a simple change of variables; nonetheless, the proof of Theorem 4.1 demands an additional layer of duality arguments to disentangle the multiplicative terms between $\mu$ and $\Sigma$ in both the objective function $\mathcal{L}$ and the divergence $\mathbb{D}$. By inspecting the expressions in (11c), one can further notice that in general the optimal solution $\mu^\star$ is distinct from the nominal mean $\widehat{\mu}$, this observation suggests that optimizing *jointly* over $(\mu, \Sigma)$ is indeed more powerful than optimizing simply over $\Sigma$ from a theoretical perspective.

## 5. Numerical Results

All experiments are run on a standard laptop with 1.4 GHz Intel Core i5 and 8GB of memory, the codes and datasets are available at `https://github.com/nian-si/bsc`.

### 5.1. Decision Boundaries

In this section, we visualize the classification decision boundaries generated by the classifiers $\mathcal{C}_{\text{nonparam}}$ proposed in Section 3 and $\mathcal{C}_{\mathcal{N}}$ proposed in Section 4 using synthetic data. To ease the exposition, we consider a two dimensional feature space $d = 2$ and the class-conditional distributions are Gaussian of the form

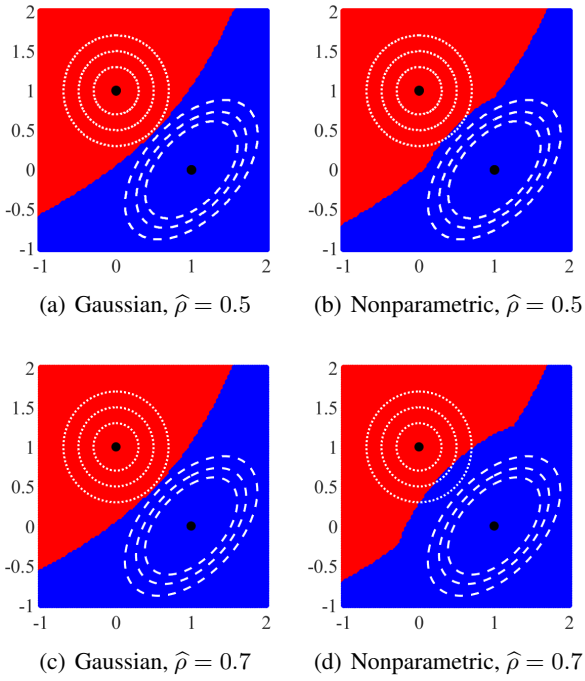
$$X|Y=0 \sim \mathcal{N}\left(\begin{bmatrix}0\\1\end{bmatrix}, I_d\right), X|Y=1 \sim \mathcal{N}\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right).$$

We sample i.i.d. data points $\{\widehat{x}_{c,i}\}_{i=1}^{n_c}$ in each $c \in \{0,1\}$ with $n_0 = n_1 = 1000$ as the training set, then estimate the nominal mean $\widehat{\mu}_c$ and the nominal covariance matrix $\widehat{\Sigma}_c$ for each class $c \in \{0,1\}$ using the sample average formula (7).

We first consider when the ambiguity sets have the same radius, i.e., $\rho_0 = \rho_1 = \widehat{\rho}$ and fix the threshold $\tau(x) = 1$ for every $x$. Figure 2 shows the optimistic Gaussian and nonparametric decision boundaries for $\widehat{\rho} \in \{0.5, 0.7\}$. We find that for optimistic Gaussian decision rule, the decision boundaries look similar across different radii; while in nonparametric case, the decision boundaries exhibit different shapes. We then consider the case with distinct radii by setting $\vec{\rho} = (\rho_0, \rho_1) = (0.1, 1.0)$ and $\vec{\rho} = (\rho_0, \rho_1) = (1.0, 0.1)$. Further, we fix the threshold to a constant $\tau(x) = \tau^\star$ for a scalar $\tau^\star \in \mathbb{R}_+$ that solves

$$\max_{\tau \geq 0} \sum_{i=1}^{n_0} \mathbf{1}\{\mathcal{R}(\widehat{x}_{0,i}) < \tau\} + \sum_{i=1}^{n_1} \mathbf{1}\{\mathcal{R}(\widehat{x}_{1,i}) \geq \tau\}, \quad (13)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The decision boundaries are plotted in Figure 3. We find the decision boundaries have different shapes for different decision rules and for different choices of radii.



