Adaptive Nonparametric Perturbations of Parametric Models with Generalized Bayes

Bohan Wu* BW2766@COLUMBIA.EDU
Department of Statistics
Columbia University
New York, NY, USA

Eli N. Weinstein* ENAWE@DTU.DK
Department of Chemistry
Technical University of Denmark
Kgs. Lyngby, Denmark

Sohrab Salehi SOHRAB.SALEHI@MSKCC.ORG
Computational Oncology, Department of Epidemiology and Biostatistics
Memorial Sloan Kettering Cancer Center
New York, NY, USA

Yixin Wang YIXINW@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor, MI, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Department of Computer Science and Department of Statistics
Columbia University
New York, NY, USA

Editor:

Abstract

Parametric Bayesian modeling offers a powerful and flexible toolbox for machine learning. Yet the model, however detailed, may still be wrong, and this can make inferences untrustworthy. In this paper we introduce a new class of semiparametric corrections for parametric Bayesian models, when the target of inference is a functional of the true data distribution. Our starting point is a fully Bayesian modeling approach, which explicitly accounts for the possibility that the parametric model is wrong. Asymptotic analysis shows that this approach is both robust to model misspecification and data efficient, achieving fast convergence when the parametric model is close to true. However, the fully Bayesian approach is limited in its practical usefulness by the challenges of conducting inference and computing a Bayes factor for a nonparametric model. We therefore propose a novel model correction based on generalized Bayes, which entirely avoids the need to compute a nonparametric Bayes factor, but preserves the robustness and efficiency of the fully Bayesian approach. We demonstrate our method by estimating causal effects of gene expression from single cell RNA sequencing data. Overall, we offer a new efficient approach to robust Bayesian inference with parametric models.

∗. Equal contribution

Keywords: Robust Bayesian learning, generalized Bayes, causality, functional inference.

1 Introduction

Parametric Bayesian modeling offers a flexible toolbox for incorporating scientific knowledge and hypotheses into probabilistic machine learning. Doing so has many benefits, such as increased data efficiency and interpretability. Yet it can also present risks, as the model remains only an approximation of reality. How can we ensure inferences are trustworthy even when the model might be wrong?

In this paper, we study this problem in the context of functional inference. While standard posterior inference focuses on the distribution of model parameters, functional inference concerns the estimation and uncertainty quantification of functionals of the population distribution. It can leverage flexible machine learning models, which learn complex population distributions. Developing accurate, calibrated and robust inference of functional estimands is fundamental to research areas such as causal inference (van der Laan and Rose, 2011; Hill, 2011; Castillo and Rousseau, 2015). However, developing Bayesian methods for functional inference remains an active challenge (Lyddon et al., 2018; Ray and van der Vaart, 2020).

In this article, we study the problem of reliable and efficient functional inference based on a parametric Bayesian model, even when the parametric model might be wrong. We are specifically concerned with the possibility that the model likelihood is misspecified. Concretely, assume the data is i.i.d. from some true underlying distribution, $x_{1:n} \overset{iid}{\sim} \mathrm{p}_0$. A parametric Bayesian model of the data is specified by a prior over an unknown parameter $\pi(\theta)$, together with a likelihood p_θ,

$$\begin{aligned} \theta &\sim \pi(\theta), \\ x_{1:n} &\overset{iid}{\sim} \mathrm{p}_\theta. \end{aligned} \tag{1}$$

The problem is that the model may not be able to describe p_0. The induced prior over the distribution of the data, $\pi(\mathrm{p}_\theta)$, only has support on a subset of all possible distributions, $\mathcal{M}_{\mathrm{pm}} = \{\mathrm{p}_\theta : \theta \in \Theta\} \subset \mathcal{P}$. If the likelihood is misspecified, such that $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, inferences drawn from the posterior may be unreliable, as no amount of data can overwhelm such an overconfident prior.

A fully Bayesian solution is to adjust the model to account for this possibility. From the perspective of de Finetti's theorem, Bayesian inference can be viewed as inference of the data distribution through a prior on probability measures. Following this perspective, rather than insist the true data distribution matches the likelihood distribution exactly, we allow for distributional perturbations, so the data is generated from an alternative distribution, $x_{1:n} \overset{iid}{\sim} \tilde{\mathrm{p}}$, instead of $x_{1:n} \overset{iid}{\sim} \mathrm{p}_\theta$. We place a prior on $\tilde{\mathrm{p}}$ that is centered at p_θ but has sufficient support that $\tilde{\mathrm{p}}$ can be any distribution in $\mathcal{P}$. The resulting *nonparametrically perturbed parametric* (NPP) Bayesian model encodes the belief that the likelihood is "probably roughly true."

This basic idea of combining a parametric model and a nonparametric component has a long history in Bayesian statistics. It can, however, be challenging to implement in large-scale machine learning settings, as it requires posterior inference over a nonparametric model. In this paper, we aim to develop a practical alternative.

First, we elucidate the key theoretical advantages of NPP models. We show that, with an appropriate choice of prior, NPP models can achieve robustness without sacrificing asymptotic efficiency. We then develop a generalized Bayes approach that imitates the NPP posterior, retaining its nice statistical properties but with more tractable computation. This *generalized NPP* (gNPP) posterior does not require any change to how the parametric model's posterior is computed, instead only needing access to samples from the predictive distribution. This efficiency is supported by asymptotic results. We demonstrate the method in simulations and with an application to gene expression in cancer.

1.1 Related Work

Robustness and efficiency are core questions in functional estimation. Perhaps the most well-established route to achieving these properties is to employ semiparametric corrections, such as influence functions, doubly robust estimators, and Neyman orthogonality (Kennedy, 2024). We show how to achieve similar properties via a very different methodological strategy, which extends beyond the context of causal inference. Specifically, consider a doubly robust estimator that uses a parametric outcome model and a nonparametric propensity model. When the outcome model is well-specified, the estimator is efficient: it converges to the estimand, the treatment effect, at a parametric rate. When the outcome model is misspecified, the estimator is robust: it still converges to the true effect, albeit at a slower rate (Bang and Robins, 2005; Antonelli et al., 2020). We show that NPP models offer the same guarantees, but are more general: they are not specific to causal functionals, and they do not require a propensity model or another influence function-based correction. We engineer gNPP posteriors to inherit these properties at reduced computational cost. Note, though, that NPP and gNPP modeling is not in conflict with influence function-based methods, and in Section 4 we explain how the two methods can be combined.

Our work sits in the broader context of robust Bayesian learning. We estimate functionals of the true data distribution p_0, following previous work on robust Bayesian modeling such as Lyddon et al. (2019); Pompe (2021); Dellaporta et al. (2022). Recent work has focused on achieving robust and efficient functional estimation by adapting semiparametric corrections to the Bayesian context (Antonelli et al., 2020; Walker, 2024; Yiu et al., 2025). Inspired by this progress, we aim for similar guarantees via alternative and more general methods.

Other robust Bayes approaches consider a different problem, in which the target of inference may not be identified as a functional of p_0 (de Finetti, 1961; Wang et al., 2017; Wang and Blei, 2018; Miller and Dunson, 2018; Bhatia et al., 2024). For example, p_0 may be corrupted by a population of outliers. The classical ϵ-contamination model posits p_0 is distorted to $(1-\epsilon)\mathrm{p}_0 + \epsilon q$ (Huber and Ronchetti, 2009). Our use of *robust* concerns reliable inference in the face of possible model misspecification, rather than data contamination.

Our work builds on the long literature on nonparametric perturbations of parametric Bayesian models. In this paper, we are interested in perturbation techniques for generic parametric models p_θ, so we focus on reviewing these methods.

Antoniak (1974) consider nonparametric perturbations with a Dirichlet process, in the form of mixture of Dirichlet processes. Berger and Guglielmi (2001) consider nonparametric perturbations with a Polya tree, but in practice their method is limited to low-dimensional Euclidean data. Lyddon et al. (2018) consider perturbations using a Dirichlet process. We

prove their approach sacrifices data efficiency, and propose an alternative that does not. Miller (2019) consider perturbations based on Dirichlet process mixture models, but they do not provide theoretical analysis. Another prominent class of examples of combining a parametric component with a nonparametric model is semiparametric Bayesian regression that involves a finite-dimensional regression parameter alongside a nonparametric function. The area has a large body of work, see e.g. Blight and Ott (1975); O'Hagan (1978); Kottas and Gelfand (2001); Berry et al. (2002); Kowal and Wu (2024) and the references therein.

Another approach that is used to perturb a parametric model is to create a mixture of different copies of the model, e.g. replacing a Gaussian with a Dirichlet process mixture of Gaussians (Lo, 1984; Escobar and West, 1995). One could consider similarly replacing a parametric model with a Dirichlet process mixture of that model. But this approach grows rapidly in theoretical and computational difficulty as the complexity of the parametric model grows.

Our posterior approximation builds on generalized Bayesian posteriors, which replace the model likelihood with an alternative loss (Jiang and Tanner, 2008; Zhang, 2006; Bissiri et al., 2016; Jewson et al., 2018; Shao et al., 2018; Miller, 2021; Knoblauch et al., 2022). Following Weinstein and Miller (2023), we construct a generalized Bayes factor between a parametric and nonparametric model by using a statistical divergence. Unlike in Weinstein and Miller (2023), we only need samples from the parametric model's posterior to estimate this generalized Bayes factor.

2 Nonparametric Perturbations of Bayesian Models

2.1 Target of inference

Suppose we observe i.i.d. data $x_{1:n} \overset{iid}{\sim} \mathrm{p}_0$. Our target estimand is the output of a functional applied to the data generating distribution, $\psi(\mathrm{p}_0)$, where $\psi : \mathcal{P} \to \mathbb{R}^s$. Here are some examples:

- **Summary statistics.** The target of inference could be the mean, in which case the functional is $\psi(\mathrm{p}_0) = \int x \, \mathrm{p}_0(x) \mathrm{d}x$. Likewise we could infer the variance, etc.

- **Loss minimizers.** The target of inference could be the minimizer of a loss or utility, $\psi(\mathrm{p}_0) = \mathrm{argmin}_\alpha \int \ell(x\,;\alpha)\mathrm{p}_0(x)\mathrm{d}x$. The loss can even be specified in terms of the parametric model p_θ, e.g. if $\ell(x\,;\alpha) = -\log \mathrm{p}_\alpha(x)$ then the target of inference corresponds to the maximum likelihood estimate (MLE) from infinite data. This is equivalent to the KL minimizer, $\psi(\mathrm{p}_0) = \mathrm{argmin}_\theta \, \mathrm{KL}\,(\mathrm{p}_0 \| \mathrm{p}_\theta)$.

- **Causal effects.** The target of inference could be a functional of p_0 describing the effect of an intervention (Kennedy, 2024). For example, consider data $x = (a, y, w)$ consisting of a treatment $a \in \{0, 1\}$, outcome y, and confounder w. The average treatment effect is,

$$\begin{aligned}\psi(\mathrm{p}_0) &= \mathbb{E}_{\mathrm{p}_0}[Y\,;\,\mathrm{do}(a=1)] - \mathbb{E}_{\mathrm{p}_0}[Y\,;\,\mathrm{do}(a=0)] \\ &= \int\int y\,[\mathrm{p}_0(y \mid a=1, w) - \mathrm{p}_0(y \mid a=0, w)]\mathrm{d}y\,\mathrm{p}_0(w)\mathrm{d}w.\end{aligned}$$

The inference challenge is that if the parametric model is misspecified, the posterior will always put zero mass on the true data distribution, and so we cannot learn the true functional, $\psi(\mathrm{p}_0)$.

2.2 Nonparametric Perturbations of Parametric Models

We will study nonparametric perturbations of parametric Bayesian models that use a specific type of prior. In particular, we consider NPP models that mix the parametric Bayesian model with a perturbed distribution,

$$\begin{aligned} \theta \sim \pi(\theta), \quad b \sim \text{Bernoulli}(\eta), \\ \begin{cases} \mathrm{p} = \mathrm{p}_\theta & \text{if } b = 1 \\ \mathrm{p} \sim \pi_{\text{pert}}(\mathrm{p} \mid \mathrm{p}_\theta) & \text{if } b = 0 \end{cases}, \\ x_{1:n} \overset{iid}{\sim} \mathrm{p}. \end{aligned} \tag{2}$$

In this model, the entire dataset $x_{1:n}$ is either drawn from the parametric model ($b = 0$) or from an alternative ($b = 1$). There are many ways to perturb. For example, $\pi_{\text{pert}}(\mathrm{p} \mid \mathrm{p}_\theta)$ could be a Dirichlet process mixture model (Miller, 2019),

$$\Sigma \sim g(\Sigma), \quad \sum_{k=1}^{\infty} w_k \delta_{\mu_k} \sim \text{DP}(\mathrm{p}_\theta, \alpha), \quad \mathrm{p} = \sum_{k=1}^{\infty} w_k \text{Normal}(\mu_k, \Sigma). \tag{3}$$

Other possibilities are Polya trees (Berger and Guglielmi, 2001) and Gaussian process density models (Adams et al., 2009). We assume the perturbation model is nonparametric, placing its support on any distribution over x, including p_θ. We do not assume that the perturbation depends on p_θ explicitly, i.e. $\pi_{\text{pert}}(\mathrm{p} \mid \mathrm{p}_\theta)$ can be replaced by a generic nonparametric model $\pi_{\text{pert}}(\mathrm{p})$.

We want to use an NPP model instead of a conventional nonparametric model because parametric models are efficient for small data. We can think of an NPP model as a nonparametric model with a special spike-and-slab prior on the data distribution, that is centered at the parametric model. The prior has a "spike" component that puts its mass just on the parametric model ($b = 1$), and a "slab" component extending over all distributions ($b = 0$). This prior is nonparametric but regularizes inferences towards the parametric model. This choice of an adaptive prior enables data-efficient learning, as we describe next.

2.3 NPP Bayesian Posterior

We now examine some of the advantages of NPP models for functional inference. We can perform functional inference with an NPP model by computing the posterior over the target functional, $\Pi(\psi(\mathrm{p}) \mid x_{1:n})$. The NPP posterior is a mixture between a parametric and nonparametric posterior,

$$\Pi\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right) = \eta_n \Pi_{\text{pm}}\left(\psi\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right) + (1 - \eta_n) \Pi_{\text{pert}}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right), \tag{4}$$

where $\Pi_{\text{pm}}\left(\psi\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right)$ is the posterior from the parametric model, $\Pi_{\text{pert}}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right)$ is the posterior from the nonparametric perturbation model, and $\eta_n := \Pi\left(b = 1 \mid x_{1:n}\right)$. This

decomposition fully separates the parametric and nonparametric components. The weight η_n adaptively trades off between the parametric and nonparametric posterior, based on the data.

NPP models provide a *robust* and *efficient* way of learning about $\psi(\mathrm{p}_0)$. To make these notions precise, we compare NPP models to two alternatives: (a) the original parametric model, given by $\theta \sim \pi(\theta), x_{1:n} \overset{iid}{\sim} \mathrm{p}_\theta$, and (b) a nonparametric model with a generic prior, $\mathrm{p} \sim \pi_{\mathrm{np}}(\mathrm{p}), x_{1:n} \overset{iid}{\sim} \mathrm{p}$, where $\pi_{\mathrm{np}}(\mathrm{p})$ has support over all distributions on $\mathcal{X}$. We use $\Pi_{\mathrm{pm}}(\cdot \mid x_{1:n})$ to denote the parametric model posterior, $\Pi_{\mathrm{np}}(\cdot \mid x_{1:n})$ to denote the nonparametric model posterior, and $\Pi(\cdot \mid x_{1:n})$ to denote the NPP model posterior. We will show that the parametric model is efficient but not robust, while the nonparametric model is robust but not efficient. Formal statements and proofs of the following results are in Section 4.

Misspecified case. We want inferences about $\psi(\mathrm{p}_0)$ to be *robust* to model misspecification, such that they always converge to the truth with enough data. The NPP model is robust: Theorem 3 shows that $\Pi\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right) \to \delta_{\psi(\mathrm{p}_0)}$ even when $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$. To see why, recall from Eq. (4) that the NPP posterior is a mixture of a parametric and nonparametric posterior. The parametric posterior will not converge to the truth, $\Pi_{\mathrm{pm}}(\psi(\mathrm{p}_\theta) \mid x_{1:n}) \not\to \delta_{\psi(\mathrm{p}_0)}$, but the nonparametric component will, $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n}) \to \delta_{\psi(\mathrm{p}_0)}$. The overall NPP posterior converges to the truth because the mixing weight η_n asymptotically places all weight on the nonparametric component: if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$ then $\eta_n \to 0$ a.s. as $n \to \infty$ (Proposition 2).

To understand the behavior of the mixing weight, we can write $\eta_n/(1-\eta_n)$ in terms of a Bayes factor that compares the marginal likelihood of the data under the parametric and nonparametric models, weighted by the prior odds $\eta/(1-\eta)$,

$$\frac{\eta_n}{1-\eta_n} = \frac{\Pi_{\mathrm{pm}}(x_{1:n})\eta}{\Pi_{\mathrm{pert}}(x_{1:n})(1-\eta)} =: \mathrm{BF}_n, \tag{5}$$

where

$$\Pi_{\mathrm{pm}}(x_{1:n}) = \int \prod_{i=1}^{n} \mathrm{p}_\theta(x_i)\pi(\theta)\mathrm{d}\theta,$$

and

$$\Pi_{\mathrm{pert}}(x_{1:n}) = \int\int \prod_{i=1}^{n} \mathrm{p}(x_i)\pi_{\mathrm{pert}}(\mathrm{p}|\mathrm{p}_\theta)\pi(\theta)\mathrm{dp}\mathrm{d}\theta.$$

When the parametric model is misspecified, the Bayes factor prefers the more expressive nonparametric model, since it can match the data distribution, while the parametric model cannot. Hence, $\mathrm{BF}_n \to 0$, so the mixing weight $\eta_n \to 0$ a.s. as $n \to \infty$ (Proposition 2).

In summary, the NPP model is robust: $\Pi(\psi(\mathrm{p}) \mid x_{1:n}) \to \delta_{\psi(\mathrm{p}_0)}$ even when when $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$. In this way, the NPP model behaves like a generic nonparametric model, for which we also expect $\Pi_{\mathrm{np}}(\psi(\mathrm{p}) \mid x_{1:n}) \to \delta_{\psi(\mathrm{p}_0)}$. A parametric model, by contrast, is not robust: in general, $\Pi_{\mathrm{pm}}(\psi(\mathrm{p}_\theta) \mid x_{1:n}) \not\to \delta_{\psi(\mathrm{p}_0)}$ when $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$.

Well-specified case. We next consider the case where the parametric model is well-specified, $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$. In this case, the NPP model converges to the truth at an efficient rate: the posterior $\Pi(\psi(\mathrm{p}) \mid x_{1:n})$ contracts at a rate $1/\sqrt{n}$ around $\psi(\mathrm{p}_0)$ when $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$. To

see this, return to the decomposition in Eq. (4). Both the parametric posterior $\Pi_{\text{pm}}(\psi(\text{p}_\theta) \mid x_{1:n})$ and the nonparametric posterior $\Pi_{\text{pert}}(\psi(\text{p}) \mid x_{1:n})$ will converge to the truth, but the parametric posterior will converge faster, at a rate $1/\sqrt{n}$, while the nonparametric component will in general converge more slowly. The NPP posterior inherits the rate of the parametric posterior because the mixing weight asymptotically places all weight on the parametric component: if $\text{p}_0 \in \mathcal{M}_{\text{pm}}$ then $\eta_n \to 1$ a.s. as $n \to \infty$ (Proposition 2).

To see why $\eta_n \to 1$ intuitively, consider again the Bayes factor, Eq. (5). Both the parametric and the nonparametric model describe the data distribution, but the parametric model uses fewer parameters than the nonparametric model: a finite number, rather than infinite. Moreover, the parametric model is nested inside the nonparametric model, since it corresponds to sampling $\text{p} = \text{p}_\theta$ from $\pi_{\text{pert}}(\text{p} \mid \text{p}_\theta)$. By Bayesian Occam's razor, the Bayes factor prefers the simpler parametric model, i.e. $\text{BF}_n \to \infty$ (Dawid, 2011; Hong and Preston, 2012; MacKay, 2003, Chap. 28). Hence $\eta_n \to 1$.

In summary, when the parametric model is correctly specified, $\text{p}_0 \in \mathcal{M}_{\text{pm}}$, the NPP posterior $\Pi(\psi(\text{p}) \mid x_{1:n})$ converges to the truth $\psi(\text{p}_0)$ at an efficient parametric rate of $1/\sqrt{n}$. So here, the NPP posterior behaves like a parametric Bayes posterior, for which we also expect a convergence rate of $1/\sqrt{n}$ by the Bernstein-von Mises theorem. In a nonparametric model, however, the convergence rate is generally slower, e.g. $\Pi_{\text{np}}(\psi(\text{p}) \mid x_{1:n})$ may converge at a rate of $1/\sqrt[3]{n}$ or worse (Ghosal et al., 2000).

Finite samples. A possible concern is that in practice there is always some amount of misspecification, and so we cannot ever expect a parametric rate. However, the Bartlett-Lindley effect says the Bayes factor will place more weight on the simpler model up until there is enough data to determine that model is wrong (Miller and Dunson, 2018). So, heuristically, we expect the NPP posterior to converge towards the truth as quickly as the parametric posterior, up until the parametric model stops providing a good description of the data.

2.4 NPP Generalized Bayesian Posterior

We have seen that NPP models offer both robustness and efficiency, but this comes at a computational cost. Computing the NPP posterior (Eq. (4)) requires computing not only the parametric model posterior, but also (a) the mixing weight η_n, and (b) the nonparametric posterior, $\Pi_{\text{pert}}(\psi(\text{p}) \mid x_{1:n})$. In this section we propose a new inference approach that replaces each term with alternatives that are efficient to compute and practical to implement, but also preserve robustness and efficiency. The aim of this *generalized NPP* (gNPP) approach is not to directly approximate the NPP, but rather to mimic its statistical behavior while easing computation.

Mixing weight. The mixing weight η_n depends on the Bayes factor BF_n (Eq. (5)), which compares the marginal likelihood of the parametric model to the marginal likelihood of a nonparametric alternative. Computing marginal likelihoods is often challenging, especially for nonparametric models. We propose to replace the Bayes factor with a generalized Bayes factor, which uses divergences instead of marginal likelihoods to evaluate the parametric model (Shao et al., 2018; Weinstein and Miller, 2023). Our approach is motivated by the observation that, asymptotically, the standard Bayes factor will converge to the posterior

expected KL divergence, $\frac{1}{n}\log \mathrm{BF}_n \to -\mathbb{E}[\mathrm{KL}\,(\mathrm{p}_0 \| \mathrm{p}_\theta) \mid x_{1:n}]$ a.s., under regularity conditions (Dawid, 2011; Shao et al., 2018; Miller, 2021; Weinstein and Miller, 2023). We consider alternative divergences, easing computation while matching the behavior of the Bayes factor.

Let $\mathrm{D}(\mathrm{p}, \mathrm{p}_0)$ denote a divergence between probability distributions, which satisfies $\mathrm{D}(\mathrm{p}, \mathrm{p}_0) = 0$ when $\mathrm{p} = \mathrm{p}_0$ and $\mathrm{D}(\mathrm{p}, \mathrm{p}_0) > 0$ when $\mathrm{p} \neq \mathrm{p}_0$. Let $\mathrm{D}_n(\mathrm{p}, \mathrm{p}_0)$ denote an estimate of $\mathrm{D}(\mathrm{p}, \mathrm{p}_0)$ based on data $x_{1:n}$ i.i.d. from p_0. Concretely, the divergences we consider are the Wasserstein distance, the maximum mean discrepancy and the kernelized Stein discrepancy, though other choices are also possible (Gretton et al., 2012; Liu et al., 2016). We use the divergence to construct a generalized Bayes factor (gBF),

$$\mathrm{gBF}_n := \Xi\left(\frac{\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)]}{\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]}(n+1)^{-r}\right)\frac{\eta}{1-\eta}. \tag{6}$$

Here, $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]$ is the expected value of the divergence under the posterior $\Pi_{\mathrm{pm}}(\theta \mid x_{1:n})$, as appears in asymptotic Bayes factor. $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)]$ is the expected divergence under the prior $\pi(\theta)$, which ensures the gBF reverts to the prior when there is no data. The rate $r > 0$ is a hyperparameter, and the function $\Xi(x) = \exp(1 - 1/x)x$ is monotonic and satisfies $\Xi(x) \to 0$ as $x \to 0$ and $\Xi(x) \to \infty$ as $x \to \infty$. Although Eq. (6) does not contain the standard Bayes factor as a special case, it is related: like the standard marginal likelihood, it is a posterior average over an estimate of the model-data mismatch. Now, as an alternative to the NPP posterior, we propose to consider a generalized posterior that uses $\hat{\eta}_n = 1/(1 + \mathrm{gBF}_n^{-1})$ in place of η_n.

The aim of this generalized Bayes method is not to directly estimate the mixing weight η_n, but rather to match its statistical behavior. When the parametric model is misspecified, $\hat{\eta}_n \to 0$, just like η_n. The reason is that $\mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0) > 0$ for all θ when $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, so $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]$ will converge to a positive value. On the other hand, when the parametric model is well-specified, $\hat{\eta}_n \to 1$, just like η_n. The reason is that the posterior will concentrate at p_0, so $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}] \to 0$. So long as the rate hyperparameter r is chosen small enough that $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}](n+1)^r \to 0$ also, we will have $\hat{\eta}_n \to 1$. So, the generalized Bayes factor is asymptotically *consistent* in detecting misspecification.

In addition to achieving the same asymptotic limits as the original Bayes factor, the generalized Bayes factor also approaches those limits at a similar rate, thanks to the choice of function $\Xi(x)$ (Section A). Before we observe data (i.e. for $n = 0$), both the Bayes factor and the generalized Bayes factor coincide with the prior odds ratio, $\mathrm{gBF}_0 = \mathrm{BF}_0 = \eta/(1-\eta)$. In the generalized Bayes factor, this is thanks to the inclusion of the prior divergence, $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)]$.

However, the key advantage of the generalized Bayes factor is computation: it requires only the ability to draw samples from the prior and posterior of the parametric model, to approximate $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)]$ and $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]$. The marginal likelihood of the parametric model does not appear. The nonparametric model does not appear at all.

As a result, an important difference with the standard Bayes factor is that the standard approach compares the parametric model's fit to the nonparametric model's fit. The generalized Bayes factor instead focuses on just checking the parametric model's fit, and does not directly trade off against an alternative. In other words, the gNPP treats the alternative as a backup in case the parametric model goes wrong.

Indeed, the generalized Bayes factor can also be understood as a form of predictive check on the parametric model (Gelman et al., 1996; Moran et al., 2024; Li and Huggins, 2024). Specifically, the term $\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]$ compares the posterior over the parametric model's predictive distribution, $\Pi_{\mathrm{pm}}(\mathrm{p}_\theta \mid x_{1:n})$, to data, using D_n as measure of model-data discrepancy.

