ESAIM: PS 21 (2017) 251–274
DOI: 10.1051/ps/2017014

ESAIM: Probability and Statistics
www.esaim-ps.org

QUALITATIVE ROBUSTNESS IN BAYESIAN INFERENCE

Houman Owhadi[1] and Clint Scovel[1]

Abstract. The practical implementation of Bayesian inference requires numerical approximation when closed-form expressions are not available. What types of accuracy (convergence) of the numerical approximations guarantee robustness and what types do not? In particular, is the recursive application of Bayes' rule robust when subsequent data or posteriors are approximated? When the prior is the push forward of a distribution by the map induced by the solution of a PDE, in which norm should that solution be approximated? Motivated by such questions, we investigate the sensitivity of the *distribution* of posterior distributions (*i.e.* of *posterior distribution*-valued random variables, randomized through the data) with respect to perturbations of the prior and data-generating distributions in the limit when the number of data points grows towards infinity.

Mathematics Subject Classification. 62F15, 62F35.

Received May 17, 2016. Revised May 5, 2017. Accepted July 21, 2017.

1. Introduction and motivations

When we apply Bayesian inference with Gaussian priors and linear observations, we do not actually compute Bayes' rule but solve the linear system whose solution is the conditional expectation, and robustness is guaranteed by that of our linear solver. This paper is motivated by robustness questions that arise in the numerical implementation of Bayes' rule for continuous systems when closed-form expressions are not available for the computation of posterior values/distributions. For example, what is the sensitivity of posterior values or the sensitivity of the distribution of posterior values to perturbations introduced by the numerical approximation of the prior? When Bayes' rule is applied in a recursive manner and approximated posterior distributions are used as prior distributions or approximated data is used in the conditioning process, such as in [2,15,16,24,34,45,52,57,61], do we have robustness guarantees on subsequent posterior distributions/values? Is it possible to numerically approximate the optimal prior/mixed strategy of a decision theory problem (arising in the continuous setting under the complete class theorem [62]) when closed form expressions are not available, and if it is, in which metric should we do so?

Figure 1 provides an illustration of partial answers to such questions, which, when combined, can be used as a map to navigate robustness/non-robustness questions/issues arising from numerical approximations. To begin, (a) [47–49] demonstrated the *brittleness* of Bayesian inference, in the sense that they show that the range of posterior values of the quantity of interest under perturbations of the prior in Prokhorov or total variation (TV) metrics is the deterministic range of the quantity of interest. Moreover, (c.1) [58] shows that

Keywords and phrases. Bayesian inference, qualitative robustness, stability, Hampel.

[1] California Institute of Technology, Computing and Mathematical Sciences, MC 9-94 Pasadena, CA 91125, USA.
owhadi@caltech.edu; clintscovel@gmail.com

Article published by EDP Sciences

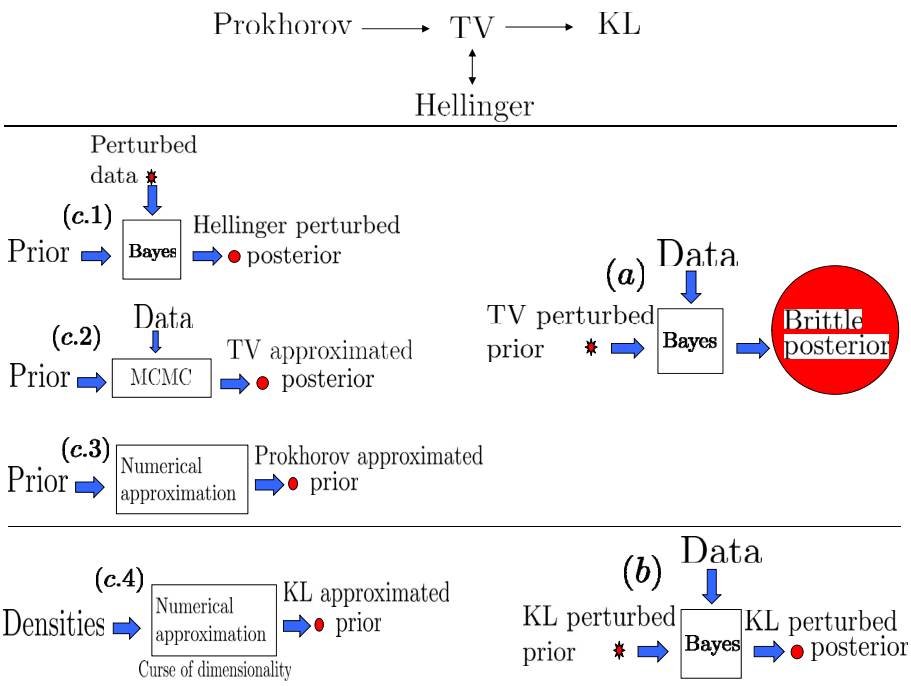


FIGURE 1. On the robustness of the Bayes' rule. The top panel illustrates relationships among probability metrics and, as in [32], a directed arrow from A to B means that (for some function h) $d_A \leq h(d_B)$. The central panel illustrates the generation of non-robustness and the bottom panel illustrates the generation of robustness.

the posterior distributions are controlled by the Hellinger metric under the application of the Bayes rule with an exact prior under the perturbation of a finite number of finite dimensional samples (see also [12]). Since the Hellinger and TV metrics are equivalent [32], (c.1) and (a) suggest that, without specific restrictions, it is, in general, not possible to offer guarantees on the robustness of recursive Bayes under perturbations of the data. Similarly (c.2) the convergence of Markov chains (on continuous state spaces) used in MCMC algorithms (such as the Metropolis algorithm) is generally analyzed with respect to the TV topology [31, 54]. However it should be noted that, according to Roberts and Rosenthal ([53], p. 10), Gibbs [31], Gelman ([29], Intro) and Madras and Sezer ([42], Intro), convergence in TV is in general not guaranteed, so that in general convergence of MCMC in any metric stronger than TV is not guaranteed. Therefore combining (c.2) and (a) suggests that, without further restrictions, it is, in general, not possible to offer guarantees on the robustness of recursive Bayes if posteriors are approximated using MCMC. Moreover, in many cases, the prior may need to undergo a numerical approximation/discretization step prior to conditioning. For example, in popular applications of Bayes' rule to stochastic PDEs [22, 64] one pushes forward the prior from the space of coefficients of the PDE to the solution space where it is conditioned. Consequently, if the PDE is numerically approximated this implies an approximation of the pushed-forward prior. Therefore (c.3) representing numerical discretization/approximation as a continuous map between two Polish spaces, it is known that the push forward of a measure under such maps is continuous in the weak topology, which is metrized by the Prokhorov distance. Therefore, combining (c.3) and (a) suggests that, without further restrictions, it is, in general, not possible to offer robustness guarantees to perturbations caused by the numerical discretization of the prior.

For positive results on the other hand, observe that (b) [35] shows that posterior values (given a finite number of data points) are robust to approximation errors of the prior measured in Kullback–Leibler divergence.

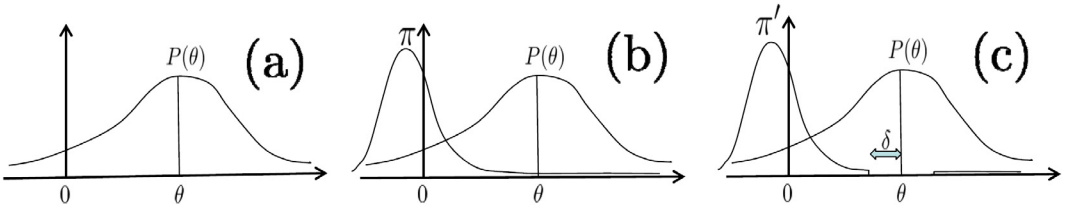


FIGURE 2. Primary mechanism for non-robustness. See Illustration 1.1.

Therefore if (c.4) the numerical approximation map is continuous in Kullback–Leibler divergence then posterior values (given a finite number of data points) are robust to that numerical discretization step. However observe that unless closed form expressions are available, one must keep track of densities to achieve the continuity of the numerical approximation map in Kullback–Leibler divergence, which is a task plagued by the curse of dimensionality.

Since the brittleness of posterior distributions and values with respect perturbations of the prior defined in TV or Prokhorov metrics is an obstacle to obtaining robustness guarantees to numerical approximation errors when closed form expressions are unavailable, it is natural to ask whether robustness could be guaranteed by considering the *distribution* of posterior distributions or posterior values generated by the random generation of the data. To answer this question we develop a framework for quantifying the sensitivity of the distribution of posterior distributions with respect to perturbations of the prior and data-generating distributions in the limit when the number of data points grows towards infinity. In this generalization of Cuevas' [19] extension of Hampel's [38] notion of *qualitative robustness* to Bayesian inference to include both perturbations of the prior and the data-generating distribution, posterior distributions are analyzed as measure-valued random variables (measures randomized through the data) and their robustness is quantified using the TV, Prokhorov, and Ky Fan metrics. Our results show that (1) the assumption that the prior has Kullback–Leibler support at the parameter value generating the data, classically used to prove consistency, can also be used to prove the non-robustness of posterior distributions with respect to infinitesimal perturbations in TV of the class of priors satisfying that assumption, (2) for a prior which has global Kullback–Leibler support on a space which is not totally bounded, we can establish non-robustness and (3) consistency, and the unstable nature of the conditions which generate it, produces non-robustness, and a careful selection of the prior is important if both properties (or their approximations) are to be achieved. The mechanisms supporting our results are different and complementary to those discovered by Hampel and developed by Cuevas. To obtain them, we derive in Section 3 a corollary to Schwartz' consistency Theorem that leads to robustness or non-robustness results depending on the topology defining the continuity of the numerical approximation map (Kullback–Leibler, TV or Prokhorov). Moreover, this corollary is further developed in Proposition A.1 to analyze the convergence of random measures in the qualitative robustness framework. More precisely, although in [35] it is shown that the Fréchet derivatives of posterior values to Kullback–Leibler perturbations of the prior may diverge to infinity with the number of data points, a simple application of Proposition A.1 implies the robustness of the distribution of posterior distributions/values under Kullback–Leibler perturbations of the prior in the limit where the number of data points goes to infinity. On the other hand, application of Proposition A.1 also suggest the lack of robustness of the distribution of posterior distributions/values to TV or Prokhorov perturbations of the prior, where we note that Prokhorov perturbations include classes of perturbations defined by generalized moment constraints, as in [6–9, 47–49].

Illustration 1.1 (Illustration of mechanisms generating non robustness)**.**

This is an informal illustration of Example 5.1. Let $X := \mathbb{R}$ and $\Theta := \mathbb{R}$ and consider the Gaussian parametric model $P : \mathbb{R} \to \mathcal{M}(\mathbb{R})$ such that the measure $P(\theta)$ is the standard Gaussian measure with mean θ, that is it has the density $\frac{1}{\sqrt{2\pi}} e^{-\frac{-|x-\theta|^2}{2}}, x \in \mathbb{R}$. We will reference Fig. 2 by indicating a), b), or c).

In our first example, we would like show that assuming the prior has Kullback−Leibler support for all $\theta \in \Theta$ does not lead to robustness. To that end, let ρ be small and a) let the data-generating distribution be $P(\theta)$ for some fixed $\theta \in \Theta$, and let us not allow any perturbations to it. One can show that any strictly positive measure, one which is strictly positive on all open sets, has full Kullback−Leibler support. By selecting b) a strictly positive prior π with a very small mass about θ, it follows that c) a measure π' which is close to π in the TV metric exists which has no support in a neighborhood of size ρ of θ. For the prior π' the corresponding posterior will remain uniformly bounded away from δ_θ at a distance of at least ρ. Moreover, since π has Kullback−Leibler support at θ, the posteriors it generates converge to δ_θ. This is the mechanism establishing the assertion of Theorem 4.1.

In our second example, we begin with a prior as in b), then select a θ so that their relationship is as in b) so that we again end up at c) in Figure 2. That is, we begin with a prior π with full Kulback-Leibler support and demonstrate non robustness to small perturbations to it, while also not allowing perturbations in the data-generating distribution. To that end, fix $\delta > 0$, let $\rho > 0$ be small, and consider a prior π which has Kullback−Leibler support for all $\theta \in \mathbb{R}$. Since $\mathbb{R}$ is not completely bounded it follows that there exists a θ such that the measure of the interval about θ of size δ is small enough that there exists a measure π' within TV distance ρ of π whose support does not intersect this interval. Selecting $P(\theta)$ as the data-generating distribution, it again follows that the posterior distributions resulting from the prior π' stay uniformly a distance ρ from δ_θ. Moreover, since π has Kullback−Leibler support at θ, the posteriors it generates converge to δ_θ. This is the mechanism establishing the assertion of Theorem 4.3.

Remark 1.2 (Instabilities in Deep Learning)**.** Are the non robustness mechanisms exhibited in [47–49] purely academic or can they be found in practical applications? In [47–49] the mechanism producing instabilities to adversarial small perturbations of the prior/model exploits the fact that the model is a finite dimensional object whose infinite co-dimension is vulnerable to adversarial perturbations using the linear dependence of the probability of the data (appearing in the denominator of the Bayes' rule) with respect to the choice of the prior. As a consequence, a perturbation of the prior can be chosen to be small (in TV metric) and, at the same time, have a large impact on posterior properties by aligning those perturbations with the events associated with the observation of the data (and those instabilities can be alleviated, at the cost of consistency, through a process of coarsening of the data [48, 49]).

