SIAM Review
Vol. 57, No. 4, pp. 566–582



# On the Brittleness of Bayesian Inference*

Houman Owhadi†
Clint Scovel‡
Tim Sullivan§

**Abstract.** With the advent of high-performance computing, Bayesian methods are becoming increasingly popular tools for the quantification of uncertainty throughout science and industry. Since these methods can impact the making of sometimes critical decisions in increasingly complicated contexts, the sensitivity of their posterior conclusions with respect to the underlying models and prior beliefs is a pressing question to which there currently exist positive and negative answers. We report new results suggesting that, although Bayesian methods are robust when the number of possible outcomes is finite or when only a finite number of marginals of the data-generating distribution are unknown, they could be generically brittle when applied to continuous systems (and their discretizations) with finite information on the data-generating distribution. If closeness is defined in terms of the total variation (TV) metric or the matching of a finite system of generalized moments, then (1) two practitioners who use arbitrarily close models and observe the same (possibly arbitrarily large amount of) data may reach opposite conclusions; and (2) any given prior and model can be slightly perturbed to achieve any desired posterior conclusion. The mechanism causing brittleness/robustness suggests that learning and robustness are antagonistic requirements, which raises the possibility of a missing stability condition when using Bayesian inference in a continuous world under finite information.

**Key words.** Bayesian inference, misspecification, robustness, uncertainty quantification, optimal uncertainty quantification, Bayesian sensitivity analysis

**AMS subject classifications.** 62F15, 62G35

**DOI.** 10.1137/130938633



The application of Bayes' theorem in the form of Bayesian inference has fueled an ongoing debate with practical consequences in science, industry, medicine, and law [21]. One commonly-cited justification for the application of Bayesian reasoning is Cox's theorem [15], which has been interpreted as stating that any "natural" extension of Aristotelian logic to uncertain contexts must be Bayesian [34]. It has now been shown that Cox's theorem as originally formulated is incomplete [28] and there is some debate about the "naturality" of the additional assumptions required for its validity [1, 20, 29, 31], e.g., the assumption that knowledge can be always represented in the form of a $\sigma$-additive probability measure that assigns to each measurable event a *single* real-valued probability.

*Received by the editors September 26, 2013; accepted for publication (in revised form) April 9, 2015; published electronically November 5, 2015. This work was supported by the Air Force Office of Scientific Research under award FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators).

http://www.siam.org/journals/sirev/57-4/93863.html

†Corresponding author. Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena CA 91125 (owhadi@caltech.edu).

‡Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena CA 91125 (clintscovel@gmail.com).

§Mathematics Institute, University of Warwick, CV4 7AL, UK (tim.sullivan@warwick.ac.uk).

However—and this is the topic of this article—regardless of the internal logic, elegance, and appealing simplicity of Bayesian reasoning, a critical question is that of the robustness of its posterior conclusions with respect to perturbations of the underlying models and priors.

For example, a frequentist statistician might ask, if the data happen to be a sequence of i.i.d. draws from a fixed data-generating distribution $\mu^\dagger$, whether or not the Bayesian posterior will asymptotically assign full mass to a parameter value that corresponds to $\mu^\dagger$. When it holds, this property is known as *frequentist consistency* of the Bayes procedure, or the *Bernstein–von Mises property*.

Alternatively, without resorting to a frequentist data-generating distribution $\mu^\dagger$, a Bayesian statistician who is also a numerical analyst might ask questions about stability and conditioning: does the posterior distribution (or the posterior value of a particular quantity of interest) change only slightly when elements of the problem setup (namely, the prior distribution, the likelihood model, and the observed data) are perturbed, e.g., as a result of observational error, numerical discretization, or algorithmic implementation? When it holds, this property is known as *robustness* of the Bayes procedure.

This paper summarizes recent results [46, 47] that give conditions under which Bayesian inference appears to be nonrobust in the most extreme fashion, in the sense that arbitrarily small changes of the prior and model class lead to arbitrarily large changes of the posterior value of a quantity of interest. We call this extreme nonrobustness "brittleness," and it can be visualized as the smooth dependence of the value of the quantity of interest on the prior breaking into a fine patchwork, in which nearby priors are associated to diametrically opposed *posterior* values. Naturally, the notion of "nearby" plays an important role, and this point will be revisited later.

Much as classical numerical analysis shows that there are "stable" and "unstable" ways to discretize a partial differential equation (PDE), these results and the wider literature of positive [8, 13, 19, 37, 38, 53, 56] and negative [3, 17, 23, 24, 35, 40] results on Bayesian inference contribute to an emerging understanding of "stable" and "unstable" ways to apply Bayes' rule in practice.

The results reported in this article show that the process of Bayesian conditioning on data at finite enough resolution is unstable (or "sensitive" as defined in [54]) with respect to the underlying distributions (under the total variation (TV) and Prokhorov metrics) and is the source of negative results similar to those caused by tail properties in statistics [2, 18]. The mechanisms causing the stability/instability of posterior predictions suggest that learning and robustness are conflicting requirements and raise the possibility of a missing stability condition when using Bayesian inference for continuous systems with finite information (akin to the Courant–Friedrichs–Lewy (CFL) stability condition when using discrete schemes to approximate continuous PDEs).

**Bayes' Theorem and Robustness.** To begin, let us consider a simple example of Bayesian reasoning in action:

Problem 1. *Consider a bag containing* 102 *coins, one of which always lands on heads, while the other* 101 *are perfectly fair. One coin is picked uniformly at random from the bag, flipped* 10 *times, and* 10 *heads are obtained. What is the probability that this coin is the unfair coin?*

The correct probability is given by applying Bayes' theorem:

$$\mathbb{P}[A|B] = \mathbb{P}[B|A]\frac{\mathbb{P}[A]}{\mathbb{P}[B]} = \frac{1}{1 + 101 \times 2^{-10}} \approx 0.91, \tag{1}$$

where $A$ is the event "the coin is the unfair coin" and $B$ is the event "10 heads are observed." If the number of coins is not known exactly and the supposedly fair coins are not exactly fair, then Bayes' theorem produces a robust inference in the following sense: if the fair coins are slightly unbalanced and the probability of getting a tail is 0.51, and if an estimate of 100 coins is used and an estimate $\frac{1}{2}$ of the fairness of the fair coins is used, then the resulting estimate $\frac{1}{1+99\times 2^{-10}}$ is still a good approximation to the correct answer. Observe also that if the prior estimate of the number of coins in the bag is grossly wrong (e.g., $10^6$), then the posterior would still be accurate in the limit of infinitely many coin flips: in this case, the Bayesian estimator is said to be *consistent*.

Do these conclusions remain true when the underlying probability space is continuous or an approximation thereof? For example, what if the random outcomes are decimal numbers—perhaps given to finite precision—rather than heads or tails?

**The General Problem and Its Bayesian Answer.**

PROBLEM 2. *Let $\mathcal{X}$ denote the space in which observations/samples take their values, and let $\mathcal{M}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$. Let $\Phi\colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a function*[1] *defining a* quantity of interest. *Let the* data-generating distribution *$\mu^\dagger \in \mathcal{M}(\mathcal{X})$ be an unknown or partially known probability measure on $\mathcal{X}$. The objective is to estimate $\Phi(\mu^\dagger)$ from the observation of $n$ i.i.d. samples from $\mu^\dagger$, which we denote by $d = (d_1, \ldots, d_n) \in \mathcal{X}^n$.*

EXAMPLE 1. *When $\mathcal{X}$ is the real line $\mathbb{R}$, a prototypical example of a quantity of interest is $\Phi(\mu) := \mu[X \geq a]$, the probability that the random variable $X$ distributed according to $\mu$ exceeds the threshold value $a$. However, the results that we report below apply to* any *prespecified quantity of interest $\Phi$.*

The Bayesian answer to this problem is to model $\mu^\dagger$'s generation of sample data as coming from a random measure on $\mathcal{X}$ and to condition $\Phi$ with respect to the observation of the $n$ i.i.d. samples. This is done by choosing a *model class* $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ and a probability measure $\pi \in \mathcal{M}(\mathcal{A})$ which we call *the prior*. This prior determines the randomness with which a representative $\mu \in \mathcal{A}$ is selected, and, for each such $\mu \in \mathcal{A}$, the generation of $n$ i.i.d. samples $d \in \mathcal{X}^n$ by randomly sampling from $\mu^n$ naturally determines a product measure on $\mathcal{A}\times\mathcal{X}^n$. The prior estimate of the quantity of interest is $\mathbb{E}_{\mu\sim\pi}[\Phi(\mu)]$ and, for an open[2] $B \subseteq \mathcal{X}^n$, the posterior estimate is defined as the conditional expectation $\mathbb{E}_{\mu\sim\pi, d\sim\mu^n}[\Phi(\mu)|d \in B]$ with respect to this product measure.

