CrossMark

Maximum a posteriori estimators as a limit of Bayes estimators

Robert Bassett[1] · **Julio Deride**[1]

Abstract Maximum a posteriori and Bayes estimators are two common methods of point estimation in Bayesian statistics. It is commonly accepted that maximum a posteriori estimators are a limiting case of Bayes estimators with 0–1 loss. In this paper, we provide a counterexample which shows that in general this claim is false. We then correct the claim that by providing a level-set condition for posterior densities such that the result holds. Since both estimators are defined in terms of optimization problems, the tools of variational analysis find a natural application to Bayesian point estimation.

1 Introduction

The purpose of this article is to relate two point estimates in Bayesian estimation: the maximum a posteriori (MAP) estimator and Bayes estimator. Both the MAP and Bayes estimator are defined in terms of optimization problems, so that any connection between MAP and Bayes estimators can be extended to a connection between corresponding optimization problems. It is commonly accepted [19, §4.1.2], [8, Thm.2.4.3], [16, §7.6.5] that *MAP estimation is the limit of Bayes estimation*. This relationship

✉ Robert Bassett
rbassett@math.ucdavis.edu

Julio Deride
jderide@math.ucdavis.edu

1   Department of Mathematics, University of California Davis, Davis, CA, USA

is appealing from a theoretical perspective because it allows MAP estimators to be subsumed by the statistical analysis and intuition of Bayes estimators. However, this assertion must be carefully studied, because it is not true in the general setting proposed in much of the literature. We apply the theory of variational analysis, a common tool used in the approximation of optimization problems, to investigate the relationship between MAP and Bayes estimators.

This paper revises the relationship between MAP and Bayes estimators, placing the theory on a solid mathematical foundation. First, we provide a counterexample to the commonly accepted notion of MAP estimators as a limit of Bayes estimators having 0-1 loss. We then provide additional conditions and resurrect the limiting relationship between the estimators by relying on the theory of variational analysis. Because each of the estimators is defined as the maximizer of a certain optimization problem, a natural setting for the analysis of this problem is the space of upper semi-continuous functions, where we use the appropriate topology for convergence of maximizers, namely *hypo-convergence*. In general, the approach in this paper is applicable to any estimator defined in terms of an optimization problem.

This paper is not the first to apply variational analytic techniques to statistical estimation. One of the earliest examples is the notion of epi-convergence in distribution, which was introduced in [23] and developed further in [17] and [13]. The majority of applications of epi-convergence in distribution [14,15,18], and general applications of variational analytic theory to statistical estimation problems [1,5,9,11,12,24], have focused on asymptotic consistency of estimators: the notion of approximating a *true* optimization problem by one constructed from a finite sample, as the size of the sample tends to infinity.

This paper differs from the literature in a number of ways. First, we consider the limit of optimization problems defined over the same measure space. Instead of contrasting the empirical solution to the true, we instead assume that the function to be optimized is changing, but that the "true" underlying measure is known a priori. In this sense, we focus more on the approximation of loss functions than on the measure over which their expected value is taken. We also focus on almost-sure convergence results, as opposed to the weaker notion of distributional convergence. Lastly, the convergence in this paper deals explicitly with the Bayesian framework and the relationship between MAP and Bayes estimators.

The rest of this paper is organized as follows. In Sect. 2, a mathematical formulation of the Bayesian point estimation framework is reviewed, including the commonly-accepted argument of the relationship between MAP and Bayes estimators. Section 3 gives variational analytic background, introducing the notion of an upper semi-continuous density and the main convergence results for the hypo-convergence topology. Section 4 provides an example where a sequence of Bayes estimators corresponding to 0–1 loss does not converge to a MAP estimator. In light of the counterexample, some condition is required for MAP estimation to be a limiting case of Bayesian estimation. The informal arguments in the literature give the misleading impression that no condition is needed. We provide a necessary condition in Sect. 5, and use it to prove positive results relating MAP and Bayes Estimators.

We conclude this section with a comment on our notation. Preserving the notation in [20], we use Greek letters ν and η to denote sequence indices. We use superscript indexing, so that x^ν is an index of the sequence x.

2 Bayesian background

In this section we review necessary Bayesian preliminaries. In point estimation, we wish to infer the value of an unknown parameter θ from an observation x.

A *parametric model* is a family of distributions $p(\cdot|\theta) : \mathscr{X} \to \mathbb{R}$ indexed by a set $\Theta \subseteq \mathbb{R}^n$. The set Θ is referred to as the parameter space, and the measurable space $\mathscr{X}$ is the called the sample space. We assume that each of these distributions is absolutely continuous with respect to Lebesgue measure; hence each $p(\cdot|\theta)$ is a probability density function.

A *Bayesian model* is a parametric model and a distribution π on Θ. The distribution π is called a *prior* on the parameter space Θ. The Bayesian point estimation problem is the following: Given an observation x of a random variable X on $\mathscr{X}$, find $\hat{\theta} \in \Theta$ such that $p(\cdot|\hat{\theta})$ is a "best choice" for the distribution of X among distributions in the parametric model. In this sense, a Bayesian point estimate is a function $\hat{\theta} : \mathscr{X} \to \Theta$ which takes observations to parameter values. Because we only refer to Bayesian point estimates in this paper, we simply refer to them as estimators.

Given a Bayesian model and an observation x, we define a posterior distribution on the parameter space Θ through Bayes' rule

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int_{z\in\Theta} p(x|z)\pi(z)\,dz} \tag{1}$$

We assume that for each $x \in \mathscr{X}$, $\int_{z\in\Theta} p(x|z)\pi(z)\,dz$ is finite and nonzero, so that (1) defines a density. By taking $\pi(\theta|x) = 0$ outside of Θ, we extend the posterior density so that it is defined on all of $\mathbb{R}^n$ for each x. Hence, without loss of generality, we assume that $\Theta = \mathbb{R}^n$.