Choice of divergence and rate. We consider several types of divergence, which prioritize different kinds of mismatch between the model and data. All of the following divergences can be approximated by Monte Carlo methods.

- **Wasserstein distance.** The Wasserstein measures the difference between two distributions as the cost of transforming one into the other by transporting probability mass. The p-Wasserstein distance is typically defined as

$$\mathrm{W}_p^p(\mathrm{p}, \mathrm{q}) := \inf_{\pi \in \Pi(\mathrm{p},\mathrm{q})} \mathbb{E}_\pi[\|X - Y\|_2^p], \tag{7}$$

  where $\Pi(\mathrm{p}, \mathrm{q})$ represents the set of all couplings between p and q, i.e. the set of all joint distributions with p and q as marginals, and $\|\cdot\|_2$ denotes the Euclidean metric on $\mathbb{R}^\kappa$. We can estimate the distance between the model and the data using m samples drawn from the model. We draw the same number of samples as data points, $m = n$, by default. Computing the Wasserstein distance between two empirical distributions reduces to solving a linear program, which can be done in $O(n^3)$ time using interior-point methods (Pele and Werman, 2009). In practice, the Euclidean cost with $p \in \{1, 2\}$ is most commonly used. The case W_1 admits the convenient Kantorovich–Rubinstein dual formulation in terms of 1-Lipschitz test functions, while W_2 enjoys favorable geometric properties (Villani, 2009). In high dimensions, however, the empirical p-Wasserstein distances suffer from the curse of dimensionality, with a slow convergence rate of $O(n^{-1/\kappa})$ even for compactly supported measures. In contrast, alternatives such as the sliced Wasserstein distance and smoothed Wasserstein distance can achieve dimension-free $O(n^{-1/2})$ rates under the same conditions (Chewi et al., 2025). For large-scale problems, entropically-regularized optimal transport (Cuturi, 2013; Peyré and Cuturi, 2019) offers scalable approximations that can be computed with iteration complexity that does not depend on κ (Carlier, 2022), and the associated Sinkhorn divergence (Genevay et al., 2018, 2019) comes with improved statistical convergence rates, $o(n^{-1/2})$ for sub-Gaussian measures and $O(n^{-1})$ for compactly supported measures (del Barrio et al., 2023).

- **Maximum mean discrepancy (MMD).** The maximum mean discrepancy focuses on the worst case difference in expected value that two distributions assign to the same function in a reproducing kernel Hilbert space (RKHS). It is given by

$$\mathrm{MMD}^2(\mathrm{p}, \mathrm{q}) := \sup_{\|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}_\mathrm{p}[f(X)] - \mathbb{E}_\mathrm{q}[f(X)]|^2 \tag{8}$$

$$= \mathbb{E}_{X,X' \sim \mathrm{p}}\left[k(X, X')\right] - 2\mathbb{E}_{X\sim\mathrm{p}, Y\sim\mathrm{q}}\left[k(X, Y)\right] + \mathbb{E}_{Y,Y'\sim\mathrm{q}}\left[k(Y, Y')\right]. \tag{9}$$

  where $\mathcal{H}_k$ is an RKHS and $\|f\|_{\mathcal{H}_k}$ is the norm of a function in that RKHS. To estimate the MMD between the model and the data, we draw samples from the model. We

again default to the same number of samples as data points, $m = n$. The MMD requires $O(n^2)$ time to compute; a linear-time approximation also exists (Gretton et al., 2012).

- **Kernelized Stein discrepancy (KSD).** The kernelized Stein discrepancy focuses on differences in the (Stein) score function between two distributions, i.e. differences in the gradient of their log densities. It is given by

$$\mathrm{KSD}^2(\mathrm{p}, \mathrm{q}) = \mathbb{E}_{X,X' \sim \mathrm{p}} \left[\Delta_{\mathrm{q,p}}(X)^T k(X, X') \Delta_{\mathrm{q,p}}(X') \right], \tag{10}$$

  where $\Delta_{\mathrm{q,p}}(x) := \nabla_x \log \mathrm{p}(x) - \nabla_x \log \mathrm{q}(x)$. To estimate the KSD between the data and the model, we evaluate the empirical average of $\Delta_{\mathrm{q,p}}(x)^T k(x, x') \Delta_{\mathrm{q,p}}(x')$ over every pair of data points (x, x'). For n data points, this requires $O(n^2)$ time to compute (Liu et al., 2016), and there also exists a near-linear time approximation (Huggins and Mackey, 2018).

Both MMD and KSD rely on a kernel, which allows the user to control which aspects of a distribution to prioritize, but it also introduces the need for hyperparameter selection. Kernel hyperparameters are absent for the Wasserstein distance; however, one must instead choose the transportation cost.

Although the MMD and the KSD are sensitive to a rescaling of the kernel, the generalized Bayes factor is scale-invariant, since in Eq. (6) we normalize the posterior expected discrepancy by the prior expected discrepancy.

Remark 1 (Choice of rate)**.** *For the MMD and KSD we have a consistent generalized Bayes factor so long as we choose a rate* $0 < r < 1/2$*, i.e. slower than the parametric rate (Theorem 5). Choosing a rate just below* $1/2$ *enables rapid convergence in the misspecified case, while still preserving consistency in the well-specified case. For Wasserstein, we need to choose a slower rate when the dimension* κ *of the data space is greater than four, namely* $r \in (0, 2/(\kappa \vee 4))$ *for* $\mathcal{X} \subseteq \mathbb{R}^\kappa$ *(Theorem 16 in Section D.3.1).*

Nonparametric posterior. Next we consider replacing the nonparametric term $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ in the NPP posterior. This term can be challenging to compute. The robustness and efficiency of the NPP model rely on the nonparametric perturbation $\pi_{\mathrm{pert}}(\mathrm{p} \mid \mathrm{p}_\theta)$ having support over all distributions on $\mathcal{X}$, so posterior computation requires integration over a very large space. On the other hand, since the mixing weight $\hat{\eta}_n$ does not depend on the nonparametric perturbation, the nonparametric component no longer needs to serve as a fully flexible density estimator; rather, it only needs to estimate the target consistently.

We propose to replace the nonparametric perturbation with a semiparametric model that can consistently estimate $\psi(\mathrm{p}_0)$. That is, we consider a model $\mathrm{p} \sim \hat{\pi}_{\mathrm{pert}}(\mathrm{p}), x_{1:n} \sim \mathrm{p}$ such that $\hat{\Pi}_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n}) \to \delta_{\psi(\mathrm{p}_0)}$. We replace the term $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ in the NPP posterior with $\hat{\Pi}_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$.

To see how this approach can simplify computation, consider data consisting of a treatment and a response, $x_i = (y_i, a_i)$, and a target functional that is the expected response to a given treatment, $\psi(\mathrm{p}) = \mathbb{E}[Y \mid a_\star] = \int y \, \mathrm{p}(y \mid a_\star) \mathrm{d}y$. An NPP model would require us to place a prior over all distributions on $\mathcal{A} \times \mathcal{Y}$. However, to estimate the target functional consistently, it is enough to use a semiparametric model $y_i \sim \mathrm{Normal}(f(a_i), 1), f \sim \pi_f(f)$,

and place a nonparametric prior on functions $f : \mathcal{A} \to \mathcal{Y}$. For example, π_f could be a Gaussian process prior or a Bayesian additive regression tree (Rasmussen and Williams, 2006; Chipman et al., 2010). These semiparametric models cannot describe all distributions over $\mathcal{A} \times \mathcal{Y}$, e.g. they cannot describe a distribution where the conditional $\mathrm{p}(y \mid a_\star)$ is not Gaussian, but they can still estimate any conditional expectation $\mathbb{E}[Y \mid a_\star]$. Moreover, the posterior is convenient to calculate, as we can rely on existing software for Gaussian processes and Bayesian additive regression trees. Finally, compared to a fully nonparametric model, semiparametric models often require less data to produce accurate inference of the target estimand.

Summary. In summary, our generalized Bayes NPP posterior is

$$\hat{\Pi}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right) := \hat{\eta}_n \Pi_{\mathrm{pm}}\left(\psi\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right) + (1 - \hat{\eta}_n)\hat{\Pi}_{\mathrm{pert}}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right), \tag{11}$$

where $\hat{\eta}_n$ comes from the generalized Bayes factor and $\hat{\Pi}_{\mathrm{pert}}$ comes from a semiparametric model. We refer to this generalized posteriors as the *generalized NPP (gNPP)* posterior. In Section 4 we prove formally that the gNPP posterior preserves the robustness and efficiency of the NPP posterior: for any p_0, $\hat{\Pi}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right) \to \delta_{\psi(\mathrm{p}_0)}$, and for $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, $\hat{\Pi}\left(\psi\left(\mathrm{p}\right) \mid x_{1:n}\right)$ converges at a rate $1/\sqrt{n}$.

The key to the gNPP is replacing the Bayes factor between parametric and nonparametric models with a generalized Bayes factor based on a divergence. In doing so, we avoid the need to compute the marginal likelihood of the parametric model or of the nonparametric perturbation. We also reduce the requirements and constraints on the perturbation model we can use: it no longer needs to be fully nonparametric, and the parametric model does not need to be nested inside it. Instead, we can use a more computationally tractable semiparametric model, as it consistently estimates the target.

3 Synthetic Data Illustration

We illustrate the behavior of NPP models and the gNPP approximation in a synthetic data setting. We follow the basic setup of Lyddon et al. (2018). Consider a Gaussian parametric model,

$$\theta \sim \mathcal{N}(0, 1), \quad x_{1:n} \overset{iid}{\sim} \mathcal{N}(\theta, 1). \tag{12}$$

We study inference in the well-specified setting where the true distribution p_0 of $x_{1:n}$ is $\mathcal{N}(0, 1)$, and in the misspecified setting, where p_0 is a *skew* normal with the same mean and variance but skew parameter 10. Details on the simulations and models are in Section B, and code for these experiments is at `https://github.com/bohanwu2000/npp`.

3.1 Polya Tree NPP

We first study an NPP model that perturbs the parametric model with a Polya tree. The advantage of Polya trees is that they admit closed form marginal likelihoods, making an accurate approximation of the Bayes factor in the NPP posterior possible. This provides a

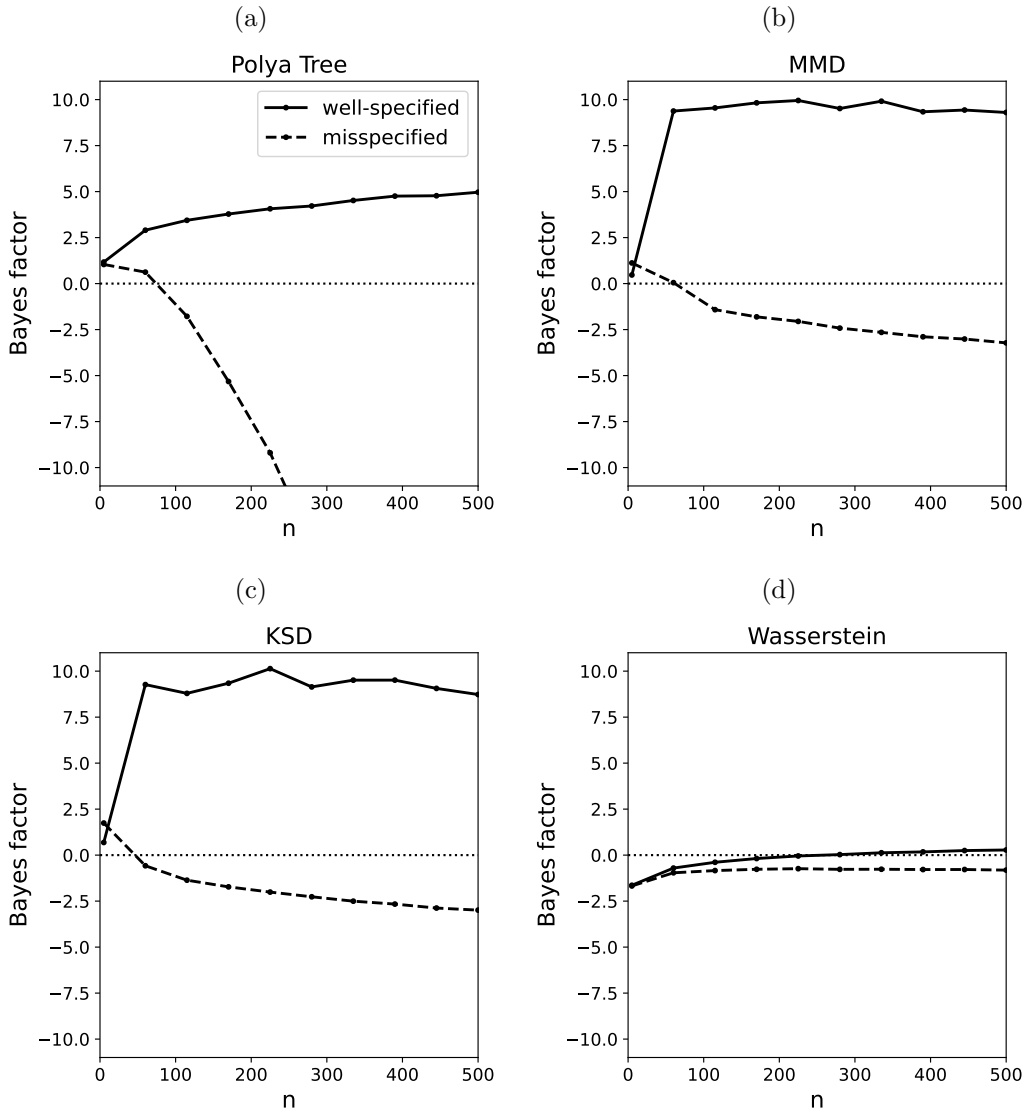


Figure 1: **Synthetic data (generalized) Bayes factors.** The log Bayes factor and log generalized Bayes factor comparing the parametric model to a nonparametric alternative, where positive values indicate the parametric model is favored. (a) NPP model with a Polya Tree. (b,c,d) gNPP models with the MMD, KSD and Wasserstein.

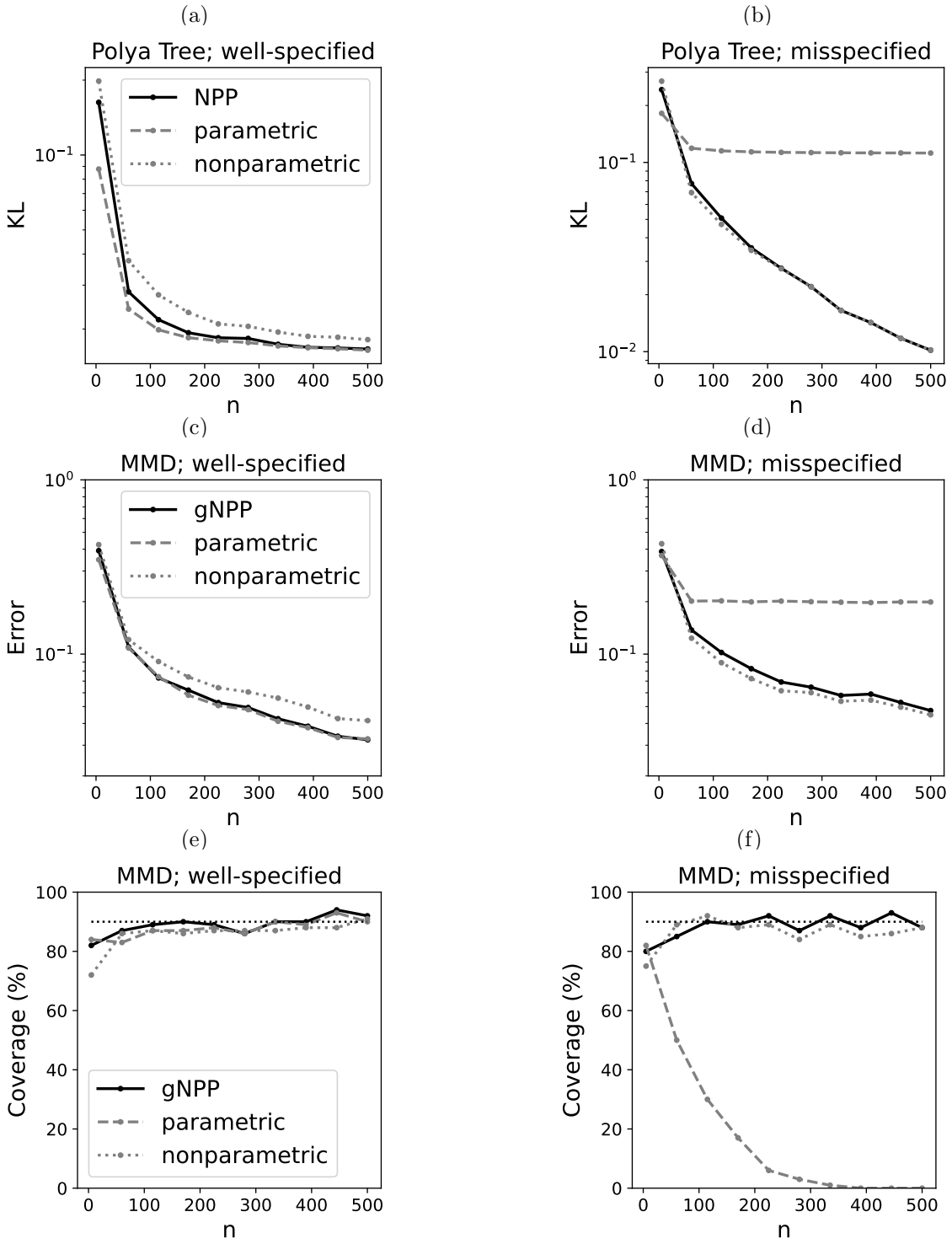


Figure 2: **Synthetic data results.** (a,b) KL divergence between the true data density and the posterior predictive of each model. (c,d) Absolute difference between the posterior mean estimate of the median and the true median, for each model, using the MMD gNPP. (e,f) Calibration of the MMD gNPP. We plot how often, across independent simulations, the posterior credible interval covers the true median. The nominal coverage is 90% (dashed).

careful check of our theoretical results. The Polya tree NPP model is,

$$\begin{aligned} &\theta \sim \mathcal{N}(0,1) \quad b \sim \text{Bernoulli}(\eta = 0.1), \\ &\begin{cases} \text{p} = \mathcal{N}(\theta, 1) & \text{if } b = 1 \\ \text{p} \sim \text{PolyaTree}(\mathcal{N}(\theta,1), h) \quad h \sim \text{Exponential}(1) & \text{if } b = 0 \end{cases}, \\ x_{1:n} &\overset{iid}{\sim} \text{p}. \end{aligned} \tag{13}$$

Here, we parameterize the Polya tree as in Berger and Guglielmi (2001); it is a distribution over distributions, with mean $\mathcal{N}(\theta, 1)$ and scale parameter h. We draw the scale parameter h from an exponential prior. We compute $\Pi_{\text{pert}}(x_{1:n})$ using the procedure in Berger and Guglielmi (2001).

We first examine convergence of the posterior distribution to the truth, measuring the KL divergence between the true distribution and the posterior predictive of the NPP model, averaged over 100 independent datasets. We compare to the performance of a parametric model (Eq. (13) given $b = 1$) and to the performance of a nonparametric model (Eq. (13) given $b = 0$).

Figure 1a shows that the NPP prefers the parametric model when the latter is well-specified, while it prefers the nonparametric model in the misspecified case. With small amounts of data ($n = 5$), the Bayes factor prefers the parametric model even if it is misspecified, and even though the prior $\eta = 0.1$ prefers the nonparametric model. This is an example of the Bartlett-Lindley effect: the parametric model is preferred even when it is misspecified since it provides a better approximation given the amount of available data.

The NPP model is efficient: in the well-specified case, it matches the faster convergence of the parametric model (Figure 2a). The NPP model is also robust: in the misspecified case, it converges to the true distribution, even though the parametric model does not (Figure 2b). Note that at very small amounts of data ($n = 5$), the parametric model provides better estimates than the nonparametric model despite its misspecification, as it offers a useful approximation of the underlying distribution. The NPP inherits some of these benefits, outperforming the nonparametric model at $n = 5$.

3.2 Generalized Bayes Approximation to the NPP

We next study the generalized Bayes approximation to the posterior (gNPP). To design our approximation, we must first choose a functional of interest. Following Lyddon et al. (2018), we focus on the median, $\psi(\text{p}) = \text{argmin}_\alpha \mathbb{E}_\text{p}[|X - \alpha|]$. Note this is a nonlinear functional of p. The parametric model assumes the mean and median of the distribution are both θ, but in the misspecified case, when the true data distribution p_0 is skew normal, the median is about -0.2. For the nonparametric component, we use a Bayesian bootstrap, a Dirichlet process with concentration 0 (Rubin, 1981). This offers consistent estimation of the target functional, and it is straightforward to draw samples from the posterior. For the generalized Bayes factor, we consider the 2-Wasserstein divergence, the MMD with an inverse multiquadric (IMQ) kernel, and the KSD with the same kernel; the IMQ, $k(x, x') = (c^2 + \|x - x'\|^2)^{-1/2}$, is a characteristic kernel that is well-suited for KSD-based inference (Gorham and Mackey, 2017). We compare the gNPP to two baselines: (a) the parametric

model alone, and (b) the Bayesian bootstrap, a nonparametric model that places no prior weight on the parametric family.

We quantify the performance of models' point estimates in terms of the absolute difference between the posterior mean of the median and the true median, $|\mathbb{E}[\psi(\mathrm{p}) \mid x_{1:n}] - \psi(\mathrm{p}_0)|$. The behavior of the gNPP follows that of the NPP. In the well-specified case, the gNPP matches the fast convergence of the parametric model (Figures 2c and 7). In the misspecified case, it converges to the true median (Figures 2d and 7). At small n we see a Bartlett-Lindley effect in the MMD and KSD generalized Bayes factors, as they prefer the misspecified but simple parametric model at $n = 5$, despite the prior $\eta = 0.1$ (Figures 1b and 1c). This explains their performance improvement over the nonparametric model at low n. Note the effect is not theoretically guaranteed to always occur, and in this case is absent from the Wasserstein gNPP (Figure 1d).

We also examined uncertainty quantification. We checked whether posterior credible sets are calibrated and achieve frequentist coverage. We computed how often, across independent simulations, the true median fell in the posterior 95% credible interval; with correct calibration, it should fall in the interval 95% of time. In the well-specified case, the gNPP is well calibrated, as is the parametric and nonparametric model (Figures 2e and 8). In the misspecified case, the parametric model is miscalibrated, but the gNPP inherits the good calibration of the nonparametric model (Figures 2f and 8).

A key hyperparameter in the gNPP is the rate r. We set $r = 0.49$ for the MMD and KSD versions, based on the intuition that r should be as large as possible to enable efficient convergence in the misspecified case, while still below $1/2$, the parametric rate, to preserve convergence in the well-specified case (Remark 1). While this default worked well for the MMD and KSD versions, in the Wasserstein version it led to a gBF that failed to favor the parametric model in the well specified case even with $n = 500$ datapoints, and even though the gBF slowly trended towards larger value (Figure 9). Instead, setting $r = 0.1$ led to improved performance by effectively discriminating between the well-specified and misspecified cases at lower values of n. So although $r = 0.49$ may be a reasonable default, problem-specific tuning can improve performance.

4 Theory

We establish the frequentist properties of NPP models and the gNPP approximation. In particular, we prove they are (a) robust, in the sense that the posterior converges to the true data distribution, and (b) efficient, in the sense that the posterior converges at a parametric rate when the underlying parametric model is well-specified. We assume a population distribution p_0 over $\mathcal{X}$ that produces i.i.d. observations $x_{1:n}$. In this section we assume, for simplicity of exposition, that $\mathcal{X}$ is an open subset of $\mathbb{R}^\kappa$, where the Lebesgue densities are well-defined. But note that this is not a requirement for implementing the NPP or the gNPP, with the Wasserstein or MMD, in practice. Finally, we assume a unique parameter θ_0 in an open set $\Theta \subseteq \mathbb{R}^d$ that minimizes the KL divergence to the population distribution, $D_{\mathrm{KL}}(\mathrm{p}_0 \parallel \mathrm{p}_\theta)$. Detailed statements and proofs of each result are in Section D.

4.1 NPP models

Model selection consistency. We first analyze the mixing weight η_n in the NPP. Proposition 2 establishes general conditions for its model selection consistency: η_n asymptotically selects the parametric model when that model is correct, and the nonparametric model otherwise. The assumptions (Assumptions 5 to 7 in Section C.2) follow the standard "prior mass and testing" framework for establishing posterior contraction rates in Bayesian nonparametrics (Ghosal and Van der Vaart, 2001). Informally, the entropy condition (Assumption 5) characterizes the complexity of local Hellinger neighborhoods in the parametric and nonparametric model classes and thereby defines the corresponding contraction rates. We further need the NPP prior to put sufficient mass in a reverse KL ball around p_0 (Assumption 6), and, when the parametric model is well-specified, the nonparametric prior should put much less mass around the truth than the parametric model (Assumption 7). The result then follows directly from Ghosal et al. (2008, Corollary 3.1).

Proposition 2 (η_n is consistent for model selection)**.** *Assume the marginal density* $\mathrm{p}_{\mathrm{pm}}(x_{1:n}) := \int_\Theta \mathrm{p}_\theta(x_{1:n}) d\Pi_{\mathrm{pm}}(\theta)$ *is well-defined. Under Assumptions 5 to 7 given in Section C.2,*

1. $\eta_n \to 1$ *a.s.* $[\mathbb{P}_0^\infty]$*, if* $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$.
2. $\eta_n \to 0$ *a.s.* $[\mathbb{P}_0^\infty]$*, if* $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$.

We check if the assumptions of Proposition 2 are satisfied for non-parametric perturbations based on a Dirichlet process and on a Dirichlet process mixture model.

Example 1 (Dirichlet process perturbations are not consistent)**.** *Consider an NPP model (Eq.* (2)*) with a Dirichlet process perturbation* $\mathrm{p} \sim \mathrm{DP}(\mathrm{p}_\theta, \alpha)$*. For this choice of perturbation, the mixing weight fails to satisfy model selection consistency. The technical assumption that fails is Assumption 6 which requires the DP prior to put sufficient mass around* $\{\mathrm{p} : D_{\mathrm{KL}}(\mathrm{p}_0 \,\|\, \mathrm{p}) < \epsilon^2\}$ *for small* ϵ*. But, since the DP prior is only supported on discrete measures, it puts zero mass on this reverse KL ball for any non-discrete* p_0*. The full result is provided in Proposition 12 in Section D.1 with a proof.*

One implication of this result is that the robust Bayes approach studied by Lyddon et al. (2018), in which a parametric model is perturbed by a Dirichlet process, is robust but not necessarily efficient.