The connection to the standard presentation of Bayesian inference in terms of a prior on a parameter space is as follows: to construct a model class $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ and a prior $\pi_0 \in \mathcal{M}(\mathcal{A})$ from a Bayesian parametric model $\mathcal{P}\colon \Theta \to \mathcal{M}(\mathcal{X})$ defined on a *parameter space* $\Theta$ equipped with a prior $p_0 \in \mathcal{M}(\Theta)$, one simply pushes forward under the map $\mathcal{P}$. That is, the model class $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ is defined by $\mathcal{A} := \mathcal{P}(\Theta)$ and the prior $\pi_0 \in \mathcal{M}(\mathcal{A})$ is defined as the push-forward $\pi_0 := \mathcal{P}p_0$ of $p_0$ by the model $\mathcal{P}$, i.e., $\pi_0(E) := p_0(\mathcal{P}^{-1}(E))$ for measurable $E \subseteq \mathcal{A}$.

**Inconsistency under Misspecification.** We now discuss the effects of misspecification on a Bayesian parametric model $\mathcal{P}\colon \Theta \to \mathcal{M}(\mathcal{X})$. It is convenient to denote such a model by $\mathcal{P}\colon \theta \mapsto \mu(\theta)$, so that the model class is $\mathcal{A} := \mathcal{P}(\Theta) = \{\mu(\theta) \mid \theta \in \Theta\}$.

[1] All spaces will be topological spaces, the term "function" will mean Borel measurable function, and "measure" will mean Borel measure.

[2] We assume $B$ to be open and of strictly positive measure to avoid problems associated with conditioning with respect to events of measure zero.

If the model class $\mathcal{P}(\Theta)$ contains the data-generating distribution $\mu^\dagger$, i.e., if there is some parameter value $\theta \in \Theta$ such that $\mu^\dagger = \mu(\theta)$, then the model is said to be *well-specified*; otherwise, it is said to be *misspecified*.

For simplicity, consider the classical case where, for each $\theta \in \Theta$, $\mu(\theta)$ has a probability density function with respect to some common reference measure on $\mathcal{X}$, that is, $\mu(\theta) = p(\cdot\,, \theta)\,\mathrm{d}x$ for some measure $\mathrm{d}x$. Then, for a prior $p_0 \in \mathcal{M}(\Theta)$, let $p_n \in \mathcal{M}(\Theta)$ denote the posterior distribution on $\Theta$ after observing the data $d$ (see, e.g., [5, p. 126]) and push forward both the prior and posterior to their corresponding measures, $\pi_0 := \mathcal{P}p_0$ and $\pi_n := \mathcal{P}p_n$, on $\mathcal{M}(\mathcal{A})$.

Now suppose that the model is well-specified and that $p_0$ gives strictly positive mass to every neighborhood of every point $\theta \in \Theta$—this assumption of "maximal open-mindedness" is commonly referred to as *Cromwell's rule* [41]. Then, when $\Theta$ is finite-dimensional, under suitable regularity conditions, the posterior value of the quantity of interest $\mathbb{E}_{\mu\sim\pi_n}\big[\Phi(\mu)\big]$ converges to $\Phi(\mu^\dagger)$ as $n \to \infty$. This convergence, which can be shown to be asymptotically normal, is commonly referred to as the *Bernstein–von Mises theorem* or *Bayesian central limit theorem* [8, 19, 38, 56]. However, for infinite-dimensional $\Theta$ and with similar regularity and strict positivity assumptions, there is a wealth of positive [13, 37, 53] and negative [3, 17, 23, 24, 35, 40] results showing that the truth or otherwise of the Bernstein–von Mises property depends sensitively on subtle topological and geometrical details.

Conversely, if the model is misspecified, then, under regularity conditions [7, 36, 37, 52], the posterior value $\mathbb{E}_{\mu\sim\pi_n}\big[\Phi(\mu)\big]$ converges as $n \to \infty$ to $\Phi\big(\mu(\theta^*)\big)$, where $\theta^*$ maximizes the expected log-likelihood function $\theta \mapsto \mathbb{E}_{\mu^\dagger}\big[\log p(\cdot\,, \theta)\big]$. If, in addition, $\mu^\dagger$ is absolutely continuous with respect to each $\mu(\theta)$ for $\theta \in \Theta$, then $\theta^*$ can also be shown to minimize the *Kullback–Leibler (KL) divergence* or *relative entropy distance* $\theta \mapsto D_{\mathrm{KL}}\big(\mu^\dagger \big\| \mu(\theta)\big)$ from $\mu^\dagger$ to $\mu(\theta)$.

EXAMPLE 2. *To illustrate this, let $\mathcal{X} = \mathbb{R}$ and consider the Gaussian model $\mu(c, \sigma)$ with mean $c$ and standard deviation $\sigma$, that is, with the probability density*

$$p(x, c, \sigma) := \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)$$

*and the expected log-likelihood*

$$\mathbb{E}_{\mu^\dagger}\big[\log p(\cdot, c, \sigma)\big] = -\int_{\mathbb{R}} \frac{(x-c)^2}{2\sigma^2}\,\mathrm{d}\mu^\dagger(x) - \log\sigma - \log\sqrt{2\pi}.$$

*If, for a data-generating distribution $\mu^\dagger$ with finite second moments, we let $c^\dagger$ denote its mean and $\sigma^\dagger$ its standard deviation, then a quick calculation shows that $\theta^* = (c^*, \sigma^*)$ maximizes the expected log-likelihood if and only if $c^* = c^\dagger$ and $\sigma^* = \sigma^\dagger$. Hence, the asymptotic Bayesian posterior estimate of $\Phi(\mu^\dagger)$ is $\Phi\big(\mu(c^\dagger, \sigma^\dagger)\big)$, irrespective of what the quantity of interest $\Phi$ might be. However, there are* many *different probability distributions $\mu$ on $\mathbb{R}$ that have the same first and second moments as $\mu^\dagger$ but have different higher-order moments, or different quantiles. Predictions of those other moments or quantiles using the Gaussian distribution $\mu(c^\dagger, \sigma^\dagger)$ can be inaccurate by orders of magnitude. A simple example is provided by the tail probability $\Phi(\mu) := \mathbb{P}_\mu\big[|X - c_\mu| \geq t\sigma_\mu\big]$, where $c_\mu$ and $\sigma_\mu$ denote the mean and standard deviation of $\mu$ and $t > 0$. Under the Gaussian model*

$$\mathbb{P}_\mu\big[|X - c_\mu| \geq t\sigma_\mu\big] = 1 + \operatorname{erf}\left(-\frac{t}{\sqrt{2}}\right),$$

*whereas the extreme cases that prove the sharpness of Chebyshev's inequality—in which the probability measure is a discrete measure with support on at most three points in* $\mathbb{R}$*—have*

$$\mathbb{P}_\mu\big[|X - c_\mu| \geq t\sigma_\mu\big] = \min\left\{1, \frac{1}{t^2}\right\}.$$

*In the case of the archetypically rare "*$6\sigma$ *event," i.e.,* $t = 6$*, the ratio between the two is approximately* $1.4 \times 10^7$*. This comparison is, of course, almost perversely extreme: it would be obvious to any observer with only moderate amounts of "Chebyshev-type" sample data that the data had been drawn from a highly non-Gaussian distribution. However, it is not inconceivable that the true distribution* $\mu^\dagger$ *has a Gaussian-looking bulk but also has tails that are significantly fatter than those of a Gaussian, and the difference may be difficult to establish using reasonable amounts of sample data; however, it is those tails that drive the occurrence of "Black Swans," catastrophically high-impact but low-probability outcomes.*

Although it is understood that Bayesian estimators can be inconsistent if the model is grossly misspecified, a pressing question is whether they have good convergence properties when the model class $\{\mu(\theta) \mid \theta \in \Theta\}$ is "close enough" to the truth $\mu^\dagger$ in an appropriate sense.

Such concerns can be traced back to Box's dictum that "essentially, all models are wrong, but some are useful" [12, p. 424] and question "how wrong do they have to be to not be useful?" [12, p. 74]. These queries are also critical because, although gross misspecification of the model can be detected before engaging in a complete Bayesian analysis [32, 61], usually one *cannot be sure* that the model is well-specified.