(a) Gaussian, $\widehat{\rho} = 0.5$ (b) Nonparametric, $\widehat{\rho} = 0.5$

(c) Gaussian, $\widehat{\rho} = 0.7$ (d) Nonparametric, $\widehat{\rho} = 0.7$

*Figure 2.* Decision boundaries for different $\widehat{\rho}$. Red/blue regions indicate the class partitions, black dots locate the mean, and white dashed ellipsoids draw the class-conditional density contours.

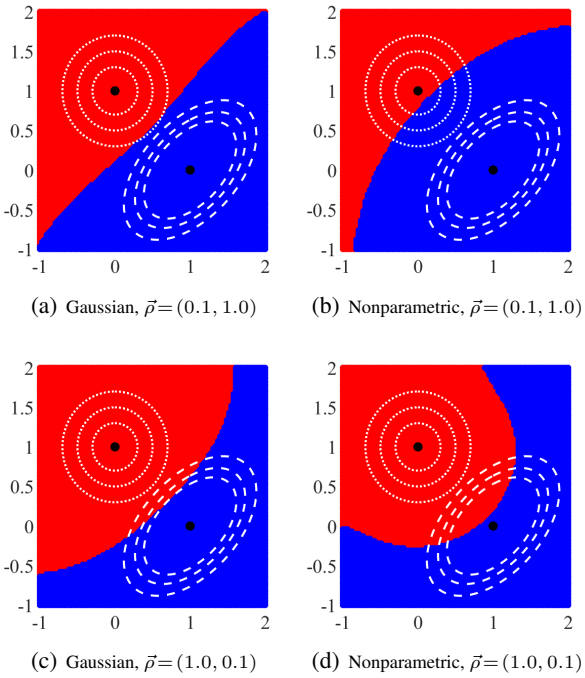
### 5.2. Real Data Experiments

In our experiments, we first compute the nominal mean by empirical average and we use the Ledoit-Wolf covariance estimator (Ledoit & Wolf, 2004) to compute a well-conditioned nominal covariance matrix. We experiment two methods of tuning the radii $\rho$ of the ambiguity sets: using cross-validation on training data, or using the quantile of the limiting distribution in Corollary 2.3. Specifically, for the second criteria, we choose

$$\rho_c = n_c^{-1} \chi^2_\alpha(d(d+3)/2) \qquad \forall c \in \{0,1\},$$

*Table 1.* Correct classification rate on the benchmark date sets. **Bold** number corresponds to the best performance in each dataset.

| DATASET | GQDA, CV | GQDA, CLT | NPQDA, CV | NPQDA, CLT | KQDA | RQDA | SQDA |
|---|---|---|---|---|---|---|---|
| AUSTRALIAN | 85.38 | 84.91 | **85.84** | 85.03 | 85.38 | 85.61 | 85.37 |
| BANKNOTE | **99.83** | 99.33 | 99.3 | **99.83** | 99.8 | 99.77 | **99.83** |
| CLIMATE MODEL | **94.81** | 89.33 | 93.92 | 90.37 | 92.59 | 94.07 | 93.92 |
| CYLINDER | 71.04 | 70.81 | **71.11** | 70.89 | **71.11** | **71.11** | 70.67 |
| DIABETIC | 75.52 | 73.49 | **76.09** | 76.30 | 75.47 | 75.00 | 74.32 |
| FOURCLASS | 77.82 | 79.26 | **80.28** | 78.98 | 78.38 | 79.07 | 78.84 |
| HABERMAN | 74.93 | 75.33 | **75.45** | **75.45** | 74.94 | 74.80 | 74.16 |
| HEART | 81.76 | 83.09 | 83.09 | 81.91 | 81.76 | 81.91 | **83.68** |
| HOUSING | 91.66 | 90.55 | 91.81 | 91.50 | 91.66 | 91.66 | **91.97** |
| ILPD | 68.84 | **69.52** | 69.25 | 68.15 | 69.18 | 67.94 | **69.52** |
| MAMMOGRAPHIC MASS | **80.39** | 80.00 | 79.90 | 79.61 | 79.95 | 80.05 | 80.24 |



