Optimistic Distributionally Robust Optimization for Nonparametric Likelihood Approximation

Viet Anh Nguyen **Soroosh Shafieezadeh-Abadeh**
École Polytechnique Fédérale de Lausanne, Switzerland
{viet-anh.nguyen, soroosh.shafiee}@epfl.ch

Man-Chung Yue
The Hong Kong Polytechnic University, Hong Kong
manchung.yue@polyu.edu.hk

Daniel Kuhn
École Polytechnique Fédérale de Lausanne, Switzerland
daniel.kuhn@epfl.ch

Wolfram Wiesemann
Imperial College Business School, United Kingdom
ww@imperial.ac.uk

Abstract

The likelihood function is a fundamental component in Bayesian statistics. However, evaluating the likelihood of an observation is computationally intractable in many applications. In this paper, we propose a non-parametric approximation of the likelihood that identifies a probability measure which lies in the neighborhood of the nominal measure and that maximizes the probability of observing the given sample point. We show that when the neighborhood is constructed by the Kullback-Leibler divergence, by moment conditions or by the Wasserstein distance, then our *optimistic likelihood* can be determined through the solution of a convex optimization problem, and it admits an analytical expression in particular cases. We also show that the posterior inference problem with our optimistic likelihood approximation enjoys strong theoretical performance guarantees, and it performs competitively in a probabilistic classification task.

1 Introduction

Bayesian statistics is a versatile mathematical framework for estimation and inference, which applications in bioinformatics [2], computational biology [47, 48], neuroscience [57], natural language processing [30, 40], computer vision [26, 31], robotics [15], machine learning [34, 53], etc. A Bayesian inference model is composed of an unknown parameter θ from a known parameter space Θ, an observed sample point x from a sample space $\mathcal{X} \subseteq \mathbb{R}^m$, a likelihood measure (or conditional density) $p(\cdot|\theta)$ over $\mathcal{X}$ and a prior distribution $\pi(\cdot)$ over Θ. The key objective of Bayesian statistics is the computation of the posterior distribution $p(\cdot|x)$ over Θ upon observing x.

Unfortunately, computing the posterior is a challenging task in practice. Bayes' theorem, which relates the posterior to the prior [49, Theorem 1.31], requires the evaluation of both the likelihood function $p(\cdot|\theta)$ and the evidence $p(x)$. Evaluating the likelihood $p(\cdot|\theta)$ at an observation $x \in \mathcal{X}$ is

Preprint. Under review.

an intractable problem in many situations. For example, the statistical model may contain hidden variables ζ, and the likelihood $p(x|\theta)$ can only be computed by marginalizing out the hidden variables $p(x|\theta) = \int p(x, \zeta|\theta)\mathrm{d}\zeta$ [38, pp. 322]. In the g-and-k model, the density function does not exist in closed form and can only be expressed in terms of the derivatives of quantile functions, which implies that $p(x|\theta)$ needs to be computed numerically for each individual observation x [23]. Likewise, evaluating the evidence $p(x)$ is intractable whenever the evaluation of the likelihood $p(x|\theta)$ is. To avoid calculating $p(x)$ in the process of constructing the posterior, the variational Bayes approach [9] maximizes the evidence lower bound (ELBO), which is tantamount to solving

$$\min_{\mathbb{Q}\in\mathcal{Q}} \ \mathrm{KL}(\mathbb{Q} \parallel \pi) - \mathbb{E}_{\mathbb{Q}}[\log p(x|\theta)], \tag{1}$$

where $\mathrm{KL}(\mathbb{Q} \parallel \pi)$ denotes the Kullback-Leibler (KL) divergence from $\mathbb{Q}$ to π. One can show that if the feasible set $\mathcal{Q}$ contains all probability measures supported on Θ, then the optimal solution $\mathbb{Q}^\star$ of (1) coincides with the true posterior distribution. Consequently, inferring the posterior is equivalent to solving the convex optimization problem (1) that depends only on the prior distribution π and the likelihood $p(x|\theta)$. There are scalable algorithms to solve the ELBO maximization problem [24], and the variational Bayes approach has been successfully applied in inference tasks [20, 21], reinforcement learning [25, 36], dimensionality reduction [39] and training deep neural networks [27]. Nevertheless, the variational Bayes approach requires both perfect knowledge and a tractable representation of the likelihood $p(x|\theta)$, which is often not available in practice.

While the likelihood $p(x|\theta)$ may be intractable to compute, we can approximate $p(x|\theta)$ from available data in many applications. For example, in the classification task where $\Theta = \{\theta_1, \ldots, \theta_C\}$ denotes the class labels, the class conditional probabilities $p(x|\theta_i)$ and the prior distribution $\pi(\theta_i)$ can be inferred from the training data, and a probabilistic classifier can be constructed by assigning x to each class randomly under the posterior distribution [8, pp. 43]. Approximating the intractable likelihood from available samples is also the key ingredient of approximate Bayesian computation (ABC), a popular statistical method for likelihood-free inference that has gained widespread success in various fields [3, 14, 54]. The sampling-based likelihood algorithm underlying ABC assumes that we have access to a simulation device that can generate N i.i.d. samples $\widehat{x}_1, \ldots, \widehat{x}_N$ from $p(\cdot|\theta)$, and it approximates the likelihood $p(x|\theta)$ by the surrogate $p_h(x|\theta)$ defined as

$$p_h(x|\theta) \ = \ \int_{\mathcal{X}} K_h\left(d(x, \widehat{x})\right) p(\widehat{x}|\theta)\mathrm{d}\widehat{x} \ \approx \ \frac{1}{N}\sum_{j=1}^{N} K_h\left(d(x, \widehat{x}_j)\right), \tag{2}$$

where K_h is a kernel function with kernel width h, $d(\cdot,\cdot)$ is a distance on $\mathcal{X}$, and the approximation is due to the reliance upon finitely many samples [43, 46].

In this paper, we propose an alternative approach to approximate the likelihood $p(x|\theta)$. We assume that the sample space $\mathcal{X}$ is countable, and hence $p(\cdot|\theta)$ is a probability mass function. We model the decision maker's nominal belief about $p(\cdot|\theta)$ by a nominal probability mass function $\widehat{\nu}_\theta$ supported on $\mathcal{X}$, which in practice typically represents the empirical distribution supported on the (possibly simulated) training samples. We then approximate the likelihood $p(x|\theta)$ by the optimal value of the following non-parametric *optimistic likelihood* problem

$$\sup_{\nu\in\mathbb{B}_\theta(\widehat{\nu}_\theta)} \ \nu(x), \tag{3}$$

where $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is a set that contains all probability mass functions in the vicinity of $\widehat{\nu}_\theta$. In the distributionally robust optimization literature, the set $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is referred to as the ambiguity set [4, 35, 56]. In contrast to the distributionally robust optimization paradigm, which would look for a worst-case measure that *minimizes* the probability of observing x among all measures contained in $\mathbb{B}_\theta(\widehat{\nu}_\theta)$, the optimistic likelihood problem (3) determines a best-case measure that *maximizes* this quantity. Thus, problem (3) is closely related to the literature on practicing optimism upon facing ambiguity, which has been shown to be beneficial in multi-armed bandit problems [12], planning [37], classification [7], image denoising [22], Bayesian optimization [11, 52], etc.

The choice of the set $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ in (3) directly impacts the performance of the optimistic likelihood approach. In the limiting case where $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ approaches a singleton $\{\widehat{\nu}_\theta\}$, the optimistic likelihood problem recovers the nominal estimate $\widehat{\nu}_\theta(x)$. Since this approximation is only reasonable when $\widehat{\nu}_\theta(x) > 0$, which is often violated when $\widehat{\nu}_\theta$ is estimated from few training samples, a strictly positive size of $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is preferred. Ideally, the shape of $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is chosen so that problem (3)

is computationally tractable and at the same time offers a promising approximation quality. We explore in this paper three different constructions of $\mathbb{B}_\theta(\widehat{\nu}_\theta)$: the Kullback-Leibler divergence [4], a description based on moment conditions [17, 33] and the Wasserstein distance [29, 41, 44, 50, 51].

The contributions of this paper may be summarized as follows.

1. We show that when $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is constructed using the KL divergence, the optimistic likelihood (3) reduces to a finite convex program, which in specific cases admits an analytical solution. However, this approach does not satisfactorily approximate $p(x|\theta)$ for previously unseen samples x.
2. We demonstrate that when $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is constructed using moment conditions, the optimistic likelihood (3) can be computed in closed form. However, since strikingly different distributions can share the same lower-order moments, this approach is often not flexible enough to accurately capture the tail behavior of $\widehat{\nu}_\theta$.
3. We show that when $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ is constructed using the Wasserstein distance, the optimistic likelihood (3) coincides with the optimal value of a linear program that can be solved using a greedy heuristics. Interestingly, this variant of the optimistic likelihood results in a likelihood approximation whose decay pattern resembles that of an exponential kernel approximation.
4. We use our optimistic likelihood approximation in the ELBO problem (1) for posterior inference. We prove that the resulting posterior inference problems under the KL divergence and the Wasserstein distance enjoy strong theoretical guarantees, and we illustrate their promising empirical performance in numerical experiments.

While this paper focuses on the non-parametric approximation of the likelihood $p(x|\theta)$, we emphasize that the optimistic likelihood approach can also be applied in the parametric setting. More specifically, if $p(\cdot|\theta)$ belongs to the family of Gaussian distributions, then the optimistic likelihood approximation can be solved efficiently using geodesically convex optimization [42].

The remainder of the paper is structured as follows. We study the optimistic likelihood problem under the KL ambiguity set, under moment conditions and under the Wasserstein distance in Sections 2–4, respectively. Section 5 provides a performance guarantee for the posterior inference problem using our optimistic likelihood. All proofs and additional material are relegated to the Appendix. In Sections 2–4, the development of the theoretical results is generic, and hence the dependence of $\widehat{\nu}_\theta$ and $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ on θ is omitted to avoid clutter.

Notation. We denote by $\mathcal{M}(\mathcal{X})$ the set of all probability mass functions supported on $\mathcal{X}$, and we refer to the support of $\nu \in \mathcal{M}(\mathcal{X})$ as $\mathrm{supp}(\nu)$. For any $z \in \mathcal{X}$, δ_z is the delta-Dirac measure at z. For any $N \in \mathbb{N}_+$, we use $[N]$ to denote the set $\{1, \ldots, N\}$. $\mathbb{1}_x(\cdot)$ is the indicator function at x, i.e., $\mathbb{1}_x(\xi) = 1$ if $\xi = x$, and $\mathbb{1}_x(\xi) = 0$ otherwise.

2 Optimistic Likelihood using the Kullback-Leibler Divergence

We first consider the optimistic likelihood problem where the ambiguity set is constructed using the KL divergence. The KL divergence is the starting point of the ELBO maximization problem (1), and thus it is natural to explore its potential in our likelihood approximation.

Definition 2.1 (KL divergence)**.** Let ν_1, ν_2 be two probability mass functions on $\mathcal{X}$ such that ν_1 is absolutely continuous with respect to ν_2. The KL divergence between ν_1 and ν_2 is defined as

$$\mathrm{KL}(\nu_1 \parallel \nu_2) \triangleq \sum_{z \in \mathcal{X}} f\left(\nu_1(z)/\nu_2(z)\right) \nu_2(z),$$

where $f(t) = t\log(t) - t + 1$.

We now consider the KL divergence ball $\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ centered at the empirical distribution $\widehat{\nu}$ with radius $\varepsilon \geq 0$, that is,

$$\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon) = \left\{\nu \in \mathcal{M}(\mathcal{X}) : \mathrm{KL}(\widehat{\nu} \parallel \nu) \leq \varepsilon\right\}. \tag{4}$$

Moreover, we assume that the nominal distribution $\widehat{\nu}$ is supported on N distinct points $\widehat{x}_1, \ldots, \widehat{x}_N$, that is, $\widehat{\nu} = \sum_{j \in [N]} \widehat{\nu}_j \delta_{\widehat{x}_j}$ with $\widehat{\nu}_j > 0 \, \forall j \in [N]$ and $\sum_{j \in [N]} \widehat{\nu}_j = 1$.

The set $\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ is not weakly compact because $\mathcal{X}$ can be unbounded, and thus the existence of a probability measure that optimizes the optimistic likelihood problem (3) over the feasible set

$\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ is not immediate. The next proposition asserts that the optimal solution exists, and it provides structural insights about the support of the optimal measure.

Proposition 2.2 (Existence of optimizers; KL ambiguity)**.** For any $\varepsilon \geq 0$ and $x \in \mathcal{X}$, there exists a measure $\nu^\star_{\mathrm{KL}} \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ such that

$$\sup_{\nu \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)} \nu(x) = \nu^\star_{\mathrm{KL}}(x) \tag{5}$$

Moreover, $\nu^\star_{\mathrm{KL}}$ is supported on at most $N+1$ points satisfying $\mathrm{supp}(\nu^\star_{\mathrm{KL}}) \subseteq \mathrm{supp}(\widehat{\nu}) \cup \{x\}$.

Proposition 2.2 suggests that the optimistic likelihood problem (5), inherently an *in*finite dimensional problem whenever $\mathcal{X}$ is infinite, can be formulated as a finite dimensional problem. The next theorem provides a finite convex programming reformulation of (5).

Theorem 2.3 (Optimistic likelihood; KL ambiguity)**.** For any $\varepsilon \geq 0$ and $x \in \mathcal{X}$,

- if $x \in \mathrm{supp}(\widehat{\nu})$, then problem (5) can be reformulated as the finite convex optimization problem

$$\sup_{\nu \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)} \nu(x) = \max \left\{ \textstyle\sum_{j \in [N]} y_j \, \mathbb{1}_x(\widehat{x}_j) : y \in \mathbb{R}^N_{++},\, \sum_{j \in [N]} \widehat{\nu}_j \log\left(\widehat{\nu}_j / y_j\right) \leq \varepsilon,\, e^\top y = 1 \right\},$$

  where e is the vector of all ones;

- if $x \notin \mathrm{supp}(\widehat{\nu})$, then problem (5) has the optimal value $1 - \exp(-\varepsilon)$.