To answer these questions we will examine the robustness of Bayesian inference by computing optimal bounds on prior and posterior values in terms of given sets of priors. Indeed, the exploration of classes of Bayesian models is one response to the concern that the choice of prior-likelihood combination could, to some degree, be arbitrary, and this forms the basis of the approach known as *robust Bayesian inference* [4, 6, 11, 58, 60]. To do so, we need some definitions.

DEFINITION 1. *For a model class* $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$*, a quantity of interest* $\Phi\colon \mathcal{A} \to \mathbb{R}$*, and a set of priors* $\Pi \subseteq \mathcal{M}(\mathcal{A})$*, let*

$$\begin{aligned}\mathcal{L}(\Pi) &:= \inf_{\pi\in\Pi} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big],\\ \mathcal{U}(\Pi) &:= \sup_{\pi\in\Pi} \mathbb{E}_{\mu\sim\pi}\big[\Phi(\mu)\big]\end{aligned}$$

*denote the* optimal lower and upper bounds on the prior values *of* $\Phi$*. For* $B$ *a nonempty open subset of the data space* $\mathcal{X}^n$*, let* $\Pi_B \subseteq \Pi$ *be the subset of priors* $\pi$ *such that the probability that* $d \in B$ *is nonzero, i.e.,* $\mathbb{P}_{\mu\sim\pi, d\sim\mu^n}[d \in B] > 0$*, and let*

$$\begin{aligned}\mathcal{L}(\Pi|B) &:= \inf_{\pi\in\Pi_B} \mathbb{E}_{\mu\sim\pi, d\sim\mu^n}\big[\Phi(\mu)\big|d \in B\big],\\ \mathcal{U}(\Pi|B) &:= \sup_{\pi\in\Pi_B} \mathbb{E}_{\mu\sim\pi, d\sim\mu^n}\big[\Phi(\mu)\big|d \in B\big]\end{aligned}$$

*denote the* optimal lower and upper bounds on the posterior values *of* $\Phi$ *given that* $d \in B$*.*

**Brittleness under Infinitesimal Perturbations.** Consider again the model $\mathcal{P}\colon \Theta \to \mathcal{M}(\mathcal{X})$, but now denote the model class by $\mathcal{A}_0 := \mathcal{P}(\Theta) = \{\mu(\theta) \mid \theta \in \Theta\}$ and

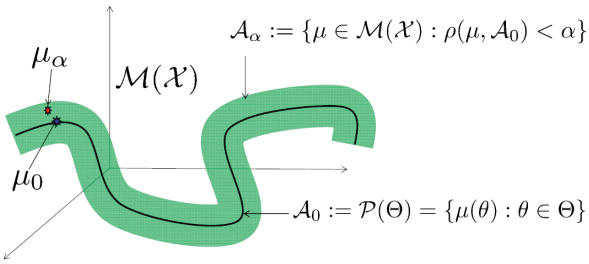



**Fig. 1** *The original model class $\mathcal{A}_0$ (black curve) is enlarged to its metric neighborhood $\mathcal{A}_\alpha$ (shaded). This procedure determines perturbations $\mu_\alpha \in \mathcal{A}_\alpha$ of the original random measure $\mu_0 \in \mathcal{A}_0$.*

the prior by $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$. To quantify perturbations in the model and define what it means for two distributions to be close to one another, we select a metric $\rho$ on $\mathcal{M}(\mathcal{X})$. As illustrated in Figure 1, for $\alpha > 0$, we enlarge the set $\mathcal{A}_0$ to its metric neighborhood $\mathcal{A}_\alpha$ and thereby naturally determine a set of priors $\Pi_\alpha \subseteq \mathcal{M}(\mathcal{A}_\alpha)$ such that the random measure $\mu_\alpha$ associated with every $\pi_\alpha \in \Pi_\alpha$ lies within distance $\alpha$ of the random measure $\mu_0$ associated with the prior $\mu_0$ and the Bayesian model $\mathcal{P}$. Then we analyze the robustness of its posteriors, as in Definition 1, with respect to these size-$\alpha$ perturbations.

To that end, suppose that $\mathcal{X}$ is metrizable and select a consistent metric d for $\mathcal{X}$. Let $\mathcal{B}(\mathcal{X})$ denote the Borel subsets of $\mathcal{X}$. We will consider two metric distances $\rho(\mu, \nu)$ between $\mu, \nu \in \mathcal{M}(\mathcal{X})$: $\rho$ will be either the *TV metric*

$$\rho_{\mathrm{TV}}(\mu, \nu) := \sup\big\{|\mu(A) - \nu(A)|\big|A \in \mathcal{B}(\mathcal{X})\big\},$$

or the *Prokhorov metric*[3]

$$\rho_{\mathrm{P}}(\mu, \nu) := \inf\left\{\varepsilon > 0 \,|\, \mu(A) \leq \nu(A^\varepsilon) + \varepsilon,\, A \in \mathcal{B}(\mathcal{X})\right\},$$

where $A^\varepsilon := \{x \in \mathcal{X} | \mathrm{d}(x, A) < \varepsilon\}$. For $\alpha > 0$, the neighborhood $\mathcal{A}_\alpha$ of $\mathcal{A}_0$ emerges naturally from the ball fibration

$$\mathcal{A}^* := \big\{(\mu_1, \mu_2) \in \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \,\big|\, \mu_1 \in \mathcal{A}_0, \rho(\mu_2, \mu_1) < \alpha\big\},$$

in the sense that if $P_0$ and $P_\alpha$ denote the projections onto the first and second components of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, then $P_0\mathcal{A}^* = \mathcal{A}_0$ and $P_\alpha\mathcal{A}^* = \mathcal{A}_\alpha$. Consequently, a natural set of priors $\Pi_\alpha \subseteq \mathcal{M}(\mathcal{A}_\alpha)$ corresponding to $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$ is defined by

$$\Pi_\alpha := \big\{\pi_\alpha \in \mathcal{M}(\mathcal{A}_\alpha) \,\big|\, \text{for some } \pi \in \mathcal{M}(\mathcal{A}^*),\, P_0\pi = \pi_0 \text{ and } P_\alpha\pi = \pi_\alpha\big\}.$$

To state our result, consider again Problem 2 and let some $x^n := (x_1, \ldots, x_n) \in \mathcal{X}^n$ be a point such that we observe $d \in B_\delta^n := \prod_{i=1}^n B_\delta(x_i)$, where $B_\delta(x) \subseteq \mathcal{X}$ is the open ball of radius $\delta$ centered on $x \in \mathcal{X}$. Using the notation of Definition 1, and $\Pi_\alpha$

[3]The TV metric is generally considered to generate too strong a topology on the space $\mathcal{M}(\mathcal{X})$ of probability measures, and the weak topology is generally considered more appropriate; see, e.g., [9]. Fortunately, when $\mathcal{X}$ is separable, this topology is metrized by the Prokhorov metric. For a thorough discussion regarding metrics on spaces of measures, see, e.g., [49].

defined above in terms of the TV or Prokhorov metric, the Brittleness Theorem 6.4 of [47] then reads as follows.[4]

THEOREM 1. *If*

$$\lim_{\delta \downarrow 0} \sup_{x \in \mathcal{X}} \sup_{\theta \in \Theta} \mu(\theta)[B_\delta(x)] = 0, \tag{2}$$

*then, for all* $\alpha > 0$*, there exists* $\delta(\alpha) > 0$ *such that for all* $0 < \delta < \delta(\alpha)$*, all* $n \in \mathbb{N}$*, and all* $x^n \in \mathcal{X}^n$*,*

$$\mathcal{L}(\Pi_\alpha | B_\delta^n) \leq \operatorname{ess\,inf}_{\pi_0}(\Phi) \quad \text{and} \quad \operatorname{ess\,sup}_{\pi_0}(\Phi) \leq \mathcal{U}(\Pi_\alpha | B_\delta^n),$$

*where* $\operatorname{ess\,inf}_{\pi_0}(\Phi) := \sup\{r \mid \pi_0[\Phi < r] = 0\}$ *and* $\operatorname{ess\,sup}_{\pi_0}(\Phi) := \inf\{r \mid \pi_0[\Phi > r] = 0\}$.