(a) Gaussian, $\vec{\rho}=(0.1, 1.0)$ (b) Nonparametric, $\vec{\rho}=(0.1, 1.0)$

(c) Gaussian, $\vec{\rho}=(1.0, 0.1)$ (d) Nonparametric, $\vec{\rho}=(1.0, 0.1)$

*Figure 3.* Decision boundaries with distinct radii. Indications are verbatim from Figure 2.

where $n_c$ is the number of training samples in class $c$ and $\chi^2_\alpha(d(d+3)/2)$ is the $\alpha$-quantile of the chi-square distribution with $d(d+3)/2$ degrees of freedom. Notice that for large degrees of freedom, the chi-square distribution concentrates around the mean, because a chi-square random variable with $k$ degrees of freedom is the sum of $k$ i.i.d. $\chi^2(1)$. The optimal asymptotic value of the radius $\rho_c$ is therefore insensitive to the choice of $\alpha$, so we select numerically $\alpha = 0.5$ in our experiments. We tune the threshold to maximize the training accuracy following (13) after computing the ratio value for each training sample. The whole procedure is summarized in Algorithm 1. In particular, this algorithm trains the parameters using only one pass over the training samples, which makes it significantly faster than the cross-validation approach. We observe empirically in most cases that the performance of classifying

**Algorithm 1** Optimistic score ratio classification

**Input:** datasets $\{\widehat{x}_{c,i}\}_{i=1}^{n_c}$ for $c \in \{0, 1\}$. A test data $x$.
Compute the nominal mean and the nominal covariance matrix
Compute the radii $\rho_c \leftarrow n_c^{-1}\chi^2_{0.5}(d(d+3)/2)$.
Compute the optimistic ratio $\mathcal{R}(\widehat{x}_{c,i})$ for every $\widehat{x}_{c,i}$
Compute the threshold $\tau^\star$ that solves (13).
**Output:** classification label $\mathbf{1}\{\mathcal{R}(x) \geq \tau\}$.

using Algorithm 1 is comparable in terms of test accuracy to classifying with cross-validating on the tuning parameters.

We test the performance of our classification rules on various datasets from the UCI repository (Dua & Graff, 2017). Specifically, we compare the following methods:

- Gaussian QDA (**GQDA**) and Nonparametric QDA (**NPQDA**): Our classifiers $\mathcal{C}_{\mathcal{N}}$ and $\mathcal{C}_{\text{nonparam}}$;
- Kullback-Leibler QDA (**KQDA**): The classifier based on KL ambiguity sets with fixed mean (Nguyen et al., 2019a);
- Regularized QDA (**RQDA**): The regularized QDA based on the linear shrinkage covariance estimator $\widehat{\Sigma}_c + \rho_c I_d$;
- Sparse QDA (**SQDA**): The sparse QDA based on the graphical lasso covariance estimator (Friedman et al., 2008) with parameter $\rho_c$.

For **GQDA** and **NPQDA**, we also compare the performance of different strategies to choose the radii $\rho$ using cross-validation (CV) and selection based on Theorem 2.2 (CLT). For all methods that need cross-validation, we randomly select 75% of the data for training and the remaining 25% for testing. The size of the ambiguity sets and the regularization parameter are selected using stratified 5-fold cross-validation. Furthermore, to promote a fair comparison, we tune the threshold for every method using (13). The performance of the classifiers is measured by the average *correct classification rate* (CCR) on the validation set. The average CCR score over 10 trials are reported in Table 1.

## Acknowledgements

We gratefully acknowledge support from the following NSF grants 1915967, 1820942, 1838576, as well as AFOSR MURI 19RT105 and the China Merchants Bank.