A common method of point estimation is through Bayes estimators. Given a loss function $L : \Theta \times \Theta \to [0, \infty)$ which quantifies cost for discrepancies in the true and estimated parameter value, a *Bayes estimator* is an estimator $\hat{\theta}_B$ which minimizes posterior expected loss.

$$\hat{\theta}_B(x) \in \underset{\theta\in\Theta}{\operatorname{argmin}} \, \mathbb{E}^{z|x}\left[L(\theta, z)\right] = \underset{\theta\in\Theta}{\operatorname{argmin}} \int_{z\in\Theta} L(\theta, z)\pi(z|x)\,dz. \tag{2}$$

The flexibility and ubiquity of (2) should not be understated. With different choices of loss functions, one can define the posterior mean, median and other quantiles [19, 2.5], as well as a variety of robust variants through expected loss minimization [7, 10.6]. For our purposes, we will focus on one particular family of loss functions, the 0–1 loss functions. The 0–1 loss function L^c is defined for any $c > 0$ as

$$L^c(\theta, z) = \begin{cases} 0 & \text{for } ||\theta - z|| < \frac{1}{c} \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

In the above and what follows, $||\cdot||$ denotes the standard Euclidean norm. Because of the topological nature of the arguments of that follow, any equivalent norm could be used instead. We focus on standard Euclidean for ease of exposition.

The rest of this paper will deal almost exclusively with the 0-1 loss, so we emphasize our notation: superscript L denotes the 0-1 loss function. We also denote by $\hat{\theta}_B^c$ the Bayes estimator associated with the 0-1 loss function L^c.

Another popular estimation procedure maximizes the posterior density directly. This defines a *maximum a posteriori estimator*, $\hat{\theta}_{MAP}$, which is given by the set of modes of the posterior distribution

$$\hat{\theta}_{MAP}(x) \in \underset{\theta}{\operatorname{argmax}}\, \pi(\theta|x).$$

This estimator can be interpreted as an analog of maximum likelihood for Bayesian estimation, where the distribution has become a posterior. A number of sources [19, §4.1.2], [8, Thm.2.4.3], [16, §7.6.5], [6, 10] claim that maximum a posteriori estimators are limits of Bayes estimators, in the following sense. Consider the sequence of 0–1 loss functions, $\{L^\nu : \mathbf{R}^n \times \mathbf{R}^n \to [0, +\infty)\}_{\nu \in \mathbf{N}}$, defined as

$$L^\nu(\theta, z) = \begin{cases} 0 & \text{for } \|\theta - z\| < \frac{1}{\nu} \\ 1 & \text{otherwise} \end{cases}, \tag{4}$$

and define $\hat{\theta}_B^\nu$ as the Bayes estimator associated to the loss function L^ν, for each ν, i.e.,

$$\hat{\theta}_B^\nu(x) \in \underset{\theta \in \Theta}{\operatorname{argmin}}\, E^{z|x}\left[L^\nu(\theta, z)\right].$$

First we translate the Bayes estimator to a maximization problem.

$$\hat{\theta}_B^\nu(x) \in \underset{\theta \in \Theta}{\operatorname{argmin}} \int_{z \in \Theta} L^\nu(\theta, z)\pi(z|x)\, dz$$

$$= \underset{\theta \in \Theta}{\operatorname{argmin}} \left(1 - \int_{\|\theta - z\| < \frac{1}{\nu}} \pi(z|x)\, dz\right)$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \int_{\|\theta - z\| < \frac{1}{\nu}} \pi(z|x)\, dz. \tag{5}$$

The claim is that the sequence $\hat{\theta}_B^\nu$ converges to $\hat{\theta}_{MAP}$. When the justification is provided, it proceeds as follows. Taking the limit as $\nu \to \infty$, we have

$$\lim_{\nu \to \infty} \hat{\theta}_B^\nu(x) \in \lim_{\nu \to \infty} \underset{\theta \in \Theta}{\operatorname{argmax}} \int_{\|\theta - z\| < \frac{1}{\nu}} \pi(z|x)\, dz$$

$$= \underset{\theta \in \Theta}{\operatorname{argmax}}\, \pi(\theta|x)$$

$$= \hat{\theta}_{MAP}(x). \tag{6}$$

This justification does not hold in general, and in fact MAP estimators are not necessarily a limiting case of Bayes estimators under the 0–1 loss indicated above. In Sect. 4, we exhibit a continuous and unimodal posterior density which is a counterexample to the claim. The problem lies in the limiting argument in (6)—the limit of maximizers is not a maximizer without additional conditions, which we establish in Sect. 5.

It is worthwhile to note what is correct about the argument in (6). Denoting by s_n the volume of the unit ball in $\mathbb{R}^n$, we have that

$$\lim_{\nu \to \infty} s_n \cdot \nu^n \cdot \int_{||\theta - z|| < \frac{1}{\nu}} \pi(z|x) \, dz = \pi(\theta|x)$$

Θ-almost everywhere by the Lebesgue differentiation theorem. This gives that a scaled version of $\int_{||\theta - z|| < \frac{1}{\nu}} \pi(z|x) \, dz$ converges pointwise a.e. to $\pi(\theta|x)$. But pointwise convergence does not guarantee convergence of maximizers. This will require the notion of hypo-convergence of functions, which we introduce in the next section.

3 Convergence of maximizers for non-random functions

3.1 General setting

This section summarizes the main results for convergence of optimization problems. The more common approach in the optimization literature is that of minimization, rather than maximization, where the theory of epi-convergence is developed for extended real-valued functions. Here, an adaptation to the maximization setting is presented, which is more natural in our estimation setting.