Recent work of Szegedy *et al.* [59], Goodfellow, Shlens and Szegedy [33], and further developed in Uličnỳ, Lundström and Byttner [60] show that deep learning image classification is unstable with respect to adversarial perturbation of test images. That is very nearby data points can produce dramatically different results, (see *e.g.* [33], p. 3) where a minuscule perturbation of an image correctly classified as a panda produces a visually identical image classified as a gibbon.

The mechanism proposed in [33] for the non-robustness of deep learning is very similar to the mechanism causing non-robustness in Bayesian inference [47–49]: deep learning networks are composed of hierarchies of linear models corresponding to scalar products taken against feature maps in a high dimensional spaces of images. A small perturbation of an image can have a large impact on classification through its alignment with feature maps.

It is also interesting to note that Gal and Ghahramani [28] and Le, Baydin, Zinkov and Wood [41] demonstrate a connection between deep learning and Bayesian inference and Patel, Nguyen and Baraniuk [50] have developed a Bayesian theory of deep learning where they propose to achieve robustness (as in) through Approximate Bayesian Computation (that is equivalent to a process of coarsening [17, 43] of the data).

2. Qualitative Robustness for Bayesian inference

Hable and Christmann [36] have recently established qualitative robustness for support vector machines. Consequently, it appears natural to inquire into the qualitative robustness of Bayesian inference. Hampel [38] introduced the notion of the qualitative robustness of a sequence of estimators and Cuevas [19] has extended Hampel's definition and his basic structural results to Polish spaces. Since the space $\mathcal{M}(\Theta)$ of priors and

posteriors equipped with the weak topology is Polish whenever Θ is, Cuevas' extension has direct applications to Bayesian inference. Boente *et al.* [11] have developed qualitative robustness for stochastic processes, Nasser *et al.* [46] for estimation, and Basu *et al.* [4] for Bayesian inference with a single sample. The notion of qualitative robustness introduce in this paper is a straightforward generalization of that introduced by Hampel [38] and developed by Cuevas [19], see also Cuevas [21]. Indeed, this version requires no introduction of loss and risk functions and concerns itself with not just expected values but with the full distribution of the effects of the randomness of the observations. It considers a fixed model so is not concerned with robustness with respect to model specification, although such considerations can easily be included. Moreover, since it is formulated with respect to classic notions in probability, it appears to us as simple, natural, flexible, without calibration issues, and easy to interpret statistically.

Metrics on spaces of measures and random variables will be important in its formulation. Our primary assumption is that both the sample space X and the parameter space Θ are Borel subsets of Polish metric spaces. It will be demonstrated that this extremely general assumption is sufficient to gives us extremely general results. However, to keep the presentation simple we will restrict our attention to the TV, Prokhorov and Ky Fan metrics (we refer to [6–9] for motivations for considering classes of priors defined in the Prokhorov metric). The necessary well-definedness and measurability considerations for such metrics and for Bayesian conditioning with a fixed measurable model on Borel subsets of Polish metric spaces is presented in Section A.1. For measurable spaces Θ and X, we write $\mathcal{M}(\Theta)$ and $\mathcal{M}(X)$ for the set of probability distributions on Θ and X respectively. Furthermore, when Θ and X are Borel subsets of Polish metric spaces we can metrize $\mathcal{M}(\Theta)$ and $\mathcal{M}(X)$ with the Prokhorov metrics $d_{Pr\Theta}$) and d_{PrX}) and having done so, we can define the the space $\mathcal{M}^2(\Theta) := \mathcal{M}(\mathcal{M}(\Theta))$ of Borel probability measures on the metric space $(\mathcal{M}(\Theta), d_{Pr\Theta})$ of Borel probability measures.

Let us fix a measurable model $P : \Theta \to \mathcal{M}(X)$. Then for a prior $\pi \in \mathcal{M}(\Theta)$, consider the measurable map

$$\bar{\pi} : X^n \to \mathcal{M}(\Theta)$$

defined by the family of posterior conditional measures

$$\bar{\pi}(x^n) := \pi_{x^n}, \quad x^n \in X^n, \tag{2.1}$$

so that its corresponding pushforward operator

$$\pi_* : \mathcal{M}(X^n) \to \mathcal{M}^2(\Theta)$$

is well-defined, where we have removed the bar over π to simplify the notation, while still emphasizing that this pushforward operator π_* corresponds to the prior π. Then the pushforward

$$\pi_*\mu^n \in \mathcal{M}^2(\Theta)$$

of the iid data-generating distribution, which is the sampling distribution of the posterior distribution π_{x^n} when $x^n \sim \mu^n$, represents how the posteriors π_{x^n} vary as a function of the sample data x^n when it is generated by i.i.d. sampling from μ. For a fixed prior $\pi \in \mathcal{M}(\Theta)$, we say that the Bayesian inference corresponding to the model P is qualitatively robust at a data-generating distribution $\mu \in \mathcal{M}(X)$ with respect to an admissible set $\mathcal{P}$ containing μ, and metrics $d_{\mathcal{M}(X)}$ and $d_{\mathcal{M}^2(\Theta)}$ on $\mathcal{M}(X)$ and $\mathcal{M}^2(\Theta)$, if for any $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\mu} \in \mathcal{P},\ d_{\mathcal{M}(X)}(\mu, \acute{\mu}) < \delta \implies d_{\mathcal{M}^2(\Theta)}(\pi_*\mu^n, \pi_*\acute{\mu}^n) < \epsilon$$

for large enough n.

On the other hand, when the data-generating distribution μ is fixed and we vary the prior π, we consider the sequence of maps

$$\pi_n : X^\infty \to \mathcal{M}(\Theta)$$

defined by

$$\pi_n(x^\infty) := \pi_{x^n}, \quad x^\infty \in X^\infty. \tag{2.2}$$

Since the projection $P_n : X^\infty \to X^n$ is continuous and $\pi_n = \bar{\pi} \circ P_n$, it follows from the measurability of $\bar{\pi} : X^n \to \mathcal{M}(\Theta)$, that π_n is measurable, and therefore the resulting sequence

$$\pi_n : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$$

is a sequence of $\mathcal{M}(\Theta)$-valued random variables. For each $\mu \in \mathcal{M}(X)$, let α_μ be a metric on the space of $\mathcal{M}(\Theta)$-valued random variables whose domain is the probability space (X^∞, μ^∞). Then, for a prior $\pi \in \mathcal{M}(\Theta)$, we say that the Bayesian inference corresponding to the model P is qualitatively robust at π with respect to an admissible set $\mathit{\Pi} \subset \mathcal{M}(\Theta)$ containing π, the metric α_μ and the metric $d_{\mathcal{M}(\Theta)}$ on $\mathcal{M}(\Theta)$, if given $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\acute{\pi} \in \mathit{\Pi},\ d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta \implies \alpha_\mu(\pi_n, \acute{\pi}_n) < \epsilon$$

for large enough n.

These two definitions can be combined in a straightforward manner to define robustness corresponding to a single prior/data-generating pair. However, to consider a larger class of distributions than a single pair along with a general class of perturbations, we let $\mathcal{Z} \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ denote the admissible set of prior/data-generating distribution pairs including allowed perturbations to be specified such that $\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) \in \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ means that $(\pi, \mu) \in \mathcal{M}(\Theta) \times \mathcal{M}(X)$ is an admissible candidate for robustness considerations and $(\acute{\pi}, \acute{\mu}) \in \mathcal{M}(\Theta) \times \mathcal{M}(X)$ is an admissible candidate for its perturbation. In particular, the projection $\mathcal{Z}_1 \subset \mathcal{M}(\Theta) \times \mathcal{M}(X)$ of $\mathcal{Z}$ onto its first component denotes the set of admissible prior/data-generating pairs. Now combining in a straightforward manner we obtain:

Definition 2.1. Let X and Θ be Borel subsets of Polish metric spaces and let $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$ be equipped with the weak topology metrized by the Prokhorov metrics d_{PrX} and $d_{Pr\Theta}$. Let $\mathcal{M}^2(\Theta) := \mathcal{M}(\mathcal{M}(\Theta))$ be the space of Borel probability measures on the metric space $(\mathcal{M}(\Theta), d_{Pr\Theta})$ of Borel probability measures on Θ equipped with its weak topology metrized by its Prokhorov metric $d_{Pr^2\Theta}$. Consider perturbation pseudometrics $d_{\mathcal{M}(X)}$, $d_{\mathcal{M}(\Theta)}$ and $d_{\mathcal{M}^2(\Theta)}$ on $\mathcal{M}(X)$, $\mathcal{M}(\Theta)$ and $\mathcal{M}^2(\Theta)$ respectively and for each $\mu \in \mathcal{M}(X)$, let α_μ be a pseudometric on the space of $\mathcal{M}(\Theta)$-valued random variables on the probability space (X^∞, μ^∞). Let $\mathcal{Z} \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ denote the admissible set of prior/data-generating distribution pairs including allowed perturbations and suppose that $P : \Theta \to \mathcal{M}(X)$ is measurable. Then the Bayesian inference corresponding to the model P is qualitatively robust with respect to $\mathcal{Z}$, if given $\epsilon_1, \epsilon_2 > 0$, there exists $\delta_1, \delta_2 > 0$ such that

$$\begin{aligned}\big((\pi, \mu), (\acute{\pi}, \acute{\mu})\big) \in \mathcal{Z}, \quad d_{\mathcal{M}(\Theta)}(\pi, \acute{\pi}) < \delta_1,\ d_{\mathcal{M}(X)}(\mu, \acute{\mu}) < \delta_2 \\ \implies d_{\mathcal{M}^2(\Theta)}(\pi_* \mu^n, \acute{\pi}_* \acute{\mu}^n) < \epsilon_1\,, \quad \alpha_\mu(\pi_n, \acute{\pi}_n) < \epsilon_2\end{aligned}$$

for large enough n.

Remark 2.2. Let us explain why we have both the Prokhorov metrics d_{PrX}, $d_{Pr\Theta}$ and the perturbation pseudometrics $d_{\mathcal{M}(X)}$, $d_{\mathcal{M}(\Theta)}$ on $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$. The primary reason for the Prokhorov metrics is that weak topology is natural for Schwartz' convergence theorem. One could also specify the Prokhorov metrics for the perturbation metrics but the proofs of our main results appear to allow the much stronger TV metric to be used. Since a small TV neighborhood is much smaller than a Prokhorov neighborhood we feel this extra expressivity warranted. Moreover, allowing the perturbation pseudometrics to be user specified will allow this definition to be used in the development of positive results, see Remark 4.6.

Finite sample versions, as introduced in Hable and Christmann ([37], Def. 2), are also available. Note that unlike Hampel and Cuevas who require "for all n" in their definitions, we follow Huber [39] and Mizera [44] in only requiring closeness "for large enough n". The results of this paper are applicable to both versions. Of course the relevance of the specific notion of qualitative robustness used depends on the perturbation metrics used. The results of this paper apply to the case when α_μ is the Ky Fan metric, metrizing convergence in probability on the space of $\mathcal{M}(\Theta)$-valued random variables with domain the probability space (X^∞, μ^∞) and the $d_{\mathcal{M}(\Theta)}$ is any metric weaker than the total variation.

3. Lorraine Schwartz' Theorem

The fundamental mechanism generating non robustness for Bayesian inference will be its consistency when combined with the unstable nature of the conditions which generate it. The breakthrough in consistency for Bayesian inference is considered to be Schwartz' theorem ([56], Thm. 6.1), so we use it as a model for consistency and the conditions sufficient to generate it. Barron, Schervish and Wasserman ([3], Intro), Wasserman ([63], p. 3) and Ghosal, Ghosh and Ramamoorthi ([30], Cor. 1) state the result for the nonparametric case. Here we consider the parametric case. To that end, observe that Schervish ([55], Thm. B.32) asserts that regular conditional probabilities exist for conditioning random variables with values in a Borel subset of a Polish space. Moreover, when the parametric model is a Markov kernel and is dominated by a σ-finite measure, then by the Bayes' Theorem for densities Schervish ([55], Thm. 1.31) we have, in addition, that the Bayes' rule for densities determines a valid family of densities for the regular conditional distributions. We say that a model $P : \Theta \to \mathcal{M}(X)$ is dominated if there exists a σ-finite Borel measure ν on X such that $P_\theta \ll \nu$, $\theta \in \Theta$.