Theorem 2.3 indicates that the determining factor in the KL optimistic likelihood approximation is whether the observation x belongs to the support of the nominal measure $\widehat{\nu}$ or not. If $x \notin \mathrm{supp}(\widehat{\nu})$, then the optimal value of (5) does not depend on x, and the KL divergence approach assigns a flat likelihood. Interestingly, in Appendix B.2 we prove a similar result for the wider class of f-divergences, which contains the KL divergence as a special case. While this flat likelihood behavior may be useful in specific cases, one would expect the relative distance of x to the atoms of $\widehat{\nu}$ to influence the optimal value of the optimistic likelihood problem, similar to the neighborhood-based intuition reflected in the kernel approximation approach. Unfortunately, the lack of an underlying metric in its definition implies that the f-divergence family cannot capture this intuition, and thus f-divergence ambiguity sets are not an attractive option to approximate the likelihood of an observation x that does not belong to the support of the nominal measure $\widehat{\nu}$.

Remark 2.4 (On the order of the measures)**.** An alternative construction of the KL ambiguity set, which has been widely used in the literature [4], is

$$\widehat{\mathbb{B}}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon) = \{\nu \in \mathcal{M}(\mathcal{X}) : \mathrm{KL}(\nu \parallel \widehat{\nu}) \leq \varepsilon\},$$

where the two measures ν and $\widehat{\nu}$ change roles. However, in this case the KL divergence imposes that all $\nu \in \widehat{\mathbb{B}}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ are absolutely continuous with respect to $\widehat{\nu}$. In particular, if $x \notin \mathrm{supp}(\widehat{\nu})$, then $\nu(x) = 0$ for all $\nu \in \widehat{\mathbb{B}}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$, and $\widehat{\mathbb{B}}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ is not able to approximate the likelihood of x in a meaningful way.

3 Optimistic Likelihood using Moment Conditions

In this section we study the optimistic likelihood problem (3) when the ambiguity set $\mathbb{B}(\widehat{\nu})$ is specified by moment conditions. For tractability purposes, we focus on ambiguity sets $\mathbb{B}_{\mathrm{MV}}(\widehat{\nu})$ that contain all distributions which share the same mean $\widehat{\mu}$ and covariance matrix $\widehat{\Sigma} \in \mathbb{S}^m_{++}$ with the nominal distribution $\widehat{\nu}$. Formally, this moment ambiguity set $\mathbb{B}_{\mathrm{MV}}(\widehat{\nu})$ can be expressed as

$$\mathbb{B}_{\mathrm{MV}}(\widehat{\nu}) = \left\{ \nu \in \mathcal{M}(\mathcal{X}) :\, \mathbb{E}_\nu[\tilde{x}] = \widehat{\mu},\, \mathbb{E}_\nu[\tilde{x}\tilde{x}^\top] = \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top \right\}.$$

The optimistic likelihood (3) over the ambiguity set $\mathbb{B}_{\mathrm{MV}}(\widehat{\nu})$ is a moment problem that is amenable to a well-known reformulation as a polynomial time solvable semidefinite program [6]. Surprisingly, in our case the optimal value of the optimistic likelihood problem is available in closed form. This result was first discovered in [32], and a proof using optimization techniques can be found in [5].

Theorem 3.1 (Optimistic likelihood; mean-variance ambiguity [5, 32])**.** Suppose that $\widehat{\nu}$ has the mean vector $\widehat{\mu} \in \mathbb{R}^m$ and the covariance matrix $\widehat{\Sigma} \in \mathbb{S}_{++}^m$. For any $x \in \mathcal{X}$, the optimistic likelihood problem (3) over the moment ambiguity set $\mathbb{B}_{\text{MV}}(\widehat{\nu})$ has the optimal value

$$\sup_{\nu \in \mathbb{B}_{\text{MV}}(\widehat{\nu})} \nu(x) = \frac{1}{1 + (x - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(x - \widehat{\mu})} \in (0, 1]. \tag{6}$$

The optimal value (6) of the optimistic likelihood problem depends on the location of the observed sample point x, and hence the moment ambiguity set captures the behavior of the likelihood function in a more realistic way than the KL divergence ambiguity set from Section 2. Moreover, the moment ambiguity set $\mathbb{B}_{\text{MV}}(\widehat{\nu})$ does not depend on any hyper-parameters that need to be tuned. However, since the construction of $\mathbb{B}_{\text{MV}}(\widehat{\nu})$ only relies on the first two moments of the nominal distribution $\widehat{\nu}$, it fails to accurately capture the tail behavior of $\widehat{\nu}$, see Appendix B.3. This motivates us to look further for an ambiguity set that faithfully accounts for the tail behavior of $\widehat{\nu}$.

4 Optimistic Likelihood using the Wasserstein Distance

We now study a third construction for the ambiguity set $\mathbb{B}(\widehat{\nu})$, which is based on the type-1 Wasserstein distance (also commonly known as the Monge-Kantorovich distance), see [55]. Contrary to the KL divergence, the Wasserstein distance inherently depends on the ground metric of the sample space $\mathcal{X}$.

Definition 4.1 (Wasserstein distance)**.** The type-1 Wasserstein distance between two measures $\nu_1, \nu_2 \in \mathcal{M}(\mathcal{X})$ is defined as

$$\mathbb{W}(\nu_1, \nu_2) \triangleq \inf_{\lambda \in \Lambda(\nu_1, \nu_2)} \mathbb{E}_\lambda \left[d(x_1, x_2) \right],$$

where $\Lambda(\nu_1, \nu_2)$ denotes the set of all distributions on $\mathcal{X} \times \mathcal{X}$ with the first and second marginal distributions being ν_1 and ν_2, respectively, and d is the ground metric of $\mathcal{X}$.

The Wasserstein ball $\mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon)$ centered at the nominal distribution $\widehat{\nu}$ with radius $\varepsilon \geq 0$ is

$$\mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon) = \{\nu \in \mathcal{M}(\mathcal{X}) : \mathbb{W}(\nu, \widehat{\nu}) \leq \varepsilon\}. \tag{7}$$

We first establish a structural result for the optimistic likelihood problem over the Wasserstein ambiguity set. This is the counterpart to Proposition 2.2 for the KL divergence.

Proposition 4.2 (Existence of optimizers; Wasserstein ambiguity)**.** For any $\varepsilon \geq 0$ and $x \in \mathcal{X}$, there exists a measure $\nu_{\mathbb{W}}^\star \in \mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon)$ such that

$$\sup_{\nu \in \mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon)} \nu(x) = \nu_{\mathbb{W}}^\star(x). \tag{8}$$

Furthermore, $\nu_{\mathbb{W}}^\star$ is supported on at most $N + 1$ points satisfying $\text{supp}(\nu_{\mathbb{W}}^\star) \subseteq \text{supp}(\widehat{\nu}) \cup \{x\}$.

Leveraging Proposition 4.2, we can show that the optimistic likelihood estimate over the Wasserstein ambiguity set coincides with the optimal value of a linear program whose number of decision variables equals the number of atoms N of the nominal measure $\widehat{\nu}$.

Theorem 4.3 (Optimistic likelihood; Wasserstein ambiguity)**.** For any $\varepsilon \geq 0$ and $x \in \mathcal{X}$, problem (8) is equivalent to the linear program

$$\sup_{\nu \in \mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon)} \nu(x) = \max \left\{ \sum_{j \in [N]} T_j : T \in \mathbb{R}_+^N, \sum_{j \in [N]} d(x, \widehat{x}_j)\, T_j \leq \varepsilon, T_j \leq \widehat{\nu}_j \; \forall j \in [N] \right\}. \tag{9}$$

The currently best complexity bound for solving a general linear program with N decision variables is $\mathcal{O}(N^{2.37})$ [13], which may be prohibitive when N is large. Fortunately, the linear program (9) can be solved to optimality using a greedy heuristics in quasilinear time.

Proposition 4.4 (Optimal solution via greedy heuristics)**.** The linear program (9) can be solved to optimality by a greedy heuristics in time $\mathcal{O}(N \log N)$.

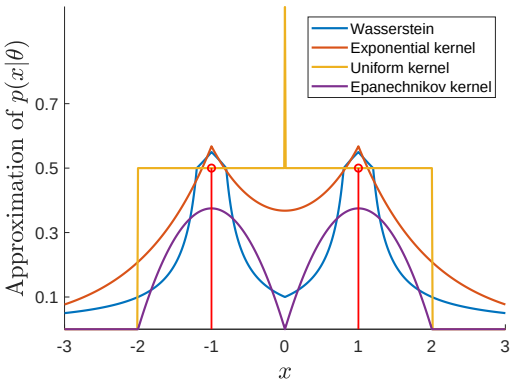


Figure 1: Comparison between the Wasserstein approximation ($\varepsilon = 0.2$) and the sample average kernel approximations ($h = 1$) of $p(x|\theta)$.

Example 4.5 (Qualitative comparison with kernel methods)**.** Let $m = 1$, $d(x, \widehat{x}) = \|x - \widehat{x}\|_1$ and $\widehat{\nu} = 0.5\delta_{-1} + 0.5\delta_1$. Figure 1 compares the approximation of $p(x|\theta)$ by the Wasserstein optimistic likelihood with those of the finite sample kernel approximations (2) with $K_h(u) = K\left(h^{-1}u\right)$, where the Kernel K is exponential with $K(y) = \exp(-y)$, uniform with $K(y) = \mathbb{1}[|y| \leq 1]$ or Epanechnikov with $K(y) = 3/4(1 - y^2)\mathbb{1}[|y| \leq 1]$. While both the uniform and the Epachnechnikov kernel may produce an approximation value of 0 when x is far away from the support of $\widehat{\nu}$, the Wasserstein approximation always returns a positive likelihood when $\varepsilon > 0$ (see Corollary A.2). Qualitatively, the Wasserstein approximation exhibits a decay pattern similar to that of the finite sample average exponential kernel approximation.

On one hand, the similarity between the optimistic likelihood over the Wasserstein ambiguity set and the exponential kernel approximation suggests that the kernel approximation can potentially be interpreted in the light of our optimistic distributionally robust optimization framework. On the other hand, and perhaps more importantly, this similarity suggests that there are possibilities to design novel and computationally efficient kernel-like approximations using advanced optimization techniques. Even though the assumption that $p(\cdot|\theta)$ is a probability mass function is fundamental for our approximation, we believe that our approach can be utilized in the ABC setting even when $p(\cdot|\theta)$ is a probability density function. We leave these ideas for future research.

Appendix B.3 illustrates further how the Wasserstein ambiguity set offers a better tail approximation of the nominal measure $\widehat{\nu}$ than the ambiguity set based on moment conditions. Interestingly, the Wasserstein approximation can also be generalized to approximate the log-likelihood of a batch of i.i.d. observations, see Appendix B.4

5 Application to the ELBO Problem

Motivated by the fact that the likelihood $p(x|\theta)$ is intractable to compute in many practical applications, we use our optimistic likelihood approximation (3) as a surrogate for $p(x|\theta)$ in the ELBO problem (1). In this section, we will focus on the KL divergence and the Wasserstein ambiguity sets, and we will impose the following assumptions.

Assumption 5.1 (Finite parameter space)**.** We assume that $\Theta = \{\theta_1, \ldots, \theta_C\}$ for some $C \geq 2$.

Assumption 5.2 (I.i.d. sampling and empirical distribution)**.** For every $i \in [C]$, we have N_i i.i.d. samples $\widehat{x}_{ij}$, $j \in [N_i]$, from the conditional probability $p(\cdot|\theta_i)$. Furthermore, each nominal distribution $\widehat{\nu}_i$ is given by the empirical distribution $\widehat{\nu}_i^{N_i} = N_i^{-1} \sum_{j \in [N_i]} \delta_{\widehat{x}_{ij}}$ on the samples $\widehat{x}_{ij}$.

Assumption 5.1 is necessary for our approach because we approximate $p(x|\theta)$ separately for every $\theta \in \Theta$. Under this assumption, the prior distribution π can be expressed by the C-dimensional vector $\pi \in \mathbb{R}_+$, and the ELBO program (1) becomes the finite-dimensional convex optimization problem

$$\mathcal{J}^{\text{true}} = \min_{q \in \mathcal{Q}} \quad \sum_{i \in [C]} q_i(\log q_i - \log \pi_i) - \sum_{i \in [C]} q_i \log p(x|\theta_i), \tag{10}$$

where by a slight abuse of notation, $\mathcal{Q}$ is now a subset of the C-dimensional simplex. Assumption 5.2, on the other hand, is a standard assumption in the nonparametric setting, and it allows us to study the statistical properties of our optimistic likelihood approximation.

We approximate $p(x|\theta_i)$ for each θ_i by the optimal value of the optimistic likelihood problem (3):

$$p(x|\theta_i) \approx \sup_{\nu_i \in \mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})} \nu_i(x) \tag{11}$$

Here, $\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})$ is the KL divergence or Wasserstein ambiguity set centered at the empirical distribution $\widehat{\nu}_i^{N_i}$. Under Assumptions 5.1 and 5.2, a surrogate model of the ELBO problem (1) is then

obtained using the approximation (11) as

$$\widehat{\mathcal{J}}_{\mathbb{B}^N} = \min_{q\in\mathcal{Q}} \sum_{i\in[C]} q_i(\log q_i - \log \pi_i) - \sum_{i\in[C]} q_i \log \left(\sup_{\nu_i \in \mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})} \nu_i(x) \right), \tag{12}$$

where we use $\mathbb{B}^N$ to denote the collection of ambiguity sets $\left\{\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})\right\}_{i=1}^C$ with $N = \sum_i N_i$.