A function $f : \mathbf{R}^n \to \overline{\mathbf{R}}$ is said to be *proper* if f is not constantly $-\infty$ and never takes the value ∞. The *effective domain* of the function f is the set

$$\text{dom } f = \{x| f(x) < \infty\},$$

and its *hypograph* is the set in $\mathbf{R}^{n+1}$

$$\text{hypo } f = \{(x, \lambda)| \lambda \leq f(x)\}.$$

The function f is called *upper semi-continuous* (usc) if its hypograph is a closed subset of $\mathbf{R}^{n+1}$. An equivalent condition is that for each $\alpha \in \mathbb{R}$ the upper level set of f

$$\text{lev}_{\geq \alpha} f = \{x \in \mathbf{R}^n| f(x) \geq \alpha\}$$

is closed. A sequential definition of upper semi-continuity can also be stated: for each $x \in \mathbf{R}^n$, and each sequence x^ν converging to x,

$$\limsup_{x^\nu \to x} f(x^\nu) \leq f(x).$$

We say that a sequence of functions f^ν *hypo-converges* to a function f if both of the following hold for each $x \in \mathscr{X}$.

$$\liminf_\nu f^\nu(x^\nu) \geq f(x) \text{ for some } x^\nu \to x$$
$$\limsup_\nu f^\nu(x^\nu) \leq f(x) \text{ for every } x^\nu \to x,$$

The notion of hypo-convergence is well-developed because of its importance in proving properties about sequences of optimization problems. An equivalent definition of hypo-convergence follows by identifying each function with its hypograph, and applying the notion of set convergence à-la Painlevé–Kuratowski. Or, one can characterize hypo-convergence via the Attouch–Wets topology on hypographs. We refer the reader to [20, Ch.7] for details on the former and [2, Ch.3] on the latter.

The following alternative characterization of hypo-convergence will be useful in the sections that follow.

Proposition 1 [20, 7.29] $f^\nu \xrightarrow{hypo} f$ *with* f *usc if and only if the following two conditions hold:*

$$\limsup_\nu \left(\sup_B f^\nu \right) \leq \sup_B f \text{ for every compact set } B \subseteq \mathbb{R}^n$$
$$\liminf_\nu \left(\sup_O f^\nu \right) \geq \sup_O f \text{ for every open set } O \subseteq \mathbb{R}^n$$

A sequence of functions f^ν is said to be *eventually level-bounded* if for each $\alpha \in \mathbb{R}$ the sequence of upper level sets $\text{lev}_{\geq \alpha} f^\nu$ is eventually bounded. That is, there exists a bounded set M, $\alpha \in \mathbb{R}$ and $m \in \mathbb{N}$ such that for all $\nu \geq m$, $\text{lev}_{\geq \alpha} f^\nu \subseteq M$.

The wide-spread adoption of hypo-convergence in optimization is largely due to the following theorem and its related extensions

Theorem 1 [20, 7.33] *Assume* $f^\nu \xrightarrow{hypo} f$, *where* f^ν *is eventually level-bounded and* $\{f^\nu, f\}$ *are usc and proper. Then*

$$\sup f^\nu \to \sup f$$

and any point which is a limit of maximizers of f^ν *maximizes* f.

This theorem is attractive because it effectively gives the convergence of maximizers for approximations to optimization problems. In our application to Bayesian point estimation, we wish to establish a similar theorem providing convergence of estimators for approximations to MAP and Bayes estimation. We do so in Sect. 5.

3.2 Upper semi-continuous densities

Before proceeding, we list the primary assumption that we use throughout the paper.

Assumption 2 For each $x \in \mathscr{X}$, the random vector θ has a density $\pi(\theta|x) : \mathbb{R}^n \to \overline{\mathbb{R}}$ (with respect to Lebesgue measure μ) which is continuous almost everywhere. In other words, for each x there is a set $C \subseteq \mathbb{R}^n$ such that $\pi(\theta|x)$ is continuous at each point in C and $\mu(\mathbb{R}^n \backslash C) = 0$.

In this framework, we allow a density $\pi(\theta|x)$ to take ∞ as a value, so as not to rule out common densities like the gamma distribution. Obviously a density cannot take ∞ as a value on a set of positive measure.

When dealing with arbitrary measurable functions, the notion of pointwise value is inherently ambiguous because measurable functions are only unique up to alteration on sets of measure zero. The added structure of continuity allows us to define the notion of pointwise value, by referring to the pointwise value of the unique continuous representative of the measurable function. We would like to generalize this to a broader class of functions.

Assumption 2 provides some minimalist structure for which we can also define an unambiguous notion of pointwise value of a function. This is necessary because without some method of defining a pointwise value of a density, the notion of maximizing the density (as required in MAP estimation) is meaningless.

For each x, take $\pi(\theta|x)$ to be its continuous version on C. On $\mathbb{R}^n \backslash C$, take $\pi(\theta|x)$ to be the upper semi-continuous envelope of $\pi(\theta|x)|_C$. That is, for any $\theta \in \mathbb{R}^n$

$$\pi(\theta|x) = \sup\{\limsup \pi(\theta^v|x) : \theta^v \to \theta, \theta^v \subset C\}. \tag{7}$$

The set on the right side of (7) is nonempty because the complement of a measure zero set is dense. Furthermore, because $\mathbb{R}^n \backslash C$ has measure zero, the integral of $\pi(\theta|x)$ over $\mathbb{R}^n$ is unchanged by taking the upper semi-continuous envelope. We refer to densities which satisfy Assumption 2 and (7) as *upper semi-continuous (usc) densities*. In the remaining sections, densities are assumed to be upper semi-continuous.

Upper semi-continuous densities are a natural class of functions to consider because they provide a large amount of flexibility while still retaining the notion of pointwise value. They contain continuous densities as a special case, but also include intuitive concepts like histograms. Many nonparametric density estimation procedures also produce functions in this family, e.g. [3,21], and kernel density estimates with piecewise continuous kernels.