Recall the Kullback–Leibler divergence K between two measures μ_1 and μ_2 defined by

$$K(\mu_1, \mu_2) := \int \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\nu} \Big/ \frac{\mathrm{d}\mu_2}{\mathrm{d}\nu}\right)\mathrm{d}\mu_1 \,,$$

where ν is any measure such that both μ_1 and μ_2 are absolutely continuous with respect to ν. It is well known that K is nonnegative, and that it is finite only if $\mu_1 \ll \mu_2$, and in that case $K(\mu_1, \mu_2) = \int \log \frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\mathrm{d}\mu_1$. From this we can define the Kullback–Leibler ball $K_\epsilon(\mu)$ of radius ϵ about $\mu \in \mathcal{M}(X)$ by $K_\epsilon(\mu) = \{\mu' \in \mathcal{M}(X) : K(\mu, \mu') \leq \epsilon\}$. For a model $P : \Theta \to \mathcal{M}(X)$, there is the pullback to a function K on Θ defined by $K(\theta_1, \theta_2) := K(P_{\theta_1}, P_{\theta_2})$ and when the model is dominated by a σ-finite measure ν, if we let $p(x|\theta) := \frac{dP_\theta}{\mathrm{d}\nu}(x), x \in X$ be a realization of the Radon-Nikodym derivative, then the pullback has the form

$$K(\theta_1, \theta_2) := \int \log \frac{p(x|\theta_1)}{p(x|\theta_2)} dP_{\theta_1}(x).$$

From this we define a Kullback–Leibler neighborhood of a point $\theta \in \Theta$ by

$$K_\epsilon(\theta) := \big\{\theta' \in \Theta : K(\theta, \theta') \leq \epsilon\big\}. \tag{3.1}$$

Let us define the set of priors $\mathcal{K}(\theta) \subset \mathcal{M}(\Theta)$ which have Kullback–Leibler *support at* θ by

$$\mathcal{K}(\theta) := \Big\{\pi \in \mathcal{M}(\Theta) : \pi\big(K_\epsilon(\theta)\big) > 0, \quad \epsilon > 0\Big\}, \tag{3.2}$$

which implicitly requires that $K_\epsilon(\theta)$ be measurable[2] for all $\epsilon > 0$. Also let $\mathcal{K} \subset \mathcal{M}(\Theta)$ denote those measures with *global* Kullback–Leibler support, that is,

$$\mathcal{K} := \cap_{\theta \in \Theta} \mathcal{K}(\theta)$$

is the set of priors which have Kullback–Leibler support at all θ.

For the nonparametric case, Barron, Schervish and Wasserman ([3], Lem. 11) demonstrate that the Kullback–Leibler neighborhoods $K_\epsilon(P_{\theta^*}) \subset \mathcal{M}(X)$ are measurable with respect to the strong topology restricted to the subspace of measures which are absolutely continuous with respect to a common σ-finite reference measure. For the parametric case, we establish the following.

Lemma 3.1. *Let X and Θ be Borel subsets of Polish spaces and suppose that the model P is measurable and dominated. Then, for each $\epsilon > 0$ and each $\theta \in \Theta$, the Kullback–Leibler neighborhood $K_\epsilon(\theta) \subset \Theta$ is measurable.*

[2] Note the change from the standard definition $K_\epsilon(\mu) = \{\mu' : K(\mu, \mu') < \epsilon\}$ to ours $K_\epsilon(\mu) = \{\mu' : K(\mu, \mu') \leq \epsilon\}$ does not affect which measures have Kullback–Leibler support, but is more convenient since then the proof of Lemma 3.1 shows that $K_\epsilon(\mu)$ is closed.

The following corollary to Schwartz' Theorem, and its implications in Proposition A.1, gives us the form of consistency that we will use in the robustness analysis. Since the σ-algebra of a Borel subset of a Polish space is countably generated, Doob's Theorem in Dellacherie and Meyer ([23], Thm. V.58) and the measurability of the dominated model P implies that a family $p(\theta), \theta \in \Theta$ of densities can be chosen to be $\mathcal{B}(X) \times \mathcal{B}(\Theta)$ measurable, so that this assumption of Schwartz' Theorem ([56], Thm. 6.1) is satisfied. We note that Dellacherie and Meyer emphasize that the countably generated condition is indispensable for Doob's Theorem to apply. Note the assumption that the map $P : \Theta \to P(\Theta)$ be open.

Corollary 3.2 (Schwartz)**.** *Let X and Θ be Borel subsets of Polish metric spaces and equip $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$ with the Prokhorov metrics. Consider an injective measurable dominated model $P : \Theta \to \mathcal{M}(X)$ such that $P : \Theta \to P(\Theta)$ is open. Then for every $\pi \in \mathcal{M}(\Theta)$ with Kullback–Leibler support at $\theta^* \in \Theta$ and for every measurable neighborhood U of θ^*, we have*

$$\pi_{x^n}(U) \to 1 \quad n \to \infty, \quad a.e.\ P_{\theta^*}^{\infty}.$$

4. Main results

Now that we have defined *qualitative robustness* for Bayesian inference and presented the consistency conditions of Schwartz' Corollary 3.2, we are now prepared for our main results. Indeed, the brittleness results of [47–49] and the non qualitative robustness results of Cuevas ([19], Thm. 7) suggest that we may obtain non qualitative robustness according to Definition 2.1 by fixing the prior and varying the data-generating distribution. However, according to Berk [5], in the misspecified case, although "there need be no convergence (in any sense)", in the limit the posterior becomes confined to a carrier set consisting of those points which are closest in terms of the Kullback–Leibler divergence. Consequently, it appears possible that a generalization of the results of Hampel ([38], Lem. 3) and Cuevas ([19], Thm. 1) which allows such a set-valued notion of consistency may be sufficient. Certainly it will require the more sophisticated notions of the continuity, or semi-continuity, of the Kullback–Leibler set-valued information projection and its dependence on the geometry of the model class $P(\Theta) \subset \mathcal{M}(X)$. Although this path will certainly be instructive and appears feasible, we instead find it simpler to obtain non qualitative robustness by fixing the data-generating distribution to be in the model class and varying the prior. In particular, we show that the inference is not robust according to Definition 2.1 when the metric α_μ is the the Ky Fan metric and the metric on $\mathcal{M}(\Theta)$ is any that is weaker than the total variation metric. It is important to note that these results do not require any misspecification. Moreover, it appears that Bayesian Inference's dependence on both the data-generating distribution and the prior leads to two complementary mechanisms generating non qualitative robustness; whereas Cuevas' result ([19], Thm. 7) utilizes consistency and the discontinuity of the infinite sample limit, this other component utilizes the non-robustness of consistency, namely that the set of consistency priors, those with Kullback–Leibler support at the data-generating distribution, is not robust.

Now let us return to our main results. For $\theta \in \Theta$, let us recall from (3.2) the set of priors $\mathcal{K}(\theta) \subset \mathcal{M}(\Theta)$ with Kullback–Leibler support at θ and, for $\rho > 0$, define a total variation uniformity $\Pi_\rho(\theta) \subset \mathcal{M}(\Theta) \times \mathcal{M}(\Theta)$ by

$$\Pi_\rho(\theta) := \{(\pi, \acute{\pi}) \in \mathcal{M}(\Theta) \times \mathcal{M}(\Theta) : \pi \in \mathcal{K}(\theta),\ d_{tv}(\pi, \acute{\pi}) < \rho\} \tag{4.1}$$

of prior pairs where the first component has Kullback–Leibler support at θ and the second component is within ρ of the first in the total variation metric. For $\theta \in \Theta$, we define an admissible set of prior/data-generating distribution pairs including allowed perturbations $\mathcal{Z}_\rho(\theta) \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ by

$$\mathcal{Z}_\rho(\theta) := \Pi_\rho(\theta) \times P_\theta \times P_\theta\,, \tag{4.2}$$

using the identification of $\big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ with $\mathcal{M}(\Theta)^2 \times \mathcal{M}(X)^2$.

Our Main Theorem shows, under the conditions of Schwartz' Corollary, that the Bayesian inference is not robust under the assumption that the prior has Kullback–Leibler support at the parameter value generating

the data. This result, along with those that follow, supports Cuevas' [20] statement that "his results suggest the possibility of proving the instability (*i.e.* the lack of qualitative robustness) for a wide class of usual Bayesian models".

Theorem 4.1. *Consider Definition 2.1 with the total variation metric $d_{\mathcal{M}(\Theta)} := d_{tv}$ on $\mathcal{M}(\Theta)$ and the Ky Fan metric α_μ on the space of $\mathcal{M}(\Theta)$-valued random variables with domain the probability space (X^∞, μ^∞). Given the conditions of Schwartz' Corollary 3.2, for all $\theta \in \Theta$ the Bayesian inference corresponding to the model P is not qualitatively robust with respect to $\mathcal{Z}_\rho(\theta)$ for all $\rho > 0$.*

Remark 4.2. Actually the proof shows more; let D denote the diameter of Θ, then for $\epsilon < \min(\frac{D}{2}, 1)$, there does not exist a $\delta > 0$ such that robustness is satisfied. Since $\min(\frac{D}{2}, 1)$ is large, either half the diameter of the space or larger than 1, we say the inference is *brittle*.

Theorem 4.1 does not assert that the Bayesian inference corresponding to the model P is not robust at any specified prior, only that it is not robust under the assumption that the prior has Kullback–Leibler support at the parameter value generating the data. To establish non-robustness at specific priors we include variation in the data-generating distribution in the model class as follows. Let $\Delta_P \subset \mathcal{M}(X) \times \mathcal{M}(X)$, defined by

$$\Delta_P = \{(P_\theta, P_\theta), \theta \in \Theta\},$$

denote the fact that we allow the data-generating distribution to vary throughout the model class but do not allow any perturbations to it. Then, for $\pi \in \mathcal{M}(\Theta)$, define the admissible set $\mathcal{Z}_\rho(\pi) \subset \big(\mathcal{M}(\Theta) \times \mathcal{M}(X)\big)^2$ by

$$\mathcal{Z}_\rho(\pi) := \pi \times B_\rho^{tv}(\pi) \times \Delta_P\,,$$

where $B_\rho^{tv}(\pi)$ is the open ball in the total variation metric.

Since the following theorem is a corollary to the theorem after it, Theorem 4.4, we do not include its proof. However, we state it here because it is the more fundamental result.

Theorem 4.3. *Consider the situation of Theorem 4.1 with Θ not totally bounded. Then if the prior π has Kullback–Leibler support for all $\theta \in \Theta$, the Bayesian inference corresponding to the model P is not qualitatively robust with respect to $\mathcal{Z}_\rho(\pi)$ for all $\rho > 0$.*

Since a metric space is totally bounded if and only if its completion is compact, when Θ is totally bounded, we assume that it is a Borel subset of a compact metric space. In this case, although Theorem 4.3 does not apply, utilizing the covering number and packing number inequalities of Kolmogorov and Tikhomirov [40], we can provide a natural *quantification of qualitative robustness*. To that end, we define covering and packing numbers. For a finite subset $\Theta' \subset \Theta$, the finite collection of open balls $\{B_\epsilon(\theta), \theta \in \Theta'\}$ is said to constitute a covering of Θ if $\Theta \subset \cup_{\theta\in\Theta'} B_\epsilon(\theta)$. For a finite set Θ' we denote its size by $|\Theta'|$. The covering numbers are defined by

$$\mathcal{N}_\epsilon(\Theta) = \min\Big\{|\Theta'| : \Theta \subset \cup_{\theta\in\Theta'} B_\epsilon(\theta)\Big\}\,,$$

that is, $\mathcal{N}_\epsilon(\Theta)$ is the smallest number of open balls of radius ϵ centered on points in Θ which covers Θ. On the other hand, a set of points $\Theta' \subset \Theta$ is said to constitute an ϵ-packing if $d(\theta_1, \theta_2) \geq \epsilon$, $\theta_1 \neq \theta_2 \in \Theta'$. The packing numbers are then defined by

$$\mathcal{M}_\epsilon(\Theta) := \max\Big\{|\Theta'| : \Theta' \text{ is an } \epsilon\text{-packing of } \Theta\Big\}.$$

Since the Kolmogorov and Tikhomirov ([40], Thm. IV) inequalities

$$\mathcal{M}_{2\epsilon}(\Theta) \leq \mathcal{N}_\epsilon(\Theta) \leq \mathcal{M}_\epsilon(\Theta) \tag{4.3}$$

are valid in the *not* totally bounded case, if we allow values of ∞, the following theorem has Theorem 4.3 as its corollary.

Theorem 4.4. *Given the conditions of Theorem 4.3 with Θ totally bounded. If the Bayesian inference corresponding to the model P is qualitatively robust with respect to $\mathcal{Z}_\rho(\pi)$ for some $\rho > 0$, then given $\epsilon_2 > 0$, we must have*

$$\delta_1 < \min\left(\frac{1}{\mathcal{N}_{2\epsilon_2}(\Theta)}, \rho\right).$$

Remark 4.5. Note that the only assumptions Theorems 4.1, 4.3 and 4.4 make on the model P is that it be measurable, dominated, and $P : \Theta \to P(\Theta)$ be open. Moreover, since the TV metric is stronger than the Prokhorov metric, the utilization of the TV metric in the statements of the theorems implies the same theorem using the Prokhorov metric instead. Moreover, Prokhorov's compactness theorem, (see *e.g.* [1], Lem. 15.21), asserting that a family of probability measure on a separable metric space is relatively compact if and only if it is tight, suggests that small TV neighborhoods are extremely small Prokhorov neighborhoods. Indeed, to make this rigorous we would somehow assert that the relevant perturbed measures form a tight family.

Remark 4.6. In infinite dimensional linear parameter spaces, the conditions of Schwartz' theorem are, in general, not sufficient to generate the stronger convergence of a Bernstein-von Mises (BvM) theorem, that is, guaranteeing that in the large sample limit the posterior distribution approximates a normal distribution. In particular, Freedman [27] has shown that a BvM theorem does not hold on a separable Hilbert space in general. However, the recent work of Castillo and Nickl [14], continuing that of [13], shows that multiscale priors of type S and H, on page 1952 in the first reference, do generate a BvM theorem on separable Hilbert space. One might ask what the above results might say under such stronger assumptions generating stronger notions of consistency like the BvM theorem. As we stated at the beginning of Section 3, the results of this paper suggest that "The fundamental mechanism generating non robustness for Bayesian inference will be its consistency when combined with unstable nature of the conditions which generate it". In particular, if the consistency rate is improved we suspect the quantitative estimates of non-robustness will also improve. Moreover, if the prior generates a BvM consistency, the above results still assert that the inference is not robust to perturbations in TV metric. Consequently, in the setting of the introduction where the standard numerical approximated posteriors are then used as priors in a new iteration, such irregular distributions may be encountered and therefore the robustness of the procedure would be suspect.