We now study the statistical properties of problem (12). We first present an asymptotic guarantee for the KL divergence. Towards this end, we define the *disappointment* as $\mathbb{P}^\infty(\mathcal{J}^{\text{true}} < \widehat{\mathcal{J}}_{\mathbb{B}^N})$.

Theorem 5.3 (Asymptotic guarantee; KL ambiguity)**.** Suppose that Assumptions 5.1 and 5.2 hold. For each $i \in [C]$, let $\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i}) = \mathbb{B}_{\text{KL}}(\widehat{\nu}_i^{N_i}, \varepsilon_i)$ for some $\varepsilon_i > 0$, and set $n \triangleq \min\{N_1, \ldots, N_C\}$. We then have

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}^\infty(\mathcal{J}^{\text{true}} < \widehat{\mathcal{J}}_{\mathbb{B}^N}) \;\leq\; -\min_{i\in[C]} \varepsilon_i \;<\; 0.$$

Theorem 5.3 shows that as the number of training samples N_i for each $i \in [C]$ grows, the disappointment decays exponentially at a rate of at least $\min_i \varepsilon_i$.

We next study the statistical properties of problem (12) when each $\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})$ is a Wasserstein ball. To this end, we additionally impose the following assumption, which essentially requires that the tail of each distribution $p(\cdot|\theta_i)$, $i \in [C]$, decays at an exponential rate.

Assumption 5.4 (Light-tailed conditional distribution)**.** For each $i \in [C]$, there exists an exponent $a_i > 1$ such that $A_i \triangleq \mathbb{E}[\exp(\|x\|^{a_i})] < \infty$, where the expectation is taken with respect to $p(\cdot|\theta_i)$.

Theorem 5.5 (Finite sample guarantee; Wasserstein ambiguity)**.** Suppose that Assumptions 5.1, 5.2 and 5.4 hold, and fix any $\beta \in (0,1)$. Assume that $m \neq 2$ and that $\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i}) = \mathbb{B}_{\mathbb{W}}(\widehat{\nu}_i^{N_i}, \varepsilon_i(\beta, C, N_i))$ for every $i \in [C]$ with

$$\varepsilon_i(\beta, C, N_i) \triangleq \begin{cases} \left(\frac{\log(k_{i1}C\beta^{-1})}{k_{i2}N_i}\right)^{1/\max\{m,2\}} & \text{if } N_i \geq \frac{\log(k_{i1})C\beta^{-1}}{k_{i2}}, \\ \left(\frac{\log(k_{i1}C\beta^{-1})}{k_{i2}N_i}\right)^{1/a_i} & \text{if } N_i < \frac{\log(k_{i1})C\beta^{-1}}{k_{i2}}, \end{cases}$$

and k_{i1}, k_{i2} are positive constants that depend on a_i, A_i and m. We then have $\mathbb{P}^N\left(\mathcal{J}^{\text{true}} < \widehat{\mathcal{J}}_{\mathbb{B}^N}\right) \leq \beta$.

Theorem 5.5 provides a finite sample guarantee for the disappointment of problem (12) under a specific choice of radii for the Wasserstein balls.

Theorem 5.6 (Asymptotic guarantee for Wasserstein)**.** Suppose that Assumptions 5.1, 5.2 and 5.4 hold. For each $i \in [C]$, let $\beta_{N_i} \in (0,1)$ be a sequence such that $\sum_{N_i=1}^\infty \beta_{N_i} < \infty$ and $\mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i}) = \mathbb{B}_{\mathbb{W}}(\widehat{\nu}_i^{N_i}, \varepsilon_i(\beta_N, C, N_i))$, where ε_i is defined as in Theorem 5.5. Then $\widehat{\mathcal{J}}_{\mathbb{B}^N} \to \mathcal{J}^{\text{true}}$ as $N_1, \ldots, N_C \to \infty$ almost surely.

Theorem 5.6 offers an asymptotic guarantee which asserts that as the numbers of training samples N_i grow, the optimal value of (12) converges to that of the ELBO problem (10).

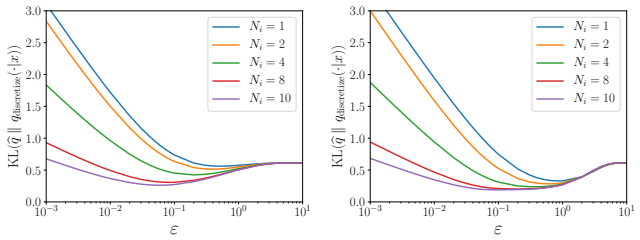
6 Numerical Experiments

We first showcase the performance guarantees from the previous section on a synthetic dataset in Section 6.1. Afterwards, Section 6.2 benchmarks the performance of the different likelihood approximations in a probabilistic classification task on standard UCI datasets. The source code, including our algorithm and all tests implemented in Python, are available from `https://github.com/sorooshafiee/Nonparam_Likelihood`.

6.1 Synthetic Dataset: Beta-Binomial Inference

We consider the beta-binomial problem in which the prior π, the likelihood $p(x|\theta)$, and the posterior distribution $q(\theta|x)$ have the following forms:

$$\pi(\theta) = \text{Beta}(\theta|\alpha, \beta), \quad p(x|\theta) = \text{Bin}(x|M, \theta), \quad q(\theta|x) = \text{Beta}(\theta|x+\alpha, M-x+\beta)$$



(a) KL divergence (b) Wasserstein distance

Figure 2: Average KL divergence between $\widehat{q}$ that solves (12) and the discretized posterior $q_{\text{discretize}}(\cdot|x)$ as a function of ε and N_i.

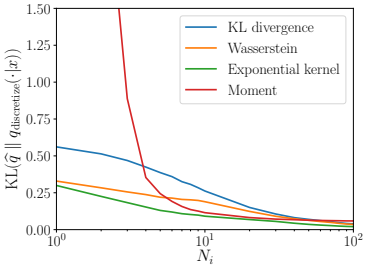


Figure 3: Optimally tuned performance of different approximation schemes with varying N_i.

We emphasize that in this setting, the posterior distribution is known in closed form, and the main goal is to study the properties of the optimistic ELBO problem (12) and the convergence of the solution of problem (12) to the true posterior distribution. We impose a uniform prior distribution π by setting $\alpha = \beta = 1$. The finite parameter space $\Theta = \{\theta_1, \ldots, \theta_C\}$ contains $C = 20$ equidistant discrete points in the range $(0, 1)$. For simplicity, we set $N_1 = \ldots = N_C$ in this experiment.

We conduct the following experiment for different training set sizes $N_i \in \{1, 2, 4, 8, 10\}$ and different ambiguity set radii ε. For each parameter setting, our experiment consists of 100 repetitions. In each repetition, we randomly generate an observation x from a binomial distribution with $M = 20$ trials and success probability $\theta_{\text{true}} = 0.6$. We then find the distribution $\widehat{q}$ that solves problem (12) using both the KL and the Wasserstein approximation. In a similar way, we find $\widehat{q}$ by solving (10), where $p(x|\theta)$ is approximated using the exponential kernel of the likelihood (2) with varying kernel width.

We evaluate the quality of the computed posteriors $\widehat{q}$ from the different approximations based on the KL divergences of $\widehat{q}$ to the true discretized posterior $q_{\text{discretize}}(\theta_i|x) \propto \text{Beta}(\theta_i|x + \alpha, M - x + \beta)$. Figures 4(a) and 4(b) depict the average quality of $\widehat{q}$ with different radii. One can readily see that the optimal size of the ambiguity set that minimizes $\text{KL}(\widehat{q} \parallel q_{\text{discretize}}(\cdot|x))$ decreases as N_i increases for both the KL and the Wasserstein approximation. Figure 3 depicts the performance of the optimally tuned approximations with different sample sizes N_i. We notice that the optimistic likelihood over the Wasserstein ambiguity set is comparable to the exponential kernel approximation.

6.2 Real World Dataset: Classification

We now consider a probabilistic classification setting with $C = 2$ classes. For each class $i = 1, 2$, we have access to N_i observations denoted by $\{\widehat{x}_{ij}\}_{j \in [N_i]}$. The nominal class-conditional probability distributions are the empirical measures, that is, $\widehat{\nu}_i = N_i^{-1} \sum_{j \in [N_i]} \delta_{\widehat{x}_{ij}}$ for $i = 1, 2$. The prior distribution π is also estimated from the training data as $\pi(\theta_i) = N_i/N$, where $N = N_1 + N_2$ is the total number of training samples. Upon observing a test sample x, the goal is to compute the posterior distribution $\widehat{q}$ by solving the optimization problem (12) using different approximation schemes. We subsequently use the posterior $\widehat{q}$ as a probabilistic classifier. In this experiment, we exclude the KL divergence approximation because $x \notin \text{supp}(\widehat{\nu}_i)$ most of the time.

In our experiments involving the Wasserstein ambiguity set, we randomly select 75% of the available data as training set and the remaining 25% as test set. We then use the training samples to tune the radii $\varepsilon_i \in \{a\sqrt{m}10^b : a \in \{1, \ldots, 9\}, b \in \{-3, -2, -1\}\}$, $i = 1, 2$, of the Wasserstein balls by a stratified 5-fold cross validation. For the moment based approximation, there is no hyper-parameter to tune, and all data is used as training set. We compare the performance of the classifiers from our optimistic likelihood approximation against the classifier selected by the exponential kernel approximation as a benchmark.

Table 1 presents the results on standard UCI benchmark datasets. All results are averages across 10 independent trials. The table shows that our optimistic likelihood approaches often outperform the exponential kernel approximation in classification tasks.

Acknowledgments We gratefully acknowledge financial support from the Swiss National Science Foundation under grant BSCGI0_157733 as well as the EPSRC grants EP/M028240/1, EP/M027856/1 and EP/N020030/1.

Table 1: Average area under the precision-recall curve for various UCI benchmark datasets. Bold numbers correspond to the best performances.

	Exponential	Moment	Wasserstein
Banknote Authentication	99.05	99.99	**100.00**
Blood Transfusion	64.91	**71.28**	68.23
Breast Cancer	97.58	**99.26**	97.99
Climate Model	**93.80**	81.94	93.40
Cylinder	76.74	75.00	**86.23**
Fourclass	99.95	82.77	**100.00**
German Credit	67.58	**75.50**	75.11
Haberman	70.82	70.20	**71.10**
Heart	78.77	**86.87**	75.86
Housing	75.62	81.89	**82.04**
ILPD	71.54	**72.95**	69.88
Ionosphere	91.02	97.05	**98.79**
Mammographic Mass	83.46	86.53	**87.86**
Pima	79.61	**82.37**	80.48
QSAR	84.44	**90.85**	90.21
Seismic Bumps	74.81	**75.68**	65.89
Sonar	85.66	83.49	**93.85**
Thoracic Surgery	54.84	**64.73**	56.32

Appendix A Proofs

A.1 Proofs of Section 2

The proof of Proposition 2.2 relies on the following auxiliary lemma, which we state first.

Lemma A.1 (Upper semicontinuity)**.** For any $x \in \mathcal{X} \subset \mathbb{R}^m$, the functional $F(\nu) = \nu(x)$ is upper semicontinuous over $\mathcal{M}(\mathcal{X})$.

Proof. We denote by $\mathbb{1}_x(\cdot)$ the indicator function at x, that is, $\mathbb{1}_x(\xi) = 1$ if $\xi = x$ and $\mathbb{1}_x(\xi) = 0$ otherwise. By definition, $F(\nu) = \int \mathbb{1}_x \mathrm{d}\nu$. Moreover, let $\{\nu_k\}_{k\in\mathbb{N}}$ be a sequence of probability measures converging weakly to $\nu \in \mathcal{M}(\mathcal{X})$. Since $\mathbb{1}_x(\cdot)$ is upper semicontinuous, the weak convergence of ν_k implies that

$$\limsup_{k\to\infty} F(\nu_k) = \limsup_{k\to\infty} \int \mathbb{1}_x \mathrm{d}\nu_k \leq \int \mathbb{1}_x \mathrm{d}\nu = F(\nu),$$

which in turn shows that the functional F is upper semicontinuous. □

Proof of Proposition 2.2. If $\varepsilon = 0$, the ball $\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ contains a singleton $\widehat{\nu}$ and the claim holds trivially. We can thus assume that $\varepsilon > 0$. Since $\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ is not necessarily weakly compact, the existence of the optimal measure $\nu^\star$ is not trivial. To show that $\nu^\star$ exists, we first establish that

$$\sup_{\nu\in\mathbb{B}_{\mathrm{KL}}(\widehat{\nu},\varepsilon)} \nu(x) = \sup_{\substack{\nu\in\mathbb{B}_{\mathrm{KL}}(\widehat{\nu},\varepsilon) \\ \mathrm{supp}(\nu)\subseteq(\widehat{\mathcal{S}}\cup\{x\})}} \nu(x), \tag{A.1}$$

where $\widehat{\mathcal{S}} = \mathrm{supp}(\widehat{\nu})$. To establish (A.1), it suffices to show that for any $\bar{\nu} \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ that assigns a non-zero probability on $\mathcal{X}\backslash(\widehat{\mathcal{S}} \cup \{x\})$, there exists $\nu' \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ satisfying $\mathrm{supp}(\nu') \subseteq \widehat{\mathcal{S}} \cup \{x\}$ such that ν' attains a higher objective value than $\bar{\nu}$, that is, $\nu'(x) > \bar{\nu}(x)$. Because $\bar{\nu}$ assigns a non-zero probability to $\mathcal{X}\backslash(\widehat{\mathcal{S}} \cup \{x\})$, we have

$$0 < \kappa \triangleq \sum_{z\in\mathcal{X}\backslash(\widehat{\mathcal{S}}\cup\{x\})} \bar{\nu}(z) \leq 1.$$