4 Counterexample

This section provides a counterexample to the following claim: Any limit of Bayes estimators $\hat{\theta}_B^{c^v}$ having 0-1 loss L^{c^v} with $c^v \to \infty$ is a MAP estimator. First, we provide the Bayesian model. Let $\Theta = \mathbb{R}$ and $\mathscr{X} = \mathbb{R}$. Take $p(x|\theta)$ to be a standard normal distribution. In other words, the parameter θ has no influence on the distribution of X. While trivial in the fact that θ does not influence X, this example will facilitate computation and illustrate the issue with the argument in (6). Consider a prior density $\pi : \mathbb{R} \to \mathbb{R}$ given by
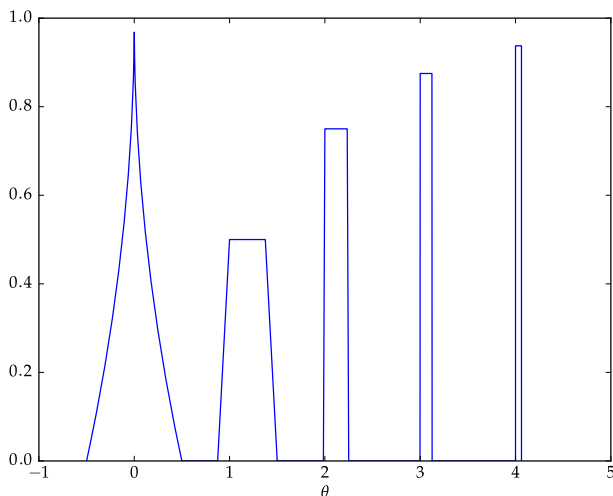
Fig. 1 Posterior density π

$$\pi(\theta) = \begin{cases} 1 - \sqrt{|2\theta|} & \theta \in \left(\frac{-1}{2}, \frac{1}{2}\right) \\ (2^n - 1)4^n \left(\theta + \frac{1}{8^n} - n\right) & \theta \in \left[n - \frac{1}{8^n}, n\right], \quad n \in \mathbb{N}_+ \\ 1 - \frac{1}{2^n} & \theta \in \left[n, n + \frac{1}{2^n} - \frac{1}{8^n}\right], \quad n \in \mathbb{N}_+ \\ (1 - 2^n)4^n \left(\theta - \left(n + \frac{1}{2^n}\right)\right) & \theta \in \left[n + \frac{1}{2^n} - \frac{1}{8^n}, n + \frac{1}{2^n}\right], \quad n \in \mathbb{N}_+ \\ 0 & \text{otherwise} \end{cases}.$$

This density is depicted in Fig. 1. One can easily check that π is indeed a continuous density function. Because θ and X are independent in this Bayesian model, equation (1) gives that the posterior is equal to the prior. Thus, for the remainder of this section, we will refer to π as the posterior distribution, and drop the dependence on x in the notation. We also comment that any choice of parametric model where altering θ does not change $p(x|\theta)$ will give rise to the same posterior distribution—the standard normal assumption is just for concreteness.

Note that the posterior density π has a unique maximum, or *posterior mode*, at $\theta = 0$, where it takes the value 1. Therefore, the MAP estimator of this Bayesian model is $\hat{\theta}_{MAP} = 0$. On the other hand, considering the 0-1 loss function L^c, we recall the equivalent definition of the associated Bayes estimator from equation (5)

$$\hat{\theta}_B^c \in \underset{\theta}{\operatorname{argmax}} \int_{||z-\theta|| < \frac{1}{c}} \pi(z)\, dz. \tag{8}$$

We next analyze the limiting behavior of the sequence of Bayes Estimators $\hat{\theta}_B^c$ when $c \to \infty$.

Consider the sequence of scalars $\{c^\nu = 2 \cdot 4^\nu : \nu \in \mathbb{N}_+\}$, and the associated family of loss functions L^{c^ν}. We will prove that for the Bayesian model that we have constructed, $\hat{\theta}_B^{c^\nu} \not\to \hat{\theta}_{MAP}$. This will show that a limit of Bayes estimators with 0-1 loss is

not necessarily a MAP estimator. In order to find the maximum in (8), we can consider the maximum on each nonzero piece of π. For the non-zero portion around the origin, the integral in (8) is obviously maximized at 0. Furthermore, for each "bump" of π, the integral is maximized (perhaps non-uniquely) at the midpoint of the interval where the density is constant. In order to show that $\hat{\theta}_B^{c^\nu} \not\to \hat{\theta}_{MAP}$, it suffices to show that for each ν there is a $\theta \notin (-1/2, 1/2)$ such that the evaluation of (8) at θ is strictly greater than the evaluation at zero. This gives that $\hat{\theta}_B^{c^\nu}$ cannot have a limit point in $(-1/2, 1/2)$, and hence cannot have 0, the MAP estimator, as a limit point. We now perform the required calculations.

(i) Evaluation of (8) at 0.

$$\int_{|z| < \frac{1}{2 \cdot 4^\nu}} 1 - \sqrt{|2z|} \, dz = 2 \int_0^{\frac{1}{2 \cdot 4^\nu}} 1 - \sqrt{2z} \, dz = \frac{1}{4^\nu} - \frac{4\sqrt{2}}{3 \cdot 2^{\frac{3}{2}}} \cdot \frac{1}{8^\nu} \qquad (9)$$

(ii) For each $\nu \in \mathbb{N}$, evaluating (8) at $\theta = 2\nu + \frac{1}{2^{2\nu+1}}$ gives

$$\int_{\left| z - \left(2\nu + \frac{1}{2^{2\nu+1}} \right) \right| < \frac{1}{2 \cdot 4^\nu}} \pi(z) \, dz$$

$$= \int_{\left[2^\nu, 2^\nu + \frac{1}{2^{2\nu}} - \frac{1}{8^{2\nu}} \right]} \pi(z) \, dz + \int_{\left[2^\nu + \frac{1}{2^\nu} - \frac{1}{8^{2\nu}}, 2^\nu + \frac{1}{2^{2\nu}} \right]} \pi(z) \, dz,$$

The first part of this sum is an integral over a constant piece of π. The second is a linear piece of π. Bounding the sum of the integrals below by the value of just the integral over the constant piece, we have that

$$\int_{\left| z - \left(2\nu + \frac{1}{2^{2\nu+1}} \right) \right| < \frac{1}{2 \cdot 4^\nu}} \pi(z) \, dz \geq \left(1 - \frac{1}{2^{2\nu}} \right) \left(\frac{1}{2^{2\nu}} - \frac{1}{8^{2\nu}} \right)$$

$$= \frac{1}{4^\nu} - \left(2 - \frac{1}{16^\nu} \right) \frac{1}{16^\nu}. \qquad (10)$$

Since (10) is strictly greater than (8) for all $\nu \geq 1$, we conclude that $\hat{\theta}_B^{c^\nu} \notin (\frac{-1}{2}, \frac{1}{2})$ for all $\nu \geq 1$. Hence the sequence of Bayes estimators $\hat{\theta}_B^{c^\nu}$ has $c^\nu \to \infty$, but does not have a MAP estimator as a limit point. This concludes the counterexample and shows that the claim is false.