Note that condition (2) is extremely weak and is satisfied for most parametric Bayesian models. Furthermore, suppose that Cromwell's rule is applied. Then, although it implies consistency if the model is well-specified, here it leads to maximal brittleness under local misspecification. More precisely, under Cromwell's rule, $\operatorname{ess\,inf}_{\pi_0}(\Phi) = \inf_{\mu \in \mathcal{A}_0} \Phi(\mu)$ and $\operatorname{ess\,sup}_{\pi_0}(\Phi) = \sup_{\mu \in \mathcal{A}_0} \Phi(\mu)$, so the conclusion of Theorem 1 becomes

$$\mathcal{L}(\Pi_\alpha | B_\delta^n) \leq \inf_{\mu \in \mathcal{A}_0} \Phi(\mu) \quad \text{and} \quad \sup_{\mu \in \mathcal{A}_0} \Phi(\mu) \leq \mathcal{U}(\Pi_\alpha | B_\delta^n).$$

In other words, the range of posterior predictions among all admissible priors is as wide as the deterministic range of the quantity of interest $\Phi$.

Note that since $\Phi$ is arbitrary, the brittleness described in Theorem 1 is not limited to a quantile or moment of $\mu$ but concerns its whole posterior distribution.

**Brittleness under Finite Information.** One response to the concern that the choices of prior and model are somewhat arbitrary [58] is to perform a sensitivity analysis over classes of priors and models. One way to specify a class $\Pi$ of admissible priors $\pi$ is to select some "features" (such as the polynomial moments, or other functionals) and specify some values, ranges, or distributions for those features. It is interesting to understand the impact of those features left unspecified, i.e., the *codimension* and not just the *dimension* of $\Pi$; while *robust Bayesian inference* [4, 6, 11, 60] has shown that posterior conclusions remain stable when $\Pi$ is finite-dimensional, our results can be interpreted as saying that brittleness ensues whenever $\Pi$ has finite codimension, regardless of how large its codimension is. It is important to note that this is in some sense the generic situation: when $\mathcal{A}$ is an infinite set, one would have to specify infinitely many features of priors $\pi \in \Pi$ to achieve a finite-dimensional $\Pi$; from a computational and epistemic standpoint, the specification of infinitely many features in finite time appears to be somewhat problematic.

To study this problem, we introduce a representation space $\mathcal{Q}$ (e.g., prototypically, $\mathbb{R}^k$) and a mapping $\Psi \colon \mathcal{A} \to \mathcal{Q}$ from the subset $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ into $\mathcal{Q}$, which can be

[4] All results of this article and those in [46, 47, 48] require some mild technical measure-theoretic and topological assumptions. For example, here it is sufficient if $\mathcal{P}(\Theta)$ is a Borel subset of a Polish space (a separable completely metrizable space). Unfortunately, $\mathcal{M}(\mathcal{X})$ is not generally separable with respect to the TV metric, and hence is not Polish. However, if $\mathcal{X}$ is Polish, then $\mathcal{M}(\mathcal{X})$ topologized by weak convergence is Polish and the Prokhorov metric provides a complete metrization of it. Consequently, when $\Theta$ is Polish, $\mathcal{X}$ is Polish, and $\mathcal{P}$ is injective and measurable with respect to the weak topology, it then follows from Suslin's Theorem that $\mathcal{P}(\Theta)$ is a Borel subset of the Polish space $\mathcal{M}(\mathcal{X})$. For a thorough investigation of such matters, illustrating the benefits of Polish spaces as the foundation for the framework, see [47].

thought of as a map to "generalized moments." Let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be a subset of the set of probability distributions on $\mathcal{Q}$ such that each distribution $\mathbb{Q} \in \mathfrak{Q}$ has its support contained in $\Psi(\mathcal{A})$. If the set $\mathfrak{Q}$ represents priors for the distribution of $\Psi(\mu), \mu \in \mathcal{A}$, then a naturally induced set of priors $\Pi$ on $\mathcal{A}$ is the pull-back $\Pi := \Psi^{-1}(\mathfrak{Q}) \subseteq \mathcal{M}(\mathcal{A})$, defined by $\Psi^{-1}(\mathfrak{Q}) := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\}$.

EXAMPLE 3. *Consider the case $\mathcal{X} = [0,1]$, $\mathcal{A} := \mathcal{M}([0,1])$, and $\Phi(\mu) = \mathbb{E}_\mu[X]$. The aim is to estimate the mean $\Phi(\mu^\dagger) = \mathbb{E}_{\mu^\dagger}[X]$ of the random variable $X$ corresponding to some unknown measure $\mu^\dagger \in \mathcal{A}$ and we observe $d = (d_1, \ldots, d_n)$, $n$ i.i.d. samples from $X$. Let $k$ be fixed and let $\Psi(\mu) = (\mathbb{E}_\mu[X], \ldots, \mathbb{E}_\mu[X^k])$ be the map to the first $k$ polynomial moments. If we write a point $q \in \mathbb{R}^k$ in terms of its coordinates $q := (q_1, \ldots, q_k)$, then $\Psi^{-1}(q)$ is exactly the set of measures $\mu \in \mathcal{M}([0,1])$ such that $\mathbb{E}_\mu[X^i] = q_i$ for $1 \le i \le k$. Now define a measure $\mathbb{Q}$ on the truncated moment space $\Psi(\mathcal{M}([0,1]) \subseteq \mathbb{R}^k$ as follows. Since the first moment $\mathbb{E}_\mu[X], \mu \in \mathcal{M}([0,1])$, ranges over the unit interval, consider the uniform measure on the unit interval in the first coordinate. Next define the conditional measure when the first coordinate is $q_1 \in [0,1]$ to be uniform on the range of the second moment $[\inf_{\mu:\,\mathbb{E}_\mu[X]=q_1} \mathbb{E}_\mu[X^2], \sup_{\mu:\,\mathbb{E}_\mu[X]=q_1} \mathbb{E}_\mu[X^2]]$. Repeat this conditioning process on the higher coordinates iteratively in the same manner. Then, the induced set of priors $\Pi := \Psi^{-1}\mathbb{Q}$ on $\mathcal{M}([0,1])$ is the set of measures $\pi$ such that, when $\mu \sim \pi$, the distribution of $(\mathbb{E}_\mu[X], \ldots, \mathbb{E}_\mu[X^k])$ is $\mathbb{Q}$.*

We now state the Brittleness Theorem 4.13 in [47] for the general case of Problem 2 and apply it to Example 3. To that end, let the model class $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$ be chosen along with a generalized moment map $\Psi\colon \mathcal{A} \to \mathcal{Q}$ to a representation space $\mathcal{Q}$. Let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be a specified set of priors on $\mathcal{Q}$ and from them determine $\Pi := \Psi^{-1}(\mathfrak{Q}) \subseteq \mathcal{M}(\mathcal{A})$ as the induced set of priors. For fixed $(x_1, \ldots, x_n) \in \mathcal{X}^n$, let $B_\delta^n := \prod_{i=1}^n B_\delta(x_i)$, where $B_\delta(x)$ is the open ball of radius $\delta$ centered on $x \in \mathcal{X}$. The following theorem gives optimal bounds on posterior values for the class of priors $\Pi$ defined above, given that the observation $d \in B_\delta^n$.

THEOREM 2. *Suppose that, for all $\gamma > 0$, there exists some $\mathbb{Q} \in \mathfrak{Q}$ such that*

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\inf_{\mu\in\Psi^{-1}(q),i=1,\ldots,n} \mu[B_\delta(x_i)]\right] = 0 \tag{3}$$

*and*

$$\mathbb{P}_{q\sim\mathbb{Q}}\left[\sup_{\mu\in\Psi^{-1}(q):\,\mu[B_\delta(x_i)]>0,i=1,\ldots,n} \Phi(\mu) > \sup_{\mu\in\mathcal{A}} \Phi(\mu) - \gamma\right] > 0\,. \tag{4}$$

*Then*

$$\mathcal{U}\big(\Pi\big|B_\delta^n\big) = \sup_{\mu\in\mathcal{A}} \Phi(\mu),$$

*with similar expressions for the lower bounds $\mathcal{L}$.*

In other words, if there is a measure $\mathbb{Q} \in \mathfrak{Q}$ such that for $\mathbb{Q}$-almost all $q \in \mathcal{Q}$, there is a $\mu \in \Psi^{-1}(q)$ which achieves an arbitrarily small mass on one of $B_\delta(x_i), i = 1, \ldots, n$, and with nonzero $\mathbb{Q}$ probability there is $\mu \in \Psi^{-1}(q)$ which almost extremizes $\Phi$ while putting positive mass on all $B_\delta(x_i), i = 1, \ldots, n$, then the range $\big[\mathcal{L}\big(\Pi\big|B_\delta^n\big), \mathcal{U}\big(\Pi\big|B_\delta^n\big)\big]$ of posterior values for $\Phi$ is exactly the "deterministic" range of $\Phi$, i.e., $\big[\inf_{\mu\in\mathcal{A}} \Phi(\mu), \sup_{\mu\in\mathcal{A}} \Phi(\mu)\big]$.