Does this mean that Bayesian inference is not qualitatively robust? Far from it. Indeed, this is the reason that we have constructed a definition of qualitative robustness for Bayesian inference in Definition 2.1 which incorporates the user specified set $\mathcal{Z}$ of an admissible set of prior/data-generating distribution pairs including allowed perturbations, along with the user specified perturbation pseudometrics $d_{\mathcal{M}(X)}$, $d_{\mathcal{M}(\Theta)}$ and $d_{\mathcal{M}^2(\Theta)}$ on $\mathcal{M}(X)$, $\mathcal{M}(\Theta)$ and $\mathcal{M}^2(\Theta)$ respectively, and for each $\mu \in \mathcal{M}(X)$, the pseudometric α_μ on the space of $\mathcal{M}(\Theta)$-valued random variables on the probability space (X^∞, μ^∞). The challenge then would be, for a class of models, to make these specifications in a way that are appropriate to some application domain, and then establish qualitative robustness. In this sense, the primary message of these theorems is that any perturbation metric on $\mathcal{M}(\Theta)$ included in the definition of $\mathcal{Z}$ that is weaker than the TV metric, such as the Prokhorov metric, does not appear to be a good choice.

As an example, let us return the setting of Castillo and Nickl ([14], Thm. 3), where a BvM theorem is available for priors of type S or H. Let us focus on class S, where we note that it is not important to know what exactly this class is for this discussion, see page 1952 of [14] for the definitions. Now, instead of Definitions 4.1 and 4.2, let us define $\mathcal{M}_S(\Theta)$ to be the priors of class S. Then define

$$\Pi_\rho^S(\theta) := \{(\pi, \hat{\pi}) \in \mathcal{M}_S(\Theta) \times \mathcal{M}_S(\Theta) : \pi \in \mathcal{K}(\theta),\ d_{tv}(\pi, \hat{\pi}) < \rho\} \tag{4.4}$$

to be the class of prior/perturbed prior pairs where the first component has Kullback−Leibler support at θ and the second component is within ρ of the first in the total variation metric. For $\theta \in \Theta$, we define an admissible set of prior/data-generating distribution pairs including allowed perturbations by

$$\mathcal{Z}_\rho^S(\theta) := \Pi_\rho^S(\theta) \times P_\theta \times P_\theta\,, \tag{4.5}$$

In this case, Theorem 4.1 does not apply since the mechanisms described in Section 5 are not available, since in this case the measures used in these mechanism cannot be obtained through small TV perturbations because, by definition, the sets of allowed perturbations must lie in $\mathcal{M}_S(\Theta)$. In this situation, although it is clear that the consistency guaranteed by the BvM Theorem guarantees the correct limit for all perturbations it is not clear these limits are obtained uniformly according to the Definition 2.1 of qualitative robustness. Although it is possible that we have qualitative robustness in this case, it is also possible that some modification of the perturbation metrics is needed to obtain it. The interaction between the quantitative knowledge of the BvM theorem and the perturbation metrics leading to qualitative robustness should be illuminating.

Let us furthermore continue this line of inquiry in the context of the setting of the introduction, where numerical approximated posteriors are then used as priors in a new iteration. If standard numerical approximations of the posteriors are used in the approximation of the posteriors, then the Definition 4.4 of $\Pi_\rho^S(\theta)$ will have to be modified to reflect this fact, but if these computations cannot protect against the TV perturbed priors that generate non robustness, then we suspect that qualitative robustness in this case may not be available. On the other hand, if the model P is such that the process of Bayesian conditioning leaves the class $\mathcal{M}_S(\Theta)$ invariant or almost invariant, and a numerical approximation scheme is used which utilizes this fact is employed, then these non-robustness mechanisms may be avoided and possibly qualitative robustness obtained. We expect such results to depend on the actual model and its interaction with the admissible set of prior/data-generating distribution pairs including allowed perturbations, along with the other user specified quantities, in particular, the numerical approximation methods of the posteriors and the quantitative information available in the BvM theorem.

5. Mechanisms generating non-robustness

For the clarity of the paper, in this subsection, we illustrate some of the mechanisms generating non qualitative robustness in Bayesian inference, which complement the mechanism discovered by Hampel ([38], Lem. 3) and Cuevas ([19], Thm. 1). These mechanisms do not utilize misspecification. Those which do are discussed in Section 5.1. The core mechanism is derived from the nature of both the assumptions and assertions of results supporting consistency. More precisely, Corollary 3.2 states that if the data-generating distribution is $\mu = P(\theta^*)$ and if the prior π attributes positive mass to every Kullback–Leibler neighborhood of θ^*, then the posterior distribution converges towards δ_{θ^*} as $n \to \infty$. The assumption that π attributes positive mass to every Kullback–Leibler neighborhood of θ^* does not require π to place a significant amount of mass around θ^*, but instead can be satisfied with an arbitrarily small amount. Therefore, if, as in Figure 3, π is a prior distribution with support centered around $\theta \neq \theta^*$, but with a very small amount of mass about θ^*, so that it satisfies the assumptions of Corollary 3.2 at θ^*, then π can be slightly perturbed into a π' with support also centered around $\theta \neq \theta^*$, but with no mass about θ^*. In this situation, although π and π' can be made arbitrarily close in total

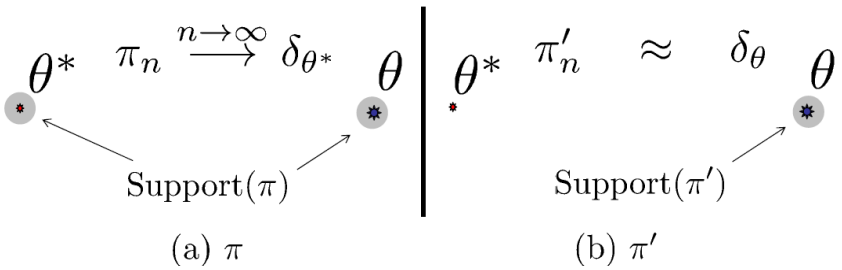
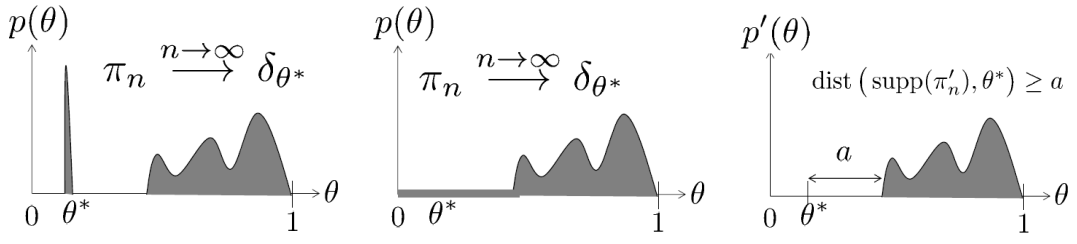


Figure 3. The data-generating distribution is $P(\theta^*)$. π' has most of its mass around θ. π is an arbitrarily small perturbation of π' so that π has Kullback–Leibler support at θ^*. Corollary 3.2 implies that π_n converges towards δ_{θ^*} while π'_n remains close to δ_θ.



(a) Density of π (b) Density of π (c) Density of π'

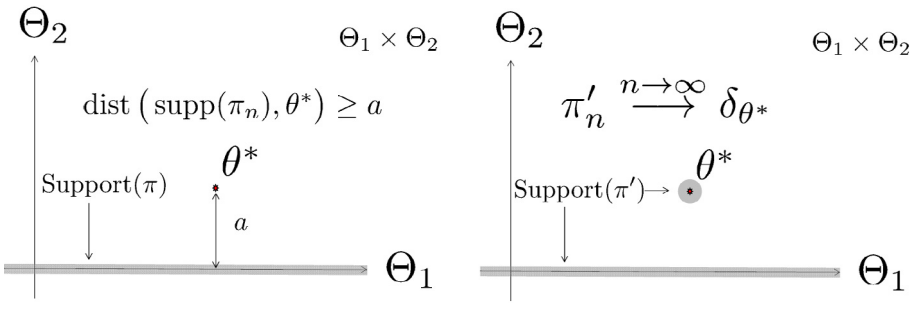
FIGURE 4. The data-generating distribution is $P(\theta^*)$. The probability density functions of π and π' are p and p' with respect to the uniform distribution on $[0,1]$. π is an arbitrarily small perturbation of π' in total variation. π_n converges towards δ_{θ^*} while the distance between the support of π'_n and θ^* remains bounded from below by $a > 0$.

variation distance, the posterior distribution of π converges towards δ_{θ^*} as $n \to \infty$, whereas that of π' remains close to δ_θ. Figure 4 gives an illustration of the same phenomenon when the parameter space Θ is the interval $[0,1]$ and the probability density functions of π and π' are p and p' with respect to the uniform measure.

Note that the mechanism illustrated in Figures 3 and 4 does not generate non qualitative robustness at *all priors* but instead for the full class of *consistency priors*, defined by the assumption of having positive mass on every Kullback–Leibler neighborhood of θ^*. One may wonder whether this non qualitative robustness can be avoided by selecting the prior π to satisfy Cromwell's rule (that is, the assumption that π gives strictly positive mass to every nontrivial open subset of the parameter space Θ). Theorem 4.3 shows that this is not the case if the parameter space Θ is not totally bounded. For example, when $\Theta = \mathbb{R}$, for all $\delta > 0$ one can find $\theta \in \mathbb{R}$ such that the mass that π places on the ball of center θ and radius one is smaller than δ, and by displacing this small amount of mass one obtains a perturbed prior π' whose posterior distribution remains asymptotically bounded away from that of π when the data-generating distribution is $P(\theta)$. Similarly if Θ is totally bounded then Theorem 4.4 places an upper bound on the size of the perturbation of the prior π that would be required as a function of the covering complexity of Θ. Note that these observations suggest that a maximally qualitatively robust prior should place as much mass as possible near all possible candidates θ for the parameter θ^* of the data-generating distribution, thereby reinforcing the notion that a maximally robust prior should have its mass spread as uniformly as possible over the parameter space.

Example 5.1. Let $X := \mathbb{R}$ and $\Theta := \mathbb{R}$ and consider the Gaussian parametric model $P : \mathbb{R} \to \mathcal{M}(\mathbb{R})$ such that the measure $P(\theta)$ is the standard Gaussian measure with mean θ, that is it has the density $\frac{1}{\sqrt{2\pi}} e^{-\frac{-|x-\theta|^2}{2}}, x \in \mathbb{R}$. This model satisfies the conditions of Schwartz' Corollary 3.2. Moreover, because of the regularity of the Gaussian measures it follows that each Kullback–Leibler neighborhood $K_\epsilon(\theta)$, defined in (3.1), contains an open interval $\mathcal{O}_\epsilon(\theta)$ about θ. Consequently, the set $\mathcal{K}(\theta)$ of measures with Kullback–Leibler support at θ defined in (3.2) contains all strictly positive measures, that is those which attribute positive mass to all open intervals.

Let us first consider the situation of Theorem 4.1. Let $\rho > 0$ be small and fix $\theta \in \Theta$ in Definition 4.2 of $\mathcal{Z}_\rho(\theta)$. In particular, $P(\theta)$ is the data-generating distribution and we do not allow any perturbations to it. The discussion above implies that the first component in the total variation uniformity $\Pi_\rho(\theta)$, defined in (4.1), contains all strictly positive measures so that the first component in the corresponding set $\mathcal{Z}_\rho(\theta)$ of admissible prior/data-generating distribution pairs including allowed perturbations, defined in (4.2), also contains all strictly positive measures. By selecting a strictly positive prior π with a very small mass about θ, it follows that a measure π' which is close to π in the TV metric exists and therefore constitutes an admissible second component in $\Pi_\rho(\theta)$, but which has no support an a neighborhood of θ. For the prior π' the corresponding posterior will remain



(a) π (b) π'

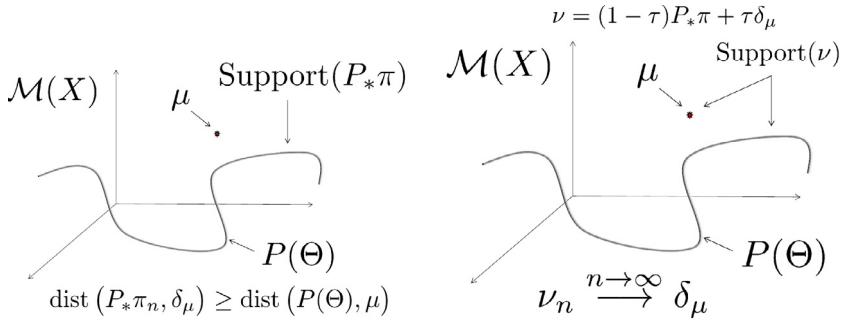
FIGURE 5. Non-robustness caused by misspecification. The parameter space of the model is Θ_1. We assume that the model $P : \Theta_1 \to \mathcal{M}(X)$ is the restriction of an injective model $\bar{P} : \Theta_1 \times \Theta_2 \to \mathcal{M}(X)$ to $\Theta_1 \times \{\theta_2 = 0\}$. The data-generating distribution is $\bar{P}(\theta^*)$ where $\theta^* := (\theta_1^*, \theta_2^*)$, with $\theta_2^* \neq 0$, so that the model P is misspecified. π satisfies Cromwell's rule. π' is an arbitrarily small perturbation of π having Kullback−Leibler support at θ^*. Corollary 3.2 implies that π'_n converges towards δ_{θ^*} while the distance between the support of π_n and θ^* remains bounded from below by $a > 0$.

uniformly bounded away from δ_θ in the TV metric and therefore the Ky Fan metric. This is the mechanism establishing the assertion of Theorem 4.1.