5 Convergence results

In this section we provide conditions on posterior distributions which guarantee that a sequence of Bayes Estimators with 0-1 loss has a MAP estimator as a limit point. In addition, we provide a partial converse by showing that each MAP estimator is a limit of approximate Bayes estimators.

Define an estimator $\hat{\theta}$ of θ to be an ϵ-*approximate Bayes estimator* with respect to loss L if for each $x \in \mathscr{X}$.

$$\int_{z \in \Theta} L(\hat{\theta}(x), z) \pi(z|x) \, dz \geq \sup_{\theta \in \Theta} \int_{z \in \Theta} L(\theta, z) \pi(z|x) \, dz - \epsilon(x).$$

Here, ϵ is a function from $\mathscr{X}$ to $\mathbb{R}$. We say that $\hat{\theta}$ is a *limit of approximate Bayes estimators* if there are sequences V^ν, $\hat{\theta}^\nu$ and ϵ^ν such that $\hat{\theta}^\nu$ converges almost surely to $\hat{\theta}$, $\hat{\theta}^\nu$ is an ϵ^ν-approximate Bayes estimator with respect to loss V^ν for each ν, and ϵ^ν converges $\mathscr{X}$-surely to 0.

As discussed in the introduction, we assume that all densities are upper semi-continuous and satisfy Assumption 2.

We begin with a deterministic result. The proof of the next lemma, which is related to the epi-convergence of mollifying approximates from [20], is included in the "Appendix A".

Lemma 1 *Assume $f : \Theta \to \overline{\mathbb{R}}$ is an upper semi-continuous density. Let s_n denote the volume of the unit ball in $\mathbb{R}^n$. Define*

$$f^\nu(\theta) := \nu^n \cdot s_n \cdot \int_{||\theta - z|| < \frac{1}{\nu}} f(z) dz.$$

Then f^ν hypo-converges to f.

We note that even though upper semi-continuous densities may take ∞ as a value, and hence need not be proper, f^ν must be proper because f integrates to one.

This lemma allows us to prove the following result.

Theorem 3 *If $\pi(\theta|x)$ is proper $\mathscr{X}$ almost surely, then any MAP estimator is a limit of approximate Bayes estimators.*

Proof Let $\hat{\theta}_{MAP}$ be a maximum a posteriori estimator of θ. By definition, $\hat{\theta}_{MAP}(x) \in \operatorname{argmax} \pi(\theta|x)$. Let x be any element in $\mathscr{X}$. For ease of notation, we will drop the explicit dependence on x by writing $\hat{\theta}_{MAP} := \hat{\theta}_{MAP}(x)$ and letting $f(\theta) := \pi(\theta|x)$. By Lemma 1, $f^\nu \xrightarrow{hypo} f$. From the definition of hypo-convergence, there is a sequence $\theta^\nu(x) \to \hat{\theta}_{MAP}(x)$ such that $\limsup_\nu f^\nu(\theta^\nu) \leq f(\hat{\theta}_{MAP})$. Also directly from the definition, $\liminf_\nu f^\nu(\theta^\nu) \geq f(\hat{\theta}_{MAP})$. Hence $\lim_\nu f^\nu(\theta^\nu) = f(\hat{\theta}_{MAP})$.

Repeating this construction pointwise for each $x \in \mathscr{X}$, we define a sequence of estimators with $\hat{\theta}^\nu(x)$ the θ^ν sequence above. Define $\epsilon^\nu(x)$ as $\sup_{\theta \in \Theta} f^\nu(\theta) - f^\nu(\theta^\nu)$. We claim that L^ν, $\hat{\theta}^\nu$, and ϵ^ν satisfy the conditions so that $\hat{\theta}_{MAP}$ is an approximate Bayes estimator. We must verify the three conditions in the definition. The first two, that $\hat{\theta}^\nu$ converges almost surely to $\hat{\theta}_{MAP}$ and that $\hat{\theta}^\nu$ is an ϵ^ν-approximate Bayes estimator, are true by construction. Lastly, we must show that $\epsilon^\nu \to 0$ almost surely. By monotonicity of the integral, we know that

$$f^\nu(\theta^\nu) \leq \sup_\theta f^\nu(\theta) \leq f(\hat{\theta}_{MAP}).$$

For each $x \in \mathcal{X}$ with $\pi(\theta|x)$ proper, we combine this inequality with the fact that $\lim_\nu f^\nu(\theta^\nu) = f(\hat{\theta}_{MAP})$ to arrive at the following

$$0 \le \epsilon^\nu(x) = \sup f^\nu - f^\nu(\theta^\nu) \le f(\hat{\theta}_{MAP}) - f^\nu(\theta^\nu) \to 0$$

Note that $\pi(\theta|x)$ must be proper for this to hold: it guarantees that $f(\hat{\theta}_{MAP}) < \infty$, and hence $f(\hat{\theta}_{MAP}) = \lim_\nu f^\nu(\theta^\nu)$ gives $f(\hat{\theta}_{MAP}) - \lim_\nu f^\nu(\theta^\nu) = 0$.