Conditions (3) and (4) are very weak, and simple dimensionality arguments suggest that they are typically satisfied if $\mathcal{Q}$ is finite-dimensional. Hence, although

Bayesian inference is robust in situations where the distributions of *all but* finitely many generalized moments of the data-generating distribution $\mu^\dagger$ are known, Theorem 2 suggests that it is brittle when the distributions of *only* finitely many generalized moments of $\mu^\dagger$ are known, while infinitely many remain unknown. As an example, it is instructive to observe how Theorem 2, applied to Example 3 in [47, Ex. 4.16], shows that if the data-generating measure has some nonatomic component, then when the number of samples $n$ is large enough and $\delta$ small enough, the optimal bounds on posterior values of $\Phi(\mu) = \mathbb{E}_\mu[X]$, given the distribution $\mathbb{Q}$ defined on its moments, are 0 and 1.

To quantify "large enough" and "small enough" and to remove the "nonatomic" requirement above, Theorem 3.1 of [46] provides a quantitative version of Theorem 2 in which the conditions of the theorem are only required to hold approximately. When applied to Example 3 with the set $\Pi := \Psi^{-1}\mathbb{Q}$ of priors generated instead by the uniform prior $\mathbb{Q}$ restricted to the truncated moment space, Theorem 3.3 of [46] establishes that, although the prior value satisfies $\mathcal{U}(\Pi) = \frac{1}{2}$, the posterior value satisfies

$$1 - 4e\Big(\frac{2k\delta}{e}\Big)^{\frac{1}{2k+1}} \leq \mathcal{U}\big(\Pi|B_\delta\big) \leq 1\,. \tag{5}$$

Consequently, regardless of the number of moment constraints $k$ and the location of a single data point, for $\delta$ smaller than an elementary known function of $k$, we have brittleness. This result also holds for arbitrary multiple samples. Remark 4.18 of [47] also suggests that brittleness would persist if the hard bound $\delta$ to specify measurement uncertainty were replaced by a level of noise with variance decreasing with $\delta$.

**Mechanism Causing Brittleness.** We will now illustrate one mechanism causing brittleness with a simple example derived from the proof of Theorem 1. In this example we are interested in estimating $\Phi(\mu^\dagger) = \mathbb{E}_{\mu^\dagger}[X]$, where $\mu^\dagger$ is an unknown distribution on the unit interval ($\mathcal{X} = [0,1]$) based on the observation of a single data point $d_1 = \frac{1}{2}$ up to resolution $\delta$ (i.e., we observe $d_1 \in B_\delta(x_1)$ with $x_1 = \frac{1}{2}$).

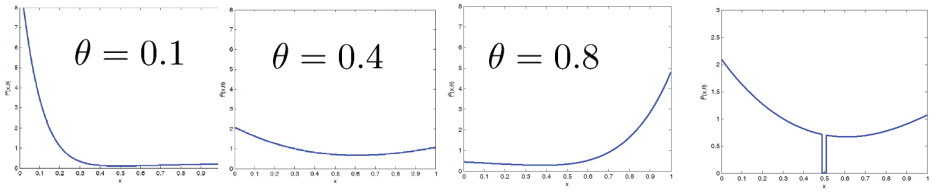
Consider the following two models $\mu^a(\theta)$ and $\mu^b(\theta)$ on the unit interval $[0,1]$, parameterized by $\theta \in (0,1)$ and with densities $f^a$ and $f^b$ given by

$$f^a(x,\theta) = (1-\theta)\big(1+\tfrac{1}{\theta}\big)(1-x)^{\frac{1}{\theta}} + \theta\big(1+\tfrac{1}{1-\theta}\big)x^{\frac{1}{1-\theta}},$$

$$f^b(x,\theta) = \begin{cases} f^a(x,\theta)\frac{1}{Z}\big(\mathbb{1}_{\{x\notin(x_1-\frac{\delta_c}{2},x_1+\frac{\delta_c}{2})\}} + 10^{-9}\mathbb{1}_{\{x\in(x_1-\frac{\delta_c}{2},x_1+\frac{\delta_c}{2})\}}\big) & \text{if } \theta < 0.999, \\ f^a(x,\theta) & \text{if } \theta \geq 0.999, \end{cases}$$

where $Z$ is a normalization constant (close to one) chosen so that $\int_{[0,1]} f^b(x,\theta)\,dx = 1$. See Figure 2 for an illustration of these densities.

Observe that the density of model $b$ is that of model $a$ besides the small gap of width $\delta_c > 0$ created around the data point for model $b$ (if $\theta < 0.999$, see Figure 2); since the data point is fixed at $x_1 = \frac{1}{2}$, the TV distance $\rho_{\mathrm{TV}}\big(\mu^a(\theta), \mu^b(\theta)\big)$ between the two models is, uniformly over $\theta \in (0,1)$, bounded by a constant times $\delta_c$. Assuming that the prior distribution on $\theta$ is the uniform distribution on $(0,1)$, observe that the prior value of the quantity of interest $\mathbb{E}_\mu[X]$ under both models ($a$ and $b$) is approximately $\frac{1}{2}$. Now, when $\theta$ is close to 1 (zero), the density of model $a$ puts most of its mass toward 1 (zero). Observe also that the density of model $b$ behaves in a similar way, with the important exception that the probability of observing the data under model $b$ is infinitesimally small for $\theta < 0.999$. Therefore, for $\delta < \delta_c$, the posterior value of the quantity of interest $\mathbb{E}_\mu[X]$ under model $a$ is $\frac{1}{2}$, whereas it is




(a) $f^a(x,\theta)$ (b) $f^b(x,\theta)$

**Fig. 2** *Illustration of the densities $f^a(x,\theta)$ of model $a$ and $f^b(x,\theta)$ of model $b$.*

close to 1 under model $b$. Observe also that a perturbed model $c$ analogous to $b$ can be constructed to lead to a posterior value close to zero.

The mechanism described here is generic and $\mu^b(\theta)$ is a simple example of what worst priors can look like after a classical Bayesian sensitivity analysis over a class of priors specified via constraints on the TV or Prokhorov distance or the distribution of a finite number of moments.

Can these worst priors be dismissed because they depend on the data? The problem with this argument is that, in the context of Bayesian sensitivity analysis, worst priors always depend on (or are preadapted to) the data. Therefore, the same argument would lead to a dismissal of Bayesian sensitivity analysis and therefore the framework of robust Bayesian inference. In some sense, the brittleness results reported here can be seen as extreme occurrences of the dilation property [59] which, in robust Bayesian inference, refers to the enlargement of optimal bounds caused by the data dependence of worst priors. Indeed, even if perturbations are quantified in KL divergence, the local sensitivity analysis (in the sense of Fréchet derivatives) of posterior values [27] shows infinite sensitivity as the number of data points goes to infinity (and this result is valid for the broader class of divergences that includes the Hellinger distance).

Can these worst priors be dismissed because they can "look unrealistic" and make the probability of observing the data very small? The problem with this argument is that these worst priors are not "isolated pathologies" but directions of instability (of Bayesian conditioning) increasing with the number of data points and the complexity of the system under investigation. We will illustrate this point with another simple example that also shows that these instabilities are the price to pay for the learning potential of Bayesian inference.

**Learning and Robustness Are Antagonistic Properties.** In this example we are interested in estimating $\Phi(\mu^\dagger) = \mu^\dagger[a,1]$ for some $a \in (0,1)$, where $\mu^\dagger$ is an unknown distribution on the unit interval ($\mathcal{X} = [0,1]$) based on the observation of $n$ data points $d_1,\ldots,d_n$ up to resolution $\delta$ (i.e., we observe $d_i \in B_\delta(x_i)$ with $x_i \in [0,1]$ for $i = 1,\ldots,n$). Our purpose is to examine the sensitivity of the Bayesian answer to this problem with respect to the choice of a particular prior. Consider the model class $\mathcal{A} := \mathcal{M}([0,1])$ and the class of priors

$$\Pi := \left\{\pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = m\right\}.$$

Observe that $\Pi$ corresponds to the assumption that $\mu^\dagger$ is the realization of a random measure on $[0,1]$ whose mean is on average $m$.