We now construct the measure ν' explicitly. Assume that $x \notin \widehat{\mathcal{S}}$. In this case, consider the discrete measure ν' supported on $\widehat{\mathcal{S}} \cup \{x\}$ given by

$$\nu'(x) = \bar{\nu}(x) + \kappa \quad \text{and} \quad \nu'(\widehat{x}_j) = \bar{\nu}(\widehat{x}_j) \quad \forall j \in [N].$$

Intuitively, ν' keeps the probability of $\bar{\nu}$ on $\widehat{\mathcal{S}}$, and it gathers the probability everywhere else and puts that mass onto x. We first show that ν' is a probability measure. Indeed, since $\kappa > 0$ and $\bar{\nu}$ is a probability measure, we have $\nu' \geq 0$. Moreover, we find

$$\sum_{z\in\mathcal{X}} \nu'(z) = \sum_{j\in[N]} \bar{\nu}(\widehat{x}_j) + \bar{\nu}(x) + \kappa = \sum_{j\in[N]} \bar{\nu}(\widehat{x}_j) + \bar{\nu}(x) + \sum_{z\in\mathcal{X}\setminus(\widehat{\mathcal{S}}\cup\{x\})} \bar{\nu}(z) = \sum_{z\in\mathcal{X}} \bar{\nu}(z) = 1,$$

where the first equality exploits the definition of $\bar{\nu}$, and the second equality follows from the definition of κ. Thus we conclude that ν' is a probability measure. We now proceed to show that ν' satisfies the KL divergence constraint. Indeed, we have

$$\begin{aligned}
\mathrm{KL}(\widehat{\nu} \,\|\, \nu') &= \sum_{z\in\mathcal{X}} f\left(\frac{\widehat{\nu}(z)}{\nu'(z)}\right)\nu'(z) \\
&= \sum_{j\in[N]} f\left(\frac{\widehat{\nu}_j}{\nu'(\widehat{x}_j)}\right)\nu'(\widehat{x}_j) + \nu'(x) && \text{(A.2a)} \\
&= \sum_{j\in[N]} f\left(\frac{\widehat{\nu}_j}{\bar{\nu}(\widehat{x}_j)}\right)\bar{\nu}(\widehat{x}_j) + \bar{\nu}(x) + \kappa && \text{(A.2b)} \\
&= \sum_{j\in[N]} f\left(\frac{\widehat{\nu}_j}{\bar{\nu}(\widehat{x}_j)}\right)\bar{\nu}(\widehat{x}_j) + \bar{\nu}(x) + \sum_{z\in\mathcal{X}\setminus(\widehat{\mathcal{S}}\cup\{x\})} f\left(\frac{\widehat{\nu}(z)}{\bar{\nu}(z)}\right)\bar{\nu}(z) && \text{(A.2c)} \\
&= \sum_{z\in\mathcal{X}} f\left(\frac{\widehat{\nu}(z)}{\bar{\nu}(z)}\right)\bar{\nu}(z) \leq \varepsilon. && \text{(A.2d)}
\end{aligned}$$

Equality (A.2a) holds because $f(0) = 1$ for the function f defined in Definition 2.1 and $\mathrm{supp}(\nu') \subseteq \widehat{\mathcal{S}} \cup \{x\}$. Equality (A.2b) follows from the construction of ν', and equality (A.2c) holds due to the definition of κ and the fact that $f(0) = 1$. Finally, the inequality in (A.2d) follows from the feasibility of $\bar{\nu}$, and it implies that $\nu' \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$. Furthermore, because $\kappa > 0$, we have $\nu'(x) = \bar{\nu}(x) + \kappa > \bar{\nu}(x)$ which asserts that $\bar{\nu}$ is strongly dominated by ν', and thus $\bar{\nu}$ cannot be an optimal measure.

Consider now the case $x \in \widehat{\mathcal{S}}$. Without loss of generality, we assume that $x = \widehat{x}_N$. In this case, it suffices to consider $\bar{\nu}$ satisfying $\bar{\nu}(\widehat{x}_N) \geq \widehat{\nu}_N$ because any $\bar{\nu}$ with $\bar{\nu}(\widehat{x}_N) < \widehat{\nu}_N$ is already dominated by the nominal measure $\widehat{\nu}$. Since $\kappa > 0$ and $\bar{\nu}(\widehat{x}_N) \geq \widehat{\nu}_N$, there must exist $K \in [N-1]$ atoms denoted without loss of generality by $\{\widehat{x}_1, \ldots, \widehat{x}_K\}$ that satisfy $\bar{\nu}(\widehat{x}_j) < \widehat{\nu}_j$ for all $k \in [K]$. Due to the continuity of the function f, there exists $\bar{\epsilon} \in (0, \kappa)$ that satisfies

$$f\left(\frac{\widehat{\nu}_N}{\bar{\nu}(\widehat{x}_N) + \bar{\epsilon}}\right)(\bar{\nu}(\widehat{x}_N) + \bar{\epsilon}) \leq f\left(\frac{\widehat{\nu}_N}{\bar{\nu}(\widehat{x}_N)}\right)\bar{\nu}(\widehat{x}_N) + \kappa.$$

We now consider the following measure ν' supported on $\widehat{\mathcal{S}}$:

$$\nu'(\widehat{x}_j) = \begin{cases} \bar{\nu}(\widehat{x}_j) + (\kappa - \bar{\epsilon}) \times (\widehat{\nu}_j - \bar{\nu}(\widehat{x}_j))/\sum_{k\in[K]}(\widehat{\nu}_k - \bar{\nu}(\widehat{x}_k)) & \forall j \in [K], \\ \bar{\nu}(\widehat{x}_j) & \forall j \in ([N-1]\setminus[K]), \\ \bar{\nu}(\widehat{x}_N) + \bar{\epsilon} & j = N. \end{cases}$$

We can verify that ν' is a probability measure supported on $\widehat{\mathcal{S}}$ and that $\nu'(\widehat{x}_N) > \bar{\nu}(\widehat{x}_N)$. Furthermore, we have

$$\begin{aligned}
\mathrm{KL}(\widehat{\nu} \,\|\, \nu') &= \sum_{j\in[N]} f\left(\frac{\widehat{\nu}_j}{\nu'(\widehat{x}_j)}\right)\nu'(\widehat{x}_j) \\
&= \sum_{j\in[K]} f\left(\frac{\widehat{\nu}_j}{\nu'(\widehat{x}_j)}\right)\nu'(\widehat{x}_j) + \sum_{j\in([N-1]\setminus[K])} f\left(\frac{\widehat{\nu}_j}{\nu'(\widehat{x}_j)}\right)\nu'(\widehat{x}_j) + f\left(\frac{\widehat{\nu}_N}{\nu'(\widehat{x}_N)}\right)\nu'(\widehat{x}_N) \\
&\leq \sum_{j\in[N]} f\left(\frac{\widehat{\nu}_j}{\bar{\nu}(\widehat{x}_j)}\right)\bar{\nu}(\widehat{x}_j) + \kappa = \mathrm{KL}(\widehat{\nu} \,\|\, \bar{\nu}) \leq \varepsilon,
\end{aligned}$$

where the first inequality follows from the definition of ν', the definition of $\bar{\epsilon}$, the fact that for any $\widehat{\nu}_j > 0$ the function $t \mapsto tf(\widehat{\nu}_j/t)$ is non-increasing in t over the domain $(0, \widehat{\nu}_j)$ and that $0 \leq$

$\bar{\nu}(\widehat{x}_j) < \nu'(\widehat{x}_j) \leq \widehat{\nu}_j$ by construction. We have thus asserted that $\bar{\nu}$ is dominated by $\nu' \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$, and we conclude that (A.1) holds.

We now consider the supremum on the right hand side of (A.1). By Lemma A.1, the objective function of (A.1) is upper semicontinuous. Furthermore, the feasible set

$$\left\{\nu \in \mathcal{M}(\mathcal{X}) : \mathrm{supp}(\nu) \subseteq (\widehat{\mathcal{S}} \cup \{x\}),\, \mathrm{KL}(\widehat{\nu} \,\|\, \nu) \leq \varepsilon\right\}$$

is weakly compact because it only contains measures supported on a finite set [1, Theorem 15.11]. By the Weierstrass maximum value theorem [1, Theorem 2.43], the supremum in (A.1) is attained and there exists $\nu^\star_{\mathrm{KL}} \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ such that

$$\sup_{\nu \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)} \nu(x) = \nu^\star_{\mathrm{KL}}(x).$$

This observation completes the proof. □

Proof of Theorem 2.3. Consider first the case when $x \in \widehat{\mathcal{S}}$, where $\widehat{\mathcal{S}} = \mathrm{supp}(\widehat{\nu})$. As a result of Proposition 2.2, the distribution that maximizes the probability at point x subject to the KL divergence constraint will be supported on at most N points from the set $\widehat{\mathcal{S}}$. The probability measures of interest thus share the form

$$\nu = \sum_{j \in [N]} y_j \delta_{\widehat{x}_j}$$

for some $y \in \mathbb{R}^N_+$, $\sum_{j\in[N]} y_j = 1$. The optimistic likelikood (5) satisfies

$$\nu^\star_{\mathrm{KL}}(x) = \sup \left\{ \sum_{j \in [N]} y_j \mathbb{1}_x(\widehat{x}_j) : y \in \mathbb{R}^N_{++},\, \sum_{j \in [N]} \widehat{\nu}_j \log\left(\frac{\widehat{\nu}_j}{y_j}\right) \leq \varepsilon,\, \sum_{j \in [N]} y_j = 1 \right\}, \tag{A.3}$$

which is a finite dimensional convex program in y.

Next, we consider the case where $x \notin \widehat{\mathcal{S}}$. To this end, for any $N \in \mathbb{N}_+$, we denote by Δ_N the simplex

$$\Delta_N \triangleq \left\{ y \in \mathbb{R}^N_+ : 0 \leq y_j \leq 1 \,\forall j \in [N],\, \sum_{j \in [N]} y_j \leq 1 \right\}. \tag{A.4}$$

The relevant measures in $\mathbb{B}_{\mathrm{KL}}(\widehat{\nu}, \varepsilon)$ then share the form

$$\nu = \sum_{j \in [N]} y_j \delta_{\widehat{x}_j} + (1 - \sum_{j \in [N]} y_j) \delta_x$$

for some $y \in \Delta_N$. In this case, the optimistic likelihood 5 evaluates to

$$\nu^\star_{\mathrm{KL}}(x) = \max_{\substack{y \in \Delta_N \\ y > 0}} \left\{ 1 - \sum_{j \in [N]} y_j : \sum_{j \in [N]} y_j f\left(\frac{\widehat{\nu}_j}{y_j}\right) - \left(1 - \sum_{j \in [N]} y_j\right) f(0) \leq \varepsilon \right\}.$$

Since f is convex, the above program is a finite convex program in y. We now show that the above optimization problem admits an analytical solution. Consider the equivalent minimization problem

$$\mathrm{OPT}^\star_{\mathrm{KL}} \triangleq \min_{\substack{y \in \Delta_N \\ y > 0}} \left\{ \sum_{j \in [N]} y_j : \sum_{j \in [N]} \widehat{\nu}_j \log \widehat{\nu}_j - \sum_{j \in [N]} \widehat{\nu}_j \log y_j \leq \varepsilon \right\}. \tag{A.5}$$

Suppose that $\varepsilon > 0$. By a standard duality argument, the above program is equivalent to

$$\begin{aligned}
\mathrm{OPT}^\star_{\mathrm{KL}} &= \inf_{\substack{y\in\Delta_N \\ y>0}} \sup_{\gamma\geq 0} \left\{ \sum_{j\in[N]} y_j + \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j \log \widehat{\nu}_j - \varepsilon - \sum_{j\in[N]} \widehat{\nu}_j \log y_j \right) \right\} && \text{(A.6a)} \\
&= \sup_{\gamma\geq 0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j \log \widehat{\nu}_j - \varepsilon \right) + \inf_{\substack{y\in\Delta_N \\ y>0}} \left\{ \sum_{j\in[N]} y_j - \gamma \sum_{j\in[N]} \widehat{\nu}_j \log y_j \right\} \right\} && \text{(A.6b)} \\
&\geq \sup_{1\geq\gamma>0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j \log \widehat{\nu}_j - \varepsilon \right) + \inf_{\substack{y\in\Delta_N \\ y>0}} \left\{ \sum_{j\in[N]} y_j - \gamma \sum_{j\in[N]} \widehat{\nu}_j \log y_j \right\} \right\} && \text{(A.6c)} \\
&= \sup_{1\geq\gamma>0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j - \varepsilon \right) - \sum_{j\in[N]} \widehat{\nu}_j \gamma \log \gamma \right\}, && \text{(A.6d)}
\end{aligned}$$

where the equality (A.6b) follows from strong duality since the Slater condition for the primal problem is satisfied. The inequality (A.6c) follows directly from the restriction of the feasible set of γ and because the objective function is continuous in γ. For any $\gamma \in (0, 1]$, the inner minimization admits the optimal solution $y^\star_j = \gamma\widehat{\nu}_j$, and this leads to the last equation (A.6d). The maximization over γ is now a convex optimization problem, and the first-order condition gives the optimal solution $\gamma^\star = \exp(-\varepsilon)$. We can thus conclude that

$$\mathrm{OPT}^\star_{\mathrm{KL}} \geq \exp(-\varepsilon).$$

By substituting the feasible solution

$$y_j = \exp(-\varepsilon)\,\widehat{\nu}_j \quad \forall j \in [N]$$

into (A.6a), we see that $\mathrm{OPT}^\star_{\mathrm{KL}} \leq \exp(-\varepsilon)$. Hence,

$$\mathrm{OPT}^\star_{\mathrm{KL}} = \exp(-\varepsilon) \quad \forall \varepsilon > 0.$$