We conclude that $\epsilon^\nu(x) \to 0$ almost everywhere, since $\pi(\theta|x)$ is proper $\mathcal{X}$ almost everywhere. This shows the third condition in the definition, so that $\hat{\theta}_{MAP}$ is a limit of approximate Bayes estimators. $\qquad\square$

The notion of an approximate Bayes estimator captures the idea that Bayes and MAP estimators are close in terms of evaluation of the objective. They need not be close in distance to each other in Θ. This point is subtle, and underlies the confusion in the incorrect claim in (6). In fact, as the counterexample in the previous section shows, Theorem 3 cannot be strengthened from approximate to true Bayes estimators without additional conditions.

We turn now to providing those conditions, in a sort of converse to Theorem 3. We turn our focus to characterizing when a sequence of Bayes estimators converges to a MAP estimator. We again begin with a deterministic result.

Theorem 4 *Assume there exists α such that $\mathrm{lev}_{\ge\alpha} f$ is bounded and has nonempty interior. Then $\mathrm{argmax} f^\nu$ is eventually nonempty, and any sequence maximizers of f^ν as $\nu \to \infty$ has a maximizer of f as a limit point.*

Proof Let f be an usc density. Assume there is an $\alpha \in \mathbb{R}$ such that the upper level set $\mathrm{lev}_{\ge\alpha} f$ is bounded and has nonempty interior. Note that $\mathrm{lev}_{\ge\alpha}$ is compact by the Heine–Borel theorem because upper level sets are closed when f is upper semicontinuous.

First, we show that there is an index ν_0 such that $\mathrm{argmax} f^\nu$ is nonempty for each $\nu > \nu_0$. Then we show that any sequence of maximizers of f^ν, has as a limit point a maximizer of f.

To show the existence of a maximizer, we first show that f^ν is upper semicontinuous with a bounded, nonempty upper level set. That f^ν attains its maximum then follows because an upper semi-continuous function attains its maximum on a compact set. We recall from Sect. 3 that our framework allows that this maximum could be ∞.

Fix $\nu \in \mathbb{N}$ and $\theta \in \mathbb{R}^n$ Let $\{\theta^\eta\}_{\eta \in \mathbb{N}}$ be any sequence that converges to θ. To show upper semicontinuity we must have

$$\limsup_\eta f^\nu(\theta^\eta) \le f^\nu(\theta),$$

but this follows from the following chain of inequalities

$$\limsup_\eta f^\nu(\theta^\eta) = \limsup_\eta s_n \cdot \nu^n \cdot \int_{||\theta^\eta - z|| < \frac{1}{\nu}} f(z)\,dz$$

$$= \limsup_\eta s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} f(z + \theta^\eta)\,dz$$

$$\leq s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} \limsup_{\eta} f(z + \theta^\eta) \, dz$$

$$\leq s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} f(z + \theta) \, dz$$

$$= f^\nu(\theta).$$

The first inequality follows from Fatou's lemma, and the second from the upper semi-continuity of f. Hence f^ν is upper semicontinuous.

The family of functions f^ν has a bounded upper level set because $\mathrm{lev}_{\geq \alpha} f$ bounded with constant M implies that $\mathrm{lev}_{\geq \alpha} f^\nu$ is bounded with constant $M + \frac{1}{\nu}$. Lastly, we show that the upper level set $\mathrm{lev}_{\geq \alpha} f^\nu$ is nonempty. Let $\theta \in \mathrm{int} \, \mathrm{lev}_{\geq \alpha} f^\nu$, and denote by δ radius such that $\mathbb{B}(\theta, \delta) \subseteq \mathrm{int} \, \mathrm{lev}_{\geq \alpha} f^\nu$. Choose $\nu_0 \geq \frac{1}{\delta}$. Then for any $\nu > \nu_0$

$$f^\nu(\theta) = s_n \cdot \nu^n \int_{||\theta - z|| < \frac{1}{\nu}} f(z) \, dz \geq s_n \cdot \nu^n \cdot \alpha \cdot \frac{1}{\nu^n} \cdot s_n = \alpha.$$

So $\theta \in \mathrm{lev}_{\geq \alpha} f^\nu$. We conclude that this level set is nonempty and bounded, so that f^ν attains its maximum.

Now let $\hat{\theta}^\nu$ be any sequence of maximizers of f^ν. For $\nu > \nu_0$, $\hat{\theta}^\nu \in \mathrm{lev}_{\geq \alpha} f^\nu$. From Lemma 1 and Proposition 1

$$\liminf_\nu f^\nu(\hat{\theta}^\nu_B) = \liminf_\nu \sup_{\mathbb{R}^n} f^\nu \geq \sup_{\mathbb{R}^n} f^\nu$$

and

$$\limsup_\nu f^\nu(\hat{\theta}^\nu)$$

$$= \limsup_\nu (\sup_{\mathbb{R}^n} f^\nu)$$

$$= \limsup_\nu \left(\sup_{\overline{\mathbb{B}}(0, M+1)} f^\nu \right)$$

$$\leq \sup_{\overline{\mathbb{B}}(0, M+1)} f$$

$$= \sup_{\mathbb{R}^n} f.$$

So $\lim f^\nu(\hat{\theta}^\nu_B) = \sup_{\mathbb{R}^n} f$.

Since $\hat{\theta}^\nu$ is eventually in $\overline{\mathbb{B}}(0, M + 1)$, which is compact, it has a convergent subsequence $\hat{\theta}^{\nu_k} \to \tilde{\theta}$. By upper semi-continuity

$$f(\tilde{\theta}) \geq \limsup f(\hat{\theta}^{\nu_k}) = \lim f(\hat{\theta}^\nu).$$

Hence $\tilde{\theta}$ maximizes f. Furthermore, by upper semi-continuity any limit point of $\hat{\theta}^\nu$ maximizes f. This proves the theorem. □

We now prove our main result about the relationship between estimators.