As in the previous example, the finite codimensional class of priors $\Pi$ leads to brittleness in the sense that the least upper bound on prior values is $\mathcal{U}(\Pi) = \frac{m}{a}$, whereas (for $\delta \ll 1/n$) the least upper bound on posterior values is the deterministic supremum of the quantity of interest (over $\mathcal{A}$), i.e., $\mathcal{U}(\Pi|B_\delta^n) = 1$. Furthermore, worst priors are obtained by selecting priors for which the probability of observing the data $\mu^n[B_\delta^n]$ is arbitrarily close to zero except when $\Phi(\mu)$ is close to its deterministic supremum.

Can this brittleness be avoided by adding a uniform constraint on the probability of observing the data in the model class? To investigate this question, let us introduce $\alpha \geq 1$ and a probability measure $\mu_0$ on $[0,1]$ with strictly positive Lebesgue density (with $\mu_0$ being the uniform measure on $[0,1]$ as a prototypical example) and consider the (new) model class

$$\mathcal{A}(\alpha) := \left\{ \mu \in \mathcal{M}[0,1] \,\middle|\, \frac{1}{\alpha}\mu_0^n[B_\delta^n] \leq \mu^n[B_\delta^n] \leq \alpha\mu_0^n[B_\delta^n] \right\} \tag{6}$$

and the (new) class of priors

$$\Pi(\alpha) := \left\{ \pi \in \mathcal{M}(\mathcal{A}(\alpha)) \,\middle|\, \mathbb{E}_{\mu\sim\pi}\big[\mathbb{E}_\mu[X]\big] = m \right\},$$

where, in (6), $B_\delta^n := \prod_{i=1}^n B_\delta(x_i)$ and $(x_1, \ldots, x_n) \in [0,1]^n$ is fixed.

Note that, for the model class $\mathcal{A}(\alpha)$, the probability of observing the data is uniformly bounded from below by $\frac{1}{\alpha}\mu_0^n[B_\delta^n]$ and from above by $\alpha\mu_0^n[B_\delta^n]$. Therefore, for $\alpha = 1$, the probability of observing the data is uniform in the model class, prior values are equal to posterior values, and the method is robust but learning is impossible. On the other hand, if $\alpha$ slightly deviates from 1, then the calculus developed in [47] (Theorems 4.8 and 4.13) gives

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(\alpha)|B_\delta^n\big) = \frac{1}{1 + \frac{1}{\alpha^2}\frac{a-m}{m}} = \frac{m}{\frac{a}{\alpha^2} + m(1 - \frac{1}{\alpha^2})}. \tag{7}$$

Note that the right-hand side of (7) is equal to $m/a$ for $\alpha = 1$ (when the probability of the data is constant on the model class) and *quickly* converges toward 1 as $\alpha$ increases. As a numerical application observe that for $a = \frac{3}{4}$ and $m = \frac{a}{2} = \frac{3}{8}$, we have $\lim_{\delta\to 0} \mathcal{U}\big(\Pi(\alpha)\big) = \frac{1}{2}$ and

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(\alpha)|B_\delta^n\big) = \frac{1}{1 + \frac{1}{\alpha^2}}.$$

Therefore, for $\alpha = 2$, we have (irrespective of the number of data points)

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(2)|B_\delta^n\big) = 0.8$$

and, for $\alpha = 10$, we have (irrespective of the number of data points)

$$\lim_{\delta\to 0} \mathcal{U}\big(\Pi(10)|B_\delta^n\big) \approx 0.99.$$

Moreover, if $\alpha$ is derived by assuming the probability of each data point to be known up to some tolerance $\gamma$, i.e., if the model class $\mathcal{A}(\alpha)$ is replaced by

$$\mathcal{A}_\gamma := \left\{ \mu \in \mathcal{M}[0,1] \,\middle|\, \frac{1}{\gamma}\mu_0[B_\delta(x_i)] \leq \mu[B_\delta(x_i)] \leq \gamma\mu_0[B_\delta(x_i)],\ i = 1, \ldots, n \right\} \tag{8}$$

for some $\gamma > 1$, then it can be shown that

$$\lim_{\delta \to 0} \mathcal{U}(\Pi | B_\delta^n) = \frac{1}{1 + \frac{1}{\gamma^{2n}}},$$

which exponentially converges toward 1 as the number $n$ of data points goes to infinity.

In conclusion, the effects of a uniform constraint on the probability of the data under finite information in the model class shows that learning ability comes at the price of loss in stability in the following sense: when $\alpha = 1$, the data is equiprobable under all measures in the model class, posterior values are equal to prior values, and the method is robust but learning is not possible. As $\alpha$ deviates from 1, the learning ability increases as robustness decreases, and when $\alpha$ is large, learning is possible but the method is brittle.

**Qualitative Robustness and Consistency.** Since the data dependence of worst priors is inherent to classical Bayesian sensitivity analysis, one might ask whether robustness could be established under finite information by leaving the strict framework of robust Bayesian inference and computing the sensitivity of posterior conclusions independently of the specific value of the data. Indeed, in the current classical Bayesian sensitivity analysis framework, given a class of priors $\Pi$ and the observation $d \in B_\delta^n(x)$, we compute

$$\sup_{\pi, \pi' \in \Pi} \left| \mathbb{E}_{\mu \sim \pi} \left[ \Phi(\mu) | d \in B_\delta^n(x) \right] - \mathbb{E}_{\mu \sim \pi'} \left[ \Phi(\mu) | d \in B_\delta^n(x) \right] \right|,$$

which corresponds to the *sensitivity of posterior values* (given the value of the data) with respect to the particular choice of prior $\pi \in \Pi$. Therefore, the interpretation of the brittleness mechanisms discussed above should be limited to the significance of such optimal bounds, which are not the sole measure of robustness of a Bayesian estimation. An alternative analysis could be to quantify the *sensitivity of the distribution of posterior values*. For instance, given a class of priors $\Pi \subset \mathcal{M}(\mathcal{X})$ over a model class $\mathcal{A} \subseteq \mathcal{M}(\mathcal{X})$, the value of

$$\sup_{\pi, \pi' \in \Pi, \nu \in \mathcal{A}} \mathbb{P}_{x \sim \nu^n} \left[ \left| \mathbb{E}_{\mu \sim \pi} \left[ \Phi(\mu) | d \in B_\delta^n(x) \right] - \mathbb{E}_{\mu \sim \pi'} \left[ \Phi(\mu) | d \in B_\delta^n(x) \right] \right| \geq \epsilon \right]$$

is the least upper bound on the probability that posterior values derived from $\pi, \pi' \in \Pi$ and randomized through an admissible candidate $\nu \in \mathcal{A}$ for the distribution of the data deviate by at least $\epsilon > 0$. This form of analysis is directly related to Hampel [30] and Cuevas' [16] notion of *qualitative robustness*, which requires closeness in *distributions of the posterior distribution* rather than in *posterior distributions*. More precisely, given a metric $\rho_2$ on $\mathcal{M}(\mathcal{M}(\mathcal{A}))$, a *qualitative* sensitivity analysis would seek to bound $\rho_2(\pi_* \nu^n, \pi'_* \nu^n)$ (over $\pi, \pi' \in \Pi$ and $\nu \in \mathcal{A}$), where $\pi_* \nu^n \in \mathcal{M}(\mathcal{M}(\mathcal{A}))$ is the distribution of the posterior distribution of the prior $\pi \in \mathcal{M}(\mathcal{A})$ when the data $d = (d_1, \ldots, d_n)$ is randomized through $\nu^n$. If, unlike Hampel and Cuevas who require "closeness for all $n$," we follow Huber [33] and Mizera [44] in only requiring closeness "for large enough $n$" (i.e., in the limit as the number of data points tends to infinity), then we obtain [45] a notion of *qualitative robustness*, where the notion of *consistency* (i.e., the property that posterior distributions convergence toward the data-generating distribution) plays an important role. Although consistency is primarily a frequentist notion, according to Blackwell and Dubins [10] and Diaconis and Freedman [17], consistency is equivalent to *intersubjective agreement*, which

means that two Bayesians will ultimately have very close predictive distributions. Fortunately, not only are there mild conditions which guarantee consistency, but the posterior distributions can be shown to contract/concentrate at an exponential rate around the data-generating distribution (see [55] for rates of contraction of posterior distributions based on Gaussian process priors) and the Bernstein–von Mises theorem goes further in providing mild conditions under which the posterior is asymptotically normal [13, 14]. The most famous of these are Doob [19], Le Cam and Schwartz [39], and Schwartz [50, Thm. 6.1].