## References

Ben-Tal, A., den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Bhagat, S., Cormode, G., and Muthukrishnan, S. Node classification in social networks. In *Social Network Data Analytics*, pp. 115–148. Springer, 2011.

Bickel, S., Brückner, M., and Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(Sep):2137–2155, 2009.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Bodendorf, F. and Kaiser, C. Detecting opinion leaders and trends in online social networks. In *Proceedings of the 2nd ACM Workshop on Social Web Search and Mining*, pp. 65–68, 2009.

Chebbi, Z. and Moakher, M. Means of Hermitian positive-definite matrices based on the log-determinant $\alpha$-divergence function. *Linear Algebra and its Applications*, 436(7):1872 – 1889, 2012.

Chérief-Abdellatif, B.-E. and Alquier, P. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. *arXiv preprint arXiv:1909.13339*, 2019.

Cox, D. R. Tests of separate families of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 105–123, 1961.

Cox, D. R. A return to an old paper: 'tests of separate families of hypotheses'. *Journal of the Royal Statistical Society: Series B*, 75(2):207–215, 2013.

Csilléry, K., Blum, M. G., Gaggiotti, O. E., and François, O. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, 25(7):410 – 418, 2010.

Delage, E. and Ye, Y. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Ghosh, A. and Basu, A. Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. *Covariate shift and local learning by distribution matching*, pp. 131–160. MIT Press, 2009.

Grünwald, P. The safe Bayesian. In *International Conference on Algorithmic Learning Theory*, pp. 169–183. Springer, 2012.

Holmes, C. and Walker, S. Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104 (2):497–503, 2017.

Iyer, A., Nath, S., and Sarawagi, S. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 530–538, 2014.

Jewson, J., Smith, J. Q., and Holmes, C. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

Knoblauch, J. Robust deep Gaussian processes. *arXiv preprint arXiv:1904.02303*, 2019.

Knoblauch, J., Jewson, J., and Damoulas, T. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, pp. 130–166, 2019.

Ledoit, O. and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365 – 411, 2004.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31*, pp. 7167–7177, 2018.

Miller, J. W. and Dunson, D. B. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125, 2019.

Mohajerin Esfahani, P. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Moreno-Torres, J., Raeder, T., Alaiz-Rodríguez, R., Chawla, N., and Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521 – 530, 2012.

Murphy, K. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29*, pp. 2208–2216, 2016.

Naraei, P., Abhari, A., and Sadeghian, A. Application of multilayer perceptron neural networks and support vector machines in classification of healthcare data. In *2016 Future Technologies Conference*, pp. 848–852. IEEE, 2016.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Calculating optimistic likelihoods using (geodesically) convex optimization. In *Advances in Neural Information Processing Systems 32*, pp. 13942–13953, 2019a.

Nguyen, V. A., Shafieezadeh-Abadeh, S., Yue, M.-C., Kuhn, D., and Wiesemann, W. Optimistic distributionally robust optimization for nonparametric likelihood approximation. In *Advances in Neural Information Processing Systems 32*, pp. 15872–15882, 2019b.

Ong, V. M. H., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28(4):971–988, 2018.

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

Rippl, T., Munk, A., and Sturm, A. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *Journal of Multivariate Analysis*, 151:90–109, 2016.

Schervish, M. J. *Theory of Statistics*. Springer, 1995.

Staib, M. and Jegelka, S. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems 32*, pp. 9134–9144, 2019.

Tomar, D. and Agarwal, S. A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5):241–266, 2013.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6(31): 187–202, 2009.

Walker, S. G. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10): 1621–1633, 2013.

Wiesemann, W., Kuhn, D., and Sim, M. Distributionally robust convex optimization. *Operations Research*, 62(6): 1358–1376, 2014.

Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466:1102–1104, 2010.

Zhao, C. and Guan, Y. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262 – 267, 2018.

Zhu, W., Miao, J., Hu, J., and Qing, L. Vehicle detection in driving simulation using extreme learning machine. *Neurocomputing*, 128:160–165, 2014.