Theorem 5 *Suppose that for each $x \in \mathscr{X}$, $\pi(\theta|x)$ satisfies the following property: There exists α such that*

$$\{\theta : \pi(\theta|x) > \alpha\}$$

is bounded and has nonempty interior. Then for any of sequence of Bayes estimators $\hat{\theta}_B^\nu$ with 0-1 loss L^ν

(i) *There is a MAP estimator $\hat{\theta}_{MAP}$ such that $\hat{\theta}_{MAP}(x)$ is a limit point of $\hat{\theta}_B^\nu(x)$ for each x.*

(ii) *Every limit point of $\hat{\theta}_B^\nu(x)$ defines a MAP estimator, in the sense that there is a MAP estimator $\hat{\theta}_{MAP}$ with $\hat{\theta}_{MAP}(x)$ equal to that limit point.*

Proof Fix $x \in \mathscr{X}$. By assumption, $\pi(\theta|x)$ has a bounded level set with nonempty interior. From the deterministic result, Theorem 4, the sequence $\hat{\theta}_B^\nu(x)$ has a limit point $\tilde{\theta}(x)$, and this limit point maximizes $\pi(\theta|x)$. Define a MAP estimator $\hat{\theta}_{MAP}(x) := \tilde{\theta}(x)$ pointwise for each $x \in \mathscr{X}$. This proves (i).

For (ii), the result follows since the method of defining a MAP estimator in the previous paragraph is valid for every limit point of $\hat{\theta}_B^\nu(x)$. □

As a consequence of the proof, we note the following: if $\hat{\theta}_B^\nu$ converges to some estimator $\hat{\theta}$ almost surely, then $\hat{\theta}$ is a MAP estimator.

We next establish related results for various shape constrained densities. Recall that a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is quasiconcave if for each $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$

$$f(\lambda x + (1 - \lambda)y) \geq \min\{f(x), f(y)\}.$$

For a quasiconcave function, local maximality implies global maximality. This shape constraint captures the notion of a probability distribution having a "single peak". It is also sometimes referred to as unimodality, but we avoid this terminology because it has multiple meanings in the literature.

Theorem 6 *Assume that for each $x \in \mathscr{X}$, $\pi(\theta|x)$ is quasiconcave. Then the conclusion of Theorem 5 holds.*

Proof For each x, since $\int \pi(z|x)\,dz = 1$, there is an $\alpha > 0$ such that $\mu(\{\theta : \pi(\theta|x) \geq \alpha\})$ has positive measure. By quasiconvexity, $\mathrm{lev}_{\geq\alpha}\,\pi(\theta|x) = \{\theta | \pi(\theta|x) \geq \alpha\}$ is convex.

Note $\mathrm{lev}_{\geq\alpha}$ having an interior point is equivalent to the set containing n affinely independent points. Suppose, towards a contradiction, that $\mathrm{lev}_{\geq\alpha}\,\pi(\theta|x)$ does not contain any interior points, then its affine hull lies in an $n - 1$ dimensional plane. This contradicts the set having positive measure. Hence $\{\theta : \pi(\theta|x)\}$ must have an interior point.

In order to apply Theorem 5 we must also show that the level set $\mathrm{lev}_{\geq\alpha}\,\pi(\theta|x)$ is bounded. Fix B to be any n-dimensional ball in $\mathrm{lev}_{\geq\alpha}\,\pi(\theta|x)$. If θ^ν were a sequence in $\mathrm{lev}_{\geq\alpha}\,\pi(\theta|x)$ such that $||\theta^\nu|| \to \infty$, then

$$\int_{z \in \Theta} \pi(z|x)\, dz \geq \int_{z \in \text{lev}_{\geq \alpha} \pi(\theta|x)} \pi(z|x)\, dz \geq \mu(\{\text{conv}\,(B \cup \{\theta^\nu\})\}) \cdot \alpha$$

Here, μ denotes Lebesgue measure. One can easily show that $\mu(\text{conv}\,(B \cup \{\theta^\nu\})) \to \infty$ when $\theta^\nu \to \infty$. This contradicts that $\int \pi(z|x)dz = 1$. Hence there cannot exist such a sequence, so the level set is bounded and the result is proven. $\qquad\square$

The following corollary about log-concave densities follows immediately. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is log-concave if for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$

$$f(\lambda x + (1 - \lambda)y) \geq f(x)^\lambda f(y)^{1-\lambda}$$

Log-concave densities have appeared in recent work of [4,22] due to their attractive computational and theoretical properties in nonparametric density estimation.

Corollary 1 *Assume that for each $x \in \mathcal{X}$, $\pi(\theta|x)$ is log-concave. Then the conclusion of Theorem 5 holds.*

Proof Log-concavity implies quasiconcavity, and the result follows immediately from the previous theorem. $\qquad\square$

A Proof of Lemma 1

In this appendix we provide the proof of Lemma 1.

Proof Let $\theta \in \mathbb{R}^n$. To show hypo-convergence, we must show that for each sequence $\theta^\nu \to \theta$, $\limsup_\nu f^\nu(\theta^\nu) \leq f(x)$ and that there exists a sequence $\theta^\nu \to \theta$ with $\liminf_\nu f^\nu(\theta^\nu) \geq f(\theta)$.

Fix $\epsilon > 0$. Since f is upper semi-continuous at θ, there is a $\delta > 0$ such that $||z - \theta|| < 2\delta$ gives $f(z) - f(\theta) < \epsilon$.

Consider any sequence $\theta^\nu \to \theta$. We have that

$$f^\nu(\theta^\nu) - f(\theta) = s_n \cdot \nu^n \cdot \int_{||\theta^\nu - z|| > \frac{1}{\nu}} (f(z) - f(\theta))dz$$

$$= s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} (f(z + \theta^\nu) - f(\theta))\, dz.$$

Choose $\nu_0 \in \mathbb{N}$ so that $||\theta - \theta^\nu|| < \delta$ and $\frac{1}{\nu} < \delta$ for each $\nu > \nu_0$. Then for any $\nu > \nu_0$,

$$s_n \cdot \nu^n \cdot \int_{||z|| < \frac{1}{\nu}} \left(f(z + \theta^\nu) - f(\theta)\right) dz \leq s_n \cdot \nu^n \cdot \epsilon \cdot \int_{||z|| < \frac{1}{\nu}} dz = \epsilon.$$

Thus $\limsup_\nu f^\nu(\theta^\nu) \leq f(\theta)$.