Example 2 (Dirichlet process normal mixture perturbations are consistent)**.** *Consider a nonparametric perturbation based on a Dirichlet process normal mixture, as described in Eqs.* (2) *and* (3)*. Unlike the Dirichlet process, this perturbation describes continuous densities over* $\mathcal{X}$*. When* p_0 *is smooth, Proposition 14 implies that the Dirichlet process normal mixtures satisfy the assumptions of Proposition 2, implying* η_n *is model selection consistent.*

Robustness and Efficiency. We now establish robustness and efficiency for the NPP posterior. Recall the target of inference is defined via a functional $\psi : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{R}^s, s < \infty$. To establish efficiency, we prove a Bernstein-von Mises (BvM) theorem: when $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, the posterior over the functional $\psi(\mathrm{p})$ is asymptotically normal, with a standard deviation proportional to $1/\sqrt{n}$. This result is closely related to semiparametric BvM theorems in

the literature, which characterize the asymptotic normality of low-dimensional functionals in nonparametric Bayesian models (Bickel and Kleijn, 2012; Rivoirard et al., 2012; Castillo and Rousseau, 2015). To establish robustness, we show that even when $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, the posterior concentrates at $\psi(\mathrm{p}_0)$.

The NPP inherits its asymptotic normality from the underlying parametric model. The standard Bernstein-von Mises theorem says the parametric posterior $\Pi_{\mathrm{pm}}(d\theta \mid x_{1:n})$ is asymptotically normal, so under smoothness conditions, the posterior over $\psi(\mathrm{p}_\theta)$ will be asymptotically normal as well.

Assumption 1 (The parametric model is asymptotically normal)**.** *There exists $\theta_0 \in \Theta$ and a positive definite matrix V_{θ_0} such that $n \to \infty$,*

(a) $\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{\mathrm{w}} \mathcal{N}\left(0, V_{\theta_0}^{-1}\right)$*, where $\xrightarrow{\mathrm{w}}$ denotes weak convergence and $\hat{\theta}_{MLE}$ is the maximum likelihood estimate of θ, and*

(b) $d_{TV}\left(\sqrt{n}\left(\theta - \hat{\theta}_{MLE}\right), \mathcal{N}\left(0, V_{\theta_0}^{-1}\right)\right) \xrightarrow{\mathbb{P}_0} 0$ *for $\theta \sim \Pi_{\mathrm{pm}}(\cdot \mid x_{1:n})$, with convergence in first and second moments in $[\mathbb{P}_0]$-probability.*

Conditions (a) and (b) assume the parametric MLE and posterior are asymptotically normal at the $\sqrt{n}$-rate. Condition (b) is the Bernstein-von Mises theorem for regular parametric models, and guarantees the frequentist coverage of posterior credible regions will asymptotically match its nominal level. The convergence in first two moments follows, for example, from convergence of the posterior to a normal distribution in the Wasserstein metric. To streamline the presentation, we write $\dot{\chi}_\theta := \nabla \chi_\theta$ and $\ddot{\chi}_\theta := \nabla^2 \chi_\theta$ for a function χ_θ of θ.

Assumption 2 (The target functional is smooth)**.** *The function $\chi_\theta := \psi(\mathrm{p}_\theta) - \psi(\mathrm{p}_{\hat{\theta}_{MLE}})$ is twice differentiable, where $\dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}$ is positive definite, $\ddot{\chi}_\theta$ is continuous at θ_0 and $\|\ddot{\chi}_{\theta_0}\|_2 < \infty$ in $[\mathbb{P}_0]$-probability.*

We now establish our main result. To describe convergence of the NPP posterior to a normal distribution, we use the bounded Lipschitz distance d_{BL}, which metrizes the topology of weak convergence (Theorem 10 and Theorem 11).

Theorem 3 (NPP models are efficient and robust)**.** *Let Proposition 2 and Assumptions 1 and 2 hold. Let $\Pi\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right)$ denote the pushforward measure of the NPP posterior through $\tilde{\psi}_n(\mathrm{p}) := \sqrt{n}(\psi(\mathrm{p}) - \psi(\mathrm{p}_{\hat{\theta}_{MLE}}))$. If $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, then*

$$d_{BL}\left(\Pi\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right)\right) \xrightarrow{\mathbb{P}_0} 0 \quad \textit{(Efficiency).} \tag{14}$$

Further assume the nonparametric perturbation $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ is consistent at $\psi(\mathrm{p}_0)$. Then if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$,

$$d_{BL}\left(\Pi\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \xrightarrow{\mathbb{P}_0} 0 \quad \textit{(Robustness).} \tag{15}$$

Hence the NPP posterior $\Pi\left(\psi(\mathrm{p}) \mid x_{1:n}\right)$ is also consistent at $\psi(\mathrm{p}_0)$.

When the parametric model is correctly specified, Theorem 3 tells us that the NPP posterior of the target estimand achieves the $\sqrt{n}$-rate and the optimal variance in the asymptotic minimax sense (Van der Vaart, 2000). When the parametric model is misspecified, the NPP posterior is still robust as it converges to the true $\psi(\mathrm{p}_0)$. Sufficient conditions for the consistency of $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ are (a) consistency of the nonparametric posterior for all p_θ, e.g. Dirichlet process normal mixture models, and (b) continuity of the functional ψ with respect to the weak topology.

The efficiency guarantee could be extended to the misspecified setting if the nonparametric posterior $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ achieves semiparametric efficiency for the target functional ψ. This could follow from a semiparametric Bernstein–von Mises theorem for the posterior of $\psi(\mathrm{p})$ around a sequence of asymptotically efficient regular estimators $\hat{\psi}_n$ in the sense of the convolution theorem (Van der Vaart, 2000, Theorem 25.20). Specifically, we require that, after centering at $\hat{\psi}_n$ and scaling by $\sqrt{n}$, the conditional distribution of $\psi(\mathrm{p})$ under $\Pi_{\mathrm{pert}}(\cdot \mid x_{1:n})$ is asymptotically Gaussian with variance equal to the semiparametric efficiency bound. Given such a result, the efficiency statement for NPP would then follow by combining: (i) the semiparametric BvM for $\Pi_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$, and (ii) consistency of the (generalized) Bayes factor, which ensures that the posterior concentrates on the appropriate limiting model. An example of an efficient nonparametric posterior is the one-step corrected posterior of Yiu et al. (2025), which achieves semiparametric efficiency under its stated assumptions.

4.2 gNPP approximation

We establish the efficiency and robustness of the gNPP approximation, proving an analogue of Theorem 3 for our generalized Bayes inference approach (Theorem 6). To do so, we first show that the generalized Bayes factor satisfies model selection consistency (Theorem 4). Our results apply to a broad class of divergences, but we verify the conditions and provide explicit contraction rates for the Wasserstein, MMD and KSD (Theorems 16, 21 and 26 in Section D.3). The contraction rates justify our guidance for setting the rate hyperparameter r.

Model selection consistency. We first analyze the model selection consistency of the mixing weight $\hat{\eta}_n$ in the gNPP. Recall $\hat{\eta}_n := 1/(1 + \mathrm{gBF}_n^{-1})$, where gBF_n depends on a posterior expected empirical divergence $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right]$ (Eq. (6)). We are interested in the asymptotic behavior of this posterior divergence, showing it correctly detects model misspecification.

To begin, we study general divergences D that are semimetrics or depend on one, in the sense that $\mathrm{D}(\mathrm{p}, \mathrm{q}) = \rho^k(\mathrm{p}, \mathrm{q})$ for $k \in \mathbb{N}$ and ρ a semimetric. The Wasserstein and MMD fit this form; the KSD is asymmetric and hence does not, so we later extend our arguments. We also assume in the general theory that the estimated divergence $\mathrm{D}_{m,n}(\mathrm{p}, \mathrm{q})$ takes the form of a *plug-in* estimator $\mathrm{D}_{m,n}(\mathrm{p}, \mathrm{q}) = \mathrm{D}\left(\mathrm{p}^m, \mathrm{q}^n\right)$, where p^m is the m-sample empirical distribution of p, and q^n is the n-sample empirical distribution of q. The empirical p-Wasserstein distance is directly given by this plug-in estimator, while the plug-in form of the MMD is its V-statistic (Gretton et al., 2012). We later extend our arguments to the MMD U-statistic, which matches the plug-in form with a vanishing error, $O(n^{-1} + m^{-1})$ (Gretton et al., 2012; Serfling, 2009).

We make regularity assumptions on the divergence.

Assumption 3 (The divergence is well-behaved)**.** *Assume* $\mathrm{D}(\mathrm{p},\mathrm{q}) = \rho^k(\mathrm{p},\mathrm{q})$ *where* $k \in \mathbb{N}$ *and* ρ *is a semimetric, and that*

(a) ρ *is continuous in the weak topology and* $\sup_{\theta\in\Theta} \rho(\mathrm{p}_\theta, \mathrm{p}_0) < \infty$*.*

(b) There exists an $M_n n^{-1/2}$ *neighborhood of* θ_0 *where* $M_n \to \infty$ *such that the mapping* $\theta \mapsto \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$ *is twice differentiable and* $\left\|\nabla^2_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})\right\|_2$ *is uniformly bounded.*

Assumption 3(a) requires the divergence to be uniformly bounded. For example, with the Wasserstein distance, the condition is satisfied if $\mathcal{X}$ is compact. For MMD, the condition is satisfied if the kernel is uniformly bounded. Assumption 3(b) requires local Lipschitz smoothness of the function $\theta \mapsto \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$ within an asymptotically vanishing ball around θ_0, where the radius is at least $n^{-\alpha/2}$, with $\alpha < 1$. This condition is typically satisfied for commonly used discrepancies under standard smoothness assumptions on the parametric family, such as those used to establish central limit theorems for minimum ρ-discrepancy estimators (e.g. Barp et al., 2019). The precise meaning of a "smooth parametric model" depends on the choice of ρ. For example, if ρ is the MMD induced by a translation-invariant kernel k, a natural notion of smoothness is to parametrize $\mathbb{P}_\theta$ as a pushforward $\mathbb{P}_\theta = T_{\theta,\#}U$ for a reference measure U. In this setting, by following the same argument of Briol et al. (2019, Theorem 2.), it suffices for Assumption 3(b) to hold if (a) k has bounded mixed derivatives up to order 2; (b) there exists a neighborhood O_n of θ_0 of radius $M_n n^{-1/2}$ such that T_θ is twice continuously differentiable in θ for all $\theta \in O_n$; and (c) the derivatives of T_θ satisfy suitable integrability/dominance conditions, e.g., $\int \sup_{\theta\in O_n} \|\nabla^i_\theta T_\theta(u)\|_2 \, dU(u) < \infty$ for $i \in \{1,2\}$. Similar conditions appear for the KSD in (Barp et al., 2019, Theorem 4 and 5) and for the Wasserstein in (Bernton et al., 2019, Theorem 2.3.).

We also assume that the empirical divergence converges to the true divergence, at least when the model likelihood is in a neighborhood of the truth.

Assumption 4 (The empirical divergence converges)**.** *There exists a sequence* $r_{m,n} \to 0$ *such that*

(a) For any sequence $\theta_n \to \theta_0$*, as* $m, n \to \infty$ *with* $m/(m+n) \to c \in (0,1)$*, we have* $\mathbb{E}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_{\theta_n}, \mathrm{p}_0\right)\right] = \mathrm{D}\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right) + O(r_{m,n})$*. The expectation is over* m *samples from* p_{θ_n} *and* n *samples from* p_0*.*

(b) There exists a function $\mathcal{V}(\theta)$ *that is finite at* θ_0 *and continuous at* θ_0 *such that for all* (m,n) *large enough,* $Var\left(\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right)\right) \le r^2_{m,n}\mathcal{V}(\theta)$*, where Var denotes the variance.*

Assumption 4 requires an asymptotic bound on the expectation of $\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right)$ and a finite-sample variance bound around θ_0, both converging at a rate $r_{m,n}$. These assumptions are local, applying only to a neighborhood of θ_0. They are also weaker than assuming uniform convergence of the mean or the variance of $\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right)$ in a neighborhood around θ_0.

When we specialize Assumption 4 to specific divergences, we can verify the assumption more closely. For example, for the MMD, the plug-in V-statistic matches the U-statistic up to an ignorable error, and for the U-statistic, Assumption 4(a) is satisfied, since the statistic is unbiased, and Assumption 4(b) follows from the variance formula of two-sample U-statistics and kernel regularity conditions.

We now study the posterior expected value of the empirical divergence, under the parametric model. To make the analysis more tractable theoretically, we assume the samples used to approximate p_0 in $\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0)$ are independent of those used to compute the posterior. This can be achieved by splitting data and cross-fitting (Chernozhukov et al., 2018). However, we expect the model selection consistency of $\hat{\eta}_n$ (Theorem 5) to remain valid even without sample splitting, and in the empirical studies we do not apply it.

Theorem 4 (The posterior expected empirical divergence converges at a rate $r_{m,n} \vee n^{-1}$)**.** *Let Assumptions 1, 3 and 4 be satisfied. As $m, n \to \infty$ with $n/(n+m) \to c \in (0,1)$, $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right]$ converges in $[\mathbb{P}_0^\infty]$-probability to $\mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0)$ at the rate of $r_{m,n} \vee n^{-1} := \max(r_{m,n}, n^{-1})$:*

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] = \mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) + O_{\mathbb{P}_0}(r_{m,n} \vee n^{-1}). \tag{16}$$

So the posterior empirical divergence converges to the true minimal divergence between the model class $\mathcal{M}_{\mathrm{pm}}$ and p_0, at a rate that is the slower of (a) $r_{m,n}$, the convergence rate of $\mathrm{D}_{m,n}$, and (b) n^{-1}.

Theorem 4 applies to general divergences. For the Wasserstein, MMD, and KSD in particular, we derive the following contraction rates (detailed theorems and conditions are in Section D.3).

1. ***p*-Wasserstein Distance.** The empirical Wasserstein divergence $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{W}_p^p(\mathrm{p}_\theta^m, \mathrm{p}_0^n) \mid x_{1:n}\right]$ converges to $\mathrm{W}_p^p\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)$ at a rate of $O_{\mathbb{P}_0}\left(n^{-2/(\kappa \vee 4)} + m^{-2/(\kappa \vee 4)}\right)$, where κ is the dimension of $\mathcal{X}$ (Section D.3.1).

2. **MMD.** The empirical MMD $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right]$, based on the U-statistic for the MMD, converges to $\mathrm{MMD}^2\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)$ at a rate of $O_{\mathbb{P}_0}\left(n^{-1/2} + m^{-1/2}\right)$ (Section D.3.2).

3. **KSD.** The empirical KSD $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right]$, based on the U-statistic for the KSD, converges to $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0})$ at a rate of $O_{\mathbb{P}_0}(n^{-1/2})$ (Section D.3.3).

For simplicity, in our implementation, we set $m = n$ for the two-sample empirical divergences (Wasserstein and MMD), so the rate $r_{m,n}$ simplifies to $r_n := r_{n,n}$. In this case, for the Wasserstein, we can take $r_n = n^{-2/(\kappa \vee 4)}$, and for the MMD and KSD, we can take $r_n = n^{-1/2}$.

We have shown that the posterior expected empirical divergence converges to the true divergence at a known rate, r_n. We can now set the rate hyperparameter r in the generalized Bayes factor to be slower than this known rate, i.e. $r \in \left(0, \frac{2}{\kappa \vee 4}\right)$ for the Wasserstein and $r \in (0, 1/2)$ for MMD and KSD. Then, the generalized Bayes factor is model selection consistent.

Theorem 5 ($\hat{\eta}_n$ is consistent for model selection)**.** *Let $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)\right]$ be bounded in $[\mathbb{P}_0^\infty]$-probability, $\eta > 0$ and assume $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] = \mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) + r_n$, where $\mathrm{D}(\cdot, \cdot)$ is a statistical divergence. Choose $r > 0$ such that $r_n(n+1)^r = o(1)$. Then, as $n \to \infty$:*

1. *$\hat{\eta}_n \to 1$ a.s. $[\mathbb{P}_0^\infty]$, if $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$.*

2. *$\hat{\eta}_n \to 0$ a.s. $[\mathbb{P}_0^\infty]$, if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$.*

Note $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)\right]$ is bounded in probability when, for example, $\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)$ is uniformly bounded or satisfies a uniform law of large numbers. Remark 25 in Section D.3.2 verifies the conditions of Theorem 5 for the model used in the synthetic study of Section 3, with the generalized Bayes factor constructed with the MMD distance.

Robustness and Efficiency. The gNPP approximation uses the generalized Bayes factor to adaptively trade off between parametric and nonparametric models. We have shown the generalized Bayes factor is model selection consistent. As a result, the gNPP approximation converges efficiently when the parametric model is well-specified, but still converges to the truth when it is misspecified.

Theorem 6 (gNPP approximations are efficient and robust)**.** *Let Assumptions 1 to 4 hold for $m = n$ and $r_n := r_{n,n}$. Assume also that $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)\right]$ is bounded in $[\mathbb{P}_0]$-probability. Let $r < 1$ be a positive constant such that $r_n(n+1)^r = o(1)$. Let $\hat{\Pi}\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right)$ denote the pushforward measure of $\tilde{\psi}_n(\mathrm{p}) := \sqrt{n}(\psi(\mathrm{p}) - \psi(\mathrm{p}_{\hat{\theta}_{MLE}}))$ onto the gNPP posterior. If $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, then*

$$d_{BL}\left(\hat{\Pi}\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right)\right) \overset{\mathbb{P}_0}{\to} 0 \quad \textit{(Efficiency).} \tag{17}$$

Furthermore, assume that the nonparametric posterior $\hat{\Pi}_{\mathrm{pert}}(\psi(\mathrm{p}) \mid x_{1:n})$ is consistent at $\psi(\mathrm{p}_0)$. Then if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$,

$$d_{BL}\left(\hat{\Pi}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \overset{\mathbb{P}_0}{\to} 0 \quad \textit{(Robustness)}, \tag{18}$$

hence the gNPP posterior $\Pi\left(\psi(\mathrm{p}) \mid x_{1:n}\right)$ is also consistent at $\psi(\mathrm{p}_0)$.

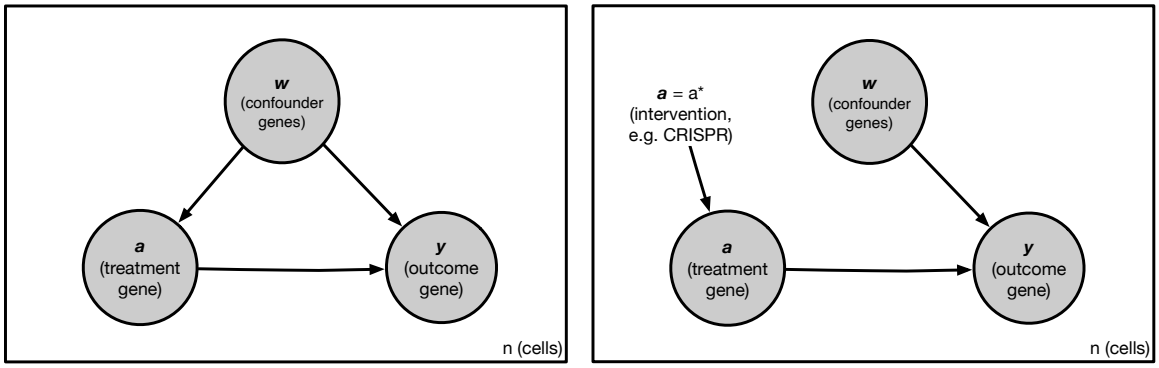
We can replace Assumptions 3 and 4 with assumptions specific to the Wasserstein, MMD and KSD, as detailed in Section D.3.

5 Application: Estimating the Effects of Gene Expression

A cell's behavior depends crucially on its genes' *expression levels*, or the number of RNA transcripts of each gene in the cell. An important question in cell biology is how the expression level of one gene affects another. Answering such questions can help scientists understand cells' behavior, and may help develop drugs which modify gene expression levels or activity.

We develop a robust method to estimate the causal effect of one gene's expression level on another's. To do so, we leverage observational data about the expression level of genes in individual cells, measured with single cell RNA sequencing (scRNAseq) (Kolodziejczyk et al., 2015). We focus on analyzing primary patient data, which is often hard to obtain and hence limited in scale. Bayesian parametric models offer data-efficient inference, but robustness is crucial given the importance of the estimates to human health.

We specify the causal effect as a functional of the data distribution. We start from a parametric Bayesian model that incorporates prior information from an auxiliary dataset, a *single cell atlas*. We robustify this parametric model by creating a gNPP approximation



(a) (b)

Figure 3: **Causal model.** To analyze the effects of gene expression, we assume this causal graphical model. (a) The initial model where all variables are observed. (b) The model after intervening on the treatment variable A. The goal is to estimate the effect of the treatment gene on the outcome gene, as highlighted in blue.

to the NPP posterior, with a BART-based nonparametric perturbation and an MMD-based generalized Bayes factor.

We apply the gNPP model to conduct an exploratory analysis of data from an ovarian cancer patient. We estimate the causal effects of genes that are potential therapeutic targets. Our results reveal that interventions on some genes may have unexpected effects in this patient.

Additional details are in Section E. Code and data can be found at `https://github.com/bohanwu2000/npp`.

5.1 Setup

Estimand. We are interested in the effect of the expression level of a treatment gene on the expression level of an outcome gene. The challenge for causal inference is that there may be confounders, factors which affect the expression of both the treatment and the outcome genes. An important source of confounding is *transcription factors*, proteins which directly modify the expression level of many other genes. To correct for this confounding, we adjust for the expression level of the transcription factor genes in each cell.

Formally, we consider the causal graph in Figure 3, where a is the treatment gene expression level, y is the outcome gene expression level, and w is a vector of expression levels for confounding genes, i.e. transcription factors. We are interested in the effect of an intervention which sets the expression level of a to $a_\star$. From Figure 3, the distribution of the outcome after the intervention is given by the backdoor adjustment, $\mathrm{p}(y \mid \mathrm{do}(a_\star)) = \int \mathrm{p}(y \mid a_\star, w)\mathrm{p}(w)\mathrm{d}w$ (Pearl, 2009).

We compare the average outcome when the treatment gene is highly expressed versus when it is unexpressed. To define "highly expressed" we take the 98th quantile of the empirical distribution of treatment expression, denoted $q_{98}(a)$. Our target functional $\psi(\mathrm{p})$

is the *average treatment effect*,

$$\begin{aligned}\psi(\mathrm{p}) = \mathrm{ATE}\,(\mathrm{p}) &:= \mathbb{E}_{\mathrm{p}}\left[Y\,;\,\mathrm{do}(a_\star = q_{98}(a))\right] - \mathbb{E}_{\mathrm{p}}\left[Y\,;\,\mathrm{do}(a_\star = 0)\right] \\ &= \int\int y\,\mathrm{p}(y \mid q_{98}(a), w)\mathrm{p}(w)\mathrm{d}y\mathrm{d}w - \int\int y\,\mathrm{p}(y \mid 0, w)\mathrm{p}(w)\mathrm{d}y\mathrm{d}w,\end{aligned} \tag{19}$$

where $\mathrm{p}(a, y, w)$ is the joint distribution over treatment, outcome and confounders.

Parametric model We first introduce a parametric Bayesian model to estimate the causal effect $\psi(\mathrm{p}_0)$. This parametric model incorporates information from a single cell atlas dataset, a large collection of scRNAseq data from donors without cancer.

Rather than using the full vector of confounding gene expression w, the model uses a low-dimensional representation $z(w)$ derived from the atlas. Specifically, we set $z(\cdot)$ to be the projection of w onto the first ten principal components of the atlas. Such low-dimensional representations of gene expression – whether derived from principal component analysis or other approaches – have been found to provide an effective summary of *cell type* and *cell state* (Lopez et al., 2018). By using representations learned from a large atlas, biologists aim to better estimate the cell type and state underlying the noisy expression levels in a small dataset (Hao et al., 2021). We apply this idea to obtain a low-dimensional representation of confounding.

Our parametric Bayesian model depends linearly on the treatment and on the low-dimensional representation of the confounders,

$$\mathrm{p}_\theta(y \mid a, w) = \mathcal{N}(c + \tau a + \gamma^\top z(w), \sigma^2), \tag{20}$$

where $\theta := [c, \tau, \gamma]$ and $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. The parameter c is an intercept, τ is the coefficient on the treatment and γ is a vector of coefficients on the confounder representation. We use an improper prior $\pi(\theta) \propto 1$ and specify $\sigma \sim \mathrm{HalfNormal}(0, 1)$. To compute the empirical divergence of the prior, we approximate the improper prior by a uniform distribution over the hypercube with slice $[-100, 100]$. We place a Dirichlet process prior on $\mathrm{p}(a, w)$ with concentration approaching zero, such that the posterior over the ATE is obtained by the Bayesian bootstrap.

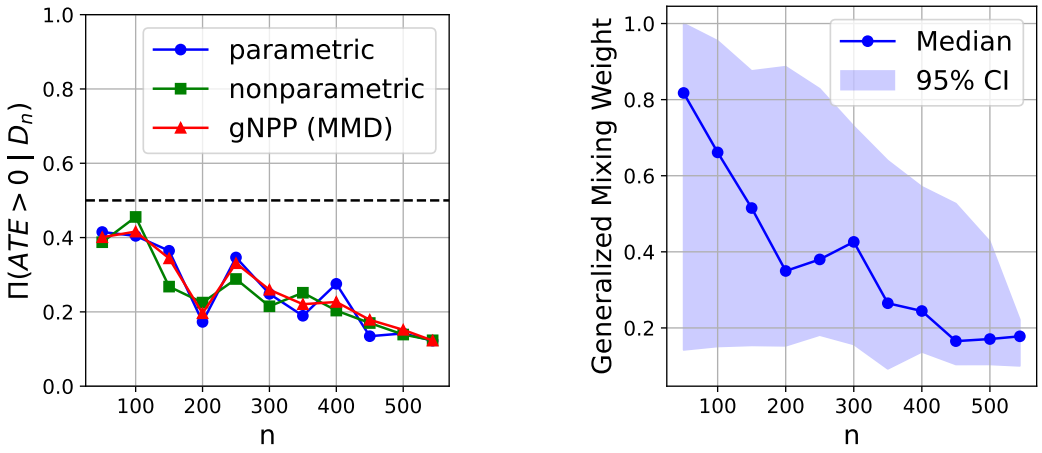
We expect the parametric model to learn efficiently even from limited numbers of cells, since the cell representation $z(w)$ is low-dimensional and hence so is the model. However, the parametric model may be misspecified, for instance if the ovarian cancer patient has a cells in a pathological state that is not present in the healthy atlas. In this case, the representation $z(w)$ may project out important variation in w, ignoring a source of confounding and leading to unreliable inferences.

gNPP model We introduce a gNPP model for robust inference. For the nonparametric model, we use a Bayesian additive regression tree (BART) with a fixed propensity score correction, an established method for nonparametric Bayesian causal inference (Hill, 2011; Hahn et al., 2020). Since single cell RNA expression data is highly non-normal, we adjust the BART model with a Yeo-Johnson transformation (Yeo and Johnson, 2000). We place a Dirichlet process prior on $\mathrm{p}(a, w)$ with concentration zero, such that the posterior over the ATE is obtained by a Bayesian bootstrap. The model does not incorporate prior information from the atlas. For the generalized Bayes factor, we use the MMD with rate $r = 0.49$.

Data and preprocessing We study an scRNAseq dataset consisting of 544 T cells from a patient with ovarian cancer (Vázquez-García et al., 2022). The atlas contains 261,611 T cells collected across 17 sites/tissues from 12 organ donors without cancer (Domínguez Conde et al., 2022). We log transform and standardize the expression data following standard practice in the field (Heumos et al., 2023). As potential confounders, we adjust for 157 transcription factors with highly variable expression. To find highly variable genes, we use the method of Stuart et al. (2019) as implemented in scanpy (Wolf et al., 2018). Briefly, the read counts are z-score normalized (using a regularized standard deviation), then genes are ranked by their empirical variance.

5.2 Results

A route to treating cancer is to intervene on a patient's immune system, such that it can better attack the cancer. One potential strategy is to modify regulatory T cells, which suppress immune attack, to be more like cytotoxic T cells, which conduct immune attack. FOXP3 is a transcription factor that makes T cells regulatory, so it may be a good drug target for achieving this goal (Revenko et al., 2022). Will decreasing the expression of FOXP3 make this ovarian cancer patient's T cells more cytotoxic? To address this question, we estimate the causal effect of FOXP3 on the expression of key genes that T cells use to clear tumors. We also consider interventions on another gene in Section E.2.

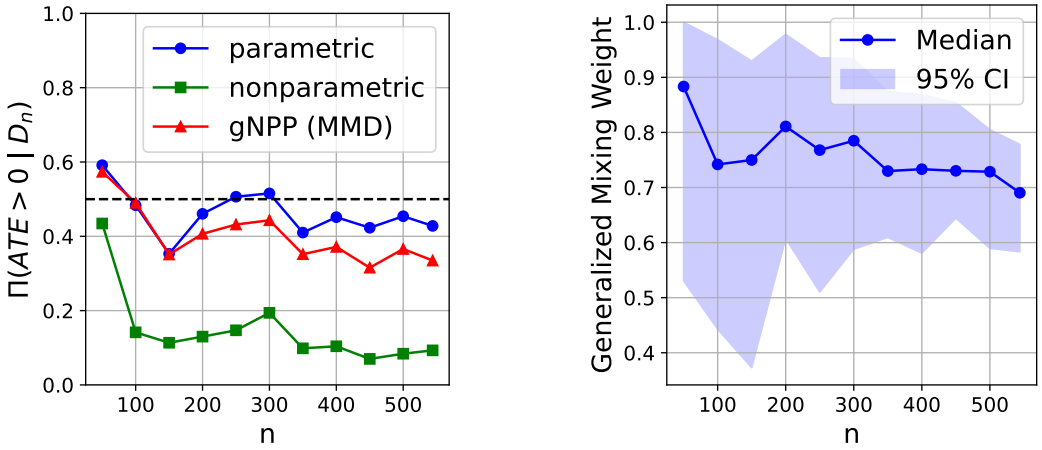


(a) (b)

Figure 4: **Effect of FOXP3 on GZMH.** a. Posterior probability of the ATE being positive under the parametric, nonparametric, and gNPP models. n denotes the size of the (subsampled) dataset. Values are the estimated median from 10 independent data subsamples and model samples. b. Generalized mixing weights, $\hat{\eta}_n$. The estimated confidence interval (CI) is across independent data subsamples and model samples.

GZMH We first consider the effect of FOXP3 on granzyme H and K (GZMH and GZMK), proteases which digests tumor proteins. For granzyme H (GZMH), under the parametric model, the posterior probability that the ATE is positive is low, 0.12 (Figure 4a). This suggests the ATE is likely negative, so an intervention that decreases FOXP3 expression will increase GZMH expression. In other words, the parametric model infers that intervening on FOXP3 will make these T cells more capable of attacking tumors. However, there are reasons to question this conclusion. The parametric model's R^2 is low, at 0.12, suggesting it explains only a small portion of the variance in the data. Moreover, its residual histogram and QQ plot suggest strong deviations from normality (Figure 11). So the parametric model may be poorly specified.

The gNPP draws inferences that are robust to misspecification. We find a small generalized mixing weights of 0.18, placing most of the posterior weight on the nonparametric model (Figure 4b). However, the posterior probability of a positive ATE under the gNPP model remains nearly unchanged, at 0.12. So, the conclusion that decreasing FOXP3 expression will increase GZMH expression is robust to possible model misspecification.

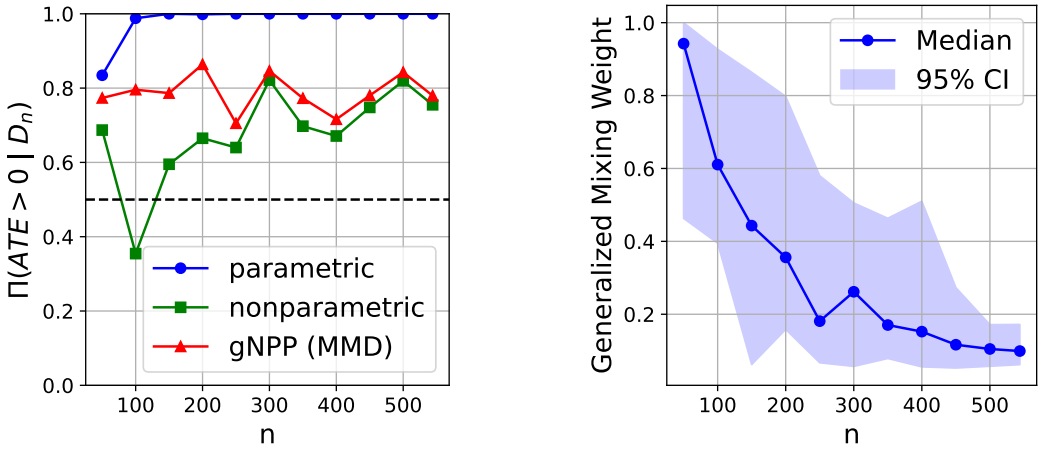


(a) (b)

Figure 5: **Effect of FOXP3 on GZMK.** a. Posterior probability of the ATE being positive under the parametric, nonparametric, and gNPP models. n denotes the size of the (subsampled) dataset. b. Generalized mixing weights, $\hat{\eta}_n$. CI: confidence interval across independent data subsamples and model samples.

GZMK We next consider the effect of FOXP3 on granzyme K (GZMK). The posterior probability that the ATE is positive, under the parametric model, is 0.43 (Figure 5a). The fact that this value is close to 0.5 suggests there is substantial uncertainty in the effect of FOXP3 on GZMK, and there may be no effect at all. The gNPP, however, revises this estimate downward to 0.33, providing more confidence that decreasing FOXP3 expression will increase GZMK expression. In this case, the gNPP still places considerable weight on the parametric model, with a generalized Bayes factor of 0.69 (Figure 5b).

We subsampled the data down to smaller numbers of cells and reran the analysis, averaging over ten independent subsamples. We found that with smaller amounts of data the gNPP model relies even more heavily on the parametric model, producing nearly the same posterior probability of a positive ATE (Figure 5a). The gNPP's estimates only begin to diverge from the parametric model at around $n = 200$ cells.



(a) (b)

Figure 6: **Effect of FOXP3 on NKG7.** a. Posterior probability of the ATE being positive under the parametric, nonparametric, and gNPP models. n denotes the size of the (subsampled) dataset. b. Generalized mixing weights, $\hat{\eta}_n$. CI: confidence interval across independent data subsamples and model samples.

NKG7 We next consider the effect of FOXP3 on the granule protein NKG7, another key component of T cells' attack. The parametric model suggests that decreasing FOXP3 expression will actually decrease NKG7, with the posterior probability of a positive ATE of 0.99 (Figure 6a). Biologically this is unexpected, since regulatory T cells do not express much NKG7 in general (Szabo et al., 2019). Correspondingly, the gNPP increases our uncertainty in the finding, revising the posterior probability down to 0.77 (Figure 6a). Here, the gNPP has strong confidence that the parametric model is wrong, with a generalized mixing weight of 0.1.

The gNPP adaptively smooths between the parametric and nonparametric posteriors, so its posterior probability of a positive ATE is quite stable here even if we sub-sample the data to below 100 datapoints (Figure 6a). The nonparametric model, by contrast, produces more variable estimates: at $n = 100$, it produces a posterior probability of 0.35, compared to the full dataset value of 0.76.

6 Discussion

We showed how adding a nonparametric perturbation to a parametric Bayesian model can robustify the model, guarding against misspecification without sacrificing data efficiency.

We then developed a generalized Bayes posterior that achieves these same properties, but enjoys more scalable computation and implementation. The basic technique is: (a) use a backup nonparametric or semiparametric model that is consistent for the target of inference and (b) mix the parametric posterior with the nonparametric posterior based on how well the parametric model describes the data. Overall, gNPP models offer a practical approach to combining scientific domain knowledge with flexible machine learning models, by regularizing inferences from a nonparametric model towards a parametric model.

The proposed gNPP posteriors have several important assumptions and limitations, which future work may overcome. First, they assume the data is independent and identically distributed (i.i.d.). In practice, data may be correlated, and the distribution may shift across time or space. An important future direction is to extend the generalized Bayes factor to handle non-i.i.d. data. Second, gNPP posteriors have several tuning parameters, including the rate r, the divergence D, and the divergence's hyperparameters, such as the choice of kernel for the MMD and the transportation cost for the Wasserstein. We have offered some basic guidance for choosing these parameters, but future work could go further to advance additional sensitivity analyses and develop techniques to tune parameters automatically and adaptively in response to data or the target estimand (Schrab et al., 2023). Third, our theoretical analysis of the gNPP has focused on its asymptotic behavior. Future theoretical and empirical work could investigate the non-asymptotic regime in depth, perhaps leading to improved designs for the generalized Bayes factor.

References

Ryan Prescott Adams, Iain Murray, and David J. C. MacKay. Nonparametric Bayesian density modeling with Gaussian processes. *arXiv preprint arXiv:0912.4896*, 2009.

Joseph Antonelli, Georgia Papadogeorgou, and Francesca Dominici. Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties. *Biometrics*, 2020.

Charles E Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, pages 1152–1174, 1974.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973, 2005.

Alessandro Barp, François Xavier Briol, Andrew B. Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

J.O. Berger and A. Guglielmi. Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, 96:174–184, 2001.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

S M Berry, R J Carroll, and D Ruppert. Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97:160–169, 2002.

Kush Bhatia, Yi-An Ma, Anca D Dragan, Peter L Bartlett, and Michael I Jordan. Bayesian robustness: A nonasymptotic viewpoint. *Journal of the American Statistical Association*, 119(546):1112–1123, 2024.

Peter J Bickel and Bas J K Kleijn. The semiparametric Bernstein–von Mises theorem. *Annals of Statistics*, 40:206–237, 2012.

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

David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, 1(2):353–355, 1973.

B. J.N. Blight and L. Ott. A Bayesian approach to model inadequacy for polynomial regression. *Biometrika*, 62, 1975.

Francois-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy, 2019.

Guillaume Carlier. On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM Journal on Optimization*, 32, 2022.

Ismaël Castillo and Judith Rousseau. A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Annals of Statistics*, 43:2353–2383, 2015.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21:C1–C68, 2018.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. *Statistical optimal transport*, volume 2364 of *Lecture Notes in Mathematics*. Springer, Cham, 2025. École d'Été de Probabilités de Saint-Flour XLIX—2019, École d'Été de Probabilités de Saint-Flour.

Hugh A Chipman, Edward I George, and Robert E McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4:266–298, 2010.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26, 2013.

A Philip Dawid. Posterior model probabilities. In Prasanta S Bandyopadhyay and Malcolm R Forster, editors, *Philosophy of Statistics*, volume 7, pages 607–630. North-Holland, 2011.

Bruno de Finetti. The Bayesian approach to the rejection of outliers. *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob*, 1, 1961.

Eustasio del Barrio, Alberto González Sanz, Jean-Michel Loubes, and Jonathan Niles-Weed. An improved central limit theorem and fast convergence rates for entropic transportation costs. *SIAM Journal on Mathematics of Data Science*, 5(3):639–669, 2023.

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. *International Conference on Artificial Intelligence and Statistics*, pages 943–970, 2022.

C Domínguez Conde, C Xu, LB Jarvis, DB Rainbow, SB Wells, T Gomes, SK Howlett, O Suchanek, K Polanski, HW King, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594):eabl5197, 2022.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.

Andrew Gelman, Xiao Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

S. Ghosal, J. Lember, and Aad W Van der Vaart. Nonparametric Bayesian model selection and averaging. *Electronic Journal of Statistics*, 2:63–89, 2008.

Subhashis Ghosal and Aad W Van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29(5):1233–1263, 2001.

Subhashis Ghosal and Aad W Van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and Aad W Van der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301, 2017.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Adityanand Guntuboyina and Bodhisattva Sen. L_1 covering numbers for uniformly bounded convex functions. *Journal of Machine Learning Research*, 23, 2012.

P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15:965–1056, 2020.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

J.L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20:217–240, 2011.

Han Hong and Bruce Preston. Nonnested model selection criteria. *unpublished*, 2005.

Han Hong and Bruce Preston. Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics*, 167(2):358–369, 2012.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, 2009.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Wenxin Jiang and Martin A Tanner. Gibbs posterior for variable selection in high-dimensional classification and data mining. *Annals of Statistics*, 36(5):2207–2231, 2008.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, pages 207–236, 2024.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular Cell*, 58(4):610–620, 2015.

Athanasios Kottas and Alan E Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96:1458–1468, 2001.

Daniel R Kowal and Bohan Wu. Monte Carlo inference for semiparametric Bayesian regression. *Journal of the American Statistical Association*, pages 1–14, 2024.

Sarita Kumari, Mohit Arora, Jay Singh, Shyam S Chauhan, Sachin Kumar, and Anita Chopra. L-selectin expression is associated with inflammatory microenvironment and favourable prognosis in breast cancer. *3 Biotech*, 11(2):38, 2021.

Jiawei Li and Jonathan H. Huggins. Calibrated model criticism using split predictive checks, 2024.

Qiang Liu, Jason D. Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. *33rd International Conference on Machine Learning*, 1, 2016.

Albert Y Lo. On a class of Bayesian nonparametric estimates: I. density estimates. *Annals of Statistics*, pages 351–357, 1984.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Ulrike Von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 2004.

Simon Lyddon, Stephen Walker, and Chris Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*, 2018.

Simon P Lyddon, C C Holmes, and S G Walker. General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106:465–478, 2019.

David JC MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.

Tudor Manole and Jonathan Niles-Weed. Sharp convergence rates for empirical optimal transport with smooth costs. *Annals of Applied Probability*, 34, 2024.

Jeffrey Miller. Flexible perturbation models for robustness to misspecification. *Unpublished*, 2019.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22:1–53, 2021.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Gemma E. Moran, David M. Blei, and Rajesh Ranganath. Holdout predictive checks for Bayesian model criticism. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 86, 2024.

A. O'Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 40, 1978.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11, 2019.

Emilia Pompe. Introducing prior information in weighted likelihood bootstrap with applications to model misspecification, 2021.

Miriana Quiroga, Pablo G Garay, Juan M. Alonso, Juan Martin Loyola, and Osvaldo A Martin. Bayesian additive regression trees for probabilistic programming, 2023.

Carl Edward Rasmussen and Christopher K I Williams. Gaussian processes for machine learning. *The MIT Press*, 2006.

Kolyan Ray and Aad van der Vaart. Semiparametric Bayesian causal inference. *Annals of Statistics*, 48:2999–3020, 2020.

Alexey Revenko, Larissa S Carnevalli, Charles Sinclair, Ben Johnson, Alison Peter, Molly Taylor, Lisa Hettrick, Melissa Chapman, Stephanie Klein, Anisha Solanki, Danielle Gattis, Andrew Watt, Adina M Hughes, Lukasz Magiera, Gozde Kar, Lucy Ireland, Deanna A Mele, Vasu Sah, Maneesh Singh, Josephine Walton, Maelle Mairesse, Matthew King, Mark Edbrooke, Paul Lyne, Simon T Barry, Stephen Fawell, Frederick W Goldberg, and A Robert MacLeod. Direct targeting of FOXP3 in tregs with AZD8701, a novel antisense oligonucleotide to relieve immunosuppression in cancer. *Journal for ImmunoTherapy of Cancer*, 10(4):e003892, 2022.

Vincent Rivoirard, Judith Rousseau, et al. Bernstein–von Mises theorem for linear functionals of the density. *Annals of Statistics*, 40:1489–1523, 2012.

Donald B Rubin. The Bayesian bootstrap. *Annals of Statistics*, pages 130–134, 1981.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24, 2023.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 2009.

Stephane Shao, Pierre E Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the Hyvärinen score: Computation and consistency. *J. Am. Stat. Assoc.*, pages 1–24, 2018.

W. Shen, S.T. Tokdar, and S. Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100:623–640, 2013.

Jake A Soloff, Adityanand Guntuboyina, and Bodhisattva Sen. Multivariate, heteroscedastic empirical Bayes via nonparametric maximum likelihood. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87:1–32, 2025.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12, 2011.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Peter A. Szabo, Hanna Mendes Levitin, Michelle Miron, Mark E. Snyder, Takashi Senda, Jinzhou Yuan, Yim Ling Cheng, Erin C. Bush, Pranay Dogra, Puspa Thapa, Donna L. Farber, and Peter A. Sims. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications*, 10, 2019.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

Aad W Van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer International Publishing, 2023.

C Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

C Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

Ignacio Vázquez-García, Florian Uhlitz, Nicholas Ceglia, Jamie L.P. Lim, Michelle Wu, Neeman Mohibullah, Juliana Niyazov, Arvin Eric B. Ruiz, Kevin M. Boehm, Viktoria Bojilova, Christopher J. Fong, Tyler Funnell, Diljot Grewal, Eliyahu Havasov, Samantha Leung, Arfath Pasha, Druv M. Patel, Maryam Pourmaleki, Nicole Rusk, Hongyu Shi, Rami Vanguri, Marc J. Williams, Allen W. Zhang, Vance Broach, Dennis S. Chi, Arnaud Da Cruz Paula, Ginger J. Gardner, Sarah H. Kim, Matthew Lennon, Kara Long Roche, Yukio Sonoda, Oliver Zivanovic, Ritika Kundra, Agnes Viale, Fatemeh N. Derakhshan, Luke Geneslaw, Shirin Issa Bhaloo, Ana Maroldi, Rahelly Nunez, Fresia Pareja, Anthe Stylianou, Mahsa Vahdatinia, Yonina Bykov, Rachel N. Grisham, Ying L. Liu, Yulia Lakhman, Ines Nikolovski, Daniel Kelly, Jianjiong Gao, Andrea Schietinger, Travis J. Hollmann, Samuel F. Bakhoum, Robert A. Soslow, Lora H. Ellenson, Nadeem R. Abu-Rustum, Carol Aghajanian, Claire F. Friedman, Andrew McPherson, Britta Weigelt, Dmitriy Zamarin, and Sohrab P. Shah. Ovarian cancer mutational processes drive site-specific immune evasion. *Nature*, 612, 2022.

Christopher D. Walker. Parametrization, prior independence, and the semiparametric Bernstein-von Mises theorem for the partially linear model, 2024.

Chong Wang and David M. Blei. A general method for robust Bayesian modeling. *Bayesian Analysis*, 13, 2018.

Yixin Wang, Alp Kucukelbir, and David M. Blei. Robust probabilistic modeling with Bayesian data reweighting. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, 2017.

Eli N Weinstein and Jeffrey W Miller. Bayesian data selection. *Journal of Machine Learning Research*, 24:1–72, 2023.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:1–5, 2018.

I. N.Kwon Yeo and Richard A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87, 2000.

In Kwon Yeo and Richard A. Johnson. A uniform strong law of large numbers for U - statistics with application to transforming to near symmetry. *Statistics and Probability Letters*, 51, 2001.

Andrew Yiu, Edwin Fong, Chris Holmes, and Judith Rousseau. Semiparametric posterior corrections. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87 (4):1025–1054, 2025.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

Appendix A. Scaling of the Generalized Bayes Factor

In Section 2, we define the generalized Bayes factor using the transformation Ξ. Here, we show that the transformed variable aligns with the typical scaling behavior of the Bayes factor (Hong and Preston, 2005; Dawid, 2011). First note,

$$-\log \mathrm{gBF}_n = \frac{\mathbb{E}\left[\mathrm{D}_n\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right]}{\mathbb{E}\left[\mathrm{D}_n\left(\mathrm{p}_\theta, \mathrm{p}_0\right)\right]}(n+1)^r - 1 + \log\left(\frac{\mathbb{E}\left[\mathrm{D}_n\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right]}{\mathbb{E}\left[\mathrm{D}_n\left(\mathrm{p}_\theta, \mathrm{p}_0\right)\right]}(n+1)^r\right).$$

Suppose the assumptions of Theorem 6 hold. Then the empirical divergence satisfies

$$\mathbb{E}\left[\mathrm{D}_n\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right] = \mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) + r_n,$$

for some rate $r_n \to 0$ such that $r_n(n+1)^r = o(1)$. In practice, for the divergences we study, $r_n = n^{-\tilde{r}}$, and we choose $0 < r < \tilde{r}$. This yields the following scaling behavior: if $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, then $\log \mathrm{gBF}_n = O((\tilde{r} - r)\log n)$; if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, then $-\log \mathrm{gBF}_n = O(n^r)$.

This aligns with the standard Bayes factor, comparing a high-dimensional model to a nested low-dimensional model, where, if p_0 is in the low-dimensional model class, then $\log \mathrm{BF}_n = O(\log n)$; otherwise $-\log \mathrm{BF}_n = O(n)$ (Hong and Preston, 2005; Dawid, 2011).

The scaling behavior of the generalized Bayes factor also reveals the tradeoffs involved in setting the hyperparameter r. Choosing smaller values of r leads to faster convergence in the well-specified case, but slower convergence in the misspecified case.

Appendix B. Details on Synthetic Experiments

In the NPP, we estimate the KL divergence between p_0 and the posterior predictive using Monte Carlo, with 1000 heldout samples from the true distribution. We use $m = n$ to estimate the divergence in the gNPP, for the MMD version and Wasserstein version. We use the IMQ kernel $k(x, y) = (c^2 + \|x - y\|_2^2)^{-0.5}$ with the bandwidth c set to the median distance between datapoints (Gorham and Mackey, 2017).

As an additional comparison, we also examined the behavior of a modified version of the gNPP that uses a point estimate of the optimal model, $\mathbb{I}(\hat{\eta}_n > 0.5)$, in place of the standard mixing weight $\hat{\eta}_n$. We find similar behavior (Figure 10).

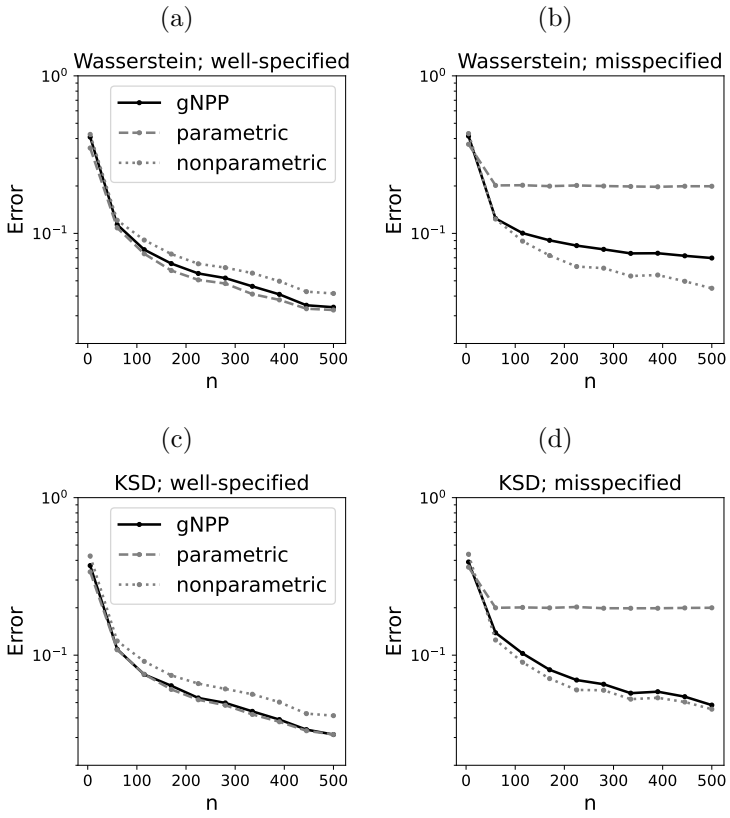


Figure 7: **Point estimation error of the gNPP.** Absolute difference between the posterior mean of the median and the true median, using the Wasserstein (a,b) and KSD (c, d) divergences, in the well-specified (a,c) and mis-specified cases (b,d).

Appendix C. Background

We review key background for our theoretical results.

C.1 Distances and Divergences

This section collect some definitions and results about statistical divergences.

Definition 7 (KL divergence)**.** *The Kullback-Leibler (KL) divergence between two probability distributions* p *and* q *is defined as*

$$D_{\mathrm{KL}}\left(\mathrm{p} \parallel \mathrm{q}\right) = \int \log\left(\frac{\mathrm{p}(t)}{\mathrm{q}(t)}\right) \mathrm{p}(t)dt.$$

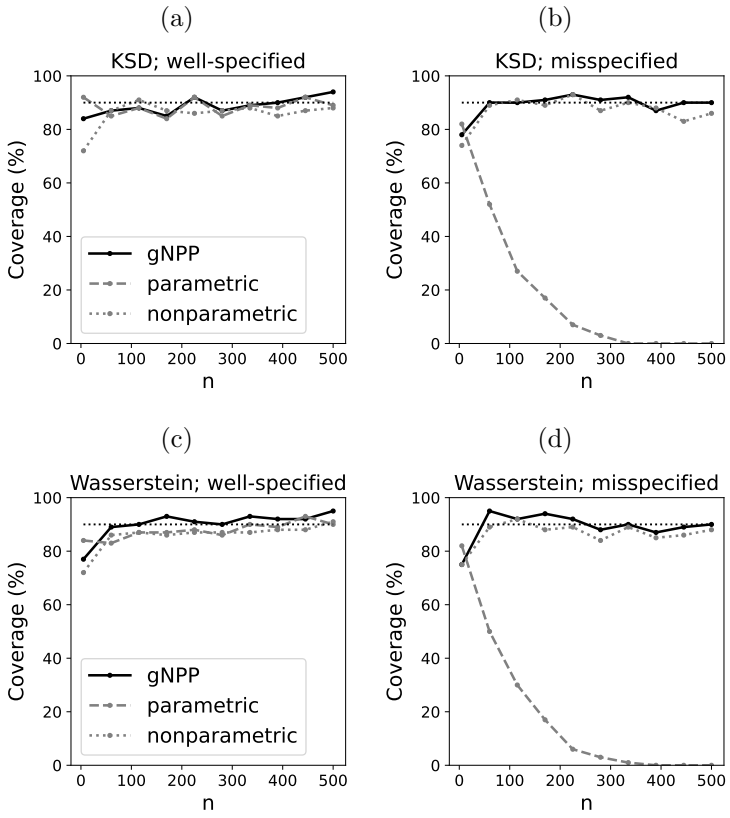


Figure 8: **Calibration of the gNPP.** We calculate how often the credible interval of the gNPP posterior includes the true median.

Definition 8 (Hellinger distance)**.** *The Hellinger distance between two probability distributions* p *and* q *is defined as*

$$d_H(\mathrm{p},\mathrm{q}) = \sqrt{\frac{1}{2}\int \left(\sqrt{\mathrm{p}(x)} - \sqrt{\mathrm{q}(x)}\right)^2 dx}.$$

Definition 9 (Total variation distance)**.** *The total variation distance between two probability distributions* p *and* q *is defined as*

$$d_{TV}(\mathrm{p},\mathrm{q}) = \frac{1}{2}\int |\mathrm{p}(x) - \mathrm{q}(x)| dx.$$

Definition 10 (Bounded Lipschitz distance)**.** *For a real-valued function* $f : \mathbb{R}^d \mapsto \mathbb{R}$*, the bounded Lipschitz norm is defined as*

$$\|f\|_{BL} = \|f\|_\infty + \|f\|_L, \tag{21}$$

where $\|f\|_\infty = \sup_x |f(x)|$ *and* $\|f\|_L = \sup_{x \neq y} \frac{|f(x)-f(y)|}{\|x-y\|_2}$.

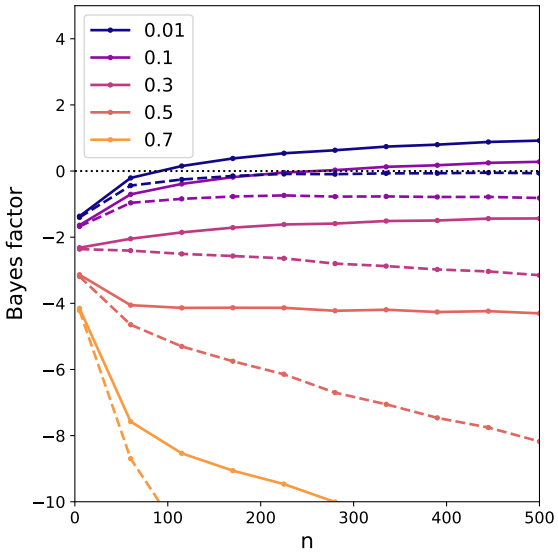


Figure 9: **Dependence of the generalized Bayes factor on the rate r.** We plot the log generalized Bayes factor with the Wasserstein distance, under different choices of rate hyperparameter r (colors), in the well-specified case (solid) and misspecified case (dashed). Values above zero indicate the parametric model is favored.

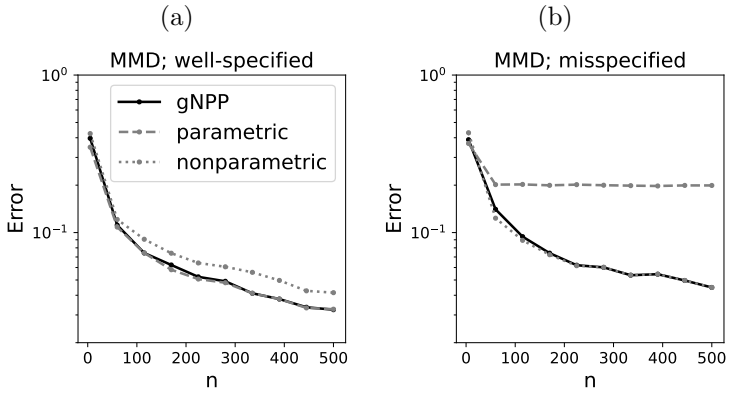


Figure 10: **Point estimation error of the gNPP with a binarized mixing weight.** Same as Figures 2c and 2d, but using a binarized mixing weight in the gNPP.

For two distributions p, q *on* $\mathbb{R}^d$*, the bounded Lipschitz metric is defined as*

$$d_{BL}(\mathrm{p}, \mathrm{q}) = \sup_{\|f\|_{BL} \leq 1} \left| \int f(x)(\mathrm{p}(x) - \mathrm{q}(x))dx \right|. \tag{22}$$

Theorem 11. *(Van der Vaart and Wellner, 2023, Theorem 1.12.4) For a sequence of distributions* p_n *and a distribution* p_0 *on* $\mathbb{R}^d$*, the following are equivalent:*

(a) p_n *converges weakly to* p_0*.*

(b) $d_{BL}(\mathrm{p}_n, \mathrm{p}_0) \to 0$*.*

C.2 Bayesian Nonparametrics

In this section we provide background on the frequentist analysis of Bayesian nonparametric models.

Notation. Let $\mathcal{X} \subseteq \mathbb{R}^d$. We denote the Hellinger distance, KL divergence, and L_2 norm between two densities $\mathrm{p}, \mathrm{q} \in \mathcal{P}(\mathcal{X})$ as $d_H(\mathrm{p}, \mathrm{q})$, $D_{\mathrm{KL}}(\mathrm{p} \parallel \mathrm{q})$, and $\|\mathrm{p} - \mathrm{q}\|_2$, respectively, as reviewed in Section C.1. We also consider a generic metric d on the space of probability densities, such as the Hellinger distance, L_1 distance, or L_2 distance for uniformly bounded densities.

The ϵ-covering numbers and ϵ-packing numbers of a metric space $(\mathcal{P}, \rho)$ denoted by $N(\epsilon, \mathcal{P}, \rho)$ and $D(\epsilon, \mathcal{P}, \rho)$ are defined as the minimal number of balls of radius ϵ needed to cover $\mathcal{P}$, and the maximal number of ϵ-separated points, respectively.

For any $\epsilon > 0$, we define the neighborhood of p_0 as follows:

$$
\begin{aligned}
B_0(\mathrm{p}_0, \epsilon, \mathcal{P}) &= \left\{\mathrm{p} \in \mathcal{P} : d_H(\mathrm{p}_0, \mathrm{p}) < \epsilon^2\right\}, \\
B_k(\mathrm{p}_0, \epsilon, \mathcal{P}) &= \left\{\mathrm{p} \in \mathcal{P} : D_{\mathrm{KL}}(\mathrm{p}_0 \parallel \mathrm{p}) < \epsilon^2, \mathbb{E}_{\mathrm{p}_0}\left[\left|\log(\mathrm{p}_0/\mathrm{p}) - D_{\mathrm{KL}}(\mathrm{p}_0 \parallel \mathrm{p})\right|^k\right] < \epsilon^k\right\}, \quad \text{for } k > 0.
\end{aligned} \tag{23}
$$

The B_0 neighborhood defines an ϵ-open ball around p_0 under the Hellinger distance. The B_k neighborhoods (for $k > 0$) consider the KL divergence and higher-order moments of the log-likelihood ratio. These definitions help us reasons about the posterior contraction rates. We keep the metric d purposefully vague, but a common choice is the Hellinger divergence.

We define the nonparametric prior $\Pi_{\mathrm{np}}(\mathrm{p})$ and $\mathcal{M}_{\mathrm{np}}$ as the support of Π_{np}.

When we say the parametric model is regular, we mean that the model is finite-dimensional and satisfies sufficient conditions for the Bernstein-von Mises theorem. Sufficient conditions can be found in Ch10 of Van der Vaart (2000), Ch 8 of Ghosal and Van der Vaart (2017) or Section 2 of Miller (2021).

Prior mass and entropy assumptions. Let $\epsilon_{n,\mathrm{pm}}, \epsilon_{n,\mathrm{np}}$ be two sequences of positive numbers tending to zero, such that $\epsilon_{n,\mathrm{pm}} < \epsilon_{n,\mathrm{np}}$.

Assumption 5 (Model Entropy)**.** *There exists constant* $C_{\mathrm{pm}}, C_{\mathrm{np}}$ *such that*

$$
\begin{aligned}
\sup_{\epsilon \geq \epsilon_{n,\mathrm{pm}}} \log N\left(\frac{\epsilon}{3}, B_0(\mathrm{p}_0, 2\epsilon, \mathcal{M}_{\mathrm{pm}}), d_H\right) &\leq C_{\mathrm{pm}} n \epsilon_{n,\mathrm{pm}}^2, \quad \textit{and} \\
\sup_{\epsilon \geq \epsilon_{n,\mathrm{np}}} \log N\left(\frac{\epsilon}{3}, B_0(\mathrm{p}_0, 2\epsilon, \mathcal{M}_{n,\mathrm{np}}), d_H\right) &\leq C_{\mathrm{np}} n \epsilon_{n,\mathrm{np}}^2.
\end{aligned} \tag{24}
$$

where $\mathcal{M}_{n,\mathrm{np}} \subseteq \mathcal{M}_{\mathrm{np}}$ *is a submodel that satisfies*

$$
\Pi_{\mathrm{np}}(\mathcal{M}_{\mathrm{np}} \backslash \mathcal{M}_{n,\mathrm{np}}) / \Pi_{\mathrm{pm}}(B_2(\mathrm{p}_0, \epsilon_{n,\mathrm{pm}}, \mathcal{M}_{\mathrm{pm}})) \leq \exp\left(-2n\epsilon_{n,\mathrm{np}}^2\right). \tag{25}
$$

Assumption 6 (Prior Mass). *For the KL neighborhood in Eq.* (23)*, the priors satisfy*

$$\begin{aligned}\Pi_{\mathrm{pm}}\left(B_2\left(\mathrm{p}_0, \epsilon_{n,\mathrm{pm}}, \mathcal{M}_{\mathrm{pm}}\right)\right) &\geq \exp\left(-n\epsilon_{n,\mathrm{pm}}^2\right), \quad \text{if } \mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}, \\ \Pi_{\mathrm{np}}\left(B_2\left(\mathrm{p}_0, \epsilon_{n,\mathrm{np}}, \mathcal{M}_{\mathrm{np}}\right)\right) &\geq \exp\left(-n\epsilon_{n,\mathrm{np}}^2\right), \quad \text{if } \mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}.\end{aligned} \tag{26}$$

Assumptions 5 and 6 establish the standard prior mass and entropy conditions necessary for posterior contraction rates (Ghosal et al., 2000). Assumption 5 defines the parametric and nonparametric rates. The second part of Assumption 6 relaxes the entropy condition to a submodel supported on most of the prior mass, in view of (Ghosal and Van der Vaart, 2017, Remark 10.4). Assumption 6 requires the parametric prior to put sufficient mass around the truth when the parametric model is correctly specified. Similarly, the nonparametric prior is required to put sufficient mass around the truth when the parametric model is incorrectly specified.

It is useful to think of $\epsilon_{n,\mathrm{pm}}, \epsilon_{n,\mathrm{np}}$ as the posterior contraction rates of the parametric and nonparametric models, respectively. When the parametric model is regular, a choice of $\epsilon_{n,\mathrm{pm}}$ is $n^{-1/2}d^{1/2}\log n$ by the Bernstein-von Mises theorem (Van der Vaart, 2000). The nonparametric contraction rate is slower by at least a logarithmic factor. For example, the posterior of a Dirichlet process mixture of normal priors contracts a rate of $n^{-1/2}(\log n)^{(d+1+1/r_0)/2}$ where d is the dimension and r_0 is some measure of smoothness of the true mixing distribution (Shen et al., 2013).

Assumption 7 (Rate Difference). *We assume that for a sufficiently large $M > 0$,*

$$\Pi_{\mathrm{np}}\left(B_0(\mathrm{p}_0, M\epsilon_{n,\mathrm{np}}, \mathcal{M}_{\mathrm{np}})\right) = o\left(\exp\left(-3n\epsilon_{n,\mathrm{pm}}^2\right)\right). \tag{27}$$

Assumption 7 establishes an upper bound for the nonparametric prior mass within a ball of radius $M\epsilon_{n,\mathrm{np}}$, which matches the prior mass lower bound in Assumption 6. As we approach the truth at a rate of $M\epsilon_{n,\mathrm{np}}$, the prior mass decreases at a rate exceeding $\exp\left(-3n\epsilon_{n,\mathrm{pm}}^2\right)$ within this neighborhood.

Appendix D. Proof of Section 4

D.1 Proofs of Section 4.1

This section contains two main results: one negative and one positive. In Proposition 12, we show that for Dirichlet process perturbations, model selection consistency can fail (Example 1). In contrast, Proposition 14 shows that for Dirichlet process normal mixture perturbations, model selection consistency holds under mild conditions.

Proposition 12. *Example 1 fails to satisfy model selection consistency as in Proposition 2.*

Proof [Proof of Proposition 12] Consider the Polya urn construction of the Dirichlet process,

$$x_1 \sim \mathrm{p}_\theta, \quad x_k \sim \frac{1}{1+h}\mathrm{p}_\theta + \frac{h}{1+h}\hat{\mathrm{p}}_{x_{1:k-1}}, \quad \text{for } k > 1, \tag{28}$$

where $\hat{\mathrm{p}}_{x_{1:k-1}}$ is the empirical distribution of $x_{1:k-1}$, and $h = 1/\alpha$ (Blackwell and MacQueen, 1973). For distinct $x_1, \cdots, x_n$, we have

$$\begin{aligned}\eta_n &= \frac{\eta \int \mathrm{p}_\theta(x_{1:n})\, d\Pi_{\mathrm{pm}}(\theta)}{(1-\eta)\int \tilde{\mathrm{p}}(x_{1:n})\, d\Pi_{\mathrm{pert}}\left(\tilde{\mathrm{p}}\right)} \\ &= \frac{\eta}{1-\eta}\frac{\int \mathrm{p}_\theta(x_{1:n})\, d\Pi_{\mathrm{pm}}(\theta)}{\left(\frac{1}{1+h}\right)^{n-1}\int \mathrm{p}_\theta(x_{1:n})\, d\Pi_{\mathrm{pm}}\left(\theta\right)} = \frac{\eta}{1-\eta}\left(\frac{1}{1+h}\right)^{1-n}.\end{aligned} \tag{29}$$

The resulting η_n does not depend on whether the parametric model is correct, and hence does not satisfy model selection consistency. ■

We next establish that for Dirichlet process mixture model perturbations, model selection consistency holds if p_0 is sufficiently smooth (Example 2). Loosely, a distribution is said to be *supersmooth* when it is a Gaussian mixture over a thin-tailed mixing distribution, and *β-smooth* when the density is thin-tailed and has sufficiently regular derivatives.

Definition 13. *A density p_0 on $\mathbb{R}^d$ is said to be*

- *supersmooth if there exists (F_0, Σ_0) such that $\mathrm{p}_0 = F_0 * \mathcal{N}(0, \Sigma_0)$ and $1 - F_0\left([-z,z]^d\right) \lesssim \exp\left(-c_0 z^{a_1}\right)$ for every $z > 0$, with $c_0 > 0$ and $a_1 \geq 2$;*
- *β-smooth if the following holds:*
  1. *The mixed partial derivative $D^k\mathrm{p}_0$ of order up to $k_+ \leq \lfloor\beta\rfloor$ satisfies*

$$\begin{aligned}&\left|(D^k\mathrm{p}_0)(x+y) - D^k\mathrm{p}_0(x)\right| \leq L(x)\exp\left(c_1\|y\|^2\right)\|y\|^{\beta-\lfloor\beta\rfloor}, \quad k_+ = \lfloor\beta\rfloor, \quad x, y \in \mathbb{R}^d, \\ &P_0\left(\frac{L + |D^k\mathrm{p}_0|}{\mathrm{p}_0}\right)^{(2\beta+\epsilon)/\beta} < \infty,\end{aligned} \tag{30}$$

     for some function $L : \mathbb{R}^d \mapsto [0,\infty)$ and $c_1, \epsilon > 0$;
  2. *For every $\|x\| > a$, $\mathrm{p}_0(x) \leq c\exp\left(-b\|x\|^r\right)$, for some $a, b, c, r > 0$.*

Proposition 14. *Assume that the true density p_0 is either supersmooth or β-smooth. Suppose there exist positive constants a_1, b_1, C_1 such that with Π_{pm}-probability at least $1 - \exp(-C_1 z^{a_1})$, the following hold:*

1. *p_θ admits a continuous and strictly positive Lebesgue density on $\mathbb{R}^d$.*
2. *$P_\theta\left([-z,z]^d\right) \geq 1 - b_1\exp(-C_1 z^{a_1})$.*

Let $\lambda_1(\Sigma) \geq \cdots \geq \lambda_d(\Sigma)$ denote the eigenvalues of Σ. Assume the prior G on the scale matrix Σ satisfies the following conditions:

1. *There exist constants $a_2, b_2, C_2 > 0$ such that, for all $s > 0$, $G(\lambda_d(\Sigma) \geq s) \leq b_2\exp(-C_2 s^{a_2})$.*
2. *There exist constants $a_3, b_3 > 0$ such that, for all $s > 0$, $G(\lambda_1(\Sigma) \leq s) \leq b_3 s^{a_3}$.*

3. *There exist constants* $a_4, a_5, b_4, C_3, C_4 > 0$ *such that, for all* $0 < s_1 \leq \cdots \leq s_d$ *and* $0 < t < 1$,

$$G\left(\bigcap_{j=1}^{d} \{s_j < \lambda_j(\Sigma) < s_j(1+t)\}\right) \geq b_4\, s_1^{a_4}\, t^{a_5}\, \exp\left(-C_3 s_d^{C_4/2}\right).$$

Then Assumptions 5 and 6 hold for the Dirichlet process normal mixture perturbation (Eqs. (2) *and* (3)*) and*

1. $\epsilon_{n,\mathrm{np}} = n^{-1/2}(\log n)^{(d-1+1/a_1)/2}$ *if* $\mathrm{p}_0 = F_0 * \mathcal{N}\left(0, \sigma_0^2 I_d\right)$ *is supersmooth with* $\sigma_0^2 > 0$*;*
2. $\epsilon_{n,\mathrm{np}} = n^{-\beta/(2\beta+d^\star)}(\log n)^{t_0}$ *if* p_0 *is* β*-smooth, where* $d^\star = d \wedge 2$ *and* $t_0 = (\beta d^\star + \beta d^\star/r + d^\star + \beta)/(2\beta + d^\star)$.

When the true density p_0 is β-Hölder smooth, Shen et al. (2013) show that Dirichlet process normal mixtures converge to p_0 at the rate $\epsilon_{n,\mathrm{np}} = n^{-\beta/(2\beta+d^\star)}(\log n)^{t_0}$ for some constant $t_0 > 0$. Under the same rate, Proposition 14 explicitly verifies Assumptions 5 and 6. To satisfy Proposition 2, Assumption 7 must also hold; for this it suffices that $\epsilon_{n,\mathrm{np}}$ lower-bound the parametric posterior contraction rate (Ghosal and Van der Vaart, 2017, Theorem 8.35). In particular, the parametric model will in general contract at the rate $\epsilon_{n,\mathrm{pm}}^2 = d \log n/n$, in which case Assumption 7 is implied by the simplified condition $\Pi_{\mathrm{np}}\left(B_0(\mathrm{P}_0, M\epsilon_{n,\mathrm{np}}, \mathcal{M}_{\mathrm{np}})\right) = o\left(n^{-3d}\right)$. This is reasonable to expect since the decay rate on the left-hand side of Assumption 7 is typically of order $\exp\left(-M' n\epsilon_{n,\mathrm{np}}^2\right)$. Unfortunately, Assumption 7 is difficult to verify directly for the Dirichlet process normal mixture, since it depends on the particular choice of priors on the mixing distribution and the scale matrix. For example, two Gaussian mixtures can be extremely close in Hellinger distance even when their mixing measures are far apart (Soloff et al., 2025).

Proof Let Π_{pert} denote the prior on p. We first verify Assumption 6.

Let E be the event

$$E := \left\{\theta :\ 1 - \mathrm{P}_\theta\left([-z,z]^d\right) \leq b_1 \exp(-C_1 z^{a_1}) \text{ for all } z > 0,\ \mathrm{p}_\theta(x) > 0 \text{ and } \mathrm{p}_\theta \text{ is continuous for all } x \in \mathbb{R}^d\right\}. \tag{31}$$

By assumption, $\Pi_{\mathrm{pm}}(E) > 0$. Conditioning on $\theta \in E$, Ghosal and Van der Vaart (2017, Proposition 9.14) implies that there exist constants $A, C > 0$ such that

$$\left(\mathrm{DP}(\mathrm{p}_\theta, \alpha) \times G\right)\left(B_2\left(\mathrm{p}_0, A\epsilon_{n,\mathrm{np}}, \mathcal{M}_{\mathrm{np}}\right)\right) \geq \exp\left(-Cn\epsilon_{n,\mathrm{np}}^2\right).$$

Integrating this lower bound over E yields, for some other constant $C' > 0$,

$$\Pi_{\mathrm{pert}}(B_2(\mathrm{p}_0, A\epsilon_{n,\mathrm{np}}, \mathcal{M}_{\mathrm{np}})) \geq \Pi_{\mathrm{pm}}(E)\, \exp\left(-Cn\epsilon_{n,\mathrm{np}}^2\right) \geq \exp\left(-C' n\epsilon_{n,\mathrm{np}}^2\right). \tag{32}$$

This verifies Assumption 6.

Now we verify Assumption 5 by following the argument in Section 9.4.4 of Ghosal and Van der Vaart (2017). Conditioning on the event E, we construct a sieve $\mathcal{M}_{n,\mathrm{np}}$ consisting of densities of the form $\mathrm{p} = F * \mathcal{N}(0, \Sigma)$, where F is a discrete mixing distribution $F = \sum_{j=1}^{\infty} w_j \delta_{z_j}$ and

$$\sum_{j=N+1}^{\infty} w_j < \epsilon^2, \qquad z_1, \ldots, z_N \in [-a, a]^d, \qquad \sigma^2 \leq \lambda_d(\Sigma) \leq \lambda_1(\Sigma) < \sigma^2(1+\epsilon^2)^n,$$

where

$$N = \frac{Cn\epsilon_{n,\mathrm{np}}^2}{\log(n\epsilon_{n,\mathrm{np}}^2)}, \quad \epsilon^2 = \frac{CN\log n}{n}, \quad a = (n\epsilon^2)^{1/a_1}, \quad \sigma^2 = (n\epsilon^2)^{-1/a_2}.$$

By Ghosal and Van der Vaart (2017, Lemma 9.15) and the definitions of N, σ^2, a, we obtain the entropy bound for the sieve:

$$\log N(\epsilon_{n,\mathrm{np}}/3,\ \mathcal{M}_{n,\mathrm{np}},\ d_H) \lesssim n\epsilon_{n,\mathrm{np}}^2. \tag{33}$$

Moreover, our assumption on Π_{pm} implies $\mathbb{E}_{\mathrm{pm}}\left[1 - \mathrm{P}_\theta\left([-a,a]^d\right)\right] \lesssim \exp(-C_1 a^{a_1})$. By part (ii) of Ghosal and Van der Vaart (2017, Lemma 9.15) and the assumptions on the prior G, we further have

$$\begin{aligned}
\Pi_{\mathrm{pert}}(\mathcal{M}_{\mathrm{np}} \backslash \mathcal{M}_{n,\mathrm{np}}) &\leq \left(\frac{2e\log\epsilon}{N}\right)^N + N\mathbb{E}_{\mathrm{pm}}\left[1 - \mathrm{P}_\theta\left([-a,a]^d\right)\right] + G\left(\lambda_1(\Sigma) \geq \sigma^2(1+\epsilon^2)^n\right) + G\left(\lambda_d(\Sigma) \leq \sigma^2\right) \\
&\lesssim \left(\frac{2e\log\epsilon}{N}\right)^N + N\exp(-C_1 a^{a_1}) + \exp(-C_2\sigma^{-2a_2}) + \sigma^{-2a_3}(1+\epsilon_{n,\mathrm{np}}^2)^{-a_3 n} \\
&\lesssim \exp\left(-C'' n\epsilon_{n,\mathrm{np}}^2\right).
\end{aligned} \tag{34}$$

for some constant $C'' > 0$. Combining Eq. (33) and Eq. (34) verifies Assumption 5. ■

Proof [Proof of Theorem 3] First note that the bounded Lipschitz metric is convex in its first argument. For $\eta \in [0,1]$ and probability measures $\mathrm{P}_0, \mathrm{P}_1, \mathrm{Q}$,

$$\begin{aligned}
&d_{BL}\left(\eta\mathrm{P}_0 + (1-\eta)\mathrm{P}_1, \mathrm{Q}\right) \\
&= \sup_{\|f\|_{BL}\leq 1}\left|\int f\, d(\eta\mathrm{P}_0 + (1-\eta)\mathrm{P}_1 - \mathrm{Q})\right| \\
&\leq \eta \sup_{\|f\|_{BL}\leq 1}\left|\int f\, d(\mathrm{P}_0 - \mathrm{Q})\right| + (1-\eta)\sup_{\|f\|_{BL}\leq 1}\left|\int f\, d(\mathrm{P}_1 - \mathrm{Q})\right| \\
&= \eta d_{BL}\left(\mathrm{P}_0, \mathrm{Q}\right) + (1-\eta) d_{BL}\left(\mathrm{P}_1, \mathrm{Q}\right).
\end{aligned} \tag{35}$$

Assumption 1(b) implies the total variation convergence

$$d_{TV}\left(\Pi_{\mathrm{pm}}\left(d\theta \mid x_{1:n}\right), \mathcal{N}\left(\hat{\theta}_{\mathrm{MLE}}, (nV_{\theta_0})^{-1}\right)\right) = O_{\mathrm{P}_0}(n^{-1/2}). \tag{36}$$

If we treat $\theta \mapsto \sqrt{n}\chi_\theta$ as a pushforward map, then by the triangle inequality,

$$\begin{aligned}
&d_{BL}\left(\Pi_{\mathrm{pm}}\left(\tilde{\psi}_n(\mathrm{p}_\theta) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1}\dot{\chi}_{\theta_0}\right)\right) \\
&\leq \underbrace{d_{BL}\left(\Pi_{\mathrm{pm}}\left(\tilde{\psi}_n(\mathrm{p}_\theta) \mid x_{1:n}\right), (\sqrt{n}\chi)_{\#}\mathcal{N}\left(\hat{\theta}_{\mathrm{MLE}}, (nV_{\theta_0})^{-1}\right)\right)}_{A_n} \\
&+ \underbrace{d_{BL}\left((\sqrt{n}\chi)_{\#}\mathcal{N}\left(\hat{\theta}_{\mathrm{MLE}}, (nV_{\theta_0})^{-1}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1}\dot{\chi}_{\theta_0}\right)\right)}_{B_n}.
\end{aligned} \tag{37}$$

Let $\tilde{\Pi}_n$ be the law of $\sqrt{n}(\theta - \hat{\theta}_{\mathrm{MLE}})$ where $\theta \sim \Pi_{\mathrm{pm}}(\cdot \mid x_{1:n})$. We upper bound A_n with the total variation bound in Assumption 1(b).

$$\begin{aligned} A_n &\leq d_{TV}\left((\sqrt{n}\chi)_{\#}\Pi_{\mathrm{pm}}\left(\cdot \mid x_{1:n}\right), (\sqrt{n}\chi)_{\#}\mathcal{N}\left(\hat{\theta}_{\mathrm{MLE}}, (nV_{\theta_0})^{-1}\right)\right) \\ &\leq d_{TV}\left(\tilde{\Pi}_n, \mathcal{N}\left(0, V_{\theta_0}^{-1}\right)\right) = o_{\mathrm{P}_0}(1). \end{aligned} \tag{38}$$

Now we bound B_n. For $\theta \sim \mathcal{N}\left(\hat{\theta}_{\mathrm{MLE}}, (nV_{\theta_0})^{-1}\right)$, we can reparametrize $\theta = \hat{\theta}_{\mathrm{MLE}} + (nV_{\theta_0})^{-1/2} Z$ for $Z \sim \mathcal{N}(0, I_d)$. By a Taylor expansion, there exists $\tilde{\theta}_n$ in an $O_{\mathrm{P}_0}(n^{-1/2})$-neighbourhood of $\hat{\theta}_{\mathrm{MLE}}$ such that

$$\chi_\theta = \chi(\hat{\theta}_{\mathrm{MLE}}) + \dot{\chi}_{\hat{\theta}_{\mathrm{MLE}}}^{\top} (nV_{\theta_0})^{-1/2} Z + n^{-1}\|V_{\theta_0}\|_2 Z^{\top} \ddot{\chi}_{\tilde{\theta}_n} Z. \tag{39}$$

We note that $\chi(\hat{\theta}_{\mathrm{MLE}}) = 0$. Since $\hat{\theta}_{\mathrm{MLE}} \xrightarrow{\mathrm{P}_0} \theta_0$, we also have $\tilde{\theta}_n \xrightarrow{\mathrm{P}_0} \theta_0$. Thus,

$$\sqrt{n}\chi_\theta = \dot{\chi}_{\hat{\theta}_{\mathrm{MLE}}}^{\top} V_{\theta_0}^{-1/2} Z + O_{\mathrm{P}_0}(n^{-1/2}) \xrightarrow{\mathrm{w}} \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^{\top} V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right). \tag{40}$$

Finally, we use the equivalence between d_{BL} and weak convergence to obtain $B_n = o_{\mathrm{P}_0}(1)$.

For p drawn from the NPP model, we have

$$d_{BL}\left(\Pi\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^{\top} V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right)\right) \leq \eta_n d_{BL}\left(\Pi_{\mathrm{pm}}\left(\tilde{\psi}_n(\mathrm{p}_\theta) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^{\top} V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right)\right) + (1 - \eta_n). \tag{41}$$

When $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, $\eta_n \to 1$ by Proposition 2, and hence

$$d_{BL}\left(\Pi\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^{\top} V_{\theta_0}^{-1} \dot{\chi}_{\theta_0}\right)\right) \xrightarrow{\mathrm{P}_0} 0.$$

When $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, $\eta_n \to 0$ by Proposition 2. Then

$$\begin{aligned} &d_{BL}\left(\Pi\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \\ &\leq \eta_n d_{BL}\left(\Pi_{\mathrm{pm}}\left(\psi(\mathrm{p}_\theta) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) + (1 - \eta_n) d_{BL}\left(\Pi_{\mathrm{pert}}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \\ &\leq \eta_n + (1 - \eta_n) d_{BL}\left(\Pi_{\mathrm{pert}}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \xrightarrow{\mathrm{P}_0} 0. \end{aligned} \tag{42}$$

■

D.2 Proofs of Section 4.2

The main goal of this section is to prove Theorem 6, which shows that gNPP approximations are efficient and robust. This builds directly on Theorem 5 and its proof, which establish consistency of the generalized Bayes factor $\hat{\eta}_n$ for model selection. Our first auxiliary result shows that the expected divergence $\mathbb{E}_{\mathrm{pm}}\left[\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}) \mid x_{1:n}\right]$ vanishes at the rate n^{-1}.

Lemma 15. *Let Assumptions 1 and 3 be satisfied. Then* $\mathbb{E}_{\mathrm{pm}}\left[\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}) \mid x_{1:n}\right] = O_{P_0}(n^{-1})$.

Proof Let $B_{M_n n^{-1/2}}(\theta_0)$ be the neighbourhood defined in Assumption 3(b). Let E_n be the event that $\hat{\theta}_{\mathrm{MLE}} \in B_{M_n n^{-1/2}}(\theta_0)$. By Assumption 1(a), $\mathrm{P}_0(E_n) \to 1$ as $n \to \infty$. For the rest of the proof, we establish the result conditional on E_n.

Let $S_\theta := \nabla_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$ and $H_\theta := \nabla^2_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$. By Taylor expansion, there exists a vector $\tilde{\theta} \in B\left(\theta_0, \|\hat{\theta}_{\mathrm{MLE}} - \theta_0\|_2\right)$ such that

$$\rho(\mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}, \mathrm{p}_{\theta_0}) \leq S_{\theta_0}^\top \left(\hat{\theta}_{\mathrm{MLE}} - \theta_0\right) + \frac{1}{2} \left\| H_{\tilde{\theta}}^{1/2} \left(\hat{\theta}_{\mathrm{MLE}} - \theta_0\right) \right\|_2^2. \tag{43}$$

We note that $\nabla_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})|_{\theta=\theta_0} = 0$ since θ_0 minimises $\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$ on the open set Θ. By the bounded Hessian assumption (Assumption 3(b)), there exists a constant C_{θ_0} such that

$$\rho(\mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}, \mathrm{p}_{\theta_0}) \leq C_{\theta_0} \left\| \hat{\theta}_{\mathrm{MLE}} - \theta_0 \right\|_2^2 = O_{\mathrm{P}_0}(n^{-1}). \tag{44}$$

Similarly, there exists $\tilde{\theta}' \in B\left(\theta_0, \|\hat{\theta}_{\mathrm{MLE}} - \theta_0\|_2\right)$ such that

$$\left\| \nabla_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})|_{\theta=\hat{\theta}_{\mathrm{MLE}}} \right\|_2 = \left\| H_{\tilde{\theta}'} (\hat{\theta}_{\mathrm{MLE}} - \theta_0) \right\|_2 \leq C_{\theta_0} \|\hat{\theta}_{\mathrm{MLE}} - \theta_0\|_2. \tag{45}$$

By Taylor expansion, we have

$$\begin{aligned} \mathbb{E}_{\mathrm{pm}}\left[\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}) \mid x_{1:n}\right] &= \rho(\mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}, \mathrm{p}_{\theta_0}) + \nabla_\theta \rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})|^\top_{\theta=\hat{\theta}_{\mathrm{MLE}}} \mathbb{E}_{\mathrm{pm}}\left[\theta - \hat{\theta}_{\mathrm{MLE}} \mid x_{1:n}\right] \\ &\quad + \mathbb{E}_{\mathrm{pm}}\left[\frac{1}{2}(\theta - \hat{\theta}_{\mathrm{MLE}})^\top \left\{ \int_0^1 (1-t)\, H_{\hat{\theta}_{\mathrm{MLE}} + t(\theta - \hat{\theta}_{\mathrm{MLE}})}\, dt \right\} (\theta - \hat{\theta}_{\mathrm{MLE}}) \mid x_{1:n}\right]. \end{aligned} \tag{46}$$

Applying the uniform bound on $\|H_\theta\|_2$ in the neighbourhood of θ_0 yields

$$\mathbb{E}_{\mathrm{pm}}\left[\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}) \mid x_{1:n}\right] = \rho(\mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}, \mathrm{p}_{\theta_0}) + \dot{\rho}^\top_{\hat{\theta}_{\mathrm{MLE}}} \mathbb{E}_{\mathrm{pm}}\left[\theta - \hat{\theta}_{\mathrm{MLE}} \mid x_{1:n}\right] + O_{\mathrm{P}_0}\left(\mathbb{E}_{\mathrm{pm}}\left[\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 \mid x_{1:n}\right]\right).$$

Given the Bernstein–von Mises theorem for $\Pi_{\mathrm{pm}}(\cdot \mid x_{1:n})$ (Assumption 1), we have

$$\mathbb{E}_{\mathrm{pm}}\left[\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 \mid x_{1:n}\right] = O_{\mathrm{P}_0}\left(\frac{d}{n}\right), \quad \text{and} \quad \left\| \mathbb{E}_{\mathrm{pm}}\left[\theta \mid x_{1:n}\right] - \hat{\theta}_{\mathrm{MLE}} \right\|_2 = O_{\mathrm{P}_0}\left(\sqrt{\frac{d}{n}}\right). \tag{47}$$

Finally, we combine the above inequalities to conclude that

$$\mathbb{E}_{\mathrm{pm}}\left[\rho(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}) \mid x_{1:n}\right] = \rho(\mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}, \mathrm{p}_{\theta_0}) + O_{\mathrm{P}_0}\left(\frac{d}{n}\right) = O_{\mathrm{P}_0}\left(\frac{d}{n}\right). \tag{48}$$

■

Proof [Proof of Theorem 4] By the triangle inequality, we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] - \mathrm{D}\left(\mathrm{p}_0, \mathrm{p}_{\theta_0}\right) \\ &\leq \mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] - \mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right] \\ &\quad + \mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right] - \mathrm{D}\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right). \end{aligned} \tag{49}$$

If $d(\cdot,\cdot)$ is a semimetric, then we can directly bound the second term by Lemma 15:

$$|\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right] - \mathrm{D}\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)| \leq \mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}\left(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}\right) \mid x_{1:n}\right] = O_{\mathrm{P}_0}\left(n^{-1}\right). \tag{50}$$

If $d(\cdot,\cdot)^{1/k}$ is a semimetric, then we apply the binomial expansion $x^k - y^k = (x-y)\sum_{j=0}^{k-1}\binom{k}{j}(x-y)^{k-1-j}y^j$ to obtain

$$\begin{aligned}
&|\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right] - \mathrm{D}\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)| \\
&= \left|\mathbb{E}_{\mathrm{pm}}\left[\left(\rho\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \rho\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)\right)\sum_{j=0}^{k-1}\binom{k}{j}\left(\rho\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \rho\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)\right)^{k-1-j}\rho\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)^j \mid x_{1:n}\right]\right| \\
&\leq C\mathbb{E}_{\mathrm{pm}}\left[\left|\rho\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \rho\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right)\right| \mid x_{1:n}\right] \\
&\leq C\mathbb{E}_{\mathrm{pm}}\left[\rho\left(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}\right) \mid x_{1:n}\right] = O_{\mathrm{P}_0}\left(n^{-1}\right),
\end{aligned} \tag{51}$$

where the inequality uses the uniform boundedness of $\rho(\cdot,\cdot)$.

Let M_n be a sequence such that $M_n \to \infty$ and $M_n = o(n^{1/2})$. To control the fluctuation of $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) \mid x_{1:n}\right]$ around $\mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)$, we decompose it into two terms inside and outside a ball of radius $M_n n^{-1/2}$:

$$\begin{aligned}
&\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] \\
&= \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\left(\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)\right) I_{B^c_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]}_{A_{m,n}} \\
&\quad + \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\left(\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)\right) I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]}_{B_{m,n}}.
\end{aligned} \tag{52}$$

Since we can write $\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) = \mathrm{D}\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right)$, the quantity $|\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)|$ is uniformly bounded. Then applying Assumption 1 yields

$$\begin{aligned}
A_{m,n} &\leq C\Pi_{\mathrm{pm}}\left(\|\theta - \theta_0\|_2 > M_n n^{-1/2} \mid x_{1:n}\right) \\
&\leq C\Pi_{\mathrm{pm}}\left(\sqrt{n}\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2 > M_n - \sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2 \mid x_{1:n}\right) \\
&\leq C\left(M_n - \sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2\right)^{-2}\mathbb{E}_{\mathrm{pm}}\left[n\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 \mid x_{1:n}\right] \\
&= O_{\mathrm{P}_0}\left(\left(M_n + O_{\mathrm{P}_0}(1)\right)^{-2}\right) O_{\mathrm{P}_0}(1) = O_{\mathrm{P}_0}\left(M_n^{-2}\right),
\end{aligned} \tag{53}$$

for some constant $C > 0$. The third line uses Chebyshev's inequality. The last line uses Assumption 1, which implies that the second moment $\mathbb{E}_{\mathrm{pm}}\left[n\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 \mid x_{1:n}\right]$ converges in P_0-probability to $\mathrm{tr}(V_{\theta_0}^{-1})$, hence $\mathbb{E}_{\mathrm{pm}}\left[n\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 \mid x_{1:n}\right] = O_{\mathrm{P}_0}(1)$. Moreover, asymptotic normality of the MLE implies that the sequence $\{\sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2\}_{n\geq 1}$ is bounded in probability, so $\sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2 = O_{\mathrm{P}_0}(1)$.

To see that $A_{m,n} = O_{\mathrm{P}_0}(n^{-1})$, suppose to the contrary that there exists a fixed m such that $n|A_{m,n}|$ is unbounded in P_0-probability. Then there exists a subsequence $\{n_k\}$ for

which $n_k|A_{m,n_k}| \to \infty$ in P_0-probability. Define

$$M_n := \begin{cases} \sqrt{n_k}\,(n_k|A_{m,n_k}|)^{-1/4}, & n = n_k, \\ 1, & \text{otherwise.} \end{cases}$$

Then $M_n = o(\sqrt{n})$ since $M_n/\sqrt{n} = 1/\sqrt{n} \to 0$ for $n \neq n_k$ and $M_{n_k}/\sqrt{n_k} = (n_k|A_{m,n_k}|)^{-1/4} \to 0$ for $n = n_k$. But for $n = n_k$, we also have $M_{n_k}^{-2} = \frac{\sqrt{|A_{m,n_k}|}}{\sqrt{n_k}}$, where

$$\frac{|A_{m,n_k}|}{M_{n_k}^{-2}} = \sqrt{n_k|A_{m,n_k}|} \to \infty,$$

so $|A_{m,n_k}|$ is not $O(M_{n_k}^{-2})$. This contradicts the assumption that $A_{m,n} = O_{\mathrm{P}_0}(M_n^{-2})$ for every $M_n = o(\sqrt{n})$.

To bound $B_{m,n}$, we take the standard approach of analysing its mean and variance with respect to the approximating samples. For the mean, we apply Fubini's theorem:

$$\begin{aligned} \mathbb{E}\left[B_{m,n}\right] &= \mathbb{E}_{\mathrm{pm}}\left[\mathbb{E}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)\right] I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right] \\ &\leq \sup_{\theta \in B_{M_n n^{-1/2}}(\theta_0)} \mathbb{E}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)\right]. \end{aligned} \tag{54}$$

Take $\theta_n \in \operatorname{argmax}_{\theta \in \bar{B}_{M_n n^{-1/2}}(\theta_0)} \mathbb{E}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_\theta, \mathrm{p}_0)\right]$. As $n \to \infty$, the sequence $\theta_n \to \theta_0$ since $\|\theta_n - \theta_0\|_2 \leq M_n n^{-1/2}$. We then apply Assumption 4(a) to conclude that

$$\mathbb{E}\left[\mathrm{D}_{m,n}\left(\mathrm{p}_{\theta_n}, \mathrm{p}_0\right) - \mathrm{D}(\mathrm{p}_{\theta_n}, \mathrm{p}_0)\right] = O(r_{m,n}).$$

Now we control the variance. First, we upper bound the variance using Jensen's inequality:

$$\mathrm{Var}\left(B_{m,n}\right) \leq \mathbb{E}_{\mathrm{pm}}\left[\mathrm{Var}\left(\mathrm{D}_{m,n}\left(\mathrm{p}_\theta, \mathrm{p}_0\right)\right) I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]. \tag{55}$$

By Assumption 4, we refine the upper bound as follows:

$$\mathrm{Var}\left(B_{m,n}\right) \leq r_{m,n}^2 \sup_{\theta \in \bar{B}_{M_n n^{-1/2}}(\theta_0)} \mathcal{V}(\theta). \tag{56}$$

As $n \to \infty$, $\sup_{\theta \in \bar{B}_{M_n n^{-1/2}}(\theta_0)} \mathcal{V}(\theta) \to \mathcal{V}(\theta_0)$ by the continuity of $\mathcal{V}$ at θ_0, and hence $\mathrm{Var}\left(B_{m,n}\right) = O(r_{m,n}^2)$. By Chebyshev's inequality, we conclude that $B_{m,n} = O_{\mathrm{P}_0}(r_{m,n})$.

Putting the bounds together, we have

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_{m,n}(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] - \mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) = O_{\mathrm{P}_0}\left(r_{m,n} \vee n^{-1}\right). \tag{57}$$

■

Proof [Proof of Theorem 5] Recall that we defined the generalized Bayes factor as

$$\mathrm{gBF}_n := \Xi\left(\frac{\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)]}{\mathbb{E}[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}]}(n+1)^{-r}\right)\frac{\eta}{1-\eta}. \tag{58}$$

The assumptions imply $r_n(n+1)^r = o(1)$ and $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0)\right] = O_{\mathbb{P}_0}(1)$, where r_n is the rate at which the expected empirical divergence converges to the true divergence (see Theorem 4 and the subsequent discussion). By the continuous mapping theorem, gBF_n scales like

$$\mathrm{gBF}_n = \Xi\left(\frac{O_{\mathbb{P}_0}(1)}{\mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0)(n+1)^r + o(1)}\right)\frac{\eta}{1-\eta}. \tag{59}$$

If $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, then $\mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) = 0$ and $\mathrm{gBF}_n \xrightarrow{\mathbb{P}_0} \infty$ by the continuous mapping theorem. Analogously, if $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, then $\mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) > 0$ and $\mathrm{gBF}_n \xrightarrow{\mathbb{P}_0} 0$.

We conclude with the desired result by applying the limit of gBF_n to $\hat{\eta}_n$. ■

Proof [Proof of Theorem 6] By Assumptions 1, 3 and 4, we have $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{D}_n(\mathrm{p}_\theta, \mathrm{p}_0) \mid x_{1:n}\right] = \mathrm{D}(\mathrm{p}_{\theta_0}, \mathrm{p}_0) + r_n$.

By the decomposition provided in Eq. (4) and Eq. (6), we have

$$d_{BL}\left(\hat{\Pi}\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1}\dot{\chi}_{\theta_0}\right)\right) \leq \hat{\eta}_n \underbrace{d_{BL}\left(\Pi_{\mathrm{pm}}\left(\tilde{\psi}_n(\mathrm{p}_\theta) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1}\dot{\chi}_{\theta_0}\right)\right)}_{\mathcal{A}_n} + (1-\hat{\eta}_n). \tag{60}$$

In the proof of Theorem 3, we showed that the convergence $\mathcal{A}_n = o_{\mathbb{P}_0}(1)$ under Assumptions 1 and 2. When $\mathrm{p}_0 \in \mathcal{M}_{\mathrm{pm}}$, $\hat{\eta}_n \xrightarrow{\mathbb{P}_0} 1$ by Theorem 5, and $d_{BL}\left(\hat{\Pi}\left(\tilde{\psi}_n(\mathrm{p}) \mid x_{1:n}\right), \mathcal{N}\left(0, \dot{\chi}_{\theta_0}^\top V_{\theta_0}^{-1}\dot{\chi}_{\theta_0}\right)\right) \xrightarrow{\mathbb{P}_0} 0$.

When $\mathrm{p}_0 \notin \mathcal{M}_{\mathrm{pm}}$, $\hat{\eta}_n \xrightarrow{\mathbb{P}_0} 0$ by Theorem 5. Then

$$\begin{aligned}
&d_{BL}\left(\hat{\Pi}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \\
&\leq \hat{\eta}_n d_{BL}\left(\Pi_{\mathrm{pm}}\left(\psi(\mathrm{p}_\theta) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) + (1-\hat{\eta}_n) d_{BL}\left(\hat{\Pi}_{\mathrm{pert}}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \qquad (61)\\
&\leq \hat{\eta}_n + (1-\hat{\eta}_n) d_{BL}\left(\hat{\Pi}_{\mathrm{pert}}\left(\psi(\mathrm{p}) \mid x_{1:n}\right), \delta_{\psi(\mathrm{p}_0)}\right) \xrightarrow{\mathbb{P}_0} 0.
\end{aligned}$$

■

D.3 Empirical Divergences

In this section, we establish the convergence rates for the empirical divergences based on Wasserstein, MMD and KSD. We introduce specific conditions for each divergence.

D.3.1 Wasserstein

We establish the convergence rate of $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{W}_p^p(\hat{\mathrm{p}}_\theta^m, \hat{\mathrm{p}}_0^n) \mid x_{1:n}\right]$ over a sample space $\mathcal{X} \subseteq \mathbb{R}^\kappa$. This implies that Theorem 6 holds for the Wasserstein divergence, by replacing Theorem 4 with Theorem 16 in the proof of Theorem 6.

Theorem 16 (Posterior expected Wasserstein convergence rate)**.** *Suppose $p \geq 1$. Let Assumptions 1, 3 and 8 be satisfied. Then*

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{W}_p^p(\hat{\mathrm{p}}_\theta^m, \hat{\mathrm{p}}_0^n) \mid x_{1:n}\right] - \mathrm{W}_p^p\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right) = O_{\mathbb{P}_0}\left(n^{-2/(\kappa \vee 4)} + m^{-2/(\kappa \vee 4)}\right). \tag{62}$$

In Assumption 8, we assume that both the parametric family and the true distribution have compact support and finite moments up to order $2p$.

Assumption 8. *The distributions* $\{\mathrm{p}_\theta\}_{\theta\in\Theta}$ *and* p_0 *satisfy*

- *(Support) The distributions are supported on a compact subset* $\mathcal{X} \subseteq \mathbb{R}^\kappa$ *and have positive densities in the interior of their respective supports.*
- *(Moments) The distributions have finite moments up to order* $2p$*. Additionally, the mapping* $\theta \mapsto \mathbb{E}_{\mathrm{p}_\theta}\left[\|X\|_2^{2p}\right]$ *is continuous at* θ_0*.*

Proofs

Lemma 17. *Let* $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{p}$ *and* $Y_1, \cdots, Y_m \overset{iid}{\sim} \mathrm{q}$ *be two independent samples with finite moments up to order* $2p$*, specifically* $\mathbb{E}_{\mathrm{p}}\left[\|X\|_2^j\right] + \mathbb{E}_{\mathrm{q}}\left[\|Y\|_2^j\right] =: M_j^j < \infty$ *for* $j \in [2p]$*. For any* $n, m > 0$*,*

$$Var\left(\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right)\right) \leq 2^{p-1} M_{2p}^{2p}\left(\frac{1}{n} + \frac{1}{m}\right)^2, \tag{63}$$

where $\hat{\mathrm{p}}^n$ *and* $\hat{\mathrm{q}}^m$ *are the empirical measures formed by the* X_i*'s and* Y_i*'s.*

Proof [Proof of Lemma 17] By the definition of w_p, we have

$$\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) = \min_{\sum_{i=1}^n w_{ij} = \frac{1}{m}, \sum_{j=1}^m w_{ij} = \frac{1}{n}} \sum_{i=1}^n \sum_{j=1}^m w_{ij}\|X_i - Y_j\|_2^p \leq \frac{\sum_{i=1}^n \sum_{j=1}^m \|X_i - Y_j\|_2^p}{nm}. \tag{64}$$

Define $Z_{ij} := \|X_i - Y_j\|_2^p$. Since each pair (X_i, Y_j) is independent, the collection $\{Z_{ij}\}$ is i.i.d. Under the bounded moment condition, the variance satisfies

$$\mathrm{Var}\left(Z_{ij}\right) \leq 2^p M_{2p}^{2p} < \infty. \tag{65}$$

Aggregating the variance shows

$$\mathrm{Var}\left(\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right)\right) \leq \frac{nm 2^p M_{2p}^{2p}}{n^2 m^2} = \frac{2^p M_{2p}^{2p}}{nm} \leq 2^{p-1} M_{2p}^{2p}\left(\frac{1}{n} + \frac{1}{m}\right)^2. \tag{66}$$

where the last inequality uses $(nm)^{-1/2} \leq \frac{1}{2}(n^{-1} + m^{-1})$. ■

Lemma 18. *If* p, q *are probability densities on a convex compact set* $\mathcal{X} \subseteq \mathbb{R}^\kappa$ *with nonempty interior, then for* $p \geq 1$*, we have*

$$\sup_{\mathrm{p},\mathrm{q}\in\mathcal{P}(\mathcal{X})} \mathbb{E}\left[\left|\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) - \mathrm{w}_p^p\left(\mathrm{p}, \mathrm{q}\right)\right|\right] = O\left(n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}\right). \tag{67}$$

for $X_1, \cdots, X_n \overset{iid}{\sim} \mathrm{p}$ *and* $Y_1, \cdots, Y_m \overset{iid}{\sim} \mathrm{q}$*.*

Proof [Proof of Lemma 18] Recall $\hat{\mathrm{p}}^n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\mathrm{q}}^m := \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}$. We note that the function $x \mapsto \|x\|_2^p$ is convex by an application of Hölder's inequality. Then, following Villani (2003, Remark 1.13), there exist Kantorovich potentials $\phi_n, \psi_m : \mathcal{X} \to \mathbb{R}$ such that

$$\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) = \mathbb{E}_{\hat{\mathrm{p}}^n}\left[\phi_n(X)\right] + \mathbb{E}_{\hat{\mathrm{q}}^m}\left[\psi_m(Y)\right]. \tag{68}$$

Let $\Phi(\mathrm{p},\mathrm{q})$ be the set of pairs $(\phi,\psi) \in L^1(\mathrm{p}) \times L^1(\mathrm{q})$ such that $\phi(x) + \psi(y) \le \|x-y\|_2^p$ for all $x, y \in \mathcal{X}$. Since p and q are compactly supported, we have $(\phi_n, \psi_m) \in \Phi(\mathrm{p},\mathrm{q})$. By Kantorovich duality, we have

$$\begin{aligned}\mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right) &\ge \mathbb{E}_{\mathrm{p}}\left[\phi_n(X)\right] + \mathbb{E}_{\mathrm{q}}\left[\psi_m(Y)\right] \\ &= \mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) + \int \phi_n(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx + \int \psi_m(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx.\end{aligned} \tag{69}$$

and

$$\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) \le \mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right) + \int \phi_n(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx + \int \psi_m(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx. \tag{70}$$

Define $\tilde{\phi}_n(x) = \phi_n(x) - \frac{L}{4}\|x\|_2^2$ and $\tilde{\psi}_m(y) = \psi_m(y) - \frac{L}{4}\|y\|_2^2$. By Manole and Niles-Weed (2024, Lemma 5), $\tilde{\phi}_n$ and $\tilde{\psi}_m$ are concave, L-Lipschitz, and uniformly bounded by L.

Let $\mathcal{F}_{L,U}(K)$ be the set of L-Lipschitz convex functions over a convex set $K \subseteq \mathbb{R}^\kappa$, where $\|f(x)\|_\infty \le U$. By Manole and Niles-Weed (2024, Lemma 5), the functions $\tilde{\phi}_n$ and $\tilde{\psi}_m$ are concave, L-Lipschitz, and uniformly bounded by L, thus $-\tilde{\phi}_n/L$ and $-\tilde{\psi}_m/L$ belong to $\mathcal{F}_{1,1}(\mathcal{X})$.

Since $\mathcal{X}$ is convex and compact, we define

$$\Delta_{n,m} := \sup_{f \in \mathcal{F}_{1,1}(\mathcal{X})} \int f(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx + \sup_{g \in \mathcal{F}_{1,1}(\mathcal{X})} \int g(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx. \tag{71}$$

Using the fact that $-\tilde{\phi}_n/L$ and $-\tilde{\psi}_m/L$ belong to $\mathcal{F}_{1,1}(\mathcal{X})$, we have

$$\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) - \mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right) \le L\Delta_{n,m} + \frac{L}{4}\int \|x\|_2^2\big((\hat{\mathrm{p}}^n(x) - \mathrm{p}(x)) + (\hat{\mathrm{q}}^m(x) - \mathrm{q}(x))\big)\,dx. \tag{72}$$

For the lower bound, there exists a pair of Kantorovich potentials $(\phi_{\mathrm{p}}, \psi_{\mathrm{q}}) \in \Phi(\mathrm{p},\mathrm{q})$ such that $\|\phi_{\mathrm{p}}\|_\infty \vee \|\psi_{\mathrm{q}}\|_\infty \le 1$ and $\mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right) = \mathbb{E}_{\mathrm{p}}\left[\phi_{\mathrm{p}}(X)\right] + \mathbb{E}_{\mathrm{q}}\left[\psi_{\mathrm{q}}(Y)\right]$. Thus,

$$\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) - \mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right) \ge \int \phi_{\mathrm{p}}(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx + \int \psi_{\mathrm{q}}(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx. \tag{73}$$

Combining the above displays, we have

$$\begin{aligned}&\mathbb{E}\left[\left|\mathrm{w}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) - \mathrm{w}_p^p\left(\mathrm{p},\mathrm{q}\right)\right|\right] \\ &\le L\mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}\left[\Delta_{n,m}\right] + \frac{L}{4}\int \|x\|_2^2\big((\hat{\mathrm{p}}^n(x) - \mathrm{p}(x)) + (\hat{\mathrm{q}}^m(x) - \mathrm{q}(x))\big)\,dx \\ &+ \mathbb{E}_{\mathrm{p}}\left[\left|\int \phi_{\mathrm{p}}(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx\right|\right] + \mathbb{E}\left[\left|\int \psi_{\mathrm{q}}(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx\right|\right].\end{aligned} \tag{74}$$

Since $\|\phi_{\mathrm{p}}\|_\infty \vee \|\psi_{\mathrm{q}}\|_\infty \leq 1$, the functional $\mathrm{p} \mapsto \mathbb{E}_{\mathrm{p}}[\phi_{\mathrm{p}}(X)]$ is a bounded linear functional. By Chebyshev's inequality, $\mathbb{P}\left(|\mathbb{E}_{\hat{\mathrm{p}}^n}[\phi_{\mathrm{p}}(X)] - \mathbb{E}_{\mathrm{p}}[\phi_{\mathrm{p}}(X)]| \geq t\right) \leq \frac{\mathrm{Var}_{\mathrm{p}}(\phi_{\mathrm{p}}(X))}{nt^2} < \frac{4}{nt^2}$. The variance bound is uniform over all distributions p on $\mathcal{X}$, thus we have

$$\sup_{\mathrm{p}} \mathbb{E}_{\mathrm{p}}\left[\left|\int \phi_{\mathrm{p}}(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx\right|\right] = O\left(n^{-1/2}\right). \tag{75}$$

Similarly, we have

$$\sup_{\mathrm{q}} \mathbb{E}_{\mathrm{q}}\left[\left|\int \psi_{\mathrm{q}}(x)(\mathrm{q}(x) - \hat{\mathrm{q}}^m(x))\,dx\right|\right] = O\left(m^{-1/2}\right), \tag{76}$$

and

$$\sup_{\mathrm{p,q}} \int \|x\|_2^2\big((\hat{\mathrm{p}}^n(x) - \mathrm{p}(x)) + (\hat{\mathrm{q}}^m(x) - \mathrm{q}(x))\big)\,dx = O\left(n^{-1/2} + m^{-1/2}\right). \tag{77}$$

As a result, we have

$$\sup_{\mathrm{p,q}} \mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}\left[\left|\mathrm{w}_p^p(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m) - \mathrm{w}_p^p(\mathrm{p},\mathrm{q})\right|\right] = L \sup_{\mathrm{p,q}} \mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}[\Delta_{n,m}] + O\left(n^{-1/2} + m^{-1/2}\right). \tag{78}$$

To upper bound $\mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}[\Delta_{n,m}]$, we note that it is a sum of expectation suprema of empirical processes indexed by convex Lipschitz functions. Thus, we can bound the expectation of suprema by applying Dudley's chaining technique in terms of metric entropy of the class $\mathcal{F}_{1,1}(\mathcal{X})$. One such result is Luxburg and Bousquet (2004, Theorem 16), which states

$$\mathbb{E}_{\mathrm{p}}\left[\sup_{f\in\mathcal{F}(\mathcal{X})} \int f(x)(\mathrm{p}(x) - \hat{\mathrm{p}}^n(x))\,dx\right] \leq 2\tau + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\tau/4}^{\sup_{f,f'\in\mathcal{F}} \|f-f'\|_{L^2(\hat{\mathrm{p}}^n)}} \sqrt{\log N(\epsilon, \mathcal{F}, L^2(\hat{\mathrm{p}}^n))}\,d\epsilon. \tag{79}$$

Since the L^2-distance is strictly smaller than the L^∞-distance,

$$N\left(\epsilon, \mathcal{F}_{1,1}(\mathcal{X}), L^2(\hat{\mathrm{p}}^n)\right) \leq N\left(\epsilon, \mathcal{F}_{1,1}(\mathcal{X}), L^\infty\right) \leq N\left(\epsilon, \mathcal{F}_{1,1}([-1,1]^\kappa), L^\infty\right), \tag{80}$$

where the last inequality uses the compactness of $\mathcal{X}$.

By Guntuboyina and Sen (2012, Theorem 1), there exists $\epsilon_0 > 0$ such that when $\epsilon \leq \epsilon_0(B+2)$,

$$\log N\left(\epsilon, \mathcal{F}_{1,1}([-1,1]^\kappa), L^\infty\right) \leq C\left(\frac{B+2}{\epsilon}\right)^{\kappa/2}, \tag{81}$$

where C is a universal constant.

Take $B = 0 \vee (\frac{2}{\epsilon_0} - 2)$. Combining the above three displays, we have

$$\begin{aligned}
\mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}[\Delta_{n,m}] &\leq 2\tau_n + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\tau_n/4}^{2} C\left(\frac{2}{\epsilon_0}\right)^{\kappa/4} \epsilon^{-\kappa/4}\,d\epsilon + 2\tau_m + \frac{4\sqrt{2}}{\sqrt{m}} \int_{\tau_m/4}^{2} C\left(\frac{2}{\epsilon_0}\right)^{\kappa/4} \epsilon^{-\kappa/4}\,d\epsilon \\
&= 2\tau_n + \frac{C'}{\sqrt{n}}\left(\frac{2}{\epsilon_0}\right)^{\kappa/4} \left|2^{1-\kappa/4} - (\tau_n/4)^{1-\kappa/4}\right| + 2\tau_m + \frac{C'}{\sqrt{m}}\left(\frac{2}{\epsilon_0}\right)^{\kappa/4} \left|2^{1-\kappa/4} - (\tau_m/4)^{1-\kappa/4}\right| \\
&\leq C''\epsilon_0^{-\kappa/4}\left(n^{-1/2} + m^{-1/2}\right) + 2\tau_n + 2\tau_m + \epsilon_0^{-\kappa/4}\left(\frac{\tau_n^{1-\kappa/4}}{\sqrt{n}} + \frac{\tau_m^{1-\kappa/4}}{\sqrt{m}}\right).
\end{aligned} \tag{82}$$

Choose $\tau_n = \epsilon_0 n^{-2/\kappa}$. Then for constants $C_{\epsilon_0,\kappa}, C'_{\epsilon_0,\kappa}$,

$$\mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}\left[\Delta_{n,m}\right] \leq C_{\epsilon_0,\kappa}\left(n^{-1/2} + m^{-1/2}\right) + C'_{\epsilon_0,\kappa}\left(n^{-2/\kappa} + m^{-2/\kappa}\right). \tag{83}$$

The constants only depend on ϵ_0 and κ, so the bound is preserved after we take the supremum over p and q. Combining the bound on $\sup_{\mathrm{p,q}} \mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}\left[\Delta_{n,m}\right]$ with Eq. (78), we get

$$\sup_{\mathrm{p,q}} \mathbb{E}_{\mathrm{p}\otimes\mathrm{q}}\left[\left|\mathrm{W}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right) - \mathrm{W}_p^p\left(\mathrm{p},\mathrm{q}\right)\right|\right] = O\left(n^{-1/2} + m^{-1/2}\right) + O\left(n^{-2/\kappa} + m^{-2/\kappa}\right). \tag{84}$$

The final statement comes from the observation that the $O\left(n^{-1/2} + m^{-1/2}\right)$ term dominates when $\kappa < 4$ and the $O\left(n^{-2/\kappa} + m^{-2/\kappa}\right)$ terms dominate when $\kappa \geq 4$. ■

Lemma 19. *Assume that* $\mathrm{diam}(\mathcal{X}) < \infty$. *Then* $\sup_{\mathrm{p,q}\in\mathcal{P}(\mathcal{X})} \mathrm{W}_p^p(\mathrm{p},\mathrm{q}) < \infty$.

Proof [Proof of Lemma 19] Under the definition of W_p, we have $\sup_{\mathrm{p,q}\in\mathbb{P}(\mathcal{X})} \mathrm{W}_p^p(\mathrm{p},\mathrm{q}) \leq \mathrm{diam}(\mathcal{X})^p < \infty$. ■

We also establish the convergence rate for the sample approximation to the W_p distance under the degenerate and non-degenerate cases.

Lemma 20. *Assume that the distributions* p_θ *and* p_0 *are supported on a compact subset* $\mathcal{X} \subseteq \mathbb{R}^\kappa$ *and have positive densities in the interior of their supports. Additionally, assume that the mapping* $\theta \mapsto \mathbb{E}_{\mathrm{p}_\theta}\left[\|X\|_2^{2p}\right]$ *is continuous at* θ_0. *Suppose that* $\lim_{m,n\to\infty} \frac{m}{m+n} = c \in (0,1)$. *Then, Assumption 4 is satisfied for* $\mathrm{D}_{m,n}(\mathrm{p},\mathrm{q}) = \mathrm{W}_p^p(\hat{\mathrm{p}}^m, \hat{\mathrm{q}}^n)$ *and* $\mathrm{D}(\mathrm{p},\mathrm{q}) = \mathrm{W}_p^p(\mathrm{p},\mathrm{q})$ *with* $r_{m,n} = n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}$.

Proof [Proof of Lemma 20] By Lemma 18, Assumption 4(a) is satisfied with $r_{m,n} = n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}$. Take $\mathcal{V}(\theta) = 2^p\mathbb{E}_{\mathrm{p}_\theta}\left[\|X\|_2^p\right] + 2^p\mathbb{E}_{\mathrm{p}_0}\left[\|Y\|_2^p\right]$. From Lemma 17, for any (m,n), we have

$$\mathrm{Var}\left(\mathrm{W}_p^p\left(\hat{\mathrm{p}}^n, \hat{\mathrm{q}}^m\right)\right) \leq \left(\frac{1}{m} + \frac{1}{n}\right)^2 \mathcal{V}(\theta). \tag{85}$$

Since $\mathcal{X}$ is compact, $\mathcal{V}(\theta_0)$ is finite and $\mathcal{V}(\theta)$ is continuous at θ_0 by assumption. Hence, Assumption 4(b) is verified with $r_{m,n} = m^{-1} + n^{-1}$. Taking $r_{m,n} = n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}$ provides an upper bound for both conditions. ■

Proof [Proof of Theorem 16] When $p \geq 1$, the W_p metric is a metric and continuous under the weak topology. Lemma 19 implies the uniform boundedness in Assumption 3. Lemma 20 verifies Assumption 4 at the rate $r_{m,n} = n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}$. This allows us to conclude from Theorem 4 that

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{W}_p^p(\hat{\mathrm{p}}_\theta^m, \hat{\mathrm{p}}_0^n) \mid x_{1:n}\right] - \mathrm{W}_p^p\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right) = O_{\mathbb{P}_0}\left(n^{-2/(\kappa\vee 4)} + m^{-2/(\kappa\vee 4)}\right). \tag{86}$$

■

D.3.2 MMD

We establish the convergence rate for $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right]$, based on the following U-statistic approximation to MMD (Gretton et al., 2012):

$$\mathrm{MMD}_U^2\left(\mathrm{p}^m, \mathrm{q}^n\right) := \frac{\sum_{i=1}^m \sum_{j\neq i} k(x_i, x_j)}{m(m-1)} + \frac{\sum_{i=1}^n \sum_{j\neq i} k(y_i, y_j)}{n(n-1)} - 2\frac{\sum_{i=1}^m \sum_{j=1}^n k(x_i, y_j)}{mn}, \tag{87}$$

with i.i.d samples $x_{1:m}$ drawn from p and $y_{1:n}$ drawn from q. We then use this to prove Theorem 6 holds for the MMD, by replacing Theorem 4 with Theorem 21 in the proof of Theorem 6. Note for large-scale applications, MMD is often preferred over Wasserstein because the computation is quadratic rather than cubic in n, and the choice of kernel offers additional flexibility and control.

Theorem 21 (Posterior expected MMD convergence rate)**.** *Let Assumptions 1, 3, 9 and 10 be satisfied. Then*

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right] - \mathrm{MMD}^2\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right) = O_{\mathbb{P}_0}\left(n^{-1/2} + m^{-1/2}\right). \tag{88}$$

Assumptions 9 and 10 impose standard regularity conditions on the RKHS and the parametric model (Gretton et al., 2012).

Assumption 9. *The RKHS kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a symmetric, positive semi-definite, characteristic kernel such that $\sup_{\theta\in\Theta} \mathbb{E}_{X,X'\sim \mathrm{p}_\theta}\left[k^2(X, X')\right] < \infty$ and $\mathbb{E}_{Y,Y'\sim \mathrm{p}_0}\left[k^2(Y, Y')\right] < \infty$.*

We require a characteristic kernel to make the MMD a valid statistical divergence (Sriperumbudur et al., 2011). The bounded second moments are sometimes referred to as the *Hilbert-Schmidt condition*, which is a sufficient assumption for the eigendecomposition of the kernel operator. Examples of kernels that satisfy Assumption 9 include the Gaussian kernel $k(x, y) = \exp\left(-\|x - y\|_2^2/\gamma^2\right)$ and the Laplace kernel $k(x, y) = \exp\left(-\|x - y\|_2/\gamma\right)$.

Assumption 10. *The mapping $\theta \mapsto \mathbb{E}_{X,X'\sim \mathrm{p}_\theta}\left[k^2(X, X')\right]$ is continuous at θ_0.*

This assumption requires that the parametric model behaves smoothly around θ_0, with respect to the RKHS geometry. In particular, it requires that the mapping $\theta \mapsto \int k^2(x, x')\, \mathrm{p}_\theta(x)\mathrm{p}_\theta(x')\, dx\, dx'$ is continuous at θ_0. This holds, for example, for models with densities p_θ that depend continuously on θ, such as $\mathrm{p}_\theta = \mathcal{N}(\theta, 1)$.

Proofs

Lemma 22 (RKHS-Cauchy-Schwarz)**.** *For any $x, y \in \mathcal{X}$, $k^2(x, y) \leq k(x, x)k(y, y)$.*

Proof Under the RKHS formalism,

$$k^2(x, y) = \langle k(x, \cdot), k(y, \cdot)\rangle^2 \leq \langle k(x, \cdot), k(x, \cdot)\rangle^2 \langle k(y, \cdot), k(y, \cdot)\rangle^2 = k(x, x)k(y, y). \tag{89}$$

■

The following lemma establishes the convergence rate of the empirical MMD, applying the results of Gretton et al. (2012).

Lemma 23. *Let Assumptions 9 and 10 be satisfied. Then Assumption 4 is satisfied for* $\mathrm{D}_{m,n}(\mathrm{p},\mathrm{q}) = \mathrm{MMD}_U^2(\mathrm{p}^m,\mathrm{q}^n)$ *and* $\mathrm{D}(\mathrm{p},\mathrm{q}) = \mathrm{MMD}^2(\mathrm{p},\mathrm{q})$ *with* $r_{m,n} = m^{-1/2} + n^{-1/2}$.

Proof [Proof of Lemma 23] Since $\mathrm{D}_{m,n}(\mathrm{p},\mathrm{q})$ is an unbiased estimate of $\mathrm{D}(\mathrm{p},\mathrm{q})$, Assumption 4(a) is satisfied for any positive sequence $r_{m,n} \to 0$. For Assumption 4(b), we need to bound the variance of the empirical MMD. First, by Eq. (87), we have

$$\begin{aligned}\mathrm{Var}\left(\mathrm{MMD}_U^2(\mathrm{p}^m,\mathrm{q}^n)\right) \le 2\,\mathrm{Var}_{X_i\overset{iid}{\sim}\mathrm{p}}\left(\frac{\sum_{i=1}^m\sum_{j\neq i}k(X_i,X_j)}{m(m-1)}\right) + 2\,\mathrm{Var}_{Y_i\overset{iid}{\sim}\mathrm{q}}\left(\frac{\sum_{i=1}^n\sum_{j\neq i}k(Y_i,Y_j)}{n(n-1)}\right)\\ + 8\,\mathrm{Var}_{X_i\overset{iid}{\sim}\mathrm{p},Y_i\overset{iid}{\sim}\mathrm{q}}\left(\frac{\sum_{i=1}^m\sum_{j=1}^n k(X_i,Y_j)}{mn}\right).\end{aligned} \tag{90}$$

The last term is simply the variance of a sum of i.i.d. variables, thus

$$\mathrm{Var}_{X_i\overset{iid}{\sim}\mathrm{p},Y_i\overset{iid}{\sim}\mathrm{q}}\left(\frac{\sum_{i=1}^m\sum_{j=1}^n k(X_i,Y_j)}{mn}\right) \le \frac{\mathbb{E}_{X\sim\mathrm{p},Y\sim\mathrm{q}}\left[k^2(X,Y)\right]}{mn}. \tag{91}$$

By Lemma 22,

$$\mathrm{Var}_{X_i\overset{iid}{\sim}\mathrm{p},Y_i\overset{iid}{\sim}\mathrm{q}}\left(\frac{\sum_{i=1}^m\sum_{j=1}^n k(X_i,Y_j)}{mn}\right) \le \frac{\mathbb{E}_{X\sim\mathrm{p}}\left[k(X,X)\right]\mathbb{E}_{Y\sim\mathrm{q}}\left[k(Y,Y)\right]}{mn}. \tag{92}$$

The first two terms in Eq. (90) are variances of one-sample U-statistics. By Serfling (2009, Section 5.2.1, Lemma A), we have the following bound:

$$\mathrm{Var}_{X_i\overset{iid}{\sim}\mathrm{p}}\left(\frac{\sum_{i=1}^m\sum_{j\neq i}k(X_i,X_j)}{m(m-1)}\right) \le \frac{2\,\mathrm{Var}_{X,X'\sim\mathrm{p}}\left(k(X,X')\right)}{m} \le \frac{2\mathbb{E}_{X,X'\sim\mathrm{p}}\left[k^2(X,X')\right]}{m}. \tag{93}$$

Analogously, we have

$$\mathrm{Var}_{Y_i\overset{iid}{\sim}\mathrm{q}}\left(\frac{\sum_{i=1}^n\sum_{j\neq i}k(Y_i,Y_j)}{n(n-1)}\right) \le 2\frac{\mathbb{E}_{Y,Y'\sim\mathrm{q}}\left[k^2(Y,Y')\right]}{n}. \tag{94}$$

Putting the bounds together and applying Jensen's inequality with some algebra, we obtain a simple bound without the cross term,

$$\mathrm{Var}\left(\mathrm{MMD}_U^2(\mathrm{p}^m,\mathrm{q}^n)\right) \le \frac{8\mathbb{E}_{X,X'\sim\mathrm{p}}\left[k^2(X,X')\right]}{m} + \frac{8\mathbb{E}_{Y,Y'\sim\mathrm{q}}\left[k^2(Y,Y')\right]}{n}. \tag{95}$$

Define

$$\mathcal{V}(\theta) := 8\left(\mathbb{E}_{X,X'\sim\mathrm{p}_\theta}\left[k^2(X,X')\right] + \mathbb{E}_{Y,Y'\sim\mathrm{p}_0}\left[k^2(Y,Y')\right]\right). \tag{96}$$

The function satisfies $\mathcal{V}(\theta_0) < \infty$ by Assumption 9 and $\mathcal{V}(\theta)$ is continuous at θ_0 by Eq. (98). Then Eq. (95) implies the bound,

$$\mathrm{Var}\left(\mathrm{MMD}_U^2(\mathrm{p}_\theta^m,\mathrm{p}_0^n)\right) \le \left(\frac{1}{m}+\frac{1}{n}\right)\mathcal{V}(\theta) \le r_{m,n}^2\mathcal{V}(\theta). \tag{97}$$

for $r_{m,n} = m^{-1/2} + n^{-1/2}$. ■

Lemma 24. *Let Assumption 9 be satisfied. Then* $\sup_{\theta\in\Theta} \mathrm{MMD}_U^2(\mathrm{p}_\theta^m, \mathrm{p}_0^n) < \infty$.

Proof This follows directly from the following representation for the MMD (Gretton et al., 2012):

$$\mathrm{MMD}^2\left(\mathrm{p},\mathrm{q}\right) = \mathbb{E}_{X,X'\sim\mathrm{p}}\left[k(X,X')\right] - 2\mathbb{E}_{X\sim\mathrm{p},Y\sim\mathrm{q}}\left[k(X,Y)\right] + \mathbb{E}_{Y,Y'\sim\mathrm{q}}\left[k(Y,Y')\right]. \tag{98}$$

By Lemma 22, we have

$$\sup_{\theta\in\Theta} \mathrm{MMD}_U^2(\mathrm{p}_\theta^m, \mathrm{p}_0^n) \leq 2\sup_{\theta\in\Theta} \mathbb{E}_{X,X'\sim\mathrm{p}_\theta}\left[k(X,X')\right] + 2\mathbb{E}_{Y,Y'\sim\mathrm{p}_0}\left[k(Y,Y')\right] < \infty. \tag{99}$$

Since the expectation of $k^2(x,x')$ is uniformly bounded in each case, $\mathrm{MMD}^2\left(\mathcal{H}_k, \mathrm{p}_\theta, \mathrm{p}_0\right)$ is uniformly bounded by Jensen's inequality. ■

Proof [Proof of Theorem 21] The divergence $\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right)$ is slightly different from the MMD between empirical distributions, $\mathrm{MMD}^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right)$. Only the latter applies in the setting of Theorem 4. But the difference between them is negligible. By Lemma 2 of Briol et al. (2019), however, this difference is bounded by a factor of $m^{-1} + n^{-1}$,

$$\sup_{\theta\in\Theta} \left|\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) - \mathrm{MMD}\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right)\right| \leq 2(m^{-1}+n^{-1}) \sup_{x,x'\in\mathcal{X}} k(x,x') = O(m^{-1}+n^{-1}). \tag{100}$$

Using this fact, we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right] - \mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right] \\ &\leq \sup_{\theta\in\Theta} \left|\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) - \mathrm{MMD}^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right)\right| = O(m^{-1}+n^{-1}). \end{aligned} \tag{101}$$

MMD is a pseudometric and continuous under the weak topology. Lemma 24 implies the uniform boundedness in Assumption 3. Lemma 20 verifies Assumption 4 with a rate of $r_{m,n} = n^{-1/2} + m^{-1/2}$. This allows us to conclude that

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{MMD}_U^2\left(\mathrm{p}_\theta^m, \mathrm{p}_0^n\right) \mid x_{1:n}\right] - \mathrm{MMD}^2\left(\mathrm{p}_{\theta_0}, \mathrm{p}_0\right) = O_{\mathbb{P}_0}\left(n^{-1/2} + m^{-1/2}\right). \tag{102}$$

Combining the two displays above yields the desired convergence rate. ■

Remark 25. *Consider the model in Section 3 with the MMD induced by an inverse multiquadric (IMQ) kernel. The parametric model is the normal mean model* $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$. *In this case, we can use Theorem 21 after verifying Assumptions 1, 3, 9 and 10. The MMD is uniformly bounded by a fixed constant, so Assumption 3(a) is satisfied. Taking the reference measure* $U = \mathcal{N}(0,1)$, *the transport map from* U *to* $\mathbb{P}_\theta$ *is* $T_\theta(u) = u + \theta$, *which is linear in* θ *and therefore has uniformly bounded first and second derivatives. Together with the discussion right after Assumption 3, this verifies Assumption 3(b). Assumption 1 holds because the normal mean model is asymptotically normal in the usual sense. For Assumption 4, this is verified in Lemma 23: the IMQ kernel is bounded, symmetric, positive semidefinite, and characteristic, and the mapping* $\theta \mapsto \int k^2(x,x')\,\mathrm{p}_\theta(x)\mathrm{p}_\theta(x')\,dx\,dx'$ *is continuous in* θ. *By Theorem 21, the MMD achieves the rate* $O_{\mathbb{P}_0}(n^{-1/2}+m^{-1/2})$; *hence, if* $m = n$, *any* $r < 1/2$ *is admissible and the generalized Bayes factor based on the MMD achieves consistency.*

D.3.3 KSD

We establish the convergence rate for $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right]$, where KSD_U^2 is a one-sample U-statistic (Liu et al., 2016):

$$\mathrm{KSD}_U^2\left(\mathrm{p}^n, \mathrm{q}\right) := \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} u_{\mathrm{q},k}(x_i, x_j), \tag{103}$$

where $x_{1:n} \overset{iid}{\sim} \mathrm{p}$ and $u_{\mathrm{q},k}(x, x') := \nabla \log \mathrm{q}(x) k(x, x') \nabla \log \mathrm{q}(x') + 2 \nabla \log \mathrm{q}(x)^\top \nabla_{x'} k(x, x') + \mathrm{tr}(\nabla_{x,x'} k(x, x'))$. This implies that Theorem 6 holds for the KSD, by replacing Theorem 4 with Theorem 26 in the proof of Theorem 6.

Theorem 26 (Posterior expected KSD convergence rate)**.** *Let Assumptions 1 and 11 to 13 be satisfied. As $n \to \infty$, $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right]$ converges in $[\mathbb{P}_0^\infty]$-probability to $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0})$ at the rate of*

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right] = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0}) + O_{\mathbb{P}_0}(n^{-1/2}). \tag{104}$$

Assumptions 11 to 13 involve regularity conditions on the kernel and the parametric model. Recall that $\Delta_{\mathrm{q,p}}(x) := \nabla_x \log \mathrm{p}(x) - \nabla_x \log \mathrm{q}(x)$

Assumption 11. *The kernel $k(x, x')$ is symmetric, integrally positive definite, uniformly bounded, and belongs to the Stein class of all continuous densities p_0 and p_θ. Additionally, $\sup_{\theta \in \Theta} \mathbb{E}_{X \sim \mathrm{p}_0}\left[\|\Delta_{\mathrm{p}_0, \mathrm{p}_\theta}(X)\|_2^2\right] < \infty$.*

The first part of Assumption 11 is satisfied when the tail of p_0 decays exponentially and k is positive definite, characteristic and uniformly bounded. The Stein class requirement ensures that KSD operates as a valid statistical divergence (Liu et al., 2016). For example, the squared-exponential kernel $k(x, y) = \exp\left(-\|x - y\|_2^2 / \gamma^2\right)$ belongs to the Stein class for smooth densities on $\mathbb{R}^d$.

The second part of Assumption 11 requires that the relative Fisher information between p_0 and p_θ be uniformly bounded over $\theta \in \Theta$. For example, if $\mathrm{p}_0 = \mathcal{N}(\theta_0, 1)$ and $\mathrm{p}_\theta = \mathcal{N}(\theta, 1)$, then the relative Fisher information scales as $O(\|\theta - \theta_0\|_2^2)$ in θ. In this case, the second part of Assumption 11 is satisfied when Θ is compact.

We also need assumptions on the regularity of the KSD.

Assumption 12. *There exists an $M_n n^{-1/2}$ neighborhood of θ_0 for some $M_n \to \infty$ such that*

(a) the mapping $\theta \mapsto \mathbb{E}_{X, X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta, k}(X, X')\right]$ is twice differentiable with bounded Hessian (in the L_2 sense).

(b) the mapping $\theta \mapsto Var_{X, X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta, k}(X, X')\right]$ is bounded and continuous at θ_0.

When p_0 and p_θ are smooth densities, Assumption 12(a) is equivalent to stating that $\theta \mapsto \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$ is twice differentiable with a bounded Hessian in a shrinking neighborhood of θ_0. Assumption 12(b) is satisfied when the parametric model is regular at θ_0.

To ensure a notion of boundedness for the empirical KSD, we require an additional assumption:

Assumption 13. *We assume that* $\sup_{x,x' \in \mathbb{R}} k(x,x') < \infty$ *and* $\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \overset{a.s.}{\to} \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$ *uniformly in* θ.

The first part of Assumption 13 requires k to be uniformly bounded, which holds for e.g. Gaussian or Laplace kernels. The second part of Assumption 13 requires more effort to verify; it essentially involves establishing a uniform law of large numbers (ULLN) for $\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta)$ across all $\theta \in \Theta$, as done in Barp et al. (2019).

Sufficient conditions for Assumption 13 includes (1) the domination of $u_{\mathrm{p}_\theta,k}(x,x')$ by an integrable and symmetric kernel $g(x,x')$, and (2) the existence of a sequence of sets where the mappings $\theta \mapsto \mathbb{E}_{X' \sim \mathrm{p}_0}[u_{\mathrm{p}_\theta,k}(x,X')]$ and $\theta \mapsto u_{\mathrm{p}_\theta,k}(x,x')$ are equicontinuous for all $x \in \mathcal{X}$ and all $(x,x') \in \mathcal{X} \times \mathcal{X}$, respectively (Yeo and Johnson, 2001). Alternatively, it suffices that the function class $\{u_{\mathbb{P}_\theta,k}\}_{\theta \in \Theta}$ is $[\mathbb{P}_0 \times \mathbb{P}_0]$-Glivenko-Cantelli.

Proofs

Lemma 27. *Let Assumptions 1, 11 and 12 be satisfied. Then* $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) \mid x_{1:n}\right] = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0}) + O_{\mathbb{P}_0}(n^{-1})$.

Proof [Proof of Lemma 27] Let $B_{M_n n^{-1/2}}(\theta_0)$ be the neighborhood defined in Assumption 12. Let E_n be the event that $\hat{\theta}_{\mathrm{MLE}} \in B_{M_n n^{-1/2}}(\theta_0)$. By Assumption 1(a), $\mathbb{P}_0(E_n) \to 1$ as $n \to \infty$. For the rest of the proof, we condition on E_n.

Let $S_\theta := \nabla_\theta \mathrm{KSD}^2(\mathrm{p}_\theta, \mathrm{p}_{\theta_0}), H_\theta := \nabla_\theta^2 \mathrm{KSD}^2(\mathrm{p}_\theta, \mathrm{p}_{\theta_0})$. By Assumption 12, there exists $L > 0$ such that

$$\sup_{\theta \in B_{M_n n^{-1/2}}(\theta_0)} \|H_\theta\|_2 \leq L. \tag{105}$$

By Taylor expanding $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$ around the MLE, we have

$$\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}) + S_{\hat{\theta}_{\mathrm{MLE}}}^T(\theta - \hat{\theta}_{\mathrm{MLE}}) + (\theta - \hat{\theta}_{\mathrm{MLE}})^T H_{\hat{\theta}_{\mathrm{MLE}}}(\theta - \hat{\theta}_{\mathrm{MLE}}) + o(\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2). \tag{106}$$

By the bonded Hessian condition, we have a sandwich inequality,

$$\left|\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}})\right| \leq L\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2 + o(\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2^2). \tag{107}$$

Applying Assumption 1(b) to the posterior expectation of the above display, we have

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) \mid x_{1:n}\right] = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}) + O_{\mathbb{P}_0}(n^{-1}). \tag{108}$$

For $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}})$, a Taylor expansion around θ_0 yields,

$$\begin{aligned} \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\hat{\theta}_{\mathrm{MLE}}}) &= \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0}) + S_{\hat{\theta}}^T(\hat{\theta}_{\mathrm{MLE}} - \theta_0) \\ &= \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0}) + LO_{\mathbb{P}_0}(\|\hat{\theta}_{\mathrm{MLE}} - \theta_0\|_2^2) = \mathrm{KSD}^2(\mathrm{p}_0 \mathrm{p}_{\theta_0}) + O_{\mathbb{P}_0}(n^{-1}). \end{aligned} \tag{109}$$

Combining the two displays above yields the desired result. ■

Lemma 28. *Let Assumption 11 be satisfied. Assume that* $\sup_{x,x' \in \mathcal{X}} k(x,x') < \infty$. *Then* $\sup_{\theta \in \Theta} \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) < \infty$.

Proof [Proof of Lemma 28] From its definition, we can upper bound the KSD by Cauchy-Schwarz,

$$\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) = \mathbb{E}_{X,X' \sim \mathrm{p}_0}\left[\Delta_{\mathrm{p}_0,\mathrm{p}_\theta}(X)^T k(X,X') \Delta_{\mathrm{p}_0,\mathrm{p}_\theta}(X')\right] \leq \sup_{x,x' \in \mathcal{X}} \left\|k(x,x')\right\|_2 \mathbb{E}_{X \sim \mathrm{p}_0}\left[\|\Delta_{\mathrm{p}_0,\mathrm{p}_\theta}(X)\|_2^2\right]. \tag{110}$$

Since $k(a,b) < C$ for all $a, b \in \mathbb{R}$, for $x, x' \in \mathcal{X}$, $\|k(x,x')\|_2 \leq \mathrm{tr}(k(x,x')) \leq \kappa C$ which uses the assumption that $k(x,x')$ is positive definite. Since $\mathbb{E}_{X \sim \mathrm{p}_0}\left[\|\Delta_{\mathrm{p}_0,\mathrm{p}_\theta}(X)\|_2^2\right]$ is uniformly bounded over Θ by Assumption 11, $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$ is uniformly bounded. ■

Lemma 29. *Let Assumptions 11 and 12 be satisfied. Then Assumption 4 is satisfied for* $\mathrm{D}_n(\mathrm{p}, \mathrm{q}) = \mathrm{KSD}_U^2(\mathrm{p}^n, \mathrm{q})$ *and* $\mathrm{D}(\mathrm{p}, \mathrm{q}) = \mathrm{KSD}^2(\mathrm{p}, \mathrm{q})$ *with* $r_n = n^{-1/2}$.

Proof [Proof of Lemma 29] (Liu et al., 2016, Theorem 3.6) shows that under Assumption 11, $\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)$ is a valid U-statistic for $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$,

$$\mathbb{E}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)\right] = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta). \tag{111}$$

Thus Assumption 4(a) holds true for any r_n.

To verify Assumption 4(b), we apply the variance bound for one-sample U-statistics (Serfling, 2009, Section 5.2.1, Lemma A) to obtain:

$$\mathrm{Var}\left(\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)\right) \leq \frac{2\,\mathrm{Var}_{X,X' \sim \mathrm{p}_0}\left(u_{\mathrm{p}_\theta,k}(X,X')\right)}{n} \leq \frac{2\mathbb{E}_{X,X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta,k}^2(X,X')\right]}{n}. \tag{112}$$

Set $\mathcal{V}(\theta) = \mathbb{E}_{X,X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta,k}^2(X,X')\right]$ and $r_n = n^{-1/2}$ provides the desired result. ■

Proof [Proof of Theorem 26] By the triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right] - \mathrm{KSD}^2\left(\mathrm{p}_0, \mathrm{p}_{\theta_0}\right) \\
&\leq \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) \mid x_{1:n}\right]}_{A_n} + \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}^2\left(\mathrm{p}_0, \mathrm{p}_\theta\right) \mid x_{1:n}\right] - \mathrm{KSD}^2\left(\mathrm{p}_0, \mathrm{p}_{\theta_0}\right)}_{R_n}.
\end{aligned} \tag{113}$$

Lemma 27 proved that $R_n = O_{\mathbb{P}_0}(n^{-1})$. To bound A_n, we bound its expectation and variance with respect to the randomness in $\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta)$

By Fubini's theorem, we exchange the integrals and apply the unbiasedness of U-statistics:

$$\mathbb{E}\left[A_n\right] = \mathbb{E}_{\mathrm{pm}}\left[\mathbb{E}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)\right] \mid x_{1:n}\right] = 0. \tag{114}$$

For the variance, we have

$$\mathrm{Var}\left[A_n\right] = \mathbb{E}_{\mathrm{pm}}\left[\mathrm{Var}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)\right] \mid x_{1:n}\right] \leq 2n^{-1}\mathbb{E}_{\mathrm{pm}}\left[\mathrm{Var}_{X,X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta,k}(X,X')\right] \mid x_{1:n}\right]. \tag{115}$$

By Assumption 12, the posterior expectation $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{Var}_{X,X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta,k}(X,X')\right] \mid x_{1:n}\right]$ is bounded in probability. Thus, by Chebyshev's bound, we conclude that $A_n = O_{\mathbb{P}_0}(n^{-1/2})$.

Let M_n be a sequence such that $M_n \to \infty$ and $M_n = o(n^{1/2})$. To control the fluctuation of $\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) \mid x_{1:n}\right]$ around $\mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)$, we decompose it into two terms inside and outside of a ball of radius $M_n n^{-1/2}$:

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta) \mid x_{1:n}\right] = \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\left(\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)\right) I_{B^c_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]}_{A_n}$$
$$+ \underbrace{\mathbb{E}_{\mathrm{pm}}\left[\left(\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)\right) I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]}_{B_n}. \tag{116}$$

To bound A_n, we use the ULLN for $\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)$. By Holder's inequality, we have

$$\begin{aligned} A_n &\leq \sup_{\theta \in \Theta} \left|\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)\right| \Pi_{\mathrm{pm}}\left(\|\theta - \theta_0\|_2 > M_n n^{-1/2} \mid x_{1:n}\right) \\ &= o_{\mathbb{P}_0}(1) \Pi_{\mathrm{pm}}\left(\|\theta - \theta_0\|_2 > M_n n^{-1/2} \mid x_{1:n}\right). \end{aligned} \tag{117}$$

Then we bound the posterior probability using Assumption 1.

$$\begin{aligned} &\Pi_{\mathrm{pm}}\left(\|\theta - \theta_0\|_2 > M_n n^{-1/2} \mid x_{1:n}\right) \\ &\leq \Pi_{\mathrm{pm}}\left(\sqrt{n}\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2 > M_n - \sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2 \mid x_{1:n}\right) \\ &\leq \inf_{\eta} \exp\left(-\eta(M_n - \sqrt{n}\|\theta_0 - \hat{\theta}_{\mathrm{MLE}}\|_2)\right) \mathbb{E}_{\mathrm{pm}}\left[\exp\left(\eta\sqrt{n}\|\theta - \hat{\theta}_{\mathrm{MLE}}\|_2\right) \mid x_{1:n}\right] \\ &= O_{\mathbb{P}_0}\left(\exp\left(-C' M_n\right)\right), \end{aligned} \tag{118}$$

for some constants C and C'. The third line applies Hoeffding's inequality. The fourth line uses the fact that weak convergence implies the convergence of characteristic functions (to a sub-Gaussian limit). Thus, if we take $M_n = n^{1/3}$, we get $A_n = O_{\mathbb{P}_0}\left(n^{-1}\right)$.

To bound B_n, we analyze its mean and variance with respect to the randomness in $\mathrm{KSD}_U^2\left(\mathrm{p}^n, \mathrm{q}\right)$. To control the mean, we apply Fubini's theorem

$$\mathbb{E}\left[B_n\right] = \mathbb{E}_{\mathrm{pm}}\left[\mathbb{E}\left[\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right) - \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_\theta)\right] I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right] = 0. \tag{119}$$

Now we control the variance. Some calculation shows

$$\mathrm{Var}\left(B_n\right) = \mathbb{E}_{\mathrm{pm}}\left[\mathrm{Var}\left(\mathrm{KSD}_U^2\left(\mathrm{p}_0^n, \mathrm{p}_\theta\right)\right) I_{B_{M_n n^{-1/2}}(\theta_0)} \mid x_{1:n}\right]. \tag{120}$$

Let $\mathcal{V}(\theta) = \mathrm{Var}_{X, X' \sim \mathrm{p}_0}\left[u_{\mathrm{p}_\theta, k}(X, X')\right]$. By Eq. (112), we refine the upper bound as follows

$$\mathrm{Var}\left(B_n\right) \leq 2n^{-1} \sup_{\theta \in \bar{B}_{M_n n^{-1/2}}(\theta_0)} \mathcal{V}(\theta). \tag{121}$$

As $n \to \infty$, $\sup_{\theta \in \bar{B}_{M_n n^{-1/2}}(\theta_0)} \mathcal{V}(\theta) \to \mathcal{V}(\theta_0)$ by the continuity of $\mathcal{V}$ at θ_0, thus $B_n = O(n^{-1/2})$ by Chebyshev's inequality.

Combining the bounds, we obtain

$$\mathbb{E}_{\mathrm{pm}}\left[\mathrm{KSD}_U^2(\mathrm{p}_0^n, \mathrm{p}_\theta) \mid x_{1:n}\right] = \mathrm{KSD}^2(\mathrm{p}_0, \mathrm{p}_{\theta_0}) + O_{\mathbb{P}_0}(n^{-1/2}). \tag{122}$$

■

Appendix E. Details on Empirical Studies

E.1 gNPP

Semiparametric model. The semiparametric model in the gNPP follows the model proposed by Hahn et al. (2020) for causal inference, with an additional transformation step to account for non-normality. The model is,

$$y_i = T_\lambda^{-1}(\mu(w_i, \hat{a}(w_i)) + \tau a_i + \epsilon_i), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{123}$$

where:

- The function $T_\lambda(\cdot)$ is the Yeo-Johnson transformation (Yeo and Johnson, 2000). It corrects for non-normality in the outcome expression level.
- The function $\mu(\cdot, \cdot)$ is a sum of piecewise constant regression trees. We place a BART prior on μ, following previous work in Bayesian causal inference (Hill, 2011).
- The function $\hat{a}(w_i)$ is a propensity model, obtained by training a five-layer neural network to predict a_i from w_i under a mean-squared loss. Hahn et al. (2020) show that including $\hat{a}(\cdot)$ in Bayesian causal inference reduces estimator bias.
- The coefficient τ determines the effect of the treatment. We place an improper flat prior on τ. For simplicity we assume the (transformed) outcome depends linearly on a_i even though it may depend nonlinearly on w_i.
- We place a half-Normal prior on the variance σ.

The Yeo-Johnson transformation, T_λ, is a monotonic function used to reduce skewness and approximate normality (Yeo and Johnson, 2000). The parameter λ is fit via maximum likelihood. The transformation is defined as:

$$T_\lambda(y) = \begin{cases} \frac{((y+1)^\lambda - 1)}{\lambda}, & \text{if } y \geq 0, \lambda \neq 0, \\ \log(y+1), & \text{if } y \geq 0, \lambda = 0, \\ \frac{-((-y+1)^{2-\lambda} - 1)}{2-\lambda}, & \text{if } y < 0, \lambda \neq 2, \\ -\log(-y+1), & \text{if } y < 0, \lambda = 2. \end{cases}$$

In the model, we assign a BART prior to μ. The BART function is represented as a sum of piecewise constant binary regression trees. Each tree T_l partitions the covariate space $\mathcal{A} \times \mathcal{X}$, with each partition element A_b assigned a parameter m_{lb}. The function $g_l(x)$ takes the value m_{lb} if $x \in A_b$ and 0 otherwise. The overall function is then $\mu(x) = \sum_{l=1}^{L} g_l(x)$.

We use 50 trees, each constrained by a prior that favors small trees and leaf parameters near zero, making them "weak learners." The prior specification follows that of Chipman et al. (2010), where the probability of a node splitting at depth h is $\eta(1+h)^{-\beta}$ with $\eta \in (0, 1)$ and $\beta \in [0, \infty)$. The splitting variable and cut-point are chosen uniformly at random. Large trees have low prior probability, with typical values $\eta = 0.95$ and $\beta = 2$. Leaf parameters m_{lb} follow independent priors $\mathcal{N}(0, \sigma_m^2)$, where $\sigma_m = \sigma_0/\sqrt{L}$. The induced marginal prior for $\mu(a, w)$ is centered at zero, with 95% of the prior mass within $\pm 2\sigma_0$.

We sample from the BART posterior via MCMC, using the PyMC-BART package (Quiroga et al., 2023). In general we found the chains were well-mixed, after a burn-in period of 1000 steps (Figure 12). Our reported results pool samples from four separately initialized chains.

To account for posterior uncertainty in $w_{1:n}$, we introduce a Bayesian bootstrap model for the distribution of $w_{1:n}$ (Rubin, 1981). The posterior of ATE(p) is obtained as the pushforward distribution under the product posterior of the conditional distribution p and the covariate distribution p_w:

$$\mathrm{p} \sim \Pi_{\mathrm{BART}}\left(\mathrm{p} \mid x_{1:n}\right), \quad \mathrm{p}_w \sim \mathrm{BB}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i}\right), \tag{124}$$

where p implicitly contains the parameters for the conditional distribution, and p_w is the distribution of w. We sample from the posterior of the ATE by simulating from the posterior predictive of Eq. (123).

Let $\Pi_{\mathrm{BART}}(\mathrm{ATE}\left(\mathrm{p}\right) \mid x_{1:n})$ be the posterior distribution of the ATE under the BART model. The gNPP posterior is then given by

$$\hat{\Pi}\left(\mathrm{ATE}\left(\mathrm{p}\right) \mid x_{1:n}\right) = \Pi_{\mathrm{pm}}\left(\mathrm{ATE}\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right)\hat{\eta}_n + \Pi_{\mathrm{BART}}(\mathrm{ATE}\left(\mathrm{p}\right) \mid x_{1:n})(1-\hat{\eta}_n),$$

where $\hat{\eta}_n$ is the generalized mixing weight based on the MMD.

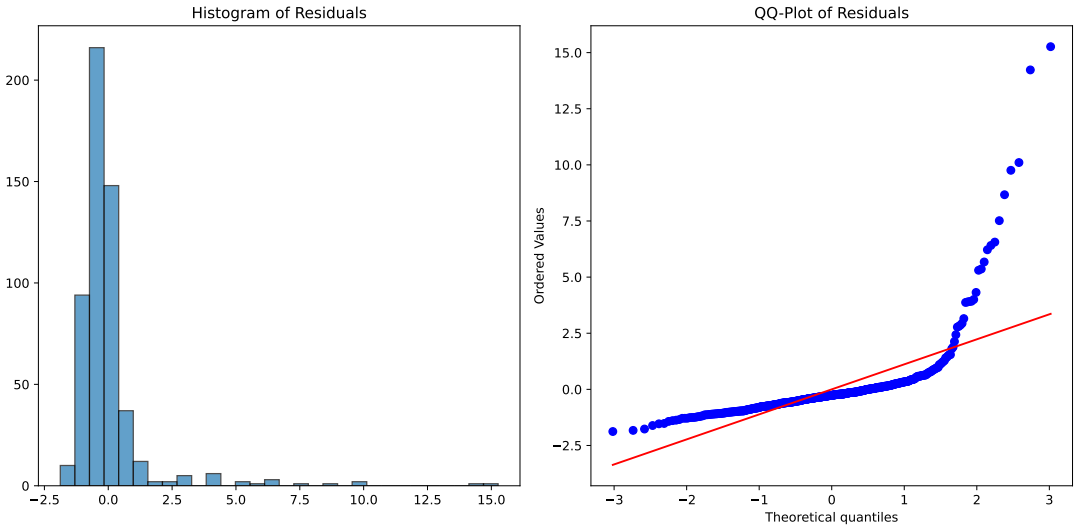


Figure 11: Diagnostic plots of the causal linear model for the effect of FOXP3 (treatment) on GZMH (outcome).

Parametric model. The parametric model assumes a linear relationship between the target gene expression (y), the treatment gene (a), and the cell-type/state representation (z). The parametric model is specified as:

$$\mathrm{p}_\theta(y \mid a, w) = \mathcal{N}\left(c + \tau a + \gamma^T z, \sigma^2\right), \tag{125}$$

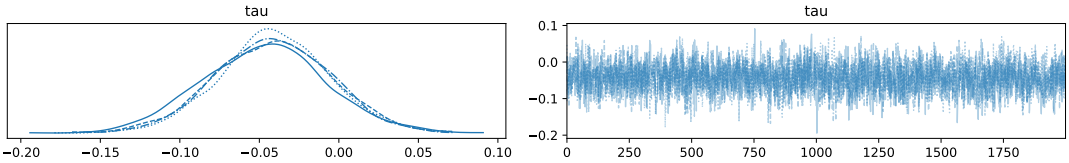


Figure 12: Traceplots of the BART model MCMC inference for the effect of FOXP3 (treatment) on GZMH (outcome), corresponding to the parameter τ. Different lines correspond to separate chains (4 total).

where c is the intercept τ is the coefficient of the treatment gene a, γ is a vector of coefficients for the confounding genes in w, and σ^2 is the noise variance.

Let $\theta \equiv [c, \tau, \gamma]$. Assuming a flat prior $\mathrm{p}(\theta) \propto 1$ and that the variables c, τ, γ are apriori independent from a and x, the posterior distribution is given by:

$$\mathrm{p}\left(\theta \mid x_{1:n}\right) \propto \mathrm{p}_\theta(y_{1:n} \mid a_{1:n}, z_{1:n}) = \mathcal{N}\left(\hat{\theta}, \hat{V}\sigma^2\right), \tag{126}$$

where

$$\begin{bmatrix} \hat{c} \\ \hat{\tau} \\ \hat{\gamma} \end{bmatrix} = \hat{\theta} \equiv (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T y, \quad \hat{V} \equiv (\tilde{X}^T\tilde{X})^{-1}, \quad \text{for } \tilde{X} \equiv \begin{bmatrix} 1 & a_1 & z_1^T \\ 1 & a_2 & z_2^T \\ \vdots & \vdots & \vdots \\ 1 & a_n & z_n^T \end{bmatrix}. \tag{127}$$

Under the parametric model (125), the CATE is constant across x so the posterior of ATE(p) simplifies to

$$\mathrm{ATE}\left(\mathrm{p}_\theta\right) = \tau(q_{98}(a) - q_0(a)), \quad \tau \sim \Pi\left(\tau \mid x_{1:n}\right).$$

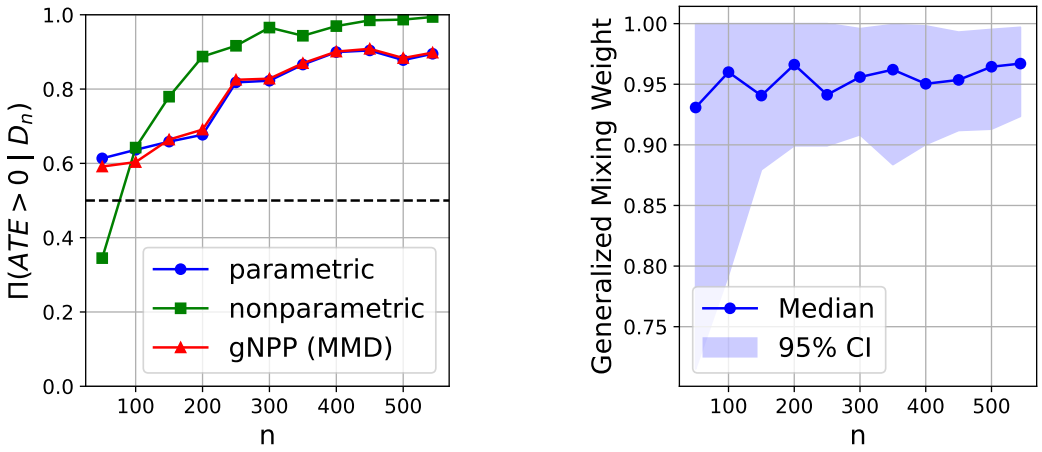
Eq. (126) shows that the joint posterior of (c, τ, γ) follows a multivariate normal distribution. After marginalization, we obtain the posterior for ATE, denoted $\Pi_{\mathrm{pm}}\left(\mathrm{ATE}\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right)$.

$$\Pi_{\mathrm{pm}}\left(\mathrm{ATE}\left(\mathrm{p}_\theta\right) \mid x_{1:n}\right) = \mathcal{N}\left(\hat{\tau}(q_{98}(a) - q_0(a)), \hat{V}_{22}(q_{98}(a) - q_0(a))^2\right), \tag{128}$$

where $\hat{V}_{22}$ is the second diagonal entry of $\hat{V}$ defined in Eq. (127).

E.2 Effect of TCF7 on SELL

Besides considering interventions on FOXP3, we also considered the effect of interventions on *TCF7* (Transcription Factor 7) on *SELL* (L-selectin). Elevated levels of SELL are associated with favorable survival outcomes in breast cancer (Kumari et al., 2021). The gNPP posterior suggests that increasing the expression of TCF7 is likely to increase the expression of SELL (Figure 13a). In this case, the generalized mixing weight places strong weight on the parametric model (Figure 13b).



(a) (b)

Figure 13: **Effect of TCF7 on SELL.** a. Posterior probability of the ATE being positive under the parametric, nonparametric, and gNPP models. n denotes the size of the (subsampled) dataset. Values are the median across 10 independent data subsamples and model samples. b. Generalized mixing weights, $\hat{\eta}_n$. The estimated confidence interval (CI) is across independent data subsamples and model samples.