Unfortunately, the conditions ensuring consistency (e.g., the condition that the prior has KL support at the parameter value generating the data[5]) are such that arbitrarily small (TV or Prokhorov) local perturbations of the prior distribution (near the data-generating distribution) may result in consistency or non-consistency, and therefore may have large impacts on the asymptotic behavior of posterior distributions [45]. A simple illustration of this mechanism is as follows [45]. Suppose that the data-generating distribution $\nu$ is at distance $\tau > 0$ from the support of the prior $\pi$. Let $\pi_1$ be a prior distribution with all of its mass on or around $\nu$ (having KL support at $\nu$). Take $\pi' := (1-\epsilon)\pi + \epsilon\pi_1$. The TV distance from $\pi'$ to $\pi$ is bounded by $\epsilon$, which can be chosen to be arbitrarily small. Furthermore, $\pi'$ inherits the KL support of $\pi_1$ at $\nu$ and by Schwartz's consistency theorem [50] its posterior distribution converges (almost surely) toward a Dirac concentrated at $\nu$ as $n \to \infty$. On the other hand, the distance between the support of the posterior distribution of $\pi$ and $\nu$ remains bounded by $\tau$. This simple example exposes a serious challenge to proving robustness in the TV metric or any weaker metric, such as those used in the convergence of MCMC.

Of course, in a parametric setting, if the parameter space $\Theta$ is compact and the model well-specified (the data generated from a parameter in that space), then choosing a prior satisfying Cromwell's rule (putting mass in the KL neighborhood of all parameters) ensures qualitative robustness (and the degree of robustness is a function of how much mass is placed in each neighborhood). However, if $\Theta$ is compact and the model is misspecified, then, even if the prior is nice and smooth, the mechanism discussed above suggests that it is not qualitatively robust (with a degree of nonrobustness corresponding to the degree of misspecification; the prior does not need to look "unrealistic" to be nonqualitatively-robust). Note also that if $\Theta$ is noncompact, then the prior cannot be qualitatively robust (because no matter how small $\epsilon$ is, one can always find a neighborhood of the parameter space with mass smaller than $\epsilon$).

In a nonparametric setting, consistent priors (such as the ones analyzed in [55] with bounds on convergence rates) remain good/natural choices when their posterior distributions can be computed. However, consistency and robustness are to some degree conflicting requirements [16, 45] from the point of view of a numerical analyst. Consider, for instance, the problem of using a sophisticated numerical Bayesian model to predict the climate where Bayes rule is applied iteratively and posterior values become prior values for the next iteration. How do we make sure that our predictions are robust, not only with respect to the choice of prior but also with respect to numerical instabilities arising in the iterative application of the Bayes rule? The nonrobustness mechanisms discussed here suggest that, unless the prior is chosen carefully, and unless we have a tight control on numerical instabilities, errors, and approximations at each step of the iteration, our final predictions might be unstable.

[5] $\pi \in \mathcal{M}(\mathcal{M}(\mathcal{X}))$ is said to have KL support at $\nu \in \mathcal{M}(\mathcal{X})$ if $\pi\{\mu \in \mathcal{M}(\mathcal{X}) \mid \int_{\mathcal{X}} \frac{d\nu}{d\mu} d\nu \leq \epsilon\}$ is strictly positive for all $\epsilon > 0$

Note that, often, these posterior distributions (which are later on used as prior distributions) are only approximated (e.g., via MCMC methods), and so how do we go about ensuring the stability of our method in such situations? The brittleness results discussed here suggest that having strong convergence of our MCMC method in TV would not be enough to ensure stability. Note in particular that although quantifying perturbations in KL ensures qualitative robustness, it would also require controlling the convergence of the MCMC method in KL or in a stronger metric.

**Conclusion and Perspectives.** It is possible that an analogy can be made between the brittleness and robustness properties of Bayesian inference and the numerical analysis of PDEs, for which many pathologies and also many necessary and/or sufficient stability conditions are known. However, in contrast to conditions such as the well-known CFL condition for PDEs, the question of the existence and nature of a stability condition when using Bayesian inference under finite information remains to be resolved. Although numerical schemes that do not satisfy the CFL condition may look grossly inadequate, the existence of such perverse examples certainly does not imply the dismissal of the necessity of a stability condition. Similarly, although one can, as in the example provided in Figure 2, exhibit grossly perverse worst priors, the existence of such priors does not invalidate the need for a study of stability conditions when using Bayesian inference under finite information. The example provided in (7) suggests that, in the framework of Bayesian sensitivity analysis, such a stability condition would depend on (i) how well the probability of the data is known or constrained in the model class, and (ii) the resolution at which the quantity of interest is conditioned upon the data. Note that the independence of the brittleness threshold $\delta_c$ from the number of data points $n$ in Theorem 1 suggests that taking $\delta$ fixed and $n \to \infty$ does not prevent brittleness in the classical Bayesian sensitivity analysis framework (it only leads to more directions of instabilities). On the other hand, for a fixed $\delta$, (5) suggests that brittleness results do not persist in that same framework when the number of moment constraints $k$ (on the class of priors) is large enough. Furthermore, taking $\delta > 0$ fixed (or discretizing space at a resolution $\delta > 0$) enables the construction of classes of qualitatively robust priors (to TV perturbations) that are nearly consistent as $n \to \infty$ (some degree of consistency is lost due to the discretization). At a higher level, the mechanisms discussed here appear to suggest that robust inference (in a continuous world under finite information) should perhaps be done with reduced/coarse models rather than highly sophisticated/complex models (with a level of "coarseness/reduction" depending on the available "finite information"). In the context of deterministic modeling versus uncertainty quantification, Stuart [53] asked, "should future increased computer resources be invested in further model resolution, or in more detailed study of uncertainty?" The results reported here suggest that the answer is the latter, at least in the context of Bayesian modeling versus robustness studies, because posterior conclusions become nonrobust if model resolution is pushed beyond a threshold defined by model uncertainties.

A close inspection of some of the cases where Bayesian inference has been successful suggests the existence of a non-Bayesian feedback loop on the evaluation of its performance [43, 51, 42]. Therefore, one natural question is whether the missing stability condition could also be derived by exiting the strictly Bayesian framework, as proposed in [21]. One example of such an approach could be using posterior predictive checking [26], [25, p. 159], whose rationale is to detect model mismatch by generating replicate data from the model, and comparing this replicate data to the original data using statistics related to the quantity of interest.

It is natural to expect that robustness and stability questions will increase in importance as Bayesian methods become more popular with the availability of computational methodologies and environments to compute the posteriors. Another strong motivation for considering Bayesian methods and investigating such questions is the complete class theorem, which, in the adversarial game theoretic setting of decision theory [57], asserts that optimal statistical estimators (leading to optimal decisions as defined by a convex loss function on a compact parameter space) live in the Bayesian class of estimators [57, 22].

**Acknowledgment.** The authors gratefully acknowledge support for this work from the Air Force Office of Scientific Research under award FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators). They thank P. Diaconis, D. Mayo, P. Stark, and L. Wasserman for stimulating discussions and relevant references and pointers. They thank the anonymous referees for valuable comments and suggestions.