To establish the second part of the hypo-convergence definition, we focus our attention on constructing a sequence that satisfies the required inequality.

Consider any $\eta \in \mathbb{N}$. Recall that f^ν is an upper semi-continuous density. Let C be the set where f is continuous. Because C is dense, for each $\nu \in \mathbb{N}$, there is a $y^\nu \in C$ such that $||y^\nu - x|| < \frac{1}{\nu}$. Furthermore, $y^\nu \in C$ means that there is a $\delta(y^\nu, \eta) > 0$ such that any $z \in \Theta$ which satisfies $||y^\nu - z|| < \delta(y^\nu, \eta)$ also has

$$\left| f(y^\nu) - f(z) \right| < \frac{1}{\eta}.$$

Here we use function notation for δ to emphasize that δ depends on both y^ν and η.

For each η, define a sequence such that

$$z^{\nu,\eta} = \begin{cases} 0 & \text{when } \frac{1}{\nu} > \delta(y^1, \eta) \\ y^1 & \text{when } \delta(y^2, \eta) \leq \frac{1}{\nu} < \delta(y^1, \eta) \\ y^2 & \text{when } \delta(y^3, \eta) \leq \frac{1}{\nu} < \min\{\delta(y^2, \eta), \delta(y^1, \eta)\} \\ y^3 & \text{when } \delta(y^4, \eta) \leq \frac{1}{\nu} < \min_{i \leq 4} \delta(y^i, \eta) \\ \vdots & \vdots \end{cases}$$

Extracting a diagonal subsequence from the sequences generated according to this procedure gives a sequence θ^ν such that $\theta^\nu \to \theta$ and $\frac{1}{\nu} < \delta(\theta^\nu, \nu)$. In particular, $|f(\theta^\nu) - f(z)| < \frac{1}{\nu}$ for z with $||\theta^\nu - z|| < \frac{1}{\nu}$.

Hence, for any $\epsilon > 0$, choosing $\nu > \frac{2}{\epsilon}$ gives

$$\left| f^\nu(\theta^\nu) - f(\theta) \right| \leq \left| f^\nu(\theta^\nu) - f(\theta^\nu) \right| + \left| f(\theta^\nu) - f(\theta) \right|$$
$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

We conclude that $\lim_\nu f^\nu(\theta^\nu) = f(\theta)$, so the result is proven. □

References

1. Attouch, H., Wets, R.J.-B.: Epigraphical Processes: Laws of Large Numbers for Random LSC Functions. Dép. des Sciences Mathématiques, Paris (1994)
2. Beer, G.: Topologies on Closed and Closed Convex Sets. Mathematics and Its Applications. Springer, New York (1993)
3. Chan, S.-O., Diakonikolas, I., Servedio, R.A., Sun, X.: Efficient density estimation via piecewise polynomial approximation. In: Proceedings of the 46th Annual ACM Symposium on Theory of Computing, pp. 604–613. ACM (2014)
4. Dümbgen, L., Rufibach, K., et al.: Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. Bernoulli **15**(1), 40–68 (2009)
5. Dupacová, J., Wets, R.: Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. Ann. Stat. **16**, 1517–1549 (1988)
6. Figueiredo, M.A.T.: Lecture Notes on Bayesian Estimation and Classification. Instituto de Telecomunicacoes-Instituto Superior Tecnico, Lisboa (2004)
7. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning, vol. 1. Springer, New York (2001)

8. Geweke, J.: Contemporary Bayesian Econometrics and Statistics, vol. 537. Wiley, New York (2005)
9. Geyer, C.J.: On the asymptotics of constrained m-estimation. Ann. Stat. **22**, 1993–2010 (1994)
10. Hoegele, W., Loeschel, R., Dobler, B., Koelbl, O., Zygmanski, P.: Bayesian estimation applied to stochastic localization with constraints due to interfaces and boundaries. Math. Probl. Eng. **2013**, 5–6 (2013)
11. King, A.J., Rockafellar, R.T.: Asymptotic theory for solutions in statistical estimation and stochastic programming. Math. Oper. Res. **18**(1), 148–162 (1993)
12. King, A.J., Wets, R.J.B.: Epi-consistency of convex stochastic programs. Stoch. Stoch. Rep. **34**(1–2), 83–92 (1991)
13. Knight, K: Epi-convergence in distribution and stochastic equi-semicontinuity. Unpublished manuscript, 37 (1999)
14. Knight, K.: Limiting distributions of linear programming estimators. Extremes **4**(2), 87–103 (2001)
15. Knight, K., Fu, W.: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356–1378 (2000)
16. Lee, P.M.: Bayesian Statistics: An Introduction. Wiley, New York (2012)
17. Pflug, G.C.: Asymptotic Dominance and Confidence for Solutions of Stochastic Programs. International Institute for Applied Systems Analysis, Laxenburg (1991)
18. Pflug, G.C.: Asymptotic stochastic programs. Math. Oper. Res. **20**(4), 769–789 (1995)
19. Robert, C.: The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer, New York (2007)
20. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Grundlehren der mathematischen Wissenschaften. Springer, Berlin (2009)
21. Royset, J.O., Wets, R.J.B.: Nonparametric density estimation via exponential epi-eplines: fusion of soft and hard information. Technical report (2013)
22. Rufibach, K.: Computing maximum likelihood estimators of a log-concave density function. J. Stat. Comput. Simul. **77**(7), 561–574 (2007)
23. Salinetti, G., Wets, R.J.-B.: On the convergence in distribution of measurable multifunctions (random sets) normal integrands, stochastic processes and stochastic infima. Math. Oper. Res. **11**(3), 385–419 (1986)
24. Shapiro, A.: Asymptotic analysis of stochastic programs. Ann. Oper. Res. **30**(1), 169–186 (1991)