Now let us consider situation of Theorem 4.3. Fix $\delta > 0$, let ρ be small, and consider a prior π which has Kullback−Leibler support for all $\theta \in \mathbb{R}$. By the discussion at the beginning, we can choose π to be any strictly positive measure. Since $\mathbb{R}$ is not completely bounded it follows that there exists a θ such that the measure of the interval about θ of size δ is small enough that there exists a measure π' within TV distance ρ of π whose support does not intersect this interval. Selecting $P(\theta)$ as the data-generating distribution in the admissible set $\mathcal{Z}_\rho(\pi)$, it again follows that the posterior distributions resulting from the prior π' stay uniformly a distance δ from δ_θ in the TV metric and therefore the Ky Fan metric. This the mechanism establishing the assertion of Theorem 4.3.

5.1. Robustness under misspecification

Although the main results of Section 4 do not utilize any model misspecification, the brittleness results of [48] suggest that misspecification should also generate non qualitative robustness. Indeed, although, one may find a prior that is both consistent and *qualitatively robust* when Θ is totally bounded and the model is well-specified, we now show how extensions of the mechanism illustrated in Figures 3 and 4 suggest that misspecification implies non *qualitative robustness*. Consider the example illustrated in Figure 5. In this example the model P is the restriction of a well specified larger model $\bar{P} : \Theta_1 \times \Theta_2 \to \mathcal{M}(X)$ to $\theta_2 = 0$. Assume that the data-generating distribution is $\bar{P}(\theta_1^*, \theta_2^*)$ where $\theta_2^* \neq 0$, so that the restricted model P is misspecified. Let π be any prior distribution on $\Theta_1 \times \{\theta_2 = 0\}$. Although π may satisfy Cromwell's rule the mechanisms presented in this paper suggest that is not *qualitatively robust* with respect to perturbed priors having support on $\Theta_1 \times \Theta_2$. Indeed, let π' be an arbitrarily small perturbation of π obtained by removing some mass from the support of π and adding that mass around θ^*. Note that π' can be chosen arbitrarily close to π while satisfying the local consistency assumption of Corollary 3.2, which implies that the posterior distributions of π' concentrate on θ^* while the posterior distributions of π remain supported on $\Theta_1 \times \{\theta_2 = 0\}$. Note that if $\bar{P}$ is interpreted as an extension of the model P, then this mechanism suggests that we can establish conditions under which Bayesian inference is not *qualitatively robust under model extension*.



(a) $P_*\pi$ (b) ν

FIGURE 6. Non-robustness caused by misspecification. The parameter space of the model is Θ. The data-generating distribution is $\mu \notin P(\Theta)$, so that the model is misspecified. ν is an arbitrarily small perturbation of $P_*\pi$ in total variation distance having non-zero mass on μ. Note that if a small mass on μ is sufficient to ensure the consistency of ν then such a mechanism also implies the non-robustness of the model with respect to misspecification since in such a case, ν_n, the posterior distribution of ν would converge towards δ_μ whereas the distance between the support of $P_*\pi_n$ and μ remains bounded from below by the distance from the model to the data-generating distribution.

Figure 6 represents a non-parametric generalization of the mechanism of Figure 5. Assume that the data-generating distribution is $\mu \notin P(\Theta)$, so that the model is misspecified. Let $\pi \in \mathcal{M}(\Theta)$ be an arbitrary prior distribution and $P_*\pi \in \mathcal{M}^2(X)$ its corresponding non-parametric prior. By removing an arbitrarily small amount of mass from $P_*\pi$ and placing it on μ one obtains an arbitrarily close prior distribution ν that is consistent with respect to the data-generating distribution μ. Therefore although $P_*\pi$ and ν may be made arbitrarily close, their posterior distributions would remain asymptotically separated by a distance corresponding to the degree of misspecification of the model (the distance from μ to $P(\Theta)$).

6. Proofs

6.1. Proof of Lemma 3.1

Dupuis and Ellis ([26], Lem. 1.4.3) assert that on a Polish space that K is lower semicontinuous in both arguments. Since the subset embedding : $X \to X'$ of a subset X of a metric space X' is isometric, when X is a Borel subset of a separable metric space X', it can be shown that the induced pushforward map $i_* : \mathcal{M}(X) \to \mathcal{M}(X')$ is isometric in the Prokhorov metrics, in particular it is continuous. Since the composition of a continuous and a lower semicontinuous function is lower semicontinuous, it follows from Dupuis and Ellis ([26], Lem. 1.4.3) that on any realization of a standard Borel space that the Kullback−Leibler divergence is lower semicontinuous in each of its arguments separately, in particular, fixing the first, it is lower semicontinuous. Therefore $K_\epsilon(P_{\theta^*}) \subset \mathcal{M}(X)$ is closed, and therefore measurable for $\epsilon > 0$. Consequently, when P is measurable, it follows that $K_\epsilon(\theta^*) \subset \Theta$ is measurable for $\epsilon > 0$.

6.2. Proof of Corollary 3.2

We seek to apply Schwartz' theorem ([56], Thm. 6.1). Since Θ and X are separable metric spaces, their Borel σ-algebras are countably generated, so Doob's Theorem ([23], Thm. V.58) and the measurability of the dominated model P implies that a family of densities can be chosen to be $\mathcal{B}(X) \times \mathcal{B}(\Theta)$ measurable, thus satisfying this requirement of ([56], Thm. 6.1). Since U is a neighborhood it follows that it contains an open

neighborhood O of θ. Since O is open and $P : \Theta \to P(\Theta)$ is open, it follows that $P(O)$ is open in $P(\Theta)$, and therefore there is an open set $V_* \subset \mathcal{M}(X)$ such that $V_* \cap P(\Theta) = P(O)$. Moreover, V_* is an open neighborhood of P_{θ^*}. Since X is a separable metric space, it follows that d_{PrX} metrizes the weak topology, and since V_* is open, it is well known (see *e.g.* [3, 30, 63]) that there exists a uniformly consistent test of P_{θ^*} against V_*^c, see Schwartz [56] for the definition of uniformly consistent test. It follows trivially that there exists a uniformly consistent test of P_{θ^*} against $V_*^c \cap P(\Theta)$. Moreover, since P is injective it follows that $O^c = P^{-1}(V_*^c)$. Therefore, there exists a uniformly consistent test of P_{θ^*} against $V_*^c \cap P(\Theta) = \{P_\theta : \theta \in O^c\}$.

Since V_* is open, it also follows that there is a Prokhorov metric ball $B_s(P_{\theta^*})$ of radius $s > 0$ about P_{θ^*} such that $B_s(P_{\theta^*}) \subset V_*$. Now consider the Kullback–Leibler ball $K_\tau(P_{\theta^*})$ for $\tau < \frac{s^2}{2}$. It follows from Csiszar, Kemperman and Kullback's [18] improvement $K \geq \frac{1}{2} d_{tv}^2$ of Pinsker's inequality and the inequality $d_{tv} \geq d_{PrX}$, that $K_\tau(P_{\theta^*}) \subset B_s(P_{\theta^*})$. Since then $K_\tau(P_{\theta^*}) \subset B_s(P_{\theta^*}) \subset V_*$ it follows that

$$P^{-1}\big(K_\tau(P_{\theta^*})\big) \subset P^{-1}(V_*) = O.$$

Consider now the Kullback–Leibler neighborhood $W_\tau(\theta^*) \subset \Theta$ of θ^* defined by pulling $K_\tau(P_{\theta^*})$ back to Θ by the model P:

$$W_\tau(\theta^*) := P^{-1}\big(K_\tau(P_{\theta^*})\big).$$

Then the previous inequality states that

$$W_\tau(\theta^*) \subset O.$$

Since the Kullback–Leibler neighborhoods are measurable in the weak topology and P is assumed measurable, it follows that $W_\tau(\theta^*)$ is measurable.

Therefore, O and $W_\tau(\theta^*)$ satisfy the assumptions of the sets V and and W in ([56], Thm. 6.1). Consequently, since by assumption, the prior π has Kullback–Leibler support, it follows that we can apply Schwartz' theorem ([56], Thm. 6.1) to obtain the assertion for O and since $U \supset O$ is measurable the assertion follows.

6.3. Proof of Theorem 4.1

Let us prove the assertion for a weaker pseudometric $\acute{\alpha}_\mu \leq \alpha_\mu$ derived from the Prokhorov metric on $d_{Pr^2\Theta}$ on $\mathcal{M}^2(\Theta)$. Since it is weaker the assertion follows. To that end, consider $\mu \in \mathcal{M}(X)$. Then for two random variables $Z, W : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$ it follows that $Z_*\mu^\infty, W_*\mu^\infty \in \mathcal{M}^2(\Theta)$, so we can define a pseudometric $\acute{\alpha}_\mu$ by

$$\acute{\alpha}_\mu(Z, W) := d_{Pr^2\Theta}(Z_*\mu^\infty, W_*\mu^\infty). \tag{6.1}$$

Since Dudley ([25], Thm. 11.3.5) asserts that

$$d_{Pr^2\Theta}(Z_*\mu^\infty, W_*\mu^\infty) \leq \alpha_\mu(Z, W), \tag{6.2}$$

we conclude that

$$\acute{\alpha}_\mu \leq \alpha_\mu. \tag{6.3}$$

For fixed π and μ and n, the $\mathcal{M}(\Theta)$-valued random variable

$$\pi_n : (X^\infty, \mu^\infty) \to \mathcal{M}(\Theta)$$

defined by $\pi_n(x^\infty) := \pi_{x^n}$ in (2.2) satisfies

$$(\pi_n)_*\mu^\infty = \pi_*\mu^n \tag{6.4}$$

where $\pi_* : \mathcal{M}(X^n) \to \mathcal{M}^2(\Theta)$ is the pushforward operator corresponding to the map $\bar{\pi} : X^n \to \mathcal{M}(\Theta)$ defined by $\bar{\pi}(x^n) := \pi_{x^n}$ in (2.1). Consequently, we obtain

$$\acute{\alpha}_\mu(\pi_n, \acute{\pi}_n) = d_{Pr^2\Theta}(\pi_*\mu^n, \acute{\pi}_*\mu^n) \tag{6.5}$$

for $\pi, \acute{\pi} \in \mathcal{M}(\Theta)$ and n fixed. From the triangle inequality we then obtain

$$\begin{aligned} d_{Pr^2\Theta}(\pi_*\mu^n, \acute{\pi}_*\acute{\mu}^n) &\leq d_{Pr^2\Theta}(\pi_*\mu^n, \acute{\pi}_*\mu^n) + d_{Pr^2\Theta}(\acute{\pi}_*\mu^n, \acute{\pi}_*\acute{\mu}^n) \\ &\leq \alpha_\mu(\pi_n, \acute{\pi}_n) + d_{Pr^2\Theta}(\acute{\pi}_*\mu^n, \acute{\pi}_*\acute{\mu}^n)\,, \end{aligned}$$

bounding the simple single term $d_{Pr^2\Theta}(\pi_*\mu^n, \acute{\pi}_*\acute{\mu}^n)$ in terms of the sum of two terms $\alpha_\mu(\pi_n, \acute{\pi}_n)$ and $d_{Pr^2\Theta}(\acute{\pi}_*\mu^n, \acute{\pi}_*\acute{\mu}^n)$ of qualitative robustness in Definition 2.1. Consequently, the assumption of the pseudo-metric $\acute{\alpha}_\mu$ amounts to Definition 2.1 with one epsilon instead of two corresponding to the metric

$$d_{Pr^2\Theta}(\pi_*\mu^n, \acute{\pi}_*\acute{\mu}^n), \quad (\pi, \acute{\pi}, \mu, \acute{\mu}) \in \mathcal{M}(\Theta)^2 \times \mathcal{M}(X)^2. \tag{6.6}$$

Moreover, non qualitative robustness with respect to this definition implies non qualitative robustness with respect to the original Definition 2.1 with the Ky Fan metric.

Now lets turn to the proof that the inference is non qualitatively robust with respect to the objective metric (6.6). Fix $\theta^* \in \Theta$ and consider another point $\theta \in \Theta$ and the Dirac mass $\delta_\theta \in \mathcal{M}(\Theta)$ situated at θ. For $\pi \in \mathcal{K}(\theta^*)$, the convex combination

$$\pi^\alpha := \alpha\pi + (1-\alpha)\delta_\theta$$

is a probability measure with Kullback–Leibler support, that is, $\pi^\alpha \in \mathcal{K}(\theta^*)$, $\alpha > 0$ and

$$d_{tv}(\pi^\alpha, \delta_\theta) \leq \alpha. \tag{6.7}$$

Therefore, it follows that

$$(\pi^\alpha, \delta_\theta) \in \Pi_\rho(\theta^*)\,, \quad \alpha < \rho\,,$$

and therefore

$$\left(\pi^\alpha, \delta_\theta, P_{\theta^*}, P_{\theta^*}\right) \in Z_\rho(\theta^*)\,, \quad \alpha < \rho\,,$$

where $\mathcal{Z}_\rho(\theta^*)$ is the admissible set defined in (4.2).