## REFERENCES

[1] S. Arnborg and G. Sjödin, *On the foundations of Bayesianism*, in Bayesian Inference and Maximum Entropy Methods in Science and Engineering (Gif-sur-Yvette, 2000), AIP Conf. Proc. 568, Amer. Inst. Phys., Melville, NY, 2001, pp. 61–71.
[2] R. R. Bahadur and L. J. Savage, *The nonexistence of certain statistical procedures in nonparametric problems*, Ann. Math. Statist., 27 (1956), pp. 1115–1122.
[3] G. Belot, *Bayesian orgulity*, Philos. Sci., 80 (2013), pp. 483–503.
[4] J. O. Berger, *The robust Bayesian viewpoint*, in Robustness of Bayesian Analyses, Stud. Bayesian Econometrics 4, North-Holland, Amsterdam, 1984, pp. 63–144.
[5] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer-Verlag, New York, 1985.
[6] J. O. Berger, *An overview of robust Bayesian analysis*, Test, 3 (1994), pp. 5–124.
[7] R. H. Berk, *Limiting behavior of posterior distributions when the model is incorrect*, Ann. Math. Statist., 37 (1966), pp. 51–58; correction, 37 (1966), pp. 745–746.
[8] S. N. Bernšteĭn, *Sobranie sochinenii. Tom* IV: *Teoriya veroyatnostei. Matematicheskaya statistika.* 1911–1946, Izdat. "Nauka", Moscow, 1964.
[9] P. Billingsley, *Convergence of Probability Measures*, 2nd ed., Wiley, New York, 1999.
[10] D. Blackwell and L. Dubins, *Merging of opinions with increasing information*, Ann. Math. Statist., 33 (1962), pp. 882–886.
[11] G. E. P. Box, *Non-normality and tests on variances*, Biometrika, 40 (1953), pp. 318–335.
[12] G. E. P. Box and N. R. Draper, *Empirical Model-Building and Response Surfaces*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley, New York, 1987.
[13] I. Castillo and R. Nickl, *Nonparametric Bernstein–von Mises theorems in Gaussian white noise*, Ann. Statist., 41 (2013), pp. 1999–2028.
[14] I. Castillo and R. Nickl, *On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures*, Ann. Statist., 42 (2014), pp. 1941–1969.
[15] R. T. Cox, *Probability, frequency and reasonable expectation*, Amer. J. Phys., 14 (1946), pp. 1–13.
[16] A. Cuevas, *Qualitative robustness in abstract inference*, J. Statist. Plann. Inference, 18 (1988), pp. 277–289.
[17] P. Diaconis and D. A. Freedman, *On the consistency of Bayes estimates*, Ann. Statist., 14 (1986), pp. 1–67.
[18] D. L. Donoho, *One-sided inference about functionals of a density*, Ann. Statist., 16 (1988), pp. 1390–1420.
[19] J. L. Doob, *Application of the theory of martingales*, in Le Calcul des Probabilités et ses Applications, Colloq. Internat. CNRS 13, Centre National de la Recherche Scientifique, Paris, 1949, pp. 23–27.
[20] M. J. Dupré and F. J. Tipler, *New axioms for rigorous Bayesian probability*, Bayesian Anal., 4 (2009), pp. 599–606.
[21] B. Efron, *Bayes' theorem in the* 21*st century*, Science, 340 (2013), pp. 1177–1178.

[22] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, Probab. Math. Statist. 1, Academic Press, New York, London, 1967.
[23] D. A. Freedman, *On the asymptotic behavior of Bayes' estimates in the discrete case*, Ann. Math. Statist., 34 (1963), pp. 1386–1403.
[24] D. A. Freedman, *On the Bernstein-von Mises theorem with infinite-dimensional parameters*, Ann. Statist., 27 (1999), pp. 1119–1140.
[25] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed., Texts in Statistical Science Series, Chapman & Hall/CRC, Boca Raton, FL, 2004.
[26] A. Gelman, X.-L. Meng, and H. Stern, *Posterior predictive assessment of model fitness via realized discrepancies*, Statist. Sinica, 6 (1996), pp. 733–807.
[27] P. Gustafson and L. Wasserman, *Local sensitivity diagnostics for Bayesian inference*, Ann. Statist., 23 (1995), pp. 2153–2167.
[28] J. Y. Halpern, *A counterexample to theorems of Cox and Fine*, J. Artificial Intelligence Res., 10 (1999), pp. 67–85.
[29] J. Y. Halpern, *Cox's theorem revisited. Technical addendum to: "A counterexample to theorems of Cox and Fine" [J. Artificial Intelligence Res.,* 10 (1999)*, pp.* 67–85*]*, J. Artificial Intelligence Res., 11 (1999), pp. 429–435.
[30] F. R. Hampel, *A general qualitative definition of robustness*, Ann. Math. Statist., 42 (1971), pp. 1887–1896.
[31] M. Hardy, *Scaled Boolean algebras*, Adv. in Appl. Math., 29 (2002), pp. 243–292.
[32] J. A. Hausman and W. E. Taylor, *A generalized specification test*, Econom. Lett., 8 (1981), pp. 239–245.
[33] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, 2nd ed., Wiley Series in Probability and Statistics, Wiley, Hoboken, NJ, 2009.
[34] E. T. Jaynes, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
[35] I. M. Johnstone, *High dimensional Bernstein–von Mises: Simple examples*, in Borrowing Strength: Theory Powering Applications: A Festschrift for Lawrence D. Brown, Inst. Math. Stat. Collect. 6, Inst. Math. Statist., Beachwood, OH, 2010, pp. 87–98.
[36] B. J. K. Kleijn and A. W. van der Vaart, *Misspecification in infinite-dimensional Bayesian statistics*, Ann. Statist., 34 (2006), pp. 837–877.
[37] B. J. K. Kleijn and A. W. van der Vaart, *The Bernstein-Von-Mises theorem under misspecification*, Electron. J. Stat., 6 (2012), pp. 354–381.
[38] L. Le Cam, *On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates*, Univ. California Publ. Statist., 1 (1953), pp. 277–329.
[39] L. Le Cam and L. Schwartz, *A necessary and sufficient condition for the existence of consistent estimates*, Ann. Math. Statist., 31 (1960), pp. 140–150.
[40] H. Leahu, *On the Bernstein–von Mises phenomenon in the Gaussian white noise model*, Electron. J. Stat., 5 (2011), pp. 373–404.
[41] D. V. Lindley, *Making Decisions*, 2nd ed., Wiley, London, 1985.
[42] D. G. Mayo, *How can we cultivate Senn's ability?*, Rationality Markets Morals, 3 (2012), pp. 14–18.
[43] D. G. Mayo and A. Spanos, *Methodology in practice: Statistical misspecification testing*, Philos. Sci., 71 (2004), pp. 1007–1025.
[44] I. Mizera, *Qualitative robustness and weak continuity: The extreme unction*, in Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurecková, Inst. Math. Stat. Collect. 7, Inst. Math. Statist., Beachwood, OH, 2010, pp. 169–181.
[45] H. Owhadi and C. Scovel, *Qualitative Robustness in Bayesian Inference*, preprint, arXiv: 1411.3984, 2014.
[46] H. Owhadi and C. Scovel, *Brittleness of Bayesian inference and new Selberg formulas*, Commun. Math. Sci., to appear (2015); arXiv:1304.7046.
[47] H. Owhadi, C. Scovel, and T. J. Sullivan, *Brittleness of Bayesian inference under finite information in a continuous world*, Electron. J. Stat., 9 (2015), pp. 1–79.
[48] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz, *Optimal uncertainty quantification*, SIAM Rev., 55 (2013), pp. 271–345.
[49] S. T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, Wiley, Chichester, UK, 1991.
[50] L. Schwartz, *On Bayes procedures*, Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4 (1965), pp. 10–26.
[51] S. Senn, *You may believe you are a Bayesian but you are probably wrong*, Rationality Markets Morals, 2 (2011), pp. 48–66.

[52] C. R. Shalizi, *Dynamics of Bayesian updating with dependent data and misspecified models*, Electron. J. Stat., 3 (2009), pp. 1039–1074.
[53] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Numer., 19 (2010), pp. 451–559.
[54] R. Tibshirani and L. A. Wasserman, *Sensitive parameters*, Canad. J. Statist., 16 (1988), pp. 185–192.
[55] A. W. van der Vaart and J. H. van Zanten, *Rates of contraction of posterior distributions based on Gaussian process priors*, Ann. Statist., 36 (2008), pp. 1435–1463.
[56] R. von Mises, *Mathematical Theory of Probability and Statistics*, H. Geiringer, ed., Academic Press, New York, 1964.
[57] A. Wald, *Statistical Decision Functions*, Wiley, New York, 1950.
[58] L. Wasserman, M. Lavine, and R. L. Wolpert, *Linearization of Bayesian robustness problems*, J. Statist. Plann. Inference, 37 (1993), pp. 307–316.
[59] L. Wasserman and T. Seidenfeld, *The dilation phenomenon in robust Bayesian inference*, J. Statist. Plann. Inference, 40 (1994), pp. 345–356.
[60] L. A. Wasserman, *Prior envelopes based on belief functions*, Ann. Statist., 18 (1990), pp. 454–464.
[61] H. White, *Maximum likelihood estimation of misspecified models*, Econometrica, 50 (1982), pp. 1–25.