Consider now the optimal value $\mathrm{OPT}^\star_{\mathrm{KL}}$ defined in (A.5) as a parametric function of the radius ε over the domain $\mathbb{R}_+$. One can show that $\mathrm{OPT}^\star_{\mathrm{KL}}$ is a continuous function over $\varepsilon \in \mathbb{R}_+$ using Berge's maximum theorem [1, Theorem 17.31]. Furthermore, the function $\exp(-\varepsilon)$ is also continuous over $\varepsilon \in \mathbb{R}_+$. We thus conclude that

$$\mathrm{OPT}^\star_{\mathrm{KL}} = \exp(-\varepsilon) \quad \forall \varepsilon \geq 0.$$

The proof for this case is completed by noticing that $\nu^\star_{\mathrm{KL}}(x) = 1 - \mathrm{OPT}^\star_{\mathrm{KL}}$. □

A.2 Proofs of Section 4

Proof of Proposition 4.2. When $\varepsilon = 0$, $\mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ is the singleton set $\{\widehat{\nu}\}$ and the claim is trivial. It thus suffices to consider $\varepsilon > 0$. Since $\mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ is weakly compact [45, Proposition 3] and the objective function in (8) is upper-semicontinuous in ν by Lemma A.1, a version of the Weierstrass maximum value theorem [1, Theorem 2.43] implies that there exists $\nu^\star \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ such that

$$\sup_{\nu\in\mathbb{B}_{\mathrm{W}}(\widehat{\nu},\varepsilon)} \nu(x) = \nu^\star_{\mathrm{W}}(x).$$

Suppose that $\bar{\nu}$ is an optimal measure that solves (8), that is, $\bar{\nu} \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ and $\bar{\nu}(x) = \nu^\star_{\mathrm{W}}(x)$. Since the ground metric distance $d(\cdot, \cdot)$ in the Wasserstein distance is continuous, there exists an optimal transport plan $\bar{\lambda}$ that maps $\widehat{\nu}$ to $\bar{\nu}$ [55, Theorem 4.1]. Since $\widehat{\nu}$ is a discrete distribution with N atoms, this optimal transport map can be characterized by N functions $\bar{\lambda}_j : \mathcal{X} \to \mathbb{R}_+$, $j \in [N]$, which satisfy the balancing constraints

$$\sum_{z\in\mathcal{X}} \bar{\lambda}_j(z) = \widehat{\nu}_j \;\forall j \in [N] \quad \text{and} \quad \sum_{j=1}^{N} \bar{\lambda}_j(z) = \bar{\nu}(z) \;\forall z \in \mathcal{X}$$

as well as the Wasserstein distance constraint

$$\sum_{j\in[N]} \sum_{z\in\mathcal{X}} d(\widehat{x}_j, z)\bar{\lambda}_j(z) \leq \varepsilon. \tag{A.7}$$

Define κ_j and η_j as

$$\kappa_j \triangleq \sum_{z\in\mathcal{X}\setminus(\widehat{\mathcal{S}}\cup\{x\})} \bar{\lambda}_j(z) \quad \text{and} \quad \eta_j \triangleq \sum_{z\in\mathcal{X}\setminus(\widehat{\mathcal{S}}\cup\{x\})} d(\widehat{x}_j, z)\bar{\lambda}_j(z) \qquad \forall j \in [N].$$

By construction, we have $0 \leq \kappa_j \leq \widehat{\nu}_j \leq 1$ and $0 \leq \eta_j$ for all $j \in [N]$. Suppose that $\bar{\nu}$ assigns non-zero probability mass on $\mathcal{X}\setminus(\widehat{\mathcal{S}} \cup \{x\})$, where $\widehat{\mathcal{S}} = \operatorname{supp}(\widehat{\nu})$. In that case, there exists $j \in [N]$ such that $\kappa_j > 0$ and $\eta_j > 0$. We will next show that $\bar{\nu}$ cannot be the optimal solution.

Assume first that $x \notin \widehat{\mathcal{S}}$, and define the transport maps $\lambda'_j : \mathcal{X} \to \mathbb{R}_+$ for $j \in [N]$ as

$$\lambda'_j(z) = \begin{cases} \bar{\lambda}_j(\widehat{x}_j) + \left(1 - \min\left\{1, \frac{\eta_j}{d(x,\widehat{x}_j)}\right\}\right)\kappa_j & \text{if } z = \widehat{x}_j, \\ \bar{\lambda}_j(\widehat{x}_k) & \text{if } z = \widehat{x}_k, k \neq j, k \in [N], \\ \bar{\lambda}_j(x) + \min\left\{1, \frac{\eta_j}{d(x,\widehat{x}_j)}\right\}\kappa_j & \text{if } z = x, \\ 0 & \text{otherwise.} \end{cases}$$

By this construction of λ'_j, we obtain

$$\sum_{z\in\mathcal{X}} \lambda'_j(z) = \sum_{z\in\mathcal{X}} \bar{\lambda}_j(z) = \widehat{\nu}_j \qquad \forall j \in [N].$$

We now construct a measure ν' explicitly using the transport map λ' from $\widehat{\nu}$ as

$$\nu'(z) = \sum_{j\in[N]} \lambda'_j(z) \qquad \forall z \in \mathcal{X}. \tag{A.8}$$

Notice that ν' is supported on $\widehat{\mathcal{S}} \cup \{x\}$, $\nu' \geq 0$ and

$$\sum_{z\in\mathcal{X}} \nu'(z) = \sum_{j\in[N]} \left(\sum_{k\in[N]} \bar{\lambda}_j(\widehat{x}_k) + \kappa_j + \bar{\lambda}_j(x) \right) = \sum_{j\in[N]} \sum_{z\in\mathcal{X}} \bar{\lambda}_j(z) = \sum_{j\in[N]} \widehat{\nu}_j = 1,$$

which further implies that ν' is a probability measure on $\mathcal{X}$. Moreover, we have

$$\begin{aligned}
\mathbb{W}(\widehat{\nu}, \nu') &\leq \sum_{j\in[N]} \sum_{k\in[N]} d(\widehat{x}_j, \widehat{x}_k)\, \lambda'_j(\widehat{x}_k) + \sum_{j\in[N]} d(\widehat{x}_j, x)\lambda'_j(x) && \text{(A.9a)} \\
&= \sum_{j\in[N]} \left(\sum_{k\in[N]} d(\widehat{x}_j, \widehat{x}_k)\bar{\lambda}_j(\widehat{x}_k) + d(\widehat{x}_j, x)\bar{\lambda}_j(x) + \min\left\{d(\widehat{x}_j, x)\kappa_j, \eta_j\kappa_j\right\} \right) \\
&\leq \sum_{j\in[N]} \left(\sum_{k\in[N]} d(\widehat{x}_j, \widehat{x}_k)\bar{\lambda}_j(\widehat{x}_k) + d(\widehat{x}_j, x)\bar{\lambda}_j(x) + \eta_j\kappa_j \right) \\
&\leq \sum_{j\in[N]} \left(\sum_{k\in[N]} d(\widehat{x}_j, \widehat{x}_k)\bar{\lambda}_j(\widehat{x}_k) + d(\widehat{x}_j, x)\bar{\lambda}_j(x) + \eta_j \right) && \text{(A.9b)} \\
&= \sum_{j\in[N]} \left(\sum_{k\in[N]} d(\widehat{x}_j, \widehat{x}_k)\bar{\lambda}_j(\widehat{x}_k) + d(\widehat{x}_j, x)\bar{\lambda}_j(x) + \sum_{z\in\mathcal{X}\setminus(\widehat{\mathcal{S}}\cup\{x\})} d(\widehat{x}_j, z)\bar{\lambda}_j(z) \right) \\
&= \sum_{j\in[N]} \sum_{z\in\mathcal{X}} d(\widehat{x}_j, z)\bar{\lambda}_j(z) \leq \varepsilon. && \text{(A.9c)}
\end{aligned}$$

Inequality (A.9a) holds because of the definition of the Wasserstein distance and the fact that $\{\lambda'_j\}_{j\in[N]}$ constitutes a feasible transportation plan from $\widehat{\nu}$ to ν'. Inequality (A.9b) holds due to the non-negativity of both η_j and κ_j and the fact that $\kappa_j \leq 1$. Inequality (A.9c) is a consequence of (A.7). The last inequality implies that $\nu' \in \mathbb{B}_{\mathbb{W}}(\widehat{\nu}, \varepsilon)$, and thus ν' is a feasible measure for the optimistic likelihood problem. Finally, we have

$$\nu'(x) = \sum_{j\in[N]} \lambda'_j(x) = \sum_{j\in[N]} \left(\bar{\lambda}_j(x) + \min\left\{1, \frac{\eta_j}{d(x,\widehat{x}_j)}\right\}\kappa_j \right) > \sum_{j\in[N]} \bar{\lambda}_j(x) = \bar{\nu}(x),$$

where the strict inequality is from the fact that there exists $j \in [N]$ such that $\kappa_j > 0$ and $\eta_j > 0$. Thus, $\nu' \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ attains a higher objective value than $\bar{\nu}$, and as a consequence $\bar{\nu}$ cannot be an optimal measure. Notice that $\mathrm{supp}(\nu') \subseteq (\widehat{\mathcal{S}} \cup \{x\})$ by construction, and thus we conclude that when $x \notin \widehat{\mathcal{S}}$, the optimal measure $\nu^\star_{\mathrm{W}}$ satisfies $\mathrm{supp}(\nu^\star_{\mathrm{W}}) \subseteq (\widehat{\mathcal{S}} \cup \{x\})$.

Assume now that $x \in \widehat{\mathcal{S}}$, and assume without loss of generality that $x = \widehat{x}_N$. Consider now the transport plan $\lambda'_j : \mathcal{X} \to \mathbb{R}_+$ for any $j \in [N]$ defined as

$$\forall j \in [N-1]: \ \lambda'_j(z) = \begin{cases} \bar{\lambda}_j(\widehat{x}_j) + \left(1 - \min\left\{1, \frac{\eta_j}{d(x, \widehat{x}_j)}\right\}\right) \kappa_j & \text{if } z = \widehat{x}_j, \\ \bar{\lambda}_j(\widehat{x}_k) & \text{if } z = \widehat{x}_k, k \neq j, k \in [N-1], \\ \bar{\lambda}_j(x) + \min\left\{1, \frac{\eta_j}{d(x, \widehat{x}_j)}\right\} \kappa_j & \text{if } z = \widehat{x}_N, \\ 0 & \text{otherwise} \end{cases}$$

and

$$\lambda'_N(z) = \begin{cases} \bar{\lambda}_N(\widehat{x}_k) & \text{if } z = \widehat{x}_k, k \in [N-1], \\ \bar{\lambda}_N(\widehat{x}'_N) + \kappa_N & \text{if } z = \widehat{x}_N, \\ 0 & \text{otherwise.} \end{cases}$$

One can readily verify that using the collection $\{\lambda'_j\}_{j \in [N]}$ to define ν' in (A.8) results in a probability measure $\nu' \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ that attains a strictly higher objective value than $\bar{\nu}$. Notice that this construction satisfies $\mathrm{supp}(\nu') \subseteq \widehat{\mathcal{S}}$, and hence we can conclude that when $x \in \widehat{\mathcal{S}}$, the optimal measure $\nu^\star_{\mathrm{W}}$ satisfies $\mathrm{supp}(\nu^\star_{\mathrm{W}}) \subseteq \widehat{\mathcal{S}}$. This completes the proof. □

Proof of Theorem 4.3. As a result of Proposition 4.2, we can restrict ourselves to probability measures that are supported on $\mathrm{supp}(\widehat{\nu}) \cup \{x\}$. Thus, it suffices to optimize over the set of discrete probability measures of the form

$$\nu = \sum_{j \in [N]} y_j \delta_{\widehat{x}_j} + \left(1 - \sum_{j \in [N]} y_j\right) \delta_x$$

for some $y \in \Delta_N$, where Δ_N is the simplex defined in (A.4). Using the Definition 4.1 of the type-1 Wasserstein distance, we can rewrite the optimistic likelihood problem over the Wasserstein ball $\mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ as the linear program

$$\sup_{\nu \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)} \nu(x) = \left\{ \begin{array}{lll} \sup & 1 - \sum_{j \in [N]} y_j & \\ \text{s.t.} & y \in \Delta_N, \ \lambda \in \mathbb{R}_+^{N \times (N+1)} & \\ & \sum_{j \in [N]} \sum_{j' \in [N]} d(\widehat{x}_j, \widehat{x}_{j'}) \lambda_{jj'} + \sum_{j \in [N]} d(\widehat{x}_j, x) \lambda_{j(N+1)} \leq \varepsilon & \\ & \sum_{j' \in [N+1]} \lambda_{jj'} = \widehat{\nu}_j & \forall j \in [N] \\ & \sum_{j \in [N]} \lambda_{jj'} = y_j & \forall j' \in [N] \\ & \sum_{j \in [N]} \lambda_{j(N+1)} = 1 - \sum_{j \in [N]} y_j. & \end{array} \right.$$

From the last constraint, we can see that maximizing $1 - \sum_{j \in [N]} y_j$ is equivalent to maximizing $\sum_{j \in [N]} \lambda_{j(N+1)}$. In particular, we thus conclude that it is optimal to set $\lambda_{jj'} = 0$ for any $j \in [N], j' \in [N]$ such that $j \neq j'$. We thus have

$$\sup_{\nu \in \mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)} \nu(x) = \left\{ \begin{array}{ll} \sup & \sum_{j \in [N]} \lambda_{j(N+1)} \\ \text{s.t.} & y \in \Delta_N, \ \lambda \in \mathbb{R}_+^{N \times (N+1)} \\ & \lambda_{jj'} = 0 \quad \forall j \in [N], j' \in [N], j \neq j' \\ & \sum_{j \in [N]} d(\widehat{x}_j, x) \, \lambda_{j(N+1)} \leq \varepsilon \\ & \lambda_{jj} + \lambda_{j(N+1)} = \widehat{\nu}_j, \quad \lambda_{jj} = y_j \qquad \forall j \in [N]. \end{array} \right.$$

By letting $T_j = \lambda_{j(N+1)}$ and eliminating the redundant components of λ, we obtain the desired reformulation. This completes the proof. □

Proof of Proposition 4.4. By a change of variables, we define the weight $\widehat{w}_j = d(\widehat{x}_j, x)\widehat{\nu}_j$ and the decision variables $z_j = \widehat{\nu}_j^{-1} T_j$ for every $j \in [N]$. The optimal value of problem (9) then coincides with the optimal value of

$$\max \left\{ \sum_{j\in[N]} \widehat{\nu}_j z_j : z \in \mathbb{R}^N_+, \sum_{j\in[N]} \widehat{w}_j z_j \leq \varepsilon, z_j \leq 1 \ \forall j \in [N] \right\}, \tag{A.10}$$

which is a continuous (or fractional) knapsack problem. The special structure of (A.10) guarantees

$$\frac{\widehat{\nu}_j}{\widehat{w}_j} = \frac{1}{d(\widehat{x}_j, x)} \quad \forall j \in [N],$$

and hence the continuous knapsack problem (A.10) admits an optimal solution $z^\star$ that can be found by sorting the support points $\widehat{x}_j$ in increasing order of distance from x and then exhausting the budget ε according to the sorted order (see [16] or [28, Proposition 17.1]). Since sorting an array of N scalars can be achieved in time $\mathcal{O}(N \log N)$, problem (A.10) can be solved efficiently, and the optimal solution $T^\star$ of (9) can be constructed from the optimal solution $z^\star$ of (A.10) by setting

$$T_j^\star = \widehat{\nu}_j z_j^\star \quad \forall j \in [N].$$

This completes the proof. □

Corollary A.2 (Comparative statics)**.** If the radius ε of the Wasserstein ball is strictly positive, then $\nu^\star_{\mathbb{W}}(x) > 0$. Moreover, if the radius satisfies $\varepsilon \geq \sum_{j\in[N]} d(x, \widehat{x}_j)\widehat{\nu}_j$, then $\nu^\star_{\mathbb{W}}(x) = 1$.

The proof of Corollary A.2 follows directly from examining the optimal value of the linear program (9) and is thus omitted.

A.3 Proofs of Section 5

In the proofs of this section, we denote by ν_i^{true} the unknown true probability measure that induces the probability mass function $p(\cdot|\theta_i)$ for each $i \in [C]$.

Proof of Theorem 5.3. Define for each $i \in [C]$ the set

$$\Phi_i \triangleq \left\{ \nu_i \in \mathcal{M}(\mathcal{X}) : \mathrm{KL}(\nu_i \parallel \nu_i^{\text{true}}) > \varepsilon_i \right\},$$

where the dependence of Φ_i on ε_i and ν_i^{true} has been made implicit. Under Assumption 5.2, the empirical measure $\widehat{\nu}_i^{N_i}$ satisfies the large deviation principle with rate function $\mathrm{KL}(\cdot \parallel \nu_i^{\text{true}})$ [18, Theorem 6.2.10]. Sanov's theorem then implies that

$$\limsup_{N_i\to\infty} \frac{1}{N_i} \log \mathbb{P}^\infty\left(\widehat{\nu}_i^{N_i} \in \Phi_i\right) \leq -\varepsilon_i < 0 \qquad \forall i \in [C]. \tag{A.11}$$

This in turn implies that there exist positive constants $\kappa_i < \infty$ such that

$$\mathbb{P}^{N_i}\left(\widehat{\nu}_i^{N_i} \in \Phi_i\right) \leq \kappa_i \exp(-N_i\varepsilon_i) \quad \text{as } N_i \to \infty.$$

We now have

$$\begin{aligned}
\mathbb{P}^\infty(\mathcal{J}^{\text{true}} \geq \widehat{\mathcal{J}}_{\mathbb{B}^N}) &\geq \mathbb{P}^\infty\left(\nu_i^{\text{true}} \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}_i^{N_i}, \varepsilon_i) \ \forall i \in [C]\right) && \text{(A.12)}\\
&= \prod_{i\in[C]} \mathbb{P}^{N_i}\left(\nu_i^{\text{true}} \in \mathbb{B}_{\mathrm{KL}}(\widehat{\nu}_i^{N_i}, \varepsilon_i)\right) && \text{(A.13)}\\
&= \prod_{i\in[C]} \left(1 - \mathbb{P}^{N_i}\left(\widehat{\nu}_i^{N_i} \in \Phi_i\right)\right) && \text{(A.14)}\\
&\geq 1 - \sum_{i\in[C]} \mathbb{P}^{N_i}\left(\widehat{\nu}_i^{N_i} \in \Phi_i\right). && \text{(A.15)}
\end{aligned}$$

Here, equality (A.13) follows from our i.i.d. assumption. Equality (A.14) follows from the fact that the event $\nu_i^{\text{true}} \in \mathbb{B}_{\text{KL}}(\widehat{\nu}_i^{N_i}, \varepsilon_i)$ is the complement of the event $\widehat{\nu}_i^{N_i} \in \Phi_i$. Inequality (A.15), finally, is due to the Weierstrass product inequality. Thus, for each i there exists $C_i < \infty$ such that as $N_i \to \infty$, we have

$$\mathbb{P}^\infty(\mathcal{J}^{\text{true}} < \widehat{\mathcal{J}}_{\mathbb{B}^N}) \leq \sum_{i\in[C]} \mathbb{P}^{N_i}\big(\widehat{\nu}_i^{N_i} \in \Phi_i\big) \leq \sum_{i\in[C]} \kappa_i \exp\big(-N_i\varepsilon_i\big) \leq \kappa C \exp\big(-n \min_{i\in[C]}\{\varepsilon_i\}\big)$$

for some $\kappa = \max_{i\in[C]} \kappa_i < \infty$. This further implies that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}^\infty(\mathcal{J}^{\text{true}} < \widehat{\mathcal{J}}_{\mathbb{B}^N}) \leq -\min_{i\in[C]}\{\varepsilon_i\} < 0.$$

This observation completes the proof. □

Proof of Theorem 5.5. If ε_i is chosen as in the statement of the theorem, then the measure concentration result for the Wasserstein distance [19, Theorem 2] implies that

$$\mathbb{P}^{N_i}\left(\mathbb{W}(\nu_i^{\text{true}}, \widehat{\nu}_i^{N_i}) \geq \varepsilon_i(\beta, C, N_i)\right) \leq \frac{\beta}{C}.$$

Thus, by applying the union bound, we obtain

$$\mathbb{P}^N\left(\mathbb{W}(\nu_i^{\text{true}}, \widehat{\nu}_i^{N_i}) \geq \varepsilon_i(\beta, C, N_i)\ \forall i\right) = \sum_i \mathbb{P}^{N_i}\left(\mathbb{W}(\nu_i^{\text{true}}, \widehat{\nu}_i^{N_i}) \geq \varepsilon_i(\beta, C, N_i)\right) \leq \beta,$$

which implies that

$$\mathbb{P}^N\left(\nu_i^{\text{true}} \in \mathbb{B}_{\mathbb{W}}\left(\widehat{\nu}_i^{N_i}, \varepsilon_i(\beta, C, N_i)\right)\ \forall i\right) \geq 1 - \beta.$$

We can now conclude that $\widehat{\mathcal{J}}_{\mathbb{B}^N} \leq \mathcal{J}^{\text{true}}$ with probability at least $1 - \beta$. □

Proof of Theorem 5.6. For every $i \in [C]$, let $\nu_i^\star \in \mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})$ be an optimal solution of the problem

$$\sup_{\nu_i \in \mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})} \nu_i(x), \tag{A.16}$$

where the dependence of $\nu_i^\star$ on the number of samples N_i has been omitted to avoid clutter. The existence of $\nu_i^\star \in \mathbb{B}_i^{N_i}(\widehat{\nu}_i^{N_i})$ is guaranteed by Proposition 4.2. By [35, Lemma 3.7], for every $i \in [C]$ it holds $(\nu_i^{\text{true}})^\infty$-almost surely that

$$\lim_{N_i\to\infty} \mathbb{W}\left(\nu_i^{\text{true}}, \nu_i^\star\right) = 0.$$

Therefore, by [55, Theorem 6.9], $\nu_i^\star$ converges to ν_i^{true} weakly as $N_i \to \infty$. Since $\mathbb{1}_x(\cdot)$ is a bounded, upper semicontinuous function, the weak continuity implies that $(\nu_i^{\text{true}})^\infty$-almost surely as $N_i \to \infty$, we have that

$$\nu_i^\star(x) \to \nu_i^{\text{true}}(x) = p(x|\theta_i). \tag{A.17}$$

Let $u^{\text{true}} \in [0,1]^C$ be the vector defined by $(u^{\text{true}})_i = p(x|\theta_i)$ for $i \in [C]$. Since $(u^{\text{true}})_i > 0$ for $i = 1, \ldots, C$, there exists $\underline{u} > 0$ such that $u^{\text{true}} \in [\underline{u}, 1]^C$. Consider the parametrized optimization problems

$$\mathcal{J}^\star(u) \triangleq \min_{q\in\mathcal{Q}} \left\{ \mathcal{J}(q,u) \triangleq \sum_{i\in[C]} q_i(\log q_i - \log \pi_i) - \sum_{i\in[C]} q_i \log u_i \right\}, \quad u \in [\underline{u}, 1]^C.$$

We observe that $\mathcal{J}(\cdot,\cdot)$ is jointly continuous on $\mathcal{Q} \times [\underline{u}, 1]^C$, $\mathcal{Q}$ is compact, and the level sets

$$\left\{ q \in \mathcal{Q} : \mathcal{J}(q,u) \leq -\sum_{i\in[C]} \pi_i \log \underline{u} \right\}$$

are non-empty and uniformly bounded over all $u \in [\underline{u}, 1]^C$. By [10, Proposition 4.4] and the discussion following its proof, $\mathcal{J}^\star(u)$ is continuous on $[\underline{u}, 1]^C$. The continuity of $\mathcal{J}^\star(\cdot)$ and the convergence (A.17) together imply that $(\nu_1^{\text{true}})^\infty \times \cdots \times (\nu_C^{\text{true}})^\infty$-almost surely, and we thus have

$$\widehat{\mathcal{J}}_{\mathbb{B}^N} = \mathcal{J}^\star((\nu_1^\star(x), \ldots, \nu_C^\star(x))) \to \mathcal{J}^\star(u^{\text{true}}) = \mathcal{J}^{\text{true}} \quad \text{as } N_1, \ldots, N_C \to \infty.$$

This observation completes the proof. □

Appendix B Additional Material

B.1 A Measure-Theoretic Derivation of the Evidence Lower Bound Problem

To keep the paper self-contained, we present in this section a derivation of the evidence lower bound (ELBO), which is a fundamental building block of the variational Bayes method.

Consider a standard Bayesian inference model where the random vector $\tilde{x}$, supported on a sample space $\mathcal{X}$, is governed by one of the distributions $\mathbb{P}_\theta$ parameterized by $\theta \in \Theta$. We assume that there exists a measure $\bar{\nu}$ on $\mathcal{X}$ such that $\mathbb{P}_\theta$ is absolutely continuous with respect to $\bar{\nu}$ for all $\theta \in \Theta$. Moreover, we denote by $f_{\tilde{x}|\theta}$ the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to $\bar{\nu}$, that is

$$f_{\tilde{x}|\theta}(x|\theta) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\bar{\nu}}(x) \quad \forall x \in \mathcal{X}.$$

Finally, we denote by π the prior measure on the parameter space Θ, while $\mathbb{P}_x$ denotes the posterior measure on Θ after observing x.

Consider an optimal solution $\mathbb{Q}^\star$ of the optimization problem

$$\mathbb{Q}^\star \in \arg\min_{\mathbb{Q}\in\mathcal{Q}} \ \mathrm{KL}(\mathbb{Q} \parallel \mathbb{P}_x),$$

where $\mathrm{KL}(\cdot \parallel \cdot)$ denotes the KL divergence defined in Definition 2.1. If the feasible set $\mathcal{Q}$ is the collection of all possible probability measures supported on Θ, then $\mathbb{Q}^\star = \mathbb{P}_x$. The objective function of this problem can be re-expressed as

$$\begin{aligned}
\mathrm{KL}(\mathbb{Q} \parallel \mathbb{P}_x) &= \int_\Theta \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}_x}\right) \mathrm{d}\mathbb{Q} && \text{(B.1a)}\\
&= \int_\Theta \log\left(\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\pi}\right) \mathrm{d}\mathbb{Q} - \int_\Theta \log\left(\frac{\mathrm{d}\mathbb{P}_x}{\mathrm{d}\pi}\right) \mathrm{d}\mathbb{Q} && \text{(B.1b)}\\
&= \mathrm{KL}(\mathbb{Q} \parallel \pi) - \int_\Theta \log\left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\bar{\nu}}(x)\right) \mathrm{d}\mathbb{Q} + \log \int_\Theta f_{\tilde{x}|\theta}(x|\theta)\mathrm{d}\pi, && \text{(B.1c)}
\end{aligned}$$

where the equality (B.1a) follows from the definition of KL divergence, and (B.1b) is due to the chain rule for the Radon-Nikodym derivatives because $\mathbb{P}_x \ll \pi$ [49, Theorem 1.31]. Equality (B.1c), finally, holds since

$$\frac{\mathrm{d}\mathbb{P}_x}{\mathrm{d}\pi}(\theta) = \frac{f_{\tilde{x}|\theta}(x|\theta)}{\int_\Theta f_{\tilde{x}|\theta}(x|\theta)\mathrm{d}\pi(\theta)} = \frac{1}{\int_\Theta f_{\tilde{x}|\theta}(x|\theta)\mathrm{d}\pi(\theta)} \cdot \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\bar{\nu}}(x),$$

where the first equality follows from Bayes' theorem [49, Theorem 1.31] and the second equality is due to the definition of $f_{\tilde{x}|\theta}$. Since the last term in (B.1c) does not involve the decision variable $\mathbb{Q}$, the measure $\mathbb{Q}^\star$ can be equivalently expressed as the optimal solution of

$$\min_{\mathbb{Q}\in\mathcal{Q}} \ \mathrm{KL}(\mathbb{Q} \parallel \pi) - \int_\Theta \log\left(\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\bar{\nu}}(x)\right) \mathrm{d}\mathbb{Q}.$$

If we define the conditional density $p(x|\theta)$ with respect to $\bar{\nu}$ of $\tilde{x}$ given the parameter θ [49, Section 1.3.1], that is,

$$p(x|\theta) = f_{\tilde{x}|\theta}(x|\theta),$$

then $\mathbb{Q}^\star$ solves

$$\min_{\mathbb{Q}\in\mathcal{Q}} \ \mathrm{KL}(\mathbb{Q} \parallel \pi) - \mathbb{E}_{\mathbb{Q}}[\log p(x|\theta)].$$

The function $p(x|\theta)$, considered as a function of the parameter θ after x has been observed, is often called the *likelihood function*. If $p(x|\theta)$ is considered as a function of x given the parameter θ, then it is often called the *conditional density*.

B.2 Generalization to f-Divergence Ambiguity Sets

In this section, we consider the class of ambiguity sets described by f-divergences, which generalizes the KL ambiguity set from Section 2.

Definition B.1 (f-divergence)**.** The f-divergence D_f between two measures ν_1 and ν_2 supported on $\mathcal{X}$ is defined as

$$D_f(\nu_1 \parallel \nu_2) = \int_{z\in\mathcal{X}} f\left(\frac{\nu_1(z)}{\nu_2(z)}\right)\nu_2(z),$$

where $f : \mathbb{R} \to \mathbb{R}$ is a convex function satisfying $f(1) = 0$. More specifically,

- If $f(t) = t\log(t) - t + 1$, then D_f is the *Kullback-Leibler divergence*.
- If $f(t) = 1 - \sqrt{t}$, then D_f is the *Hellinger distance*.
- If $f(t) = (t-1)^2$, then D_f is the *Pearson's χ^2-divergence*.
- If $f(t) = |t-1|$, then D_f is the *total variation distance*.

We now consider the f-divergence ball $\mathbb{B}_f(\widehat{\nu}, \varepsilon)$ of radius $\varepsilon \geq 0$, which contains all probability measures in the neighborhood of $\widehat{\nu}$ as measured by the f-divergence:

$$\mathbb{B}_f(\widehat{\nu}, \varepsilon) \triangleq \{\nu \in \mathcal{M}(\mathcal{X}) : D_f(\widehat{\nu} \parallel \nu) \leq \varepsilon\} \tag{B.2}$$

Moreover, we assume that the nominal distribution $\widehat{\nu}$ is supported on N distinct points $\widehat{x}_1, \ldots, \widehat{x}_N$, that is, $\widehat{\nu} = \sum_{j\in[N]} \widehat{\nu}_j \delta_{\widehat{x}_j}$ with $\widehat{\nu}_j > 0\, \forall j \in [N]$ and $\sum_{j\in[N]} \widehat{\nu}_j = 1$.

In analogy to Section 2, we first provide a generalized version of Proposition 2.2.

Corollary B.2 (Existence of optimizers; f-divergence ambiguity)**.** For any $\varepsilon \geq 0$ and $x \in \mathcal{X}$, there exists a measure $\nu_f^\star \in \mathbb{B}_f(\widehat{\nu}, \varepsilon)$ such that

$$\sup_{\nu\in\mathbb{B}_f(\widehat{\nu},\varepsilon)} \nu(x) = \nu_f^\star(x). \tag{B.3}$$

Moreover, $\nu_f^\star$ is supported on at most $N+1$ points satisfying $\mathrm{supp}(\nu_f^\star) \subseteq \mathrm{supp}(\widehat{\nu}) \cup \{x\}$.

The proof of Corollary B.2 follows from the proof of Proposition 2.2 and thus it is omitted.

Theorem B.3 (Optimistic likelihood; f-divergence ambiguity)**.** Suppose that $\widehat{\nu} = \sum_{j\in[N]} \widehat{\nu}_j \delta_{\widehat{x}_j}$. For any data point $x \in \mathcal{X}$, the optimization problem in (B.3) can be reformulated as a finite convex program. Moreover, if $x \neq \widehat{x}_j$ for all $j \in [N]$, then:

1. If D_f is the Hellinger distance, then for any $\varepsilon \in [0, 1]$, we have $\nu^\star_{\text{Hellinger}}(x) = 1 - (1-\varepsilon)^2$.
2. If D_f is the Pearson's χ^2-divergence, then for any $\varepsilon \in \mathbb{R}_+$, we have $\nu^\star_{\chi^2}(x) = 1 - (1+\varepsilon)^{-1}$.
3. If D_f is the total variation distance, then for any $\varepsilon \in \mathbb{R}_+$, we have $\nu^\star_{\text{TV}}(x) = \varepsilon/2$.

Proof of Theorem B.3. The reformulation as a convex program follows directly from the first part of the proof of Theorem 2.3 using the general function f, and it is thus omitted. We now proceed to consider the case when $x \notin \widehat{\mathcal{S}}$, and we derive the optimal value $\nu_f^\star(x)$ for each divergence f.

1. **Hellinger distance.** Following the same approach as in the proof of Theorem 2.3, we employ the definition of the Hellinger distance to obtain the equivalent minimization problem

$$\text{OPT}^\star_{\text{Hellinger}} = \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j : \sum_{j\in[N]} \widehat{\nu}_j - \sum_{j\in[N]} \sqrt{\widehat{\nu}_j}\sqrt{y_j} \leq \varepsilon \right\}.$$

Suppose that $\varepsilon \in (0,1]$. Using a duality argument, we have

$$\begin{aligned}
\mathrm{OPT}^{\star}_{\text{Hellinger}} &= \min_{y\in\Delta_N} \max_{\gamma\geq 0} \left\{ \sum_{j\in[N]} y_j + \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j - \sum_{j\in[N]} \sqrt{\widehat{\nu}_j}\sqrt{y_j} - \varepsilon \right) \right\} \\
&= \max_{\gamma\geq 0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j - \varepsilon \right) + \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j - \gamma \sum_{j\in[N]} \sqrt{\widehat{\nu}_j}\sqrt{y_j} \right\} \right\} \\
&\geq \sup_{2\geq\gamma>0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j - \varepsilon \right) + \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j - \gamma \sum_{j\in[N]} \sqrt{\widehat{\nu}_j}\sqrt{y_j} \right\} \right\} \\
&= \sup_{2\geq\gamma>0} \left\{ \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j - \varepsilon \right) - \frac{\gamma^2}{4} \sum_{j\in[N]} \widehat{\nu}_j \right\},
\end{aligned}$$

where we have used the optimal solution $y^{\star}_j = \gamma^2 \widehat{\nu}_j/4$ to arrive at the last equation. The supremum over γ admits the optimal solution $\gamma^{\star} = 2\,(1-\varepsilon)$. We can thus show that

$$OPT^{\star}_{\text{Hellinger}} \geq (1-\varepsilon)^2 \quad \forall \varepsilon \in (0,1].$$

The rest of the proof is analogous to the proof of Theorem 2.3.

2. **Pearson's χ^2-divergence.** By definition of the divergence, we obtain

$$\mathrm{OPT}^{\star}_{\chi^2} = \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j : \sum_{j\in[N]} \widehat{\nu}_j^2 y_j^{-1} - \sum_{j\in[N]} \widehat{\nu}_j \leq \varepsilon \right\}.$$

Suppose that $\varepsilon > 0$. Using a duality argument, we have

$$\begin{aligned}
\mathrm{OPT}^{\star}_{\chi^2} &= \min_{y\in\Delta_N} \max_{\gamma\geq 0} \left\{ \sum_{j\in[N]} y_j + \gamma \left(\sum_{j\in[N]} \widehat{\nu}_j^2 y_j^{-1} - \sum_{j\in[N]} \widehat{\nu}_j - \varepsilon \right) \right\} \\
&= \max_{\gamma\geq 0} \left\{ -\gamma \left(\sum_{j\in[N]} \widehat{\nu}_j + \varepsilon \right) + \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j + \gamma \sum_{j\in[N]} \widehat{\nu}_j^2 y_j^{-1} \right\} \right\} \\
&\geq \sup_{1\geq\gamma>0} \left\{ -\gamma \left(\sum_{j\in[N]} \widehat{\nu}_j + \varepsilon \right) + \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j + \gamma \sum_{j\in[N]} \widehat{\nu}_j^2 y_j^{-1} \right\} \right\} \\
&= \sup_{1\geq\gamma>0} \left\{ -\gamma \left(\sum_{j\in[N]} \widehat{\nu}_j + \varepsilon \right) + 2\sqrt{\gamma} \sum_{j\in[N]} \widehat{\nu}_j \right\},
\end{aligned}$$

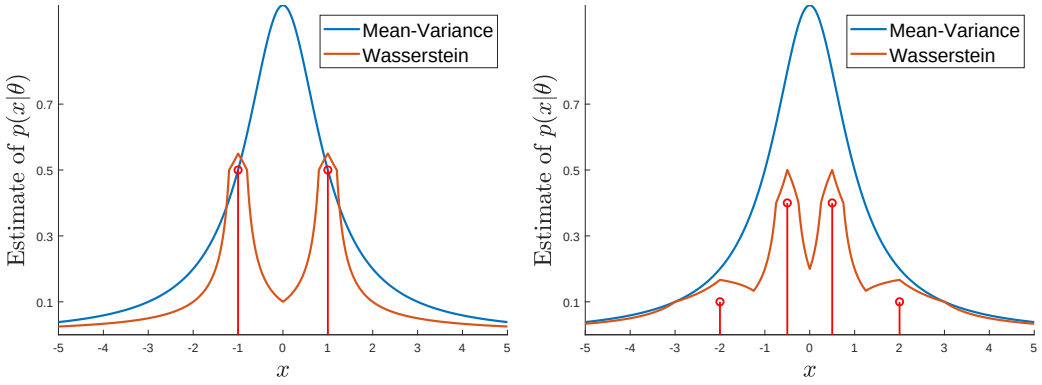
where we have used the optimal solution $y^{\star}_j = \sqrt{\gamma}\widehat{\nu}_j$ to arrive at the last equation. The supremum over γ admits the optimal solution $\gamma^{\star} = (1+\varepsilon)^{-2}$, which implies that

$$\mathrm{OPT}^{\star}_{\chi^2} \geq (1+\varepsilon)^{-1} \quad \forall \varepsilon > 0.$$

The rest of the proof is analogous to the proof of Theorem 2.3.

3. **Total variation distance.** We have

$$\mathrm{OPT}^{\star}_{\text{TV}} = \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j : \sum_{j\in[N]} |\widehat{\nu}_j - y_j| + 1 - \sum_{j\in[N]} y_j \leq \varepsilon \right\}.$$



(a) Nominal measure $\widehat{\nu}^{(1)}$ (b) Nominal measure $\widehat{\nu}^{(2)}$

Figure 4: Approximations of the likelihood $p(x|\theta)$ under two different nominal measures. The approximation offered by the mean-variance ambiguity set is the same for both $\widehat{\nu}^{(1)}$ and $\widehat{\nu}^{(2)}$. In contrast, the approximation offered by the Wasserstein ambiguity set produces a fatter tail under the nominal measure $\widehat{\nu}^{(2)}$, whose support is more spread out.

For any $\varepsilon \geq 0$, the optimal solution $y^\star$ satisfies $y_j^\star \leq \widehat{\nu}_j$, and thus we have

$$\begin{aligned}\text{OPT}_{\text{TV}}^\star &= \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j : \sum_{j\in[N]} (\widehat{\nu}_j - y_j) + 1 - \sum_{j\in[N]} y_j \leq \varepsilon \right\} \\ &= \min_{y\in\Delta_N} \left\{ \sum_{j\in[N]} y_j : 2 - 2\sum_{j\in[N]} y_j \leq \varepsilon \right\} = 1 - \frac{\varepsilon}{2},\end{aligned}$$

which finishes the proof for the total variation distance.

These observations complete the proof. □

B.3 Comparison of Moment and Wasserstein Ambiguity Sets

In this section, we empirically demonstrate that the approximation using the Wasserstein ambiguity set can capture the tail behavior of the nominal distribution $\widehat{\nu}$ better than the approximation using the moment ambiguity set. To this end, consider the two univariate discrete nominal measures

$$\widehat{\nu}^{(1)} = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1 \qquad \text{and} \qquad \widehat{\nu}^{(2)} = 0.1\delta_{-2} + 0.4\delta_{-\frac{1}{2}} + 0.4\delta_{\frac{1}{2}} + 0.1\delta_2.$$

Notice that both $\widehat{\nu}^{(1)}$ and $\widehat{\nu}^{(2)}$ share the same mean 0 and the same variance 1, and thus we find that

$$\sup_{\nu\in\mathbb{B}_{\text{MV}}(\widehat{\nu}^{(1)})} \nu(x) = \sup_{\nu\in\mathbb{B}_{\text{MV}}(\widehat{\nu}^{(2)})} \nu(x) \qquad \forall x \in \mathcal{X}.$$

However, if we use the Wasserstein ambiguity set $\mathbb{B}_{\text{W}}(\cdot)$, then in general we have

$$\sup_{\nu\in\mathbb{B}_{\text{W}}(\widehat{\nu}^{(1)},\varepsilon)} \nu(x) \neq \sup_{\nu\in\mathbb{B}_{\text{W}}(\widehat{\nu}^{(2)},\varepsilon)} \nu(x).$$

Figure 4 illustrates the approximations $p(x|\theta)$ offered by the optimal value of the optimistic likelihood problem (3) over these two ambiguity sets. If we choose $\widehat{\nu}^{(2)}$ as the nominal measure, we would expect the true distribution $p(\cdot|\theta)$ to be more spread out than when we choose $\widehat{\nu}^{(1)}$. Nevertheless, this structural information is discarded by the moment ambiguity set, and the optimal value of the optimistic likelihood problem is the same for $\widehat{\nu}^{(1)}$ and $\widehat{\nu}^{(2)}$. In contrast, the Wasserstein ambiguity set produces a fatter tail under the nominal measure $\widehat{\nu}^{(2)}$ than under $\widehat{\nu}^{(1)}$, which better reflects the information contained in the nominal distribution.

Interestingly, if $x = 0$, then we have

$$\sup_{\nu \in \mathbb{B}_{\text{MV}}(\widehat{\nu}^{(1)})} \nu(0) = \sup_{\nu \in \mathbb{B}_{\text{MV}}(\widehat{\nu}^{(2)})} \nu(0) = 1.$$

Indeed, consider the family of discrete measures $\{\nu_k\}_{k \in \mathbb{N}_+}$ defined as

$$\nu_k = \left(1 - \frac{1}{k^2}\right) \delta_0 + \frac{1}{2k^2} \left(\delta_k + \delta_{-k}\right) \qquad \forall k \in \mathbb{N}_+.$$

By construction, ν_k has mean 0 and variance 1, and thus $\{\nu_k\}_{k \in \mathbb{N}_+}$ belong to $\mathbb{B}_{\text{MV}}(\widehat{\nu}^{(1)})$ and attain the optimal value of 1 asymptotically.

B.4 Approximation of the Log-Likelihood for Multiple Observations

In many cases, the update of the posterior is carried out after observing a batch of L i.i.d. samples $x_1^L \triangleq \{x_1, \ldots, x_L\}$. In this case, the log-likelihood of the data x_1^L can be written as

$$\log p(x_1^L|\theta) = \log \prod_{\ell \in [L]} p(x_\ell|\theta) = \sum_{\ell \in [L]} \log p(x_\ell|\theta).$$

When $p(\cdot|\theta)$ is intractable, we propose the optimistic log-likelihood approximation

$$\log p(x_1^L|\theta) \approx \sup_{\nu \in \mathbb{B}_\theta(\widehat{\nu}_\theta)} \sum_{\ell \in [L]} \log \nu(x_\ell) \tag{B.4}$$

for some ambiguity set $\mathbb{B}_\theta(\widehat{\nu}_\theta)$ defined below. Note that the optimistic log-likelihood approximation (B.4) follows the spirit of the optimistic likelihood approximation (3).

Because the log function attains $-\infty$ at 0, we need to restrict ourselves to a subset of $\mathcal{M}(\mathcal{X})$ over which the objective function of (B.4) is well-defined. For any batch data x_1^L, we denote by $\mathcal{M}_{x_1^L}(\mathcal{X})$ the set of measures supported on $\mathcal{X}$ with positive mass at any $x_\ell \in x_1^L$, that is,

$$\mathcal{M}_{x_1^L}(\mathcal{X}) = \{\nu \in \mathcal{M}(\mathcal{X}) : \nu(x_\ell) > 0 \; \forall \ell \in [L]\}.$$

We first establish the upper semicontinuity of the objective function in (B.4).

Lemma B.4 (Upper semicontinuity)**.** For any batch data x_1^L, the functional $G(\nu) = \sum_{\ell \in [L]} \log \nu(x_\ell)$ is upper semicontinuous over $\mathcal{M}_{x_1^L}(\mathcal{X})$.

Proof. Let $\{\nu_k\}_{k \in \mathbb{N}_+}$ be a sequence of probability measures in $\mathcal{M}_{x_1^L}(\mathcal{X})$ converging weakly to $\nu \in \mathcal{M}_{x_1^L}(\mathcal{X})$. We have

$$\limsup_{k \to \infty} G(\nu_k) = \limsup_{k \to \infty} \sum_{\ell \in [L]} \log \nu_k(x_\ell) = \sum_{\ell \in [L]} \log \left(\limsup_{k \to \infty} \nu_k(x_\ell)\right) \leq \sum_{\ell \in [L]} \log \nu(x_\ell) = G(\nu),$$

where the first and last equalities are from the definition of G, the second equality is from the continuity of the log function over $\mathcal{M}_{x_1^L}(\mathcal{X})$, and the inequality is due to the upper semicontinuity of the function $F(\nu) = \nu(x)$ established in Lemma A.1. This completes the proof. □

Given batch data x_1^L, we now consider the Wasserstein ambiguity set centered at the nominal distribution $\widehat{\nu}$,

$$\mathbb{B}_{\text{W}}(\widehat{\nu}, \varepsilon) = \{\nu \in \mathcal{M}_{x_1^L}(\mathcal{X}) : \mathbb{W}(\nu, \widehat{\nu}) \leq \varepsilon\},$$

where the dependence on θ and x_1^L has been made implicit to avoid clutter.

Theorem B.5 (Optimistic log-likelihood; Wasserstein ambiguity)**.** Suppose that Assumption **??** holds. For any batch data x_1^L and radius $\varepsilon > 0$, the optimistic log-likelihood problem (B.4) under the Wasserstein ball $\mathbb{B}_{\text{W}}(\widehat{\nu}, \varepsilon)$ is equivalent to the finite convex program

$$\sup_{\nu \in \mathbb{B}_{\text{W}}(\widehat{\nu}, \varepsilon)} \sum_{\ell \in [L]} \log \nu(x) = \left\{ \begin{array}{ll} \max & \displaystyle\sum_{\ell \in [L]} \log \left(\sum_{j \in [N]} T_{j\ell} \right) \\ \text{s.t.} & T \in \mathbb{R}_+^{N \times L}, \displaystyle\sum_{\substack{j \in [N] \\ \ell \in [L]}} d(\widehat{x}_j, x_\ell) T_{j\ell} \leq \varepsilon \\ & \displaystyle\sum_{\ell \in [L]} T_{j\ell} \leq \widehat{\nu}_j \qquad \forall j \in [N]. \end{array} \right. \tag{B.5}$$

Proof. We first combine the fact that the logarithm is strictly increasing with the proof of Proposition 4.2 to show that there is an optimal measure $\nu^\star_{\mathrm{W}}$ that is supported on $\mathrm{supp}(\nu^\star_{\mathrm{W}}) \subseteq \mathrm{supp}(\widehat{\nu}) \cup x_1^L$, a finite set of cardinality $N+L$. Notice that the existence of this optimal measure is guaranteed by the upper semicontinuity of the objective function established in Lemma B.4 and the weak compactness of $\mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ established in [45, Proposition 3]. The details of this step are omitted for brevity.

Since the optimal measure is supported on $\mathrm{supp}(\widehat{\nu}) \cup x_1^L$, it suffices to consider measures of the form

$$\nu = \sum_{j\in[N]} y_j \delta_{\widehat{x}_j} + \sum_{\ell\in[L]} z_\ell \delta_{x_\ell}$$

for some $y \in \mathbb{R}^N_+$, $z \in \mathbb{R}^L_+$ satisfying $\sum_{j\in[N]} y_j + \sum_{\ell\in[L]} z_\ell = 1$. Using the Definition 4.1 of the type-1 Wasserstein distance, we can rewrite the optimistic log-likelihood problem over the Wasserstein ball $\mathbb{B}_{\mathrm{W}}(\widehat{\nu}, \varepsilon)$ as the convex program

$$\begin{array}{lll}
\sup & \displaystyle\sum_{\ell\in[L]} \log(z_\ell) & \\
\text{s.t.} & y \in \mathbb{R}^N_+,\ z \in \mathbb{R}^L_+,\ \lambda \in \mathbb{R}^{N\times(N+L)}_+ & \\
& \displaystyle\sum_{j\in[N]}\sum_{j'\in[N]} d(\widehat{x}_j, \widehat{x}_{j'})\lambda_{jj'} + \sum_{j\in[N]}\sum_{\ell\in[L]} d(\widehat{x}_j, x_\ell)\lambda_{j(N+\ell)} \le \varepsilon & \\
& \displaystyle\sum_{j'\in[N+L]} \lambda_{jj'} = \widehat{\nu}_j & \forall j \in [N] \\
& \displaystyle\sum_{j\in[N]} \lambda_{jj'} = y_j & \forall j' \in [N] \\
& \displaystyle\sum_{j\in[N]} \lambda_{jj'} = z_{j'-N} & \forall j' \in [N+L]\backslash[N] \\
& \sum_{j\in[N]} y_j + \sum_{\ell\in[L]} z_\ell = 1. &
\end{array}$$

By letting $T_{j\ell} = \lambda_{j(N+\ell)}$ and eliminating the redundant components of λ, we obtain the desired reformulation. This completes the proof. □

References

[1] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

[2] P. Baldi, S. Brunak, and F. Bach. *Bioinformatics: The Machine Learning Approach*. MIT Press, 2001.

[3] M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

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

[5] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

[6] D. Bertsimas and J. Sethuraman. Moment problems and semidefinite optimization. In *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pages 469–509. Springer, 2000.

[7] J. Bi and T. Zhang. Support vector classification with input data uncertainty. In *Advances in Neural Information Processing Systems*, pages 161–168, 2005.

[8] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

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

[10] J. F. Bonnans and A. Shapiro. *Perturbation Analysis of Optimization Problems*. Springer, 2013.

[11] E. Brochu, V. M. Cora, and N. De Freitas. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[12] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

[13] M. B. Cohen, Y. T. Lee, and Z. Song. Solving linear programs in the current matrix multiplication time. *arXiv preprint arXiv:1810.07896*, 2018.

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

[15] M. Cummins and P. Newman. FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665, 2008.

[16] G. B. Dantzig. Discrete-variable extremum problems. *Operations Research*, 5(2):266–277, 1957.

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

[18] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2010.

[19] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[20] S. Gao, G. Ver Steeg, and A. Galstyan. Variational information maximization for feature selection. In *Advances in Neural Information Processing Systems*, pages 487–495, 2016.

[21] N. S. Gorbach, S. Bauer, and J. M. Buhmann. Scalable variational inference for dynamical systems. In *Advances in Neural Information Processing Systems*, pages 4806–4815, 2017.

[22] G. A. Hanasusanto, V. Roitch, D. Kuhn, and W. Wiesemann. Ambiguous joint chance constraints under mean and dispersion information. *Operations Research*, 65(3):751–767, 2017.

[23] M. A. Haynes, H. MacGillivray, and K. Mengersen. Robustness of ranking and selection rules using generalised g-and-k distributions. *Journal of Statistical Planning and Inference*, 65(1):45–66, 1997.

[24] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[25] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

[26] N. Jojic and B. J. Frey. Learning flexible sprites in video layers. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 199–206, 2001.

[27] D. P. Kingma, T. Salimans, and M. Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

[28] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms*. Springer, 2007.

[29] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. *INFORMS TutORials in Operations Research*, 2019.

[30] P. Liang, S. Petrov, M. Jordan, and D. Klein. The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[31] A. C. Likas and N. P. Galatsanos. A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 52(8):2222–2233, 2004.

[32] A. W. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4):1001–1014, 1960.

[33] K. L. Mengersen, P. Pudlo, and C. P. Robert. Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences*, 110(4):1321–1326, 2013.

[34] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

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

[36] S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2125–2133, 2015.

[37] R. Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

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

[39] S. Nakajima, R. Tomioka, M. Sugiyama, and S. D. Babacan. Perfect dimensionality recovery by variational Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 971–979, 2012.

[40] T. Naseem, H. Chen, R. Barzilay, and M. Johnson. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, 2010.

[41] V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. Distributionally robust inverse covariance estimation: The Wasserstein shrinkage estimator. *arXiv preprint arXiv:1805.07194*, 2018.

[42] V. A. Nguyen, S. Shafieezadeh-Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. Calculating optimistic likelihoods using (geodesically) convex optimization. In *Advances in Neural Information Processing Systems*, 2019.

[43] M. Park, W. Jitkrittum, and D. Sejdinovic. K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 398–407, 2016.

[44] G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[45] A. Pichler and H. Xu. Quantitative stability analysis for minimax distributionally robust risk optimization. *To appear in Mathematical Programming*, 2019.

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

[47] A. Raj, M. Stephens, and J. K. Pritchard. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589, 2014.

[48] G. Sanguinetti, N. D. Lawrence, and M. Rattray. Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781, 2006.

[49] M. J. Schervish. *Theory of Statistics*. Springer, 1995.

[50] S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

[51] A. Sinha, H. Namkoong, and J. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proceedings of International Conference on Learning Representations*, 2018.

[52] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of International Conference on Machine Learning*, pages 1015–1022, 2010.

[53] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244, 2001.

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

[55] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

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

[57] M. W. Woolrich, T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith. Multilevel linear modelling for FMRI group analysis using Bayesian inference. *Neuroimage*, 21(4):1732–1747, 2004.