For the prior π^α, let $\pi_n^\alpha : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n^\alpha(x^\infty) := \pi_{x^n}^\alpha$, $x^\infty \in X^\infty$, denote the corresponding sequence of posterior random variables, and let $(\pi_n^\alpha)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$ denote its induced sequence of laws. On the other hand, for the prior δ_θ, it is easy to see that $(\delta_\theta)_{x^n} = \delta_\theta$, $x^n \in X^n$, so that if we denote the corresponding sequence of posterior random variables by δ_θ^n, then $(\delta_\theta^n)_* P_{\theta^*}^\infty = (\delta_\theta)_* P_{\theta^*}^n = \delta_{\delta_\theta}$.

Since the assumptions of Schwartz' Corollary 3.2 are satisfied and π^α has Kullback–Leibler support at θ^*, we can apply the assertion (A.4) of Proposition A.1

$$P_{\theta^*}^\infty \left\{ d_{Pr\Theta}(\pi_n^\alpha, \delta_{\theta^*}) > \epsilon \right\} \to 0 \quad n \to \infty\,,$$

for $\epsilon > 0$. To complete the proof we simply use the fact that convergence in law to a Dirac mass is equivalent to convergence in probability to a constant random variable, that is use the equivalent assertion (A.5) of Proposition A.1

$$d_{Pr^2\Theta}\left((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right) \to 0 \quad n \to \infty. \tag{6.8}$$

Now the proof is very simple. Indeed, from the triangle inequality we have

$$d_{Pr^2\Theta}\left((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_\theta}\right) \geq d_{Pr^2\Theta}\left(\delta_{\delta_{\theta^*}}, \delta_{\delta_\theta}\right) - d_{Pr^2\Theta}\left((\pi_n^\alpha)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right)$$

and, by two applications of Proposition A.4, we have

$$\begin{aligned} d_{Pr^2\Theta}\left(\delta_{\delta_{\theta^*}}, \delta_{\delta_\theta}\right) &= \min\left(d_{Pr\Theta}\left(\delta_{\theta^*}, \delta_\theta\right), 1\right) \\ &= \min\left(\min\left(d(\theta^*, \theta), 1\right), 1\right) \\ &= \min\left(d(\theta^*, \theta), 1\right). \end{aligned}$$

Therefore, since $(\delta_\theta^n)_* P_{\theta^*}^\infty = \delta_{\delta_\theta}$, the convergence (6.8) implies that

$$d_{Pr^2\Theta}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, (\delta_\theta^n)_* P_{\theta^*}^\infty\Big) \to \min\big(d(\theta^*,\theta), 1\big), \quad n \to \infty.$$

Finally, since $d_{Pr\Theta} \le d_{tv}$, it follows from (6.7) that

$$d_{Pr\Theta}(\pi^\alpha, \delta_\theta) \le \alpha.$$

Then, for any $\delta > 0$, if we restrict α so that $\alpha < \min(\delta, \rho)$, it follows that $d_{tv}(\pi^\alpha, \delta_\theta) < \rho$ and $d_{Pr\Theta}(\pi^\alpha, \delta_\theta) < \delta$, so that

$$\big(\pi^\alpha, \delta_\theta, P_{\theta^*}, P_{\theta^*}\big) \in \mathcal{Z}_\rho(\theta^*), \tag{6.9}$$

$$d_{Pr\Theta}(\pi^\alpha, \delta_\theta) < \delta. \tag{6.10}$$

Let $D := \sup\{d(\theta_1,\theta_2) : \theta_1, \theta_2 \in \Theta\}$ denote the diameter of Θ. Then it follows from the triangle inequality that, for any $\epsilon > 0$, there exists a $\theta \in \Theta$ such that $d(\theta^*, \theta) \ge \frac{D}{2} - \epsilon$. Consequently, for any $\bar\epsilon < \min(\frac{D}{2}, 1)$, no matter how small δ is, there is an $\alpha > 0$ such that, in addition to (6.9) and (6.10), we have

$$d_{Pr^2\Theta}\Big((\pi_n^\alpha)_* P_{\theta^*}^\infty, (\delta_\theta^n)_* P_{\theta^*}^\infty\Big) > \bar\epsilon,$$

for large enough n. Consequently, the assertion is proved.

6.4. Proof of Theorem 4.4

As in the proof of Theorem 4.1, we establish the assertion with respect to the modified form of qualitative robustness defined by (6.6), and since this form is weaker it implies the assertion. It follows from the definition of the packing numbers that, for $\epsilon > 0$, there is a packing $\{\theta_i, i = 1, .., \mathcal{M}_{2\epsilon}(\Theta)\}$ and therefore the collection of open balls $B_\epsilon(\theta_i)$, $i = 1, .., \mathcal{M}_{2\epsilon}(\Theta)$ is a disjoint union. Denoting $\mathcal{N}_{2\epsilon} := \mathcal{N}_{2\epsilon}(\Theta)$ and $\mathcal{M}_{2\epsilon} := \mathcal{M}_{2\epsilon}(\Theta)$, we therefore obtain

$$\begin{aligned} 1 &= \pi(\Theta) \\ &\ge \pi\big(\cup_{i=1}^{\mathcal{M}_{2\epsilon}} B_\epsilon(\theta_i)\big) \\ &= \sum_{i=1}^{\mathcal{M}_{2\epsilon}} \pi\big(B_\epsilon(\theta_i)\big) \\ &\ge \mathcal{M}_{2\epsilon} \min_{i=1,\mathcal{M}_{2\epsilon}} \pi\big(B_\epsilon(\theta_i)\big). \end{aligned}$$

Consequently, since (4.3) implies $\mathcal{M}_{2\epsilon} \ge \mathcal{N}_{2\epsilon}$, there exists a point $\theta^* \in \Theta$ such that

$$\pi\big(B_\epsilon(\theta^*)\big) \le \frac{1}{\mathcal{N}_{2\epsilon}}. \tag{6.11}$$

Let $B_\epsilon := B_\epsilon(\theta^*)$ denote the open ball about θ^* and let B_ϵ^c denote its complement. Let $\pi^\epsilon \in \mathcal{M}(\Theta)$, defined by

$$\pi^\epsilon(B) := \frac{\pi(B_\epsilon^c \cap B)}{\pi(B_\epsilon^c)}, \; B \in \mathcal{B}(\Theta),$$

denote the normalization of the restriction of π to B_ϵ^c which, by the inequality (6.11), is well defined. Since $\pi = \pi(B_\epsilon^c)\pi^\epsilon + \pi|_{B_\epsilon}$ it follows that $\pi - \pi^\epsilon = \pi|_{B_\epsilon} - \pi(B_\epsilon)\pi^\epsilon$ so that we obtain

$$d_{tv}(\pi^\epsilon, \pi) \le \pi(B_\epsilon) \le \frac{1}{\mathcal{N}_{2\epsilon}}$$

from which we obtain

$$d_{Pr\Theta}(\pi^\epsilon, \pi) \leq \frac{1}{\mathcal{N}_{2\epsilon}}. \tag{6.12}$$

In particular, when $\frac{1}{\mathcal{N}_{2\epsilon}} < \rho$, we obtain

$$\pi^\epsilon \in B_\rho^{tv}(\pi)$$

and therefore

$$\left(\pi, \pi^\epsilon, P_{\theta^*}, P_{\theta^*}\right) \in Z_\rho(\pi)\,.$$

That is, when $\frac{1}{\mathcal{N}_{2\epsilon}} < \rho$, the point $\left(\pi, \pi^\epsilon, P_{\theta^*}, P_{\theta^*}\right) \in Z_\rho(\pi)$.

For the prior π^ϵ, let $\pi_n^\epsilon : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n^\epsilon(x^\infty) := \pi_{x^n}^\epsilon$, $x^\infty \in X^\infty$, denote the corresponding sequence of posterior random variables, and let $(\pi_n^\epsilon)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$ denote its induced sequence of laws. Since the assumptions of Schwartz' Corollary 3.2 are satisfied and π has Kullback−Leibler support at θ^*, we can apply the assertion (A.5) of Proposition A.1 to the sequence of posterior laws $(\pi_n)_* P_{\theta^*}^\infty$ corresponding to π:

$$d_{Pr^2\Theta}\left((\pi_n)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right) \to 0 \quad n \to \infty. \tag{6.13}$$

From the triangle inequality we have

$$d_{Pr^2\Theta}\left((\pi_n)_* P_{\theta^*}^\infty, (\pi_n^\epsilon)_* P_{\theta^*}^\infty\right) \geq d_{Pr^2\Theta}\left((\pi_n^\epsilon)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right) - d_{Pr^2\Theta}\left((\pi_n)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right), \tag{6.14}$$

so to lower bound the lefthand side it is sufficient in the limit to lower bound the first term on the right. To that end, we use a quantitative version of the partial converse ([25], Thm. 11.3.5) of convergence in probability implies convergence in law, valid when the convergence in law is to a Dirac mass. Indeed, if we denote the Ky Fan metric determined from the measure $P_{\theta^*}^\infty$ by α_{θ^*}, Lemma A.3 asserts that

$$d_{Pr^2\Theta}\left((\pi_n^\epsilon)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right) = \alpha_{\theta^*}(\pi_n^\epsilon, \delta_{\theta^*}). \tag{6.15}$$

To evaluate the Ky Fan distance on the righthand side, first observe that since π^ϵ has support contained in the closed set B_ϵ^c, it follows from Schervish ([55], Thm. 1.31) that $\pi_{x^n}^\epsilon$ also has support contained in B_ϵ^c a.e $P_{\theta^*}^n$. Therefore, if we define $B_0 := \{\theta^*\}$ and $B_r := B_r(\theta^*)$, it follows that $B_0^r = B_r$, so that

$$\pi_{x^n}^\epsilon(B_0^r) = 0\,, \quad a.e.\ P_{\theta^*}^n, \quad r < \epsilon$$

and

$$(\delta_{\theta^*})_{x^n}(B_0) = 1\,, \quad a.e.\ P_{\theta^*}^n.$$

It follows from Lemma A.2 that

$$d_{Pr\Theta}\left(\pi_{x^n}^\epsilon, (\delta_{\theta^*})_{x^n}\right) \geq \min(\epsilon, 1) \quad a.e.\ P_{\theta^*}^\infty\,,$$

and, since $\epsilon \leq 1$, we obtain

$$P_{\theta^*}^\infty\left(d_{Pr\Theta}\left(\pi_{x^n}^\epsilon, (\delta_{\theta^*})_{x^n}\right) \geq \epsilon\right) = 1.$$

Therefore, by the definition (A.2) of the Ky Fan metric, we obtain $\alpha_{\theta^*}(\pi_n^\epsilon, \delta_{\theta^*}) \geq \epsilon$ and, by the identity (6.15), we conclude that

$$d_{Pr^2\Theta}\left((\pi_n^\epsilon)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}}\right) \geq \epsilon.$$

Consequently, from the triangle inequality (6.14) and the convergence (6.13), we conclude, for any $\acute{\epsilon} > 0$, that for large enough n we have

$$d_{Pr^2\Theta}\left((\pi_n)_* P_{\theta^*}^\infty, (\pi_n^\epsilon)_* P_{\theta^*}^\infty\right) \geq \epsilon - \acute{\epsilon}. \tag{6.16}$$

Consequently, if this Bayesian inference is qualitatively robust, then for $\epsilon > 0$, it follows from (6.16) and (6.12) that $\delta < \frac{1}{\mathcal{N}_{2\epsilon}}$. The requirement that perturbations be admissible, that is determine members in $\mathcal{Z}_\rho(\pi)$, implies that $\delta < \rho$.

Appendix A.

A.1. Metrics on spaces of measures and random variable and the measurability of conditioning

Metrics on spaces of measures and random variables is a well studied field, see *e.g.* Rachev *et al.* [51] and Gibbs and Su [32], but to keep the presentation simple, here we will restrict our attention to the total variation, Prokhorov and Ky Fan metrics (we refer to [6–9] for motivations for considering classes of priors defined in the Prokhorov metric). For measurable spaces Θ and X, we write $\mathcal{M}(\Theta)$ and $\mathcal{M}(X)$ for the set of probability distributions on Θ and X respectively. In this general setting, we can metrize the spaces of measures $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$ using total variation. This latter metrization makes $\mathcal{M}(\Theta)$ into a topological space whose Borel structure can be used to define the space $\mathcal{M}^2(\Theta)$ of probability measures on $\mathcal{M}(\Theta)$, which can also be metrized using the total variation. However, the separability of these spaces will be extremely useful for us and, in general, these spaces will not be separable under the total variation metric. Recall that for a metric space (S, d), the Prokhorov metric d_{Pr} on the space $\mathcal{M}(S)$ of Borel probability measures is defined by

$$d_{Pr}(\mu_1, \mu_2) := \inf\left\{\epsilon : \mu_1(A) \leq \mu_2(A^\epsilon) + \epsilon,\ A \in \mathcal{B}(S)\right\}, \quad \mu_1, \mu_2 \in \mathcal{M}(S)\,, \tag{A.1}$$

where

$$A^\epsilon := \{x' \in S : d(x, x') < \epsilon \text{ for some } x \in A\}\,.$$

According to Dudley ([25], Thm. 11.3.1), the Prokhorov metric is a metric on $\mathcal{M}(S)$. Consequently, when Θ and X are metric, we can also metrize the spaces of measures $\mathcal{M}(\Theta)$ and $\mathcal{M}(X)$ with the Prokhorov metrics, and having done so we can define the space $\mathcal{M}^2(\Theta) := \mathcal{M}(\mathcal{M}(\Theta))$ of Borel probability measures on the metric space $(\mathcal{M}(\Theta), d_{Pr\Theta})$ of Borel probability measures on Θ and metrize it with the Prokhorov metric $d_{Pr^2\Theta}$. Furthermore, when (S, d) is a separable metric space, Dudley ([25], Thm. 11.3.3) asserts that the Prokhorov metric metrizes weak convergence and Aliprantis and Border ([1], Thm. 15.12) asserts that the metric space $(\mathcal{M}(S), d_{Pr})$ is separable.

Therefore, when X and Θ are separable metric spaces, the Prokhorov metrics d_{PrX} and $d_{Pr\Theta}$ metrize weak convergence in $\mathcal{M}(X)$ and $\mathcal{M}(\Theta)$ respectively and both metric spaces $(\mathcal{M}(X), d_{PrX})$ and $(\mathcal{M}(\Theta), d_{Pr\Theta})$ are separable. Consequently, when Θ is a separable metric space, $(\mathcal{M}(\Theta), d_{Pr\Theta})$ is a separable metric space and therefore $(\mathcal{M}^2(\Theta), d_{Pr^2\Theta})$ is a separable metric space.

The separability of $(\mathcal{M}(\Theta), d_{Pr\Theta})$ is sufficient to define the Ky Fan metric on a space of $(\mathcal{M}(\Theta)$-valued random variables. Indeed, for a separable metric space S, probability space (Ω, Σ, P), and two S-valued random variables $Z : \Omega \to S$ and $W : \Omega \to S$, the Ky Fan distance between Z and W, (see *e.g.* Dudley [25], p. 289), is defined as

$$\alpha(Z, W) := \inf\left\{\epsilon \geq 0 : P(d(Z, W) > \epsilon) \leq \epsilon\right\}. \tag{A.2}$$

By Dudley ([25], Thm. 9.2.2), the Ky Fan metric is a metric on the space of S-valued random variables from (Ω, Σ, P) and metrizes convergence in probability for them. Consequently, when Θ is a separable metric space and $\mathcal{M}(\Theta)$ is metrized with the Prokhorov metric $d_{Pr\Theta}$, the Ky Fan metric α of (A.2) metrizes the space of $\mathcal{M}(\Theta)$-valued random variables $Z : (X^\infty, \mu^\infty) \to (\mathcal{M}(\Theta), d_{Pr\Theta})$ for each $\mu \in \mathcal{M}(X)$. Since this family of metrics depends on the measure μ we indicate this dependence by writing α_μ. Moreover, when Θ and X are separable metric spaces, their Borel σ-algebras are countably generated, which is required to apply Doob's Theorem to assert that a dominated measurable model has a jointly measurable family of densities, which is required in the consistency theorem of Schwartz which we will need.

When Θ and X are Borel subsets of Polish metric spaces, they are separable metric spaces so the above applies. Let us now show that the assumption also facilitates the *measurability* of Bayesian conditioning that will be needed to define its qualitative robustness. To that end, from now on let us place as default the weak topologies on $\mathcal{M}(X)$, $\mathcal{M}(\Theta)$ and $\mathcal{M}^2(\Theta)$ and metrize them with the Prokhorov metrics d_{PrX}, $d_{Pr\Theta}$ and $d_{Pr^2\Theta}$. This is primarily to obtain well-defined Bayesian conditioning while at the same time applicability of Schwartz's

consistency theorem. We will also place other metric structures on $\mathcal{M}(X)$, $\mathcal{M}(\Theta)$ and $\mathcal{M}^2(\Theta)$ to quantify the size of perturbations and indicate them with the notation $d_{\mathcal{M}(X)}$, $d_{\mathcal{M}(\Theta)}$ and $d_{\mathcal{M}^2(\Theta)}$. Consider a measurable model $P : \Theta \to \mathcal{M}(X)$. Since Aliprantis and Border ([1], Thm. 15.13) implies that the map $\mathcal{M}(X) \to \mathbb{R}$ defined by $\mu \mapsto \mu(A)$ is Borel measurable for all $A \in \mathcal{B}(X)$, it follows that P corresponds to a Markov kernel. Consider a prior $\pi \in \mathcal{M}(\Theta)$. Then since Θ is assumed to be a Borel subset of a Polish space, it follows from Schervish ([55], Thm. B.46) that there exists a family $\pi_x, x \in X$ of conditional probability measures generated by the model P such that the map $x \mapsto \int_\Theta f d\pi_x, x \in X$ is $\mathcal{B}(X)$-measurable for all bounded and measurable functions $f : \Theta \to \mathbb{R}$. Note that after the proof Schervish mentions that such a family of conditional measures is not unique. Since both Θ and X are separable and metrizable, it then follows from Aliprantis and Border ([1], Thm. 19.7) that the resulting map $x \mapsto \pi_x$ from X to $\mathcal{M}(\Theta)$ is measurable. For multiple samples, it is clear that X^n is a Borel subset of the n-th power of the ambient Polish space of X. By Billingsly's ([10], Thm. 2.8) characterization of weak convergence on product spaces it follows that the injection $\mathcal{M}(X) \to \mathcal{M}(X^n)$ defined by $\mu \mapsto \mu^n$ is continuous, so that it follows that $P^n : \Theta \to \mathcal{M}(X^n)$, defined by $P^n(\theta) = P(\theta)^n, \theta \in \Theta$, is measurable and therefore, by the same arguments as above, we obtain a family of multisample conditional measures $\pi_{x^n}, x^n \in X^n$ such that the resulting map

$$\bar{\pi} : X^n \to \mathcal{M}(\Theta)$$

defined by the determination of the posteriors

$$\bar{\pi}(x^n) := \pi_{x^n}, \quad x^n \in X^n \tag{A.3}$$

is measurable. Therefore, its corresponding pushforward operator

$$\pi_* : \mathcal{M}(X^n) \to \mathcal{M}^2(\Theta)$$

is well-defined, where we have removed the bar over π to simplify the notation, but still emphasize that this pushforward operator π_* corresponds to the prior π. Then to consider how the posteriors π_{x^n} vary as a function of the sample data x^n when it is generated by i.i.d. sampling from μ, since $\mathcal{M}^n(X) \subset \mathcal{M}(X^n)$ it follows that $\mu^n \in \mathcal{M}(X^n)$ so we can utilize the pushforward operator π_* to define

$$\pi_* \mu^n \in \mathcal{M}^2(\Theta)$$

the sampling distribution of the posterior distribution π_{x^n} when $x^n \sim \mu^n$.

A.2. Schwartz' Theorem and the convergence of random measures

It will be useful to express the assertion of Corollary 3.2 and some of its consequences in terms of the convergence of measures and random measures. To that end, recall the notation $\mathcal{M}^2(\Theta) := \mathcal{M}(\mathcal{M}(\Theta))$, and consider the corresponding sequence of random variables $\pi_n : (X^\infty, P_{\theta^*}^\infty) \to \mathcal{M}(\Theta)$, defined by $\pi_n(x^\infty) := \pi_{x^n}$, $x^\infty \in X^\infty$, and its induced sequence of laws $(\pi_n)_* P_{\theta^*}^\infty \in \mathcal{M}^2(\Theta)$. Note especially that $\delta_{\delta_{\theta^*}}$ is the Dirac mass in $\mathcal{M}^2(\Theta)$ situated at the Dirac mass δ_{θ^*} in $\mathcal{M}(\Theta)$ situated at θ^*.

Proposition A.1. *The assertion of Corollary 3.2 is equivalent to*

$$\pi_{x^n} \mapsto \delta_{\theta^*} \quad a.e.\ P_{\theta^*}^\infty,$$

where $\mapsto$ is weak convergence. This in turn implies that

$$P_{\theta^*}^\infty \Big\{ d_{Pr\Theta}(\pi_n, \delta_{\theta^*}) > \epsilon \Big\} \to 0 \quad n \to \infty, \tag{A.4}$$

for $\epsilon > 0$, which is equivalent to

$$d_{Pr^2\Theta}\Big((\pi_n)_* P_{\theta^*}^\infty, \delta_{\delta_{\theta^*}} \Big) \to 0 \quad n \to \infty, \tag{A.5}$$

where $d_{Pr^2\Theta}$ is the Prokhorov metric on $\mathcal{M}^2(\Theta)$ defined with respect to the Prokhorov metric $d_{Pr\Theta}$ on $\mathcal{M}(\Theta)$.

Proof. Let $\mathcal{O}$ denote the open sets in Θ and $\mathcal{O}_{\theta^*} \subset \mathcal{O}$ denote the open neighborhoods of θ^*. Then, under the conditions of Corollary 3.2, for $O \in \mathcal{O}_{\theta^*}$, it follows that

$$\pi_{x^n}(O) \to 1 \quad n \to \infty, \quad \text{a.e. } P_{\theta^*}^{\infty}.$$

Since $\delta_{\theta^*}(O) = 1$, $O \in \mathcal{O}_{\theta^*}$ and $\delta_{\theta^*}(O) = 0$, $O \in \mathcal{O} \setminus \mathcal{O}_{\theta^*}$ it easily follows that

$$\liminf_n \pi_{x^n}(O) \geq \delta_{\theta^*}(O), \ \forall O \in \mathcal{O}, \quad a.e.\ P_{\theta^*}^{\infty}.$$

which, by the Portmanteau theorem ([25], Thm. 11.1.1), is equivalent to

$$\pi_{x^n} \mapsto \delta_{\theta^*} \quad \text{a.e. } P_{\theta^*}^{\infty}.$$

where $\mapsto$ denotes weak convergence.

Now consider the corresponding sequence of random variables $\pi_n : (X^{\infty}, P_{\theta^*}^{\infty}) \to \mathcal{M}(\Theta)$, defined by $\pi_n(x^{\infty}) := \pi_{x^n}$, $x^{\infty} \in X^{\infty}$, and its induced sequence of laws $(\pi_n)_* P_{\theta^*}^{\infty} \in \mathcal{M}^2(\Theta)$. Then $\pi_{x^n} \mapsto \delta_{\theta^*}$ *a.e.* $P_{\theta^*}^{\infty}$ is equivalent to

$$\pi_n \mapsto \delta_{\theta^*} \quad \text{a.s. } P_{\theta^*}^{\infty}.$$

Since Θ is a separable metric space it follows that $\mathcal{M}(\Theta)$ equipped with the Prokhorov metric is a separable metric space. Since a.s. convergence implies convergence in probability for random variables with values in a separable metric space, it follows that

$$\pi_n \mapsto \delta_{\theta^*} \text{ in } P_{\theta^*}^{\infty} - \text{probability},$$

that is,

$$P_{\theta^*}^{\infty}\left\{ d_{Pr\Theta}(\pi_n, \delta_{\theta^*}) > \epsilon \right\} \to 0 \quad n \to \infty.$$

Since $\mathcal{M}(\Theta)$ is a separable metric space it follows that $\mathcal{M}^2(\Theta)$ equipped with the Prokhorov metric is also a separable metric space. Therefore, since on separable metric spaces convergence in probability to a constant valued random variable is equivalent to the weak convergence of the corresponding set of laws to the Dirac mass situated at that value, (see *e.g.* Dudley [25], Prop. 11.1.3), it follows that the convergence in probability, $\pi_n \to \delta_{\theta^*}$ in $P_{\theta^*}^{\infty}$ – probability, is equivalent to the corresponding convergence of laws

$$(\pi_n)_* P_{\theta^*}^{\infty} \mapsto \delta_{\delta_{\theta^*}} \quad n \to \infty.$$

Finally, since the Proprokhorov metric $d_{Pr^2\Theta}$ on $\mathcal{M}^2(\Theta)$ metrizes the weak topology on $\mathcal{M}^2(\Theta) = \mathcal{M}(\mathcal{M}(\Theta))$, it follows that the latter is equivalent to

$$d_{Pr^2\Theta}\left((\pi_n)_* P_{\theta^*}^{\infty}, \delta_{\delta_{\theta^*}} \right) \to 0 \quad n \to \infty. \qquad \square$$

A.3. Some Prokhorov Geometry

We establish a basic mechanism to bound from below the Prokhorov distance between two measures based on the values of the measures on the neighborhood of a single set.

Lemma A.2. *Let Z be a metric space and consider the space $\mathcal{M}(Z)$ of Borel probability measures equipped with the Prokhorov metric. Consider $\mu \in \mathcal{M}(Z)$ and suppose that there exists a set $B \in \mathcal{B}(Z)$ and $\alpha, \delta \geq 0$ such that*

$$\mu(B^{\epsilon}) \leq \delta, \quad \epsilon < \alpha.$$

Then, for any $\mu' \in \mathcal{M}(Z)$, we have

$$d_{Pr}(\mu, \mu') \geq \min\left(\alpha, \mu'(B) - \delta\right).$$

Proof. If $d_{Pr}(\mu_1,\mu_2) \geq \alpha$ the assertion is proved, so let us assume that $d_{Pr}(\mu_1,\mu_2) < \alpha$. Then, denoting $d^* := d_{Pr}(\mu_1,\mu_2)$, it follows from the assumption that $\mu(A^{d^*}) \leq \delta$, so that

$$\begin{aligned}\mu'(A) &\leq \mu(A^{d^*}) + d^* \\ &\leq \delta + d^*\end{aligned}$$

from which we conclude that $\mu'(A) - \delta \leq d^*$. Therefore, either $d_{Pr}(\mu_1,\mu_2) \geq \alpha$ or $d_{Pr}(\mu_1,\mu_2) \geq \mu'(A) - \delta$, proving the assertion. □

Lemma A.3. *Let S be a separable metric space. Then, for an S-valued random variable X we have*

$$\alpha(X,s) = d_{Pr}(\mathcal{L}(X),\delta_s)$$

where α is the Ky Fan metric and s denotes the random variable with constant value s.

Proof. Let us denote $\alpha := \alpha(X,s)$ and $\rho := d_{Pr}(\mathcal{L}(X),\delta_s)$. Define the set $B_0 := \{s\}$ and $B_r := B_r(s), r > 0$ and observe that $B_0^r = B_r, r > 0$. Therefore, by the definition of ρ we have

$$\mathcal{L}(s)(B_0) \leq \mathcal{L}(X)(B_0^\rho) + \rho$$

and since $\mathcal{L}(s)(B_0) = 1$ we obtain

$$\mathcal{L}(X)(B_0^\rho) \geq 1 - \rho$$

from which we obtain $P(d(X,s) \geq \rho) \leq \rho$. Since this implies that

$$P(d(X,s) > \rho) \leq P(d(X,s) \geq \rho) \leq \rho$$

we conclude that $\rho \leq \alpha$. Since Dudley ([25], Thm. 11.3.5) asserts that $\alpha \leq \rho$, the assertion follows. □

Proposition A.4.

$$d_{Pr}(\delta_{x_1},\delta_{x_2}) = \min\big(1, d(x_1,x_2)\big)$$

Proof. Consider the set $B := \{x_1\}$. Then since $B^\epsilon = B_\epsilon(x_1)$, it follows that for $\epsilon < d(x_1,x_2)$ that $x_2 \notin B^\epsilon$. Consequently, since $\delta_{x_1}(B) = 1$, the inequality

$$\delta_{x_1}(B) \leq \delta_{x_2}(B^\epsilon) + \epsilon$$

requires either $\epsilon \geq 1$ or $x_2 \in B^\epsilon$ which implies that $\epsilon \geq d(x_1,x_2)$. Consequently, $d_{Pr}(\delta_{x_1},\delta_{x_2}) \geq \min\big(1, d(x_1,x_2)\big)$. To obtain equality, suppose that $d_{Pr}(\delta_{x_1},\delta_{x_2}) > d(x_1,x_2)$. Then, for any d' which satisfies $d_{Pr}(\delta_{x_1},\delta_{x_2}) > d' > d(x_1,x_2)$ there exists a measurable set B such that

$$\delta_{x_1}(B) > \delta_{x_2}(B^{d'}) + d'$$

Consequently, $x_1 \in B$, but $d' > d(x_1,x_2)$ implies that $x_2 \in B^{d'}$, which implies the contradiction $1 > 1 + d'$. □

Acknowledgements. The authors gratefully acknowledges this work supported by the Air Force Office of Scientific Research and the DARPA EQUiPS Program under awards number FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators) and number FA9550-16-1-0054 (Computational Information Games). The authors also gratefully acknowledge the thoughtful comments and concerns expressed by the referees, which resulted in subtantial improvement in the both the substance and the style of the paper.

References

[1] C.D. Aliprantis and K.C. Border, Infinite Dimensional Analysis: A Hitchhiker's Guide. Springer, Berlin, 3rd edition (2006).

[2] E. Atkins, M. Morzfeld and A. J. Chorin, Implicit particle methods and their connection with variational data assimilation. *Monthly Weather Rev.* **141** (2013) 1786–1803.

[3] A. Barron, M.J. Schervish and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** (1999) 536–561.

[4] S. Basu, S.R. Jammalamadaka and W. Liu. Stability and infinitesimal robustness of posterior distributions and posterior quantities. *J. Stat. Plann. Inference* **71** (1998) 151–162.

[5] R.H. Berk, Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* **37** (1966) 51–58; correction, ibid **37** (1966) 745–746.

[6] B. Betrò, Numerical treatment of Bayesian robustness problems. *Internat. J. Approx. Reason.* **50** (2009) 279–288.

[7] B. Betrò and A. Guglielmi. Numerical robust Bayesian analysis under generalized moment conditions. In Bayesian robustness (Rimini, 1995), volume 29 of *IMS Lecture Notes Monogr. Ser.* 3–20. With a discussion by Elías Moreno and a rejoinder by the authors. Inst. Math. Statist., Hayward, CA (1996).

[8] B. Betrò and A. Guglielmi, Methods for global prior robustness under generalized moment conditions. In Robust Bayesian analysis, Vol. 152 of *Lecture Notes in Statist.* Springer, New York (2000) 273–293.

[9] B. Betrò, F. Ruggeri and M. Męczarski, Robust Bayesian analysis under generalized moments conditions. *J. Stat. Plann. Inference* **41** (1994) 257–266.

[10] P. Billingsley, Convergence of Probability Measures. Wiley, New York, 2nd edition (1999).

[11] G. Boente, R. Fraiman and V.J. Yohai, Qualitative robustness for stochastic processes. *Ann. Statist.* **46** (1987) 1293–1312.

[12] T. Bui-Thanh and O. Ghattas, An analysis of infinite dimensional Bayesian inverse shape acoustic scattering and its numerical approximation. *SIAM/ASA J. Uncertain. Quantif.* **2** (2014) 203–222.

[13] I. Castillo and R. Nickl, Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** (2013) 1999–2028.

[14] I. Castillo and R. Nickl, On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** (2014) 1941–1969.

[15] A.J. Chorin, M. Morzfeld and X. Tu, Implicit particle filters for data assimilation. *Commun. Appl. Math. Comput. Sci.* **5** (2010) 221–240.

[16] A. J. Chorin and X. Tu, Implicit sampling for particle filters. *Proc. National Acad. Sci.* **106** (2009) 17249–17254.

[17] K. Csilléry, M.G.B. Blum, O.E. Gaggiotti and O. François, Approximate Bayesian computation (abc) in practice. *Trends Ecol. Evol.* **25** (2010) 410–418.

[18] I. Csiszár, I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* (1975) 146–158.

[19] A. Cuevas. Qualitative robustness in abstract inference. *J. Statist. Plan. Inference* **18** (1988) 277–289.

[20] A. Cuevas, Comment on 'Bounds on posterior expectations for density bounded class with constant bandwidth' by Sivaganesan. *J. Statist. Plan. Inference* **40** (1994) 340–343.

[21] A. Cuevas González, Una definición de robustez cualitativa en inferencia Bayesiana. *Trabajos de Estadística y de Investigación Operativa* **35** (1984) 170–186.

[22] M. Dashti and A.M. Stuart, Uncertainty quantification and weak approximation of an elliptic inverse problem. *SIAM J. Numer. Anal.* **49** (2011) 2524–2542.

[23] C. Dellacherie and P.-A. Meyer, Probabilities and Potential. B. Vol. 72 of North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam (1982).

[24] A. Doucet, N. De Freitas and N. Gordon, An introduction to sequential Monte Carlo methods. In Sequential Monte Carlo Methods in Practice. Springer (2001) 3–14.

[25] R.M. Dudley, Real Analysis and Probability. Vol. 74 of *Cambridge Studies in Advanced Mathematics*. Revised reprint of the 1989 original. Cambridge University Press, Cambridge (2002).

[26] P. Dupuis and R.S. Ellis, A Weak Convergence Approach to the Theory of Large Deviations. Vol. 902. John Wiley and Sons (2011).

[27] D. Freedman, On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** (1999) 1119–1140.

[28] Y. Gal and Z. Ghahramani, Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Inter. Conf. Machine Learning PLMR* **48** (2016) 1050–1059.

[29] A. Gelman, Inference and monitoring convergence. In *Markov chain Monte Carlo in practice*. Springer (1996) 131–143.

[30] S. Ghosal, J.K. Ghosh and R.V. Ramamoorthi, Consistency issues in Bayesian nonparametrics. *Stat. Textbooks Monogr.* **158** (1999) 639–668.

[31] A.L. Gibbs, Convergence in the Wasserstein metric for Markov chain Monte Carlo algorithms with applications to image restoration. *Stoch. Models* **20** (2004) 473–492.

[32] A.L. Gibbs and F.E. Su, On choosing and bounding probability metrics. *Inter. Statist. Rev.* **70** (2002) 419–435.

[33] I.J. Goodfellow, J. Shlens and C. Szegedy, Explaining and harnessing adversarial examples. *Inter. Confer. Learning Representations*. Preprint arXiv:1412.6572 (2015).

[34] N.J. Gordon, D.J. Salmond and A.F.M. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation. Radar and Signal Processing. *IEE Proc. F* **140** (1993) 107–113. IET.

[35] P. Gustafson and L. Wasserman, Local sensitivity diagnostics for Bayesian inference. *Ann. Statist.* **23** (1995) 2153–2167.
[36] R. Hable and A. Christmann, On qualitative robustness of support vector machines. *J. Multivariate Anal.* **102** (2011) 993–1007.
[37] R. Hable and A. Christmann, Robustness versus consistency in ill-posed classification and regression problems. In *Classification and Data Mining*. Springer (2013) 27–35.
[38] F.R. Hampel, A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, pages 1887–1896, 1971.
[39] P.J. Huber and E.M. Ronchetti, *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley and Sons Inc., Hoboken, NJ, 2nd edition (2009).
[40] A.N. Kolmogorov and V.M. Tikhomirov, ε-entropy and ε-capacity of sets in function spaces. *Trans. Amer. Math. Soc.* **17** (1961) 277–364.
[41] T.A. Le, A.G. Baydin, R. Zinkov and F. Wood, Using synthetic data to train neural networks is model-based reasoning. *arXiv preprint arXiv:1703.00868* (2017).
[42] N. Madras and D. Sezer, Quantitative bounds for Markov chain convergence: Wasserstein and total variation distances. *Bernoulli* **16** (20108) 82–908.
[43] J.W. Miller and D.B. Dunson, Robust bayesian inference via coarsening. *arXiv preprint arXiv:1506.06101* (2015).
[44] I. Mizera, Qualitative robustness and weak continuity: the extreme unction. *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurecková* **1** (2010) 169.
[45] M. Morzfeld and A.J. Chorin, Implicit particle filtering for models with partial noise, and an application to geomagnetic data assimilation. *Nonlinear Processes in Geophysics* **19** (2012).
[46] M. Nasser, N.A. Hamzah and Md.A. Alam, Qualitative robustness in estimation. *Pakistan J. Statist. Oper. Res.* **8** (2012) 619–634.
[47] H. Owhadi and C. Scovel, Brittleness of Bayesian inference and new Selberg formulas. *Commun. Math. Sci.* **13** (2013) 75.
[48] H. Owhadi, C. Scovel and T.J. Sullivan, Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Statist.* **9** (2015) 1–79.
[49] H. Owhadi, C. Scovel and T.J. Sullivan, On the brittleness of Bayesian inference. *SIAM Rev.* **57** (2015) 566–582.
[50] A.B. Patel, M.T. Nguyen and R. Baraniuk, A probabilistic framework for deep learning. In *Advances in Neural Information Processing Systems* (2016) 2558–2566.
[51] S.T. Rachev, L.B. Klebakov, S.V. Stoyanov and F.J. Fabozzi, *The Methods of Distances in the Theory of Probability and Statistics*. Springer, New York (2013).
[52] P. Rebeschini and R. Van Handel, Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability* **25** (2015) 2809–2866.
[53] G.O. Roberts and J.S. Rosenthal, Markov-chain Monte Carlo: Some practical implications of theoretical results. *Canadian J. Statist.* **26** (1998) 5–20.
[54] G.O. Roberts and J.S. Rosenthal, General state space Markov chains and MCMC algorithms. *Probab. Surveys* **1** (2004) 20–71.
[55] M.J. Schervish, *Theory of Statistics*. Springer (1995).
[56] L. Schwartz, On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4** (1965) 10–26.
[57] A. Smith, A. Doucet, N. de Freitas and N. Gordon, *Sequential Monte Carlo Methods in Practice*. Springer Science & Business Media (2013).
[58] A.M. Stuart, Inverse problems: a Bayesian perspective. *Acta Numer.* **19** (2010) 451–559.
[59] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow and R. Fergus, Intriguing properties of neural networks. In *International Conference on Learning Representations* (2014). http://arxiv.org/abs/1312.6199.
[60] M. Uličnỳ, J. Lundström and S. Byttner, Robustness of deep convolutional neural networks for image recognition. In *Inter. Symp. Intelligent Comput. Syst.* Springer (2016) 16–30.
[61] P.J. Van Leeuwen, Particle filtering in geophysical systems. *Monthly Weather Rev.* **137** (2009) 4089–4114.
[62] A. Wald, *Statistical Decision Functions*. John Wiley and Sons Inc., New York, NY (1950).
[63] L. Wasserman, Asymptotic properties of nonparametric Bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics*. Springer (1998) 293–304.
[64] A.D. Woodbury and T.J. Ulrych, A full-bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Res.* **36** (2000) 2081–2093.