Bayesian Analysis (2019) **14**, Number 1, pp. 235–260

On Bayesian Oracle Properties

Wenxin Jiang*, and Cheng Li†

Abstract. When model uncertainty is handled by Bayesian model averaging (BMA) or Bayesian model selection (BMS), the posterior distribution possesses a desirable "oracle property" for parametric inference, if for large enough data it is nearly as good as the oracle posterior, obtained by assuming unrealistically that the true model is known and only the true model is used. We study the oracle properties in a very general context of quasi-posterior, which can accommodate non-regular models with cubic root asymptotics and partial identification. Our approach for proving the oracle properties is based on a unified treatment that bounds the posterior probability of model mis-selection. This theoretical framework can be of interest to Bayesian statisticians who would like to theoretically justify their new model selection or model averaging methods in addition to empirical results. Furthermore, for non-regular models, we obtain nontrivial conclusions on the choice of prior penalty on model complexity, the temperature parameter of the quasi-posterior, and the advantage of BMA over BMS.

MSC 2010 subject classifications: 62E99, 62F15.

Keywords: Bayesian model selection, consistency, model averaging, oracle property, cubic root asymptotics, partial identification.

1 Introduction

The terminology of frequentist *oracle property* was first introduced in Fan and Li (2001) for a frequentist penalization method in model selection, by which statistical inferences "work as well as if the correct submodel were known." Thereafter the oracle property has become a popular concept in the statistics literature. On the other hand, analogs of such an oracle property have not been widely studied in the Bayesian context, with the exception of a few recent works in special model setups (Ishwaran and Rao 2011, Castillo, Schmidt-Hieber, and van der Vaart 2015, Li and Jiang 2016, etc.)

In this paper, we define different versions of Bayesian oracle properties in a general framework with quasi-posteriors and present a systematic way to study them by bounding the probability of model mis-selection. In particular, we are interested in the interplay between several different subjects: Bayesian model averaging (BMA), Bayesian model selection (BMS) based on the *Maximum-A-Posteriori* (MAP) model, and Bayesian posterior inference based on the unknown true model (i.e. the oracle model). We reveal some surprisingly simple and general relations between these different topics, and discuss their applications in non-regular models with cubic root asymptotics and partial identification.

*Shandong University (Taishan Scholar Adjunct Professor) and Northwestern University, wjiang@northwestern.edu

†Corresponding author, National University of Singapore, stalic@nus.edu.sg

https://doi.org/10.1214/18-BA1097

We first introduce the basic notation we will use throughout this paper. Let $\mathbf{D}$ be the observed data with sample size n. Let M be a generic model index, and the "true model" M^* be a possible value of M which is related to the data generating mechanism. In Bayesian model averaging and model selection, we always consider a countable sequence of models $\{M_j\}$ indexed by $j = 1, 2, \ldots$, among which is the true model M^*. A prior probability $\pi(M_j)$ is assigned to each model M_j. Then each model M_j proposes a different prior density $\pi(\theta|M_j)$ for the parameter θ, supported on a parameter space Θ_j, which can possibly overlap. The full parameter space is $\Theta = \cup_{j\geq 1}\Theta_j$. The overall prior distribution with density $\pi(\theta)$ is given by

$$\pi(\theta) = \sum_{j\geq 1} \pi(\theta|M_j)\pi(M_j), \quad \text{for } \theta \in \Theta.$$

Given the model M_j and its proposed parameter θ, let $p(\mathbf{D}\,|\theta, M_j)$ be the likelihood function. Then the posterior density of θ through Bayesian model averaging (BMA) is given by

$$\pi(\theta|\,\mathbf{D}) \propto \sum_{j\geq 1} p(\mathbf{D}\,|\theta, M_j)\pi(\theta|M_j)\pi(M_j), \quad \text{for } \theta \in \Theta.$$

Throughout the paper, we use Π to denote the underlying probability measure associated with density π.

Below we explain why the Bayesian version of oracle properties is desirable for dimension reduction in standard regular models, why the more general quasi-Bayesian framework is useful, and why our work will be of interest to the community of Bayesian statisticians.

1.1 Bayesian oracle property is desirable for dimension reduction

Consider a simple example of linear regression with known error variance, where $y \sim N(\sum_{j=1}^{p} x_j\theta_j, 1)$, and $N(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2. Suppose that there exists an unknown true model M^*, in which only the first p^* components of $\theta = (\theta_1, \ldots, \theta_p)^\top$ are nonzero. Suppose we consider these nested candidate models $M_1, \ldots, M_p$, where the first j components of θ are nonzero if θ comes from the model M_j. Given an observed independent and identically distributed (i.i.d.) sample $\mathbf{D} = \{(y_i, x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n\}$, the BMA involves using the posterior

$$\pi\left(\theta|\,\mathbf{D}\right) \propto \sum_{j=1}^{p} e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{\ell=1}^{j} x_{i\ell}\theta_\ell)^2} \pi\left(\theta|M_j\right)\pi\left(M_j\right).$$

When $p \ll n$, we can set the prior $\pi(\theta|M_j)$ to be a component-wise independent product of normal priors, and $\pi(M_j) = 1/p$ as a uniform prior.

For this simple example, the Bayesian oracle property can be roughly described as

$$\pi(\theta|\,\mathbf{D}) \approx \pi(\theta|\,\mathbf{D}, M^*) \propto e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \sum_{\ell=1}^{p^*} x_{i\ell}\theta_\ell)^2} \pi\left(\theta|M^*\right),$$

which is the posterior based on the true model M^*, as if we knew the truth M^*. This approximation can be in the sense of total variation norm, or in some other sense depending on what is regarded as meaningful. This kind of result is desirable for automatic dimension reduction and variance reduction. If $p = 10$ but $p^* = 1$, then the mean squared error for estimating the mean function $\sum_{j=1}^{p} x_j\theta_j$ can be reduced from about $10/n$ when using the full model M_p, to about $1/n$ when using BMA. Such advantage of BMA in dimension reduction and better prediction error has been empirically noticed in a variety of applications, such as in Li and Jiang (2016) in the context of Bayesian generalized method of moments. When $p \gg n$, such dimension reduction through BMA is almost indispensable for any useful statistical inference, and has been widely studied in the literature with sparsity-inducing priors (Johnson and Rossell 2012, Liang et al. 2013, Castillo, Schmidt-Hieber, and van der Vaart 2015, etc.)

1.2 It is useful to extend consideration to quasi-posteriors

Our current paper extends the standard BMA to the general case of a *quasi-posterior*, where

$$\pi(\theta|\,\mathbf{D}) \propto \sum_{j\geq 1} e^{-\lambda R_n(\mathbf{D}\,|\theta,M_j)}\pi\left(\theta|M_j\right)\pi\left(M_j\right).$$

Here, the likelihood function $p(\mathbf{D}\,|\theta, M_j)$ is replaced by $e^{-\lambda R_n(\mathbf{D}\,|\theta,M_j)}$, where R_n is a empirical risk function of the data under the model M_j and the parameter θ. The scaling parameter $\lambda > 0$ can depend on the sample size n, which is analogous to the inverse temperature in statistical physics. Typically $\lambda \propto n$, as in the usual Bayesian posterior where $-\lambda R_n(\theta)$ is the log likelihood function. However in general we allow λ to increase with n at any rate.

This quasi-posterior framework is very useful since it does not need to make as much assumptions on the data generation mechanism as is needed to have a true likelihood function. Although the quasi-posterior originates from other fields such as machine learning and econometrics, research on quasi-posterior from statisticians has been increasing in recent years. It has been applied to problems such as partial likelihood in Cox regression, model-free clustering (Bissiri, Holmes, and Walker 2016), and clinically important difference (Syring and Martin 2017). The latter involves an interesting case of quasi-posterior with general polynomial convergence rates. The current paper will give two more applications of quasi-posteriors, one incorporating model averaging to cube-root asymptotics, another allowing partial identification.

1.3 Why our study may be of interest to Bayesian statistics

Since the Bayesian oracle property is a desirable property for BMA, one naturally hopes that it holds and would like to prove it for some well-established or new methods (see e.g., Li and Jiang 2016 for Bayesian generalized method of moments, Ishwaran and Rao 2011 for spike and lab linear regression). Our current paper shows that it is widely valid in the regular cases for general quasi-posteriors, as long as the model selection consistency holds. This will be useful for Bayesian researchers who invent a new method and would like to go one step further to provide a theoretical justification, in addition to empirical results.

Under the quasi-posterior framework, the more interesting cases are those nonregular models, in which the extremum estimators related to R_n may have nonstandard convergence rates, or the parameters are only partially identified. In such situations, we will show that the Bayesian oracle property does not always hold, and its most straightforward definition may not be always useful. Precaution is needed on how to define a useful oracle property, on how to choose the complexity penalty in the prior, on how to choose the inverse temperature of the quasi-posterior, and on how to choose between BMA and BMS. From the two examples we study, we find that the answers to the aforementioned questions are highly nontrivial, which could be of interest to Bayesian statisticians.

Example 1. Cubic-root asymptotics

Let $Y = I(Z > 0)$ be an observed binary response variable with a latent variable Z related to the utility of the binary choice between $Y = 0$ and $Y = 1$, where $I(\cdot)$ denotes the indicator function. Z can be modeled as a linear combination of an observed vector of predictors X. Given an i.i.d. sample $\mathbf{D} = \{(Y_i, X_i) : i = 1, \ldots, n\}$, one can minimize the empirical risk $R_n(\theta) = -n^{-1} \sum_{i=1}^{n} Y_i I(X_i^\top \theta \geq 0)$. Manski (1975) discovers that the minimization of $R_n(\theta)$ leads to consistent estimation of θ, when the median of Z is proportional to $X^\top \theta$, without any other distributional assumption on Z such as being normal or logistic. This motivates research on quasi-posteriors (e.g. Jun, Pinkse, and Wan 2015) using $e^{-\lambda R_n(\theta)}$ to play the role of the likelihood function, whose posterior means consistently estimate θ in a robust way, without additional distributional assumptions on the data. The exponent function $-\lambda R_n(\theta)$ in this example is discontinuous and its minimizers can converge at a rate of $n^{-1/3}$. This is just one example of many similar cases where cubic-root asymptotics appear.

Our study on BMA allows models with various subsets of X components and proves the oracle property, where the asymptotic behavior of the quasi-posterior from BMA is the same as if the true subset of X components were known. In particular, our study in Section 4 shows several nontrivial results in the presence of cubic-root asymptotics:

1. *On choice of inverse temperature λ:* The standard choice of λ in the likelihoods of regular models is not very useful since it causes the limiting distribution of the posterior mean to be a nonstandard distribution. The BMA has a more useful oracle property when λ growers slower than $n^{2/3}$ and faster than $n^{2/5}$.

2. *On which oracle property is useful or not useful for the quasi-posterior:* The oracle property on the quasi-posterior distribution itself is not so useful as a more carefully defined oracle property of the quasi-posterior mean. This is due to the well known result that asymptotically the quasi-posterior distribution may have the correct centering location but the wrong spread. See, e.g., Chernozhukov and Hong (2003) who show that the quasi-posterior distributions can give consistent parameter estimates but with wrong standard errors. Therefore, for the purpose of statistical inference, it is more meaningful to consider the mean of the quasi-posterior, rather than the whole quasi-posterior distribution.

Example 2. Partial identification

Consider the example of interval censored data, where an unobservable random variable Y lies in the interval $[L, U]$, and both L and U are observable random variables. The goal is to estimate $\theta = \mathrm{E}(Y)$. Given an i.i.d. sample $\mathbf{D} = \{(L_i, U_i) : i = 1, \dots, n\}$, one can use the risk function $R_n(\theta) = [(\bar{U} - \theta)_+]^2 + [(\theta - \bar{L})_+]^2$ (Chernozhukov, Hong, and Tamer 2007), where $(a)_+ \equiv \max\{a, 0\}$, and $\bar{L}$ and $\bar{U}$ are sample averages of L and U. The minimizer of $R_n(\theta)$ can be the entire non-singleton set $[\bar{L}, \bar{U}]$. A quasi-posterior approach based on this R_n is studied in Wan (2013). If there exist different prior beliefs in the location of θ, then one can further perform BMA over these different models. A different approach is provided in the example of Section 5, where we use the framework in Moon and Schorfheide (2012) with a reduced-form parameter and a structural parameter. How to properly define BMA and BMS in such partially identified models is very subtle. Through our effort in finding suitable definitions of Bayesian oracle properties and finding the conditions for them to hold, we obtain several nontrivial results in Section 5 and in a supplementary material (Jiang and Li, 2019), which we believe are of interest to Bayesian statisticians:

1. *On the formulation of Bayesian oracle properties:* The "true" model needs to be carefully defined. Partial identification can lead to multiple models that achieve the same minimal risk and are qualified to be the "true model" simultaneously. In our simple example above, any model that assigns a uniform prior for θ in a closed interval can minimize R_n to be zero, as long as this closed interval has non-empty intersection with $[\bar{L}, \bar{U}]$. Therefore, it makes more sense to group all such minimum-risk models to form a combined true model in the definition of Bayesian oracle properties, instead of defining the true model as the minimum-risk model with the lowest model complexity.

2. *On prior choice of complexity penalty:* In the partial identification problem, it is not wise to artificially penalize the model complexity in the prior, in order to favor the simplest minimum-risk model and make it the unique large sample limit in the posterior. In the simple interval censoring example above, suppose that $\mathrm{E}(L) = -0.1$, $\mathrm{E}(U) = 0.3$, and the true parameter is $\theta^* = \mathrm{E}(Y) = 0.1$. Suppose that one model is given by $\theta \in \{0\}$, i.e. it proposes a singleton prior at $\theta = 0$, while the other models do not propose singletons. Then this singleton model achieves the minimum risk zero for R_n asymptotically since $0 \in [\mathrm{E}(L), \mathrm{E}(U)]$, but it gives the wrong parameter value since $\mathrm{E}(Y) \neq 0$. Therefore, any penalization through the model priors to favor this simplest but wrong model could lead to misleading inference from the quasi-posterior distribution.

3. *On BMA versus BMS:* In the presence of partial identification, the oracle property does not hold for the BMS in general. The BMS picks only one of the possibly many minimum-risk models, which may miss the true parameter, as already explained in our first point before. Hence, BMS is not so reliable as BMA, whose limiting quasi-posterior distribution usually includes all those minimum-risk models compatible with the observed data.

In addition to these qualitative guidances on practice, our study also has a number of virtues in theoretical contribution, which are summarized in a technical report Jiang and Li (2015).

1.4 Related works

Bayesian oracle property under model averaging has been considered in the linear model setup by Ishwaran and Rao (2011) and Castillo, Schmidt-Hieber, and van der Vaart (2015). In contrast, our paper is more general in the sense that it does not assume linear models. Hong and Preston (2012) addressed post selection prediction with possibly nonnested models. Li and Jiang (2016) considered Bayesian generalized method of moments with increasing dimensionality. However, both works assume a regular asymptotic behavior with identifiability and $\sqrt{n}$ asymptotics. The current paper, on the other hand, allows partial identification and cubic-root asymptotics, which entails nonstandard limiting posterior distributions.

We also note that the relationships studied by Hong and Preston (2012) are somewhat different from ours: they relate the point prediction from BMA to the frequentist post-selection predictor, while we study the total variation distance between the entire distributions of the BMA posterior and the oracle posterior given the true model. In this sense, their work and our work are complementary to each other from different perspectives.

1.5 Organization of the paper

The rest of the paper is organized as follows. In Section 2 we introduce three types of Bayesian oracle properties for Bayesian model averaging, MAP model selection, and the posterior mean. Section 3 outlines how one can achieve these Bayesian oracle properties in a general quasi-Bayesian framework. These general approaches are then applied to the examples of cubic root asymptotics in Section 4 and partially identified models in Section 5. Section 6 summarizes the paper with some discussions. Section 7 contains the proofs of the propositions. All other technical details and proofs are included in a supplementary material.

We introduce some useful notation. For two n-dependent sequences $\{a_n\}$ and $\{b_n\}$, $a_n \prec b_n$ and $b_n \succ a_n$ denote the relation $\lim_{n\to\infty} a_n/b_n = 0$. $a_n \preceq b_n$ and $b_n \succeq a_n$ denote that a_n/b_n is bounded by constant. $a_n \asymp b_n$ is equivalent to $a_n \preceq b_n$ and $b_n \preceq a_n$. We use $I(\cdot)$ to denote the indicator function. We use $o_p(1)$ and $O_p(1)$ to denote the orders under the probability measure of $\mathbf{D}$ as the sample size n increases to infinity.

2 Bayesian oracle properties

2.1 Bayesian model averaging

The first property we define here is the global model selection consistency.

Property O1. $\pi(\theta|\,\mathbf{D})$ *satisfies the global model selection consistency, if* $1-\pi(M^*|\,\mathbf{D}) = o_p(1)$.

The global model selection consistency says that the true model M^* has posterior probability converging to 1 as the sample size increases to infinity. The consistency holds for the regular parametric model under the Bayesian framework, based on the standard Bayesian information criterion theory (BIC, Schwartz 1978). It also holds for general high dimensional regression models under certain priors that induce sparsity (Johnson and Rossell 2012, Liang et al. 2013, etc.).

For any (data-dependent) measurable event A, we are interested in the difference between two probabilities

$$|\Pi(A|\,\mathbf{D}) - \Pi(A|M^*, \mathbf{D})|\,,$$

where $\Pi(A|M^*, \mathbf{D})$ is the probability of A under the "oracle" posterior distribution, pretending that the true model M^* is known, whereas $\Pi(A|\,\mathbf{D}) = \sum_{j\geq 1} \pi(M_j|\,\mathbf{D}) \cdot \Pi(A|M_j, \mathbf{D})$ is the mixed posterior distribution via model averaging, allowing possibilities of all models which are weighted by the model posterior probabilities $\pi(M_j|\,\mathbf{D})$ for $j = 1, 2, \ldots$.

Property O2. *$\pi(\theta|\,\mathbf{D})$ satisfies the oracle property for Bayesian model averaging, if $\sup_{A\in\mathcal{F}} |\Pi(A|\,\mathbf{D}) - \Pi(A|M^*, \mathbf{D})| = o_p(1)$ where $\mathcal{F}$ is the set of all measurable events.*

This defines an oracle property for Bayesian model averaging, which basically says that any posterior inference based on model averaging is asymptotically equivalent to the oracle posterior inference based on only the true model. It turns out that one can establish the following fundamental inequality.

Proposition 1.

$$\sup_{A\in\mathcal{F}} |\Pi(A|\,\mathbf{D}) - \Pi(A|M^*, \mathbf{D})| \leq 1 - \pi(M^*|\,\mathbf{D}),$$

where $\mathcal{F}$ is the set of all measurable events.

This proposition reveals a deep relation between three quantities: the model averaging posterior $\pi(\theta|\,\mathbf{D})$, the oracle posterior $\pi(\theta|M^*, \mathbf{D})$, and the posterior probability of the true model $\pi(M^*|\,\mathbf{D})$. The total variation distance between the model averaging posterior and the oracle posterior is bounded above by the posterior probability of missing the true model. A direct consequence of Proposition 1 is the relation between the global model selection consistency (Property O1) and the oracle property for Bayesian model averaging (Property O2).

Theorem 1. *The global model selection consistency (Property O1) implies the oracle property for Bayesian model averaging (Property O2).*

Therefore, as the sample size increases to infinity, if the true model has posterior probability converging to 1, then the limiting behavior of the posterior distribution under model averaging is the same in total variation norm as the oracle posterior pretending to have known the true model. This kind of oracle property is similar in essence to the frequentist oracle property of Fan and Li (2001) but is more general.

To fully appreciate the generality of Theorem 1, we emphasize that in the current general context, we do not require the oracle posterior $\pi(\theta|M^*, \mathbf{D})$ to satisfy the parametric Bernstein von Mises theorem (BvM), i.e. converging to a normal limiting distribution asymptotically at the rate of $n^{-1/2}$. The most attractive aspect of Fan and Li's oracle property is that the inferential results "work as well as if the correct submodel were known" (see the abstract of Fan and Li 2001). This aspect has already been fully captured by Property O2 and there is no need to impose any additional restrictions on the oracle posterior $\pi(\theta|M^*, \mathbf{D})$. Our relaxation makes it possible to include many nonstandard models where a parametric BvM type result does not hold, such as the (quasi-)posteriors with discontinuous (quasi-)likelihoods which is characterized by the cubic root asymptotics (see, e.g., Jun, Pinkse, and Wan 2015), and the partially-identifying posterior distributions with the $O(1)$ rate asymptotics (see, e.g., Moon and Schorfheide 2012).

2.2 MAP (maximum a posteriori) model selection

As an alternative to Bayesian model averaging, one could select only one MAP model that has the maximum posterior probability. We would like to establish similar results to Theorem 1 for MAP model selection. Suppose $\widehat{M}$ is any MAP model choice, so that $\pi(\widehat{M}|\,\mathbf{D}) = \max_{j\geq 1} \pi(M_j|\,\mathbf{D})$. We are interested in the total variation distance between the posterior $\pi(\theta|\widehat{M}, \mathbf{D})$ based on the MAP model, and the oracle posterior $\pi(\theta|M^*, \mathbf{D})$ based on the true model M^*. We hope that inference based on the MAP model choice $\widehat{M}$ is almost as good as if based on the true model M^*.

Property O3. $\pi(\theta|\,\mathbf{D})$ *satisfies the oracle property for MAP model selection, if* $\sup_{A\in\mathcal{F}} |\Pi(A|\widehat{M}, \mathbf{D}) - \Pi(A|M^*, \mathbf{D})| = o_p(1)$ *where* $\mathcal{F}$ *is the set of all measurable events.*

Based on this definition, we have the following proposition.

Proposition 2. *The maximal total variation distances among any of the three posteriors* $\Pi(\cdot|\,\mathbf{D})$, $\Pi(\cdot|\widehat{M}, \mathbf{D})$, *and* $\Pi(\cdot|M^*, \mathbf{D})$, *are at most twice the posterior probability of missing the true model* $2[1 - \pi(M^*|\,\mathbf{D})]$.

A direct consequence of this proposition is

Theorem 2. *The global model selection consistency (Property O1) implies the oracle property for MAP model selection (Property O3).*

2.3 Mean oracle property

In some situations the (quasi-)posterior $\pi(\theta|\,\mathbf{D})$ itself is either not of main interest or does not have any valid interpretation, but the posterior mean $\mathrm{E}(\theta|\,\mathbf{D}) = \int_\Theta \theta d\pi(\theta|\,\mathbf{D})$ for some parameter θ is still of interest, which may have a well understood limiting distribution that can be used for inference on θ. This can happen for quasi-posteriors when its credible region does not have asymptotically correct coverage probability. One example is the Bayesian quantile regression with a quasi-likelihood constructed from the check function. The generalized information criterion is violated and the quasi-posterior has no valid interpretation (Chernozhukov and Hong 2003), but the posterior mean can be used as a convenient frequentist estimator for the quantile regression coefficients.

Another example is the Laplace version of the least median of squares estimator (Jun, Pinkse, and Wan 2011). In this case, it is desirable to have a version of *Bayesian oracle property for the posterior mean: If we make inference based on the overall posterior mean, it is as if we were making inference based on the posterior mean conditional on the true model only.*

To achieve such oracle inference for the mean for a posterior distribution $\pi(\cdot)$, it is usually not sufficient to only have the relation $\| \mathrm{E}(\theta|\,\mathbf{D}) - \mathrm{E}(\theta|M^*, \mathbf{D})\| = o_p(1)$, because $\mathrm{E}(\theta|\,\mathbf{D})$ and $\mathrm{E}(\theta|M^*, \mathbf{D})$ may both converge to a true parameter θ^* but with different convergence rates. A more proper version of mean oracle property is defined as follows.

Property O4. $\pi(\theta|\,\mathbf{D})$ *satisfies the mean oracle property, if* $\| E(\theta|\,\mathbf{D}) - E(\theta|M^*, \mathbf{D})\| = o_p(1) \cdot \| E(\theta|M^*, \mathbf{D}) - \theta^*\|$.

In other words, we require that the difference between posterior means from Bayesian model averaging and the oracle is of higher order compared to the posterior bias under the oracle posterior. This will guarantee that $\mathrm{E}(\theta|\,\mathbf{D}) - \theta^*$ and $\mathrm{E}(\theta|M^*, \mathbf{D}) - \theta^*$ are approximately the same, and not merely both converging to zero.

A useful relation which can be applied to achieve the mean oracle property is

$$\mathrm{E}(\theta|\,\mathbf{D}) - \mathrm{E}(\theta|M^*, \mathbf{D}) = \sum_{j\geq 1, M_j \neq M^*} \pi(M_j|\,\mathbf{D}) \left[\mathrm{E}(\theta|M_j, \mathbf{D}) - \mathrm{E}(\theta|M^*, \mathbf{D})\right]. \qquad (1)$$

The mean oracle property holds if there is a fixed number of model candidates and for every model $M_j \neq M^*$, $\pi(M_j|\,\mathbf{D})\| \mathrm{E}(\theta|M_j, \mathbf{D}) - \mathrm{E}(\theta|M^*, \mathbf{D})\| = o_p(1)\| \mathrm{E}(\theta|M^*, \mathbf{D}) - \theta^*\|$. Each product in the sum of (1) can be made small enough for different reasons. For example, consider the standard variable selection problem in linear models. For those models that miss nonzero parameters, $\pi(M_j|\,\mathbf{D})$ is typically exponentially small. For the models that do not miss nonzero parameters but include redundant parameters, $\mathrm{E}(\theta|M_j, \mathbf{D}) - \theta^*$ is typically of the same order as $\mathrm{E}(\theta|M^*, \mathbf{D}) - \theta^*$, and therefore $\mathrm{E}(\theta|M_j, \mathbf{D}) - \mathrm{E}(\theta|M^*, \mathbf{D})$ is also of the same order as $\mathrm{E}(\theta|M^*, \mathbf{D}) - \theta^*$; then it is sufficient to have $\pi(M_j|\,\mathbf{D}) = o_p(1)$. The method described here will be applied to a nonstandard example with cubic-root asymptotics in Section 4.

2.4 Applications

There has been extensive work in Bayesian model selection consistency, especially the global model selection consistency (Property O1). All these results can be readily extended to the oracle property for Bayesian model averaging (Property O2) and for MAP model selection (Property O3). Whenever there are already known results on the limiting distribution of the oracle posterior $\pi(\theta|M^*, \mathbf{D})$ under the true model M^*, the limiting distribution automatically applies to $\pi(\theta|\,\mathbf{D})$ from model averaging by Theorem 1 and to $\pi(\theta|\widehat{M}, \mathbf{D})$ from model selection by Theorem 2.

The most well known example is the regular finite dimensional models, where BvM type results hold and the posterior distribution of finite dimensional parameters converges in total variation norm to the normal limit at the parametric rate of $n^{-1/2}$.

See for example, Section 10.2 of van der Vaart (1998) for finite dimensional parametric models, and Shen (2002) for nonparametric and semiparametric models. Consequently, in combination with the classic BIC theory from Schwartz (1978), one can derive the global model selection consistency (Property O1) for such finite dimensional cases (see for example Wasserman 2000 Equation 42), and our theorems suggest that the posterior inference based on model averaging or model selection is also equivalent to the inference under the limiting normal distribution given the (unknown) true model. When the model is regular and high dimensional, exactly the same equivalence holds as long as a BvM type result can be established for the low dimensional true model M^*, with properly chosen sparsity inducing priors, such as the priors used in Johnson and Rossell (2012) and Liang et al. (2013).

In this paper, we are interested in applications of the Bayesian oracle property under a more general Bayesian framework than the regular parametric models. We extend the likelihood-based posterior to the general quasi-posterior, in which the likelihood function is replaced by a quasi-likelihood based on a risk function. We propose two ways to achieve the Bayesian oracle properties in Section 3.1 and 3.2 respectively, with two applications: The first application is to the cubic root asymptotics where the convergence rate is not the standard parametric rate $n^{-1/2}$. The second application is to partially identified models where the posterior distribution has a nonstandard limit and a BvM type result does not hold.

3 Quasi-posterior with general risk

We will work under the general framework of a (quasi-)posterior where we can derive general bounds on the mis-selection probability $1 - \pi(M^* | \mathbf{D})$. As discussed in Section 1.2, we consider the quasi-posterior distribution

$$\pi(\theta | \mathbf{D}) = \frac{e^{-\lambda R_n(\theta)} d\pi(\theta)}{\int_\Theta e^{-\lambda R_n(\theta)} d\pi(\theta)}, \tag{2}$$

where $\pi(\theta)$ is the prior density and R_n is an empirical risk function dependent on both the parameter θ and the data $\mathbf{D}$. Related to $R_n(\theta)$ is a theoretical risk function $R(\theta)$, which is typically the large sample limit of $R_n(\theta)$. The scaling parameter $\lambda > 0$ can depend on n and increase with n at any rate, which is analogous to the inverse temperature in statistical physics.

We describe what a true model and a true parameter mean. This is not always clear in the context of quasi-posteriors. Since our quasi-posterior is related to an empirical risk $R_n(\theta)$, which usually has a theoretical risk $R(\theta)$ as its large sample limit, we will treat the minimizer of $R(\theta)$ over the entire parameter space Θ as our true parameter θ^*. We will define a minimum-risk model to be *a model whose prior support includes* θ^*. Situations can be complicated in that there may be multiple minimum-risk models. Conventional wisdom suggests defining the true model M^* as the simplest minimum-risk model that has the lowest dimension of the prior support. If needed, we can also group multiple minimum-risk models together as a composite true model with a mixture prior. A later Section 5 uses this approach to handle partial identification, where the minimizer of $R(\theta)$ is not a singleton and some variation is needed in defining the true model.

In the following, we consider two methods of bounding $1 - \pi(M^* | \mathbf{D})$, the quasi-posterior probability of mis-selecting the true model. Our results from previous sections have shown that bounding this mis-selection probability can lead to various oracle properties. We will make an assumption of finitely many models for simplicity.

3.1 Bounding the mis-selection probability: Extending the BIC approximation for quasi-posterior

In the classical BIC approach (Schwartz 1978), a complexity penalty arises indirectly from approximating an integral in the posterior calculation. Suppose that the parameter space Θ_j is finite dimensional for any $j \geq 1$ and the dimension $d_j = \dim(\Theta_j)$ is bounded. Let Θ^* be the parameter space of M^* and $d^* = \dim(\Theta^*)$. The prior probabilities $\pi(M_j)$ are all assumed to be of order 1 and will not affect the asymptotic behavior. Suppose that the risk functions $R(\theta)$ and $R_n(\theta)$ only depend on the value of θ and do not depend on the model index M_j. For convenience we assume that $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$ is the unique minimizer of $R(\theta)$. We can extend the BIC approximation to general quasi-posteriors and bound the posterior mis-selection probability.

Proposition 3. *Consider the following assumptions:*

(i) The total number of models is bounded above by a constant integer, and all models have a positive prior probability;

(ii) For any minimum-risk model M_j that satisfies $\inf_{\Theta_j} R(\theta) = R(\theta^*)$ (which implies $\theta^* \in \Theta_j$), the integral in the posterior model probability satisfies a BIC type approximation

$$-\ln \int_{\Theta_j} e^{-\lambda R_n(\theta)} d\pi(\theta | M_j) = \lambda R_n(\theta^*) + \frac{d_j \ln \lambda}{2} + O_p(1); \tag{3}$$

(iii) For any minimum-risk model $M_j \neq M^*$, $d_j \geq d^* + 1$;

(iv) For any non-minimum-risk model M_j with $\inf_{\Theta_j} R(\theta) - R(\theta^*) \equiv \gamma_j > 0$, we have $\gamma_j \succeq 1$ and $S_n(\theta) = o_p(1/\lambda)$ uniformly over $\theta \in \Theta_j$, where $S_n(\theta) = [R_n(\theta) - R(\theta)] - [R_n(\theta^*) - R(\theta^*)]$;

(v) $\lambda \to \infty$ as $n \to \infty$.

Then Bayesian oracle properties O1, O2, and O3 hold under the assumptions (i)–(v).

Although the approach outlined in this subsection is still mathematically a BIC approximation, it is somewhat more general, in that it accommodates non-likelihood based quasi-posterior and an arbitrary scaling λ that may increase at a different rate than n. Furthermore, one can establish the Bayesian mean oracle property O4 with the following assumptions in addition to the assumptions (i)–(iv): Θ is compact; the scaling parameter λ grows polynomially in n; for any minimum-risk model M_j, $\|\mathrm{E}(\theta|M_j,\mathbf{D})-\theta^*\| = O_p(1)\cdot\|\mathrm{E}(\theta|M^*,\mathbf{D})-\theta^*\|$; and $\|\mathrm{E}(\theta|M^*,\mathbf{D})-\theta^*\| \succeq \epsilon_n$, where $\epsilon_n = o(1)$ is polynomial in n. The extension of BIC in Proposition 3 as well as the aforementioned additional assumptions for Property O4 can be applied to the example with nonstandard cubic-root asymptotics in Section 4.

3.2 Bounding the mis-selection probability: Assumption-free upperbound for quasi-posterior

In the later example with partial identification (Section 5), the BIC approximation (which uses a local approximation of the theoretical risk R near its minimum) will no longer work. We will apply the following assumption-free upper bound on the mis-selection probability $1-\pi(M^*|\,\mathbf{D})$, which does not require $\arg\min_{\theta\in\Theta} R(\theta)$ to be a singleton, and can therefore be applied to situations with partial identification.

Proposition 4. *(Model selection with quasi-posterior) The mis-selection probability is upper bounded by*

$$\ln[1-\pi(M^*|\,\mathbf{D})] \le -0.5\lambda(\gamma - r - 2|u|)$$

where

$$\gamma = \inf_{\theta\in\Theta, M\neq M^*} R(\theta) - \inf_{\theta\in\Theta} R(\theta),$$
$$r = -\lambda^{-1}\ln\int_\Theta e^{-\lambda[R(\theta)-\inf_{\theta\in\Theta}R(\theta)]}\pi(\theta)d\theta,$$
$$u = -(2\lambda)^{-1}\ln\int_\Theta e^{-2\lambda\left[(R_n(\theta)-R(\theta))-\int_{\theta\in\Theta}(R_n(\theta)-R(\theta))\pi_\infty(\theta)d\theta\right]}\pi_\infty(\theta)d\theta,$$

and

$$\pi_\infty(\theta) = \frac{e^{-\lambda R(\theta)}\pi(\theta)d\theta}{\int_{\theta\in\Theta} e^{-\lambda R(\theta)}\pi(\theta)d\theta}$$

is the limiting version of the quasi-posterior $\pi(\theta|\,\mathbf{D})$, in which the theoretical risk R is used in place of the empirical risk R_n.

This assumption-free bound uses three quantities: γ (*gap*), which differentiates the best possible risks achievable by model M^* and by other models; and r (*excess*), which is a nonstochastic term related to the excess risk $R(\theta)-\inf_{\theta\in\Theta}R(\theta)$ which we will bound later; $|u|$ (*noise*), which is a stochastic noise term determined by the difference $R_n(\theta)-R(\theta)$. This assumption-free bound is only useful when $\gamma > r+2|u| > 0$. We show in the following how it is possible to make $r+2|u| = o_p(\gamma)$, such that $1-\pi(M^*|\,\mathbf{D})$ can be exponentially small in $\lambda\gamma$ and decreases very quickly with sample size n.

The noise term u measures the difference $R_n(\theta)-R(\theta)$ on the support of the limiting posterior. We can use the simplest uniform bound

$$|u| \le 2\sup_{\theta\in\Theta}|R_n(\theta)-R(\theta)|\,.$$

By using uniform large deviation, this will typically lead to $u = O_p(\ln n/\sqrt{n})$. The nonstochastic term r can be bounded by $r = O(\ln\lambda/\lambda)$ if $R(\theta)$ allows a Laplace approximation.

In general, without assuming a Laplace approximation for $R(\theta)$, the rate $r = O(\ln\lambda/\lambda)$ can be derived by the inequality

$$\begin{aligned} r &= -\lambda^{-1}\ln\int_{\theta\in\Theta} e^{-\lambda[R(\theta)-\inf_{\theta\in\Theta}R(\theta)]}\pi(\theta)d\theta \\ &\leq \inf_{a>0}\left[a - \frac{1}{\lambda}\ln\Pi\left(\left\{\theta : R(\theta) - \inf_{\theta\in\Theta}R(\theta) < a\right\}\right)\right], \end{aligned} \tag{4}$$

and choosing $a = \ln\lambda/\lambda$. Detailed argument is similar to the remarks after Proposition 1 in Li, Jiang, and Tanner (2014).

Therefore, if $\gamma \succ r + 2|u|$ and $\gamma \succ \ln n/\lambda$, then $1 - \pi(M^*|\,\mathbf{D}) = \pi(M \neq M^*|\,\mathbf{D}) \prec e^{-\ln n} = 1/n \to 0$ as $n \to \infty$, and we achieve the global model selection consistency (Property O1). Therefore the oracle properties O2 and O3 also hold true. The above bound for $1 - \pi(M^*|\,\mathbf{D})$ may also be used to prove the mean oracle property O4 with the help of (1).

4 Cubic root asymptotics

Suppose that we observe i.i.d. data $\mathbf{D} = \{D_1, \ldots, D_n\}$, and the parameter of interest is $\theta \in \Theta \subseteq \mathbb{R}^p$, whose true value θ^* is the unique solution to the optimization problem $\min_{\theta\in\Theta} Eg(D_1,\theta)$ for some known criterion function g and the expectation is taken with respect to the true underlying distribution of D_1. Let $R(\theta) = \mathrm{E}\, g(D_1,\theta)$ be the theoretical risk and $R_n(\theta) = n^{-1}\sum_{i=1}^n g_i(\theta)$ be the empirical risk where $g_i(\theta)$ is a shorthand for $g(D_i,\theta)$. Instead of the parametric rate $n^{-1/2}$, the frequentist extremum estimator which minimizes $R_n(\theta)$ may have a slower $n^{-1/3}$ convergence rate when g is discontinuous in θ. For example, if one predicts a binary variable Y_i with a vector of continuous predictors $(X_{0,i}, X_i)^\top \in \mathbb{R}^{p+1}$, the maximum score estimator (Manski 1975) minimizes $R_n(\theta)$ with $g_i(\theta) = -Y_i I(X_i^\top\theta - X_{0,i} \geq 0)$, which asymptotically can have a $n^{-1/3}$ convergence rate. Here we assume that the variable $X_{0,i}$ is always selected and its coefficient is -1 to ensure the identification of θ^*. Other applications of the cubic root asymptotics include shorth estimation, least median of squares estimator, isotonic regression, quantile regression with interval censoring, etc. See Kim and Pollard (1990) and Jun, Pinkse, and Wan (2015) for more examples.

We consider the quasi-posterior defined in (2) using the empirical risk function $R_n(\theta) = n^{-1}\sum_{i=1}^n g_i(\theta)$. For the ease of presentation, we only consider the "theta class" in Jun, Pinkse, and Wan (2015). The Laplace type estimator of θ discussed in Jun, Pinkse, and Wan (2015) is the posterior mean of (2). The standard model/variable selection in this cubic root problem assumes that the true parameter θ^* could possibly lie in a lower dimensional space $\Theta \cap \mathbb{R}^{p^*}$ with $1 \leq p^* \leq p$. For example, for the maximum score estimator, our goal is to select only the relevant predictors in X and we set the θ coefficients of all irrelevant components of X to be zero. Then a model

M_j in this context is defined as a coordinate subspace of $\Theta \cap \mathbb{R}^p$. The maximum number of possible models in $\Theta \cap \mathbb{R}^p$ is $2^p - 1$. The true model M^* is defined to be the lowest dimensional coordinate subspace that contains the true parameter θ^* such that all components of θ^* in M^* are nonzero. We assume that the prior density has the decomposition $\pi(\theta|M_j)\pi(M_j)$, where $\pi(\theta|M_j)$ is a continuous density on $\Theta_j \equiv \Theta \cap \mathbb{R}^{d_j}$, d_j is the dimension of M_j, and $\sum_{j=1}^{2^p-1} \pi(M_j) = 1$ give the discrete probabilities for all models.

We make the following assumptions on the model and the prior.

(C1) Θ is compact. θ^* is an interior point of $\Theta \subseteq \mathbb{R}^p$ with p being a constant dimension. $\min_{j \in M^*} |\theta_j^*| \geq c_\theta$, where θ_j^* for $j \in M^*$ denotes the jth nonzero component of θ^* and $c_\theta > 0$ is a constant.

(C2) For all $\theta \neq \theta^*$, $R(\theta) > R(\theta^*)$.

(C3) $R(\theta) = \mathrm{E}\, g(D_1, \theta)$ is three times continuously differentiable in Θ. Let $V = \partial_{\theta\theta^\top} R(\theta^*)$ be the second derivative matrix of $R(\theta)$ evaluated at $\theta = \theta^*$. Then V is positive definite with eigenvalues bounded from below and above by positive constants.

(C4) For any $t, s \in \mathbb{R}^p$, the function $H(t,s) = \lim_{a \to +\infty} a\, \mathrm{E}[g_1(\theta^* + t/a) g_1(\theta^* + s/a)]$ exists and is always positive.

(C5) $\pi(\theta|M_j)$ is continuously differentiable for all $\theta \in \Theta_j$ and all models M_j. $\pi(\theta|M_j)$ and $\partial_\theta \pi(\theta|M_j)$ are uniformly bounded from above by constant for all $\theta \in \Theta_j$ and all models M_j. For all models M_j that satisfy $M_j \supseteq M^*$, $\pi(\theta^*|M_j)$ is uniformly bounded from below by a positive constant. $\pi(M_j)$ is bounded from above and below by positive constants for all models M_j.

Similar to Jun, Pinkse, and Wan (2015), we make the following assumptions on the envelope function of $g(D_1, \theta)$. These assumptions depend on the inverse temperature parameter λ in the quasi-posterior (2). Let $g^\circ(D_1, t) = \lambda^{1/4}[g(D_1, \theta^* + t/\sqrt{\lambda}) - g(D_1, \theta^*)] / (\|t\| + 1)$. Let $\mathcal{G}_n = \{g^\circ(D_1, t) : t \in \mathbb{R}^p\}$.

(C6) There exists an envelope function $G(\cdot)$ such that $\sup_{t \in \mathbb{R}^p} |g^\circ(D_1, t)| \leq G(D_1)$ almost surely under the distribution of D_1. Furthermore, $\mathrm{E}[G^2(D_1)] < \infty$ and $\lim_{n \to \infty} \mathrm{E}[G^2(D_1) I(G(D_1) > c\sqrt{n})] = 0$ for any $c > 0$.

(C7) For any $0 < \epsilon_n = o(1)$, $\sup_{t,s \in \mathbb{R}^p, \|t-s\| \leq \epsilon_n} \mathrm{E}[g^\circ(D_1, t) - g^\circ(D_1, s)]^2 = o(1)$.

(C8) Let $\mathcal{N}(\epsilon, \mathcal{G}_n, L_2(P))$ be the L_2-covering number for $\mathcal{G}_n$ with respect to the probability measure P. Then for every sequence $0 < \epsilon_n = o(1)$,

$$\sup_{P^*} \int_0^{\epsilon_n} \sqrt{\log\left[\mathcal{N}\left(\epsilon \|G(D_1)\|_{P^*}, \mathcal{G}_n, L_2(P^*)\right)\right]} d\epsilon = o(1),$$

where $\sup_{P^*}$ is the supremum taken over all finitely discrete probability measures P^* with $\|G(D_1)\|_{P^*} = \sqrt{\mathrm{E}_{P^*}[G^2(D_1)]} > 0$.

(C1) assumes the standard beta-min condition on θ^* to distinguish its nonzero and zero components. We use the constant lower bound c_θ for technical convenience, as it could be replaced by a rate slowly decreasing to zero that depends on the growth rate of λ. (C2)–(C4) and (C6)–(C8) are similar to the conditions used in Jun, Pinkse, and Wan (2015), which leads to the cubic root behavior of the frequentist extremum estimator that minimizes $R_n(\theta)$. (C5) contains mild conditions on the model selection prior. The essential requirement is that every plausible model should have positive prior probabilities, and the prior mass around the true parameter θ^* should not be too small.

Theorem 3. *Suppose (C1)–(C8) hold with λ satisfying $n^{2/5} \prec \lambda \prec n^{2/3}$. Then the global model selection consistency (Property O1), the Bayesian model averaging oracle property (Property O2), the MAP model selection oracle property (Property O3), and the mean oracle property (Property O4) all hold for the quasi-posterior $\pi(\theta|\,\mathbf{D})$ in (2).*

In Theorem 3 we restrict the growth rate of λ to be between $n^{2/5}$ and $n^{2/3}$. The main reason is that with such λ, the limiting distributions of both the quasi-posterior and the posterior mean will be normal with mean zero, even under a model selection setup with our condition (C5) on the prior. The contribution of our mean oracle property basically says that the asymptotics of the posterior mean from Jun, Pinkse, and Wan (2015), who did not consider model selection but assumed the true model to be known, still remains valid as if the true model were known when we have a pool of candidate models with an unknown true model.

The conclusion of Theorem 3 follows from the BIC type approximation in Case (iii) of Theorem 1 in Jun, Pinkse, and Wan (2015) together with our approach in Section 3.1. A heuristic argument is as follows. The exponent in the quasi-posterior (2) has the decomposition $\lambda R_n(\theta) = \lambda[R(\theta) - R(\theta^*)] + \lambda S_n(\theta)$ with $S_n(\theta)$ defined in Proposition 3. Although $R_n(\theta)$ is discontinuous in θ, $R(\theta)$ is continuously differentiable in θ by (C3). As a result, for any model M that includes the true model M^* as a submodel (including M^* itself), we have a quadratic approximation $\lambda[R(\theta)-R(\theta^*)] \asymp \lambda\|\theta-\theta^*\|^2$. Meanwhile it can be shown that the $S_n(\theta)$ term has a Gaussian process limit and is about the order $O_p(n^{-1/2}\|\theta-\theta^*\|^{1/2})$. Therefore the nonstochastic term of $\lambda[R(\theta)-R(\theta^*)]$ will dominate the stochastic term $\lambda S_n(\theta)$ if $\lambda n^{-1/2}\|\theta - \theta^*\|^{1/2} \prec \lambda\|\theta - \theta^*\|^2 \asymp 1$ in the asymptotics, which leads to $\lambda \prec n^{2/3}$ and $\|\theta - \theta^*\| \asymp \lambda^{-1/2}$. Hence the BIC approximation in Proposition 3 works for the minimum-risk models. For any wrong model M that misses at least one component of M^*, it follows from the aforementioned relations that $S_n(\theta) \asymp n^{-1/2}\lambda^{-1/2} \prec 1/\lambda$, which implies that the integral in (5) is $O_p(1)$. Hence these models will have exponentially small posterior probabilities in λ. The other condition $\lambda \succ n^{2/5}$ in Theorem 3 is required to eliminate the asymptotic bias of the posterior mean. See the comments after Theorem 1 of Jun, Pinkse, and Wan (2015). As a result, the global model selection consistency and the Bayesian oracle properties (Properties O1–O4) hold true following the argument in Section 3.1.

The slowly growing λ in Theorem 3 can overcome the discontinuity in the empirical risk $R_n(\theta)$ with a smoothing effect and justifies the BIC type approximations. The posterior convergence rate is $\lambda^{-1/2}$ from the BIC approximation discussed above, which is slower than $n^{-1/3}$ due to the condition on λ. The posterior mean has a different convergence rate of $n^{-1/2}\lambda^{1/4}$ (see Jun, Pinkse, and Wan 2015 Theorem 1 (iii)), which is faster than $n^{-1/3}$.

In this cubic root example, although the limiting distribution of $\pi(\theta|\,\mathbf{D})$ in (2) is normal, the quasi-posterior itself typically does not have the usual Bayesian interpretation even in the asymptotic sense of Chernozhukov and Hong (2003). Therefore, the MAP model selection oracle property (Property O3) and the model averaging oracle property (Property O2) are not meaningful, since the quasi-Bayesian inference based on the true model may still be invalid. However, the mean oracle property (Property O2) can be very useful because the posterior mean can converge faster than $n^{-1/3}$ to a limiting normal distribution, under the choice of λ in Theorem 3. The normal limit allows us to use various tools such as bootstraps or subsampling to construct asymptotically valid confidence intervals for the posterior mean estimator. Hence statistical inference based on the posterior mean estimator can be more advantageous than that based on the frequentist extremum estimator whose limiting distribution is the Chernoff's distribution (Kim and Pollard 1990).

5 Partial identification

In econometric and statistical literatures, there exist two different approaches to handle partial identification. One aims for more informative inference about the partially identified point parameter θ by incorporating prior information (see, e.g., Poirier 1998, Moon and Schorfheide 2012, Gustafson 2015). Another aims for more robust inference about the fully identified identification region Ω (see, e.g., Wan 2013, Kline and Tamer 2016, and Chen, Christensen, and Tamer 2016). The current paper follows the first approach.

In this section, we apply Bayesian model averaging to a situation with partial identification as described in Moon and Schorfheide (2012), who showed that the limiting posterior is nonstandard. The posterior contraction rate for a structural parameter for interest is typically of order 1, instead of the classical order $n^{-1/2}$, due to partial identification. For example, (4) of Moon and Schorfheide (2012) provides a simple example where the limiting posterior for the structural parameter of interest is uniform over an non-shrinking interval. Despite such nonstandard limiting behavior with partial identification, our machinery in Section 3.2 (based on bounding the mis-selection probability) can be used to study the oracle properties under Bayesian model averaging, which uses a conservative approach to preserve all submodels that are compatible with the data.

5.1 A simple example

This example is similar to the simple example in Moon and Schorfheide (2012). We add the aspect of model selection or model averaging, and make a small variation that a quasi-likelihood is used instead of a real likelihood. Suppose we are interested in a structural parameter $\omega = \mathrm{E}\,Y$, where $Y \in [0,4]$ is the GPA of a college student. However, the GPA is sometimes only known to fall in some interval. For simplicity, assume only its integer part $Z = \lfloor Y \rfloor$ of the GPA is observed. The fractional part $U = Y - Z$ is unobserved. We define $\mathrm{E}\, Z = \phi$, which is called the reduced-form parameter which is identified by the observed data Z. We will call the "combined" parameter $\theta = (\omega, \phi)$. Note that $Z \in [Y - 1, Y]$, and therefore $\phi = \mathrm{E}\, Z \in [\mathrm{E}\, Y - 1, \mathrm{E}\, Y] = [\omega - 1, \omega]$.

In Bayesian approach, the relation between ϕ and ω is described by a conditional prior distribution $\pi(\phi|\omega)$ such as const × $I(\{\phi \in [\omega - 1, \omega] \cap [0, 4]\})$. This conditional prior will be assumed to be the same for all models that we will consider, since we are interested in model selection or model averaging on the structural parameter ω only. Each candidate models M_j, indexed by $j = 1, 2, \ldots$ and weighted by $\pi(M_j)$, proposes a different prior $\pi(\omega|M_j)$ for the structural parameter ω. So the joint prior for the combined parameter θ and M_j is

$$\pi(\theta, M_j) = \pi(M_j)\pi(\omega|M_j)\pi(\phi|\omega).$$

This way, we can convert the model selection problem for the structural parameter ω to a model selection problem with the combined parameter θ. This is for a technical reason to apply the framework of Section 3 in establishing the oracle properties with Bayesian model averaging, later in Section 5.2.

We will introduce some related concepts first for a very simple example, where $j = 1, 2$, $\pi(M_j) = 1/2$, $\pi(\omega|M_1) = \delta_3(\omega)$ is a point mass supported on $W_1 = \{3\}$, proposing mean GPA to be 3, and $\pi(\omega|M_2) = 0.25I(\{\omega \in [0, 4]\})$ is a prior supported on $W_2 = [0, 4]$, proposing no restriction on the mean GPA. This can be regarded as a simplified version of the example in the supplementary material, where Figure 1 illustrates prior densities for more than two candidate models, the first two of them being the same as the current models with $j = 1, 2$.

The observed data Z is integer valued and nonnormal. However, we can use a normal *quasi*-likelihood based on $\bar{Z}$, (the observed sample average of Z), which is typically asymptotically normal iid data: $\sqrt{n/\hat{v}}(\bar{Z} - \phi) \to N(0, 1)$ as $n \to \infty$, where n is the sample size, and $\hat{v}$ is a consistent estimate of $v = \mathrm{var}(Z)$. Then the corresponding quasi-posterior has the form $\pi(\theta, M_j) \propto e^{-\lambda R_n(\theta)}\pi(\theta, M_j)$, where $\lambda = n$ and $R_n(\theta) = 0.5\hat{v}^{-1}(\bar{Z} - \phi)^2$ is an empirical risk derived from asymptotic normality. The corresponding theoretical risk is $R(\theta) = 0.5v^{-1}(\mathrm{E}\, Z - \phi)^2$, minimized at $\phi = \mathrm{E}\, Z$.

The model here is partially identified, since the quasi-likelihood $e^{-\lambda R_n(\theta)}$ only depends on the reduced-form parameter ϕ. The data can only identify ϕ. Given ϕ, the structural parameter ω can still be anywhere from the prior support of $\pi(\omega|\phi) \propto \sum_j \pi(M_j)\pi(\omega|M_j)\pi(\phi|\omega)$, which is supported on $\Omega(\phi) = [\phi, \phi + 1] \cap [0, 4]$. Here $\Omega(\phi)$ is called the *identification region* for ω given ϕ. This is related to the minimizer of the theoretical risk of R when R is regarded as a function of $\theta = (\omega, \phi)$, even if it depends really on ϕ only. Suppose R has a unique minimizer $\phi = \phi^*$ (the "true" ϕ), then attaching all possible ω values in $\Omega(\phi^*)$, we have

$$\arg\min_\theta R(\theta) = \{\phi^*\} \times \Omega(\phi^*).$$

Suppose the true $\phi^* = 3.6$. Then the identification region for ω is $\Omega(3.6) = [3.6, 4]$, and $\arg\min_\theta R(\theta) = \{3.6\} \times [3.6, 4]$.

Model M_1 is "incompatible" with data, in the sense that its prior cannot reach the minimum theoretical risk for $R(\theta)$. The proposed prior on ω does not allow $\phi = \phi^*$, the risk minimizer and the true ϕ. In other words, the prior support of $\pi(\theta|M_1) = \pi(\phi|\omega)\pi(\omega|M_1)$ does not intersect $\arg\min_\theta R(\theta) = \{\phi^*\} \times \Omega(\phi^*)$, since the support of $\pi(\omega|M_1)$ is $\{3\}$, which does not intersect with $\Omega(\phi^*) = [3.6, 4]$.

Model M_2 is "compatible" with data, in the sense that its prior can reach the minimum theoretical risk R. The proposed prior on ω does allow $\phi = \phi^*$, the risk minimizer and the true ϕ.) In other words, the prior support of $\pi(\theta|M_2) = \pi(\phi|\omega)\pi(\omega|M_2)$ intersects $\arg\min_\theta R(\theta) = \{\phi^*\} \times \Omega(\phi^*)$, since the support of $\pi(\omega|M_2)$ is $[0, 4]$, which intersects with $\Omega(\phi^*) = [3.6, 4]$.

This simple example will be generalized in the next Section 5.2, where there can be more than two model candidates and the quasi-posterior can also involve more than two parameters. We hope that with Bayesian model averaging, incompatible models can have small posterior probability asymptotically, so that the posterior from model averaging will be as good as the oracle posterior, which assumes that we knew beforehand and had only used those models that are compatible with data.

5.2 Bayesian model averaging and oracle properties with partial identification

We first derive oracle properties for model selection and BMA in the general framework of quasi-posterior as defined in (2). Later we will consider the special case of partial identification described in Moon and Schorfheide (2012).

Define the index set $\mathcal{J}_0 = \{j \geq 1 : \inf_{\theta\in\Theta_j} R(\theta) = \inf_{\theta\in\Theta} R(\theta)\}$, which includes all model indexes under which the global minimum risk can be reached. These models will be called "compatible models". With partial identification, it is important to allow all compatible models in consideration, and not to exclusively favor one compatible model, even if it is the simplest model with the lowest model complexity. An alternative approach could use a dimensional penalty to favor the simplest compatible model, but this could miss true values of the parameter θ due to partial identification, as discussed in an earlier technical report Jiang and Li (2015) Section 6.6.2. Another example that illustrates this kind of subtlety is described as a technical detail in a supplementary material of the current paper.

In response to this subtlety with partial identification, we will group all the compatible models together to form our "true" model $M^* = \{M_j : j \in \mathcal{J}_0\}$. Then $\pi(\theta, M^*|\,\mathbf{D}) \propto e^{-\lambda R_n(\theta)} \sum_{j\in\mathcal{J}_0} \pi(\theta, M_j)$. The resulting joint prior on θ and M_j can be rewritten as $\pi(\theta, M^*) = \sum_{j\in\mathcal{J}_0} \pi(\theta, M_j) = \pi(M^*)\pi(\theta|M^*)$, where $\pi(M^*) = \sum_{j\in\mathcal{J}_0} \pi(M_j)$, and $\pi(\theta|M^*) = \sum_{j\in\mathcal{J}_0} \pi(M_j)\pi(\theta|M_j)/\sum_{j\in\mathcal{J}_0} \pi(M_j)$ is a mixture prior for θ conditional on the composite true model M^*.

All incompatible models are indexed by $j \in \mathcal{J}_1$. For incompatible models, we assume the quantity $\gamma = \inf_{j\in\mathcal{J}_1} \inf_{\theta\in\Theta_j} R(\theta) - \inf_{\theta\in\Theta} R(\theta)$ to be a positive constant, which holds true if there is a fixed number of candidate models. This γ is exactly the same γ used in Proposition 4. We can derive an upper bound for the posterior mis-selection probability $1 - \pi(M^* | \mathbf{D})$ (where $M^* = \{M_j : j \in \mathcal{J}_0\}$) as exponentially small in λ from Proposition 4, which leads to the Bayesian oracle properties O1 and O2 in Section 2. The oracle posterior here is still $\pi(\theta | M^*, \mathbf{D})$, conditional on compatible models only.

We make the following assumptions.

(A1) $\lambda \succ 1$ as $n \to \infty$.

(A2) $\sup_{\theta \in \Theta} |R_n(\theta) - R(\theta)| = o_p(1)$ as $n \to \infty$.

(A3) $\Pi(\{\theta : R(\theta) - \inf_{\theta \in \Theta} R(\theta) < a\}) > 0$ for any small $a > 0$.

(A4) $\gamma = \inf_{j \in \mathcal{J}_1} \inf_{\theta \in \Theta_j} R(\theta) - \inf_{\theta \in \Theta} R(\theta)$ is a positive constant.

Assumption (A1) is true when $\lambda \propto n$. When R_n is a sample average of independently and identically distributed data, we can take the theoretical risk in (A2) to be the expectation $R(\theta) = \mathrm{E}\, R_n(\theta)$ over the true distribution of the randomly generated data. Then (A2) can be satisfied due to a uniform law of large numbers, which holds, e.g., when the entire parameter space Θ is compact and the risk functions are stochastically equicontinuous (see, e.g, Newey 1991).[1] Assumption (A4) is true when the number of candidate models is fixed. Regarding (A3), suppose the prior support Θ is compact and contains a risk minimizer of R in its interior. Then a small enough neighborhood of this risk minimizer will have positive prior π and can have risk $R(\theta)$ being arbitrarily close to the minimum risk, if $R(\theta)$ is continuous in θ.

We can summarize the analysis above formally in the following theorem.

Theorem 4. *Assume that (A1)–(A4) hold and $M^* = \{M_j : j \in \mathcal{J}_0\}$. Then the total variation distance between the distributions $\pi(\theta | \mathbf{D})$ and $\pi(\theta | M^*, \mathbf{D})$ is $o_p(1)$ as the sample size $n \to \infty$, i.e., the global model selection consistency (Property O1) and the Bayesian model averaging oracle property (Property O2) both hold.*

The proof of Theorem 4 shows $\pi(\mathcal{J}_1) = o_p(1)$ by applying Proposition 4. Therefore even though it is impossible to point identify the minimizer of the theoretical risk, we can still have a similar form of Bayesian oracle properties by selecting all the compatible models. As a result, the posterior inference based on model averaging is asymptotically equivalent to the posterior inference based on only those compatible models weighted by their priors.

The above Theorem 4 is very general. Moon and Schorfheide (2012) considered a special case where $-\lambda R_n(\theta)$ is the log likelihood function. Also, the "combined" parameter can be decomposed as $\theta = (\omega, \phi)$, where ω is a structural parameter of interest and ϕ is a reduced-form parameter that is identified by data. The candidate models impose different priors on the structural parameter ω, so that $\pi(\theta, M_j) = \pi(M_j)\pi(\theta | M_j) = \pi(M_j)\pi(\omega | M_j)\pi(\phi | \omega)$. A simple example of this kind of parametrization and the corresponding prior distribution is described in Section 5.1, using a quasi-likelihood derived

[1] Assumption (A2) may also be satisfied when R_n is not an average itself, but is a function of some sample averages, such as is easy too verify for the example in Section 5.1. In fact it is easy to check that all conditions are valid for that example assuming that $\mathrm{var}(Z) > 0$.

from asymptotic normality. For such situations when only the structural parameter ω is of primary interest, the BMA oracle Property O2 for the marginal posterior on ω also holds:

Corollary 1. *Under the assumptions made for Theorem 4, the BMA oracle Property O2 holds marginally for the structural parameter of interest ω, i.e., $\int |\pi(\omega| \mathbf{D}) - \pi(\omega|M^*, \mathbf{D})| d\omega = o_p(1)$, if ω is a sub-vector of the combined parameter θ.*

So far we have discussed Property O1 (for global model selection consistency) and Property O2 (for the oracle property with BMA). There is an important exception here: Property O3 for MAP model selection is not guaranteed in this partially identified model. This is because here the true model M^* is effectively the set of all compatible models which is possibly a nonsingleton, and the proof of Proposition 2 does not go through. When there are two or more compatible models, the MAP model selection may only choose one compatible model and neglect all the other ones. Posterior inference based on the MAP model may be different from using the oracle posterior given all the compatible models and may end up missing the true value of a point parameter. We will describe this as a technical detail with a simple example in a supplementary material.

Regarding the mean oracle Property O4, we conjecture that it usually holds for the structural parameter of interest, as will be discussed as some additional technical details in the supplementary material.

6 Discussion

We have established a fundamental relation between three different topics: Bayesian model averaging, model selection consistency, and oracle performance in posterior distribution. The relatively basic property of model selection consistency is shown to imply a seemingly more advanced distributional result, the oracle property. The result is very simple and general. Unlike some previous Bayesian oracle properties discussed in special cases such as Ishwaran and Rao (2011), and Castillo, Schmidt-Hieber, and van der Vaart (2015), who consider linear models, and Hong and Preston (2012) and Li and Jiang (2016) who consider identifiable models with standard limiting distributions, the current work is completely free from any restriction on the type of prior or (quasi-)likelihood function used, or even from any restriction on the limiting distribution of the oracle posterior.

For applications, we considered two classes of models with nonstandard limiting distributions studied in Moon and Schorfheide (2012) and Jun, Pinkse, and Wan (2015). They involve partial identifiability or nonstandard rates of convergence, but we can still show the Bayesian oracle properties, which suggest that Bayesian model averaging can be applied to their methods and work well for Bayesian inference of the unknown point parameter. On the other hand, we suspect that model selection based on MAP may not be reliable for the partial identification example and may miss reasonable models (see a discussion after Corollary 1).

When the model is misspecified, the model that minimizes the theoretical risk R plays the role of the true model in our theory. Our oracle property will imply that the quasi-posterior based on BMA will converge to the quasi-posterior based on the minimum risk model, asymptotically. Grünwald and van Ommen (2014) discovered suboptimal predictive performance when a homoscedastic linear model is misspecified. Their numerical experiments seem to indicate that the performance of BMA still converges to the performance of the true model eventually, albeit with a much larger sample size compared to the correctly specified case. This indicates a much slower convergence speed of BMA when the models are misspecified. Our current paper only addresses the limiting distributional behavior of BMA and BMS, but not their convergence speed. As a possible future work, we may consider extending our theory in Section 3.1 to study the convergence speed in the presence of model misspecification and how the convergence depends on the temperature parameter, as discussed in Grünwald and van Ommen (2014).

Given the success of the frequentist oracle properties studied by Fan and Li (2001), we expect that the Bayesian version should also have applications in a wide variety of situations, in addition to the examples discussed in this paper. For example, the relationships described in Section 2 and 3 can be generalized to models with increasing or high dimensions, and potentially to other nonstandard model selection problems with appropriate conditions on the priors. For instance, Drton and Plummer (2017) have developed a generalized version of BIC type approximation for the class of singular models (such as factor models), where the posterior model probability does not allow a quadratic approximation and results in an extra $\ln\ln n$ term in the BIC approximation (3). Our Bayesian oracle properties may also apply to these singular models. In addition, in the context with partial identification, our paper only considered inference about the partially identified point parameter, following the approach of, e.g., Poirier (1998), Moon and Schorfheide (2012), and Gustafson (2015). It may also be of interest to consider inference about the fully identified "set parameter", following the approach of, e.g., Wan (2013), Kline and Tamer (2016), and Chen, Christensen, and Tamer (2016), and develop similar oracle properties for Bayesian model selection or model averaging.

7 Proofs of the propositions

Proof of Proposition 1. For any event A and B, we have

$$\Pi(A|\,\mathbf{D}) = \Pi(A|B,\mathbf{D})\Pi(B|\,\mathbf{D}) + \Pi(A|B^c,\mathbf{D})\Pi(B^c|\,\mathbf{D}),$$
$$\Pi(A|B,\mathbf{D}) = \Pi(A|B,\mathbf{D})\Pi(B|\,\mathbf{D}) + \Pi(A|B,\mathbf{D})\Pi(B^c|\,\mathbf{D}).$$

Therefore

$$|\Pi(A|\,\mathbf{D}) - \Pi(A|B,\mathbf{D})| = |\Pi(A|B^c,\mathbf{D}) - \Pi(A|B,\mathbf{D})|\Pi(B^c|\,\mathbf{D}) \le \Pi(B^c|\,\mathbf{D})$$

for any A. Taking supremum over all event A and setting event $B = \{M = M^*\}$ lead to the proof. (Note that Castillo, Schmidt-Hieber, and van der Vaart 2015 used a double-sized upper bound in proving their Theorem 6 in the context of Bayesian linear regression.) □

Proof of Proposition 2. The MAP choice $\widehat{M}$ satisfies $\pi(\widehat{M}|\,\mathbf{D}) \ge \pi(M^*|\,\mathbf{D})$ by definition. In the proof of Proposition 1 above, we can replace M^* by $\widehat{M}$ and obtain that $\sup_A |\Pi(A|\,\mathbf{D}) - \Pi(A|\widehat{M},\mathbf{D})| \le 1 - \pi(\widehat{M}|\,\mathbf{D})$. The right hand side is at most $1 - \pi(M^*|\,\mathbf{D})$

since $\pi(\widehat{M}|\,\mathbf{D}) \geq \pi(M^*|\,\mathbf{D})$. Now combining this with the result of Proposition 1 using the triangle inequality leads to the conclusion. □

Proof of Proposition 3. Let $p(M_j) = \int_{\Theta_j} e^{-\lambda R_n(\theta)} d\pi(\theta|M_j)$. Under the assumption (ii), we have that for any minimum-risk model $M_j \neq M^*$,

$$\begin{aligned} -\ln p(M_j) &= \lambda R_n(\theta^*) + \frac{d_j \ln\lambda}{2} + O_p(1), \\ -\ln p(M^*) &= \lambda R_n(\theta^*) + \frac{d^* \ln\lambda}{2} + O_p(1). \end{aligned} \tag{5}$$

Taking the difference between these two equations gives

$$-\ln p(M_j)/p(M^*) = \frac{(d_j - d^*)\ln\lambda}{2} + O_p(1).$$

Due to the assumptions (iii) and (v), $O_p(1)$ is negligible compared to the $\ln\lambda$ term. Therefore, for any minimum-risk model M_j, there exists a constant $C_1 > 0$ such that

$$p(M_j)/p(M^*) \leq C_1 \lambda^{-(d_j - d^*)/4} \leq C_1 \lambda^{-1/4}. \tag{6}$$

We notice that from (5), $p(M^*) = \exp[-\lambda R_n(\theta) - d^* \ln\lambda/2 + O_p(1)]$. For any non-minimum-risk model M_j, we can use the assumption (iv) to obtain that for some constant $C_2 > 0$,

$$\begin{aligned} p(M_j) &= \int_{\Theta_j} e^{-\lambda S_n(\theta) - \lambda[R(\theta) - R(\theta^*)] - \lambda R_n(\theta^*)} d\pi(\theta|M_j) \\ &\leq \int_{\Theta_j} e^{-\lambda S_n(\theta)} d\pi(\theta|M_j) \cdot e^{-\lambda\gamma_j} \cdot p(M^*) e^{\frac{d^* \ln\lambda}{2} + O_p(1)} \\ &\leq C_2 \lambda^{d^*/2} e^{-\lambda\gamma_j} p(M^*). \end{aligned} \tag{7}$$

Since $\gamma_j \succeq 1$ and $\max_{j\geq 1} d_j$ is upper bounded by constant, the exponential rate $e^{-\lambda\gamma_j}$ dominates the polynomial rate $\lambda^{d^*/2}$. Furthermore, from the assumption (i), we also have that the prior ratio $\pi(M_j)/\pi(M^*)$ is lower and upper bounded by constants for any model M_j. Therefore, from (6) and (7), we have that

$$\begin{aligned} & 1 - \pi(M^*|\,\mathbf{D}) \\ &= \frac{\sum_{M_j \neq M^*} \pi(M_j) p(M_j)}{\sum_{M_j \neq M^*} \pi(M_j) p(M_j) + \pi(M^*) p(M^*)} = \frac{\sum_{M_j \neq M^*} \frac{\pi(M_j)}{\pi(M^*)} \frac{p(M_j)}{p(M^*)}}{\sum_{M_j \neq M^*} \frac{\pi(M)}{\pi(M^*)} \frac{p(M)}{p(M^*)} + 1} \\ &\leq 1 - \Big[\sum_{M_j \neq M^* \text{ and } M_j \text{ is minimum-risk}} \frac{\pi(M_j)}{\pi(M^*)} C_1 \lambda^{-1/4} \\ &\quad + \sum_{M_j \text{ is non-minimum-risk}} \frac{\pi(M_j)}{\pi(M^*)} C_2 \lambda^{d^*/2} e^{-\lambda\gamma_j} + 1 \Big]^{-1} \\ &= o_p(1). \end{aligned}$$

Therefore the global model selection consistency (Property O1) is proved. By Propositions 1 and 2, the Bayesian oracle properties for BMA (Property O2) and BMS (Property O3) also hold. □

Proof of Proposition 4. First it is clear that on all models not equal to M^*, $R(\theta) - \inf_{\theta\in\Theta} R(\theta) \geq \inf_{\theta\in\Theta, M\neq M^*} R(\theta) - \inf_{\theta\in\Theta} R(\theta) = \gamma$. Hence

$$1 - \pi(M^* |\, \mathbf{D}) \leq \Pi\left(\left\{\theta : R(\theta) \geq \inf_{\theta\in\Theta} R(\theta) + \gamma\right\} \Big|\, \mathbf{D}\right). \tag{8}$$

We then show that for any bounded measurable function $h(\theta)$,

$$\ln \mathrm{E}[h(\theta)|\, \mathbf{D}] \leq \frac{1}{2} \ln \mathrm{E}_\infty[h^2(\theta)] - \lambda u, \tag{9}$$

where $\mathrm{E}[h(\theta)|\, \mathbf{D}] = \int_{\theta\in\Theta} h(\theta)\pi(\theta|\, \mathbf{D})d\theta$, $\mathrm{E}_\infty[h^2(\theta)] = \int_{\theta\in\Theta} h^2(\theta)\pi_\infty(\theta)d\theta$, u is defined as in Proposition 4. To see why (9) holds true, we recall the definitions of a quasi-posterior $\pi(\theta|\, \mathbf{D})$ and its "limiting posterior" $\pi_\infty(\theta)$:

$$\begin{aligned}\mathrm{E}[h(\theta)|\, \mathbf{D}] &= \frac{\int_{\theta\in\Theta} e^{-\lambda R_n(\theta)} h(\theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta} e^{-\lambda R_n(\theta)}\pi(\theta)d\theta} \\ &= \frac{\int_{\theta\in\Theta} e^{-\lambda[R_n(\theta)-R(\theta)]} h(\theta)\pi(\theta)d\theta}{\int_{\theta\in\Theta} e^{-\lambda[R_n(\theta)-R(\theta)]}\pi(\theta)d\theta}.\end{aligned}$$

Then we apply the Jensen's inequality to the denominator and apply the Cauchy-Schwarz inequality to the numerator to obtain that

$$\begin{aligned}\mathrm{E}[h(\theta)|\, \mathbf{D}] &\leq \frac{\sqrt{\int_{\theta\in\Theta} e^{-2\lambda[R_n(\theta)-R(\theta)]}\pi_\infty(\theta)d\theta}\sqrt{\int_{\theta\in\Theta} h^2(\theta)\pi_\infty(\theta)d\theta}}{e^{-\lambda\int_{\theta\in\Theta}[R_n(\theta)-R(\theta)]\pi_\infty(\theta)d\theta}} \\ &= \sqrt{\int_{\theta\in\Theta} e^{-2\lambda\left[(R_n(\theta)-R(\theta))-\int_{\theta\in\Theta}(R_n(\theta)-R(\theta))\pi_\infty(\theta)d\theta\right]}\pi_\infty(\theta)d\theta}\sqrt{\int_{\theta\in\Theta} h^2(\theta)\pi_\infty(\theta)d\theta},\end{aligned}$$

which leads to (9). Then we take $h = I(A)$ for a measurable set A and obtain that

$$\ln \Pi(A|\, \mathbf{D}) \leq \frac{1}{2}\ln \Pi_\infty(A) - \lambda u. \tag{10}$$

Set $A = \{\theta : R(\theta) - \inf_{\theta\in\Theta} R(\theta) \geq \gamma\}$ and use the definition of r in Proposition 4:

$$\begin{aligned}\Pi_\infty(A) &= \frac{\int_{\theta\in\Theta} e^{-\lambda[R(\theta)-\inf_{\theta\in\Theta} R(\theta)]} I(A)\pi(\theta)d\theta}{\int_{\theta\in\Theta} e^{-\lambda[R(\theta)-\inf_{\theta\in\Theta} R(\theta)]}\pi(\theta)d\theta} \\ &= \int_{\theta\in\Theta} e^{-\lambda[R(\theta)-\inf_{\theta\in\Theta} R(\theta)-r]} I(A)\pi(\theta)d\theta \leq e^{-\lambda(\gamma-r)}.\end{aligned}$$

Then applying this upper bound of $\Pi_\infty(A)$ to (10) and using (8) leads to the proof. □

Supplementary Material

Supplement to "On Bayesian Oracle Properties" (DOI: 10.1214/18-BA1097SUPP; .pdf). We provide the technical proofs of Theorem 3, Theorem 4, and Corollary 1, as well as additional discussion on the Bayesian oracle properties for partial identification.

References

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of Royal Statistical Society: Series B*, 78: 1103–1130. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 237

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43: 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 235, 237, 240, 254, 255

Chen, X., Christensen, T., and Tamer, E. (2016). "MCMC confidence sets for identified sets." *Cowles Foundation Discussion Paper*, No. 2037. ArXiv:1605.00499. 250, 255

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation." *Journal of Econometrics*, 115: 293–346. MR1984779. doi: https://doi.org/10.1016/S0304-4076(03)00100-3. 238, 242, 250

Chernozhukov, V., Hong, H., and Tamer, E. (2007). "Estimation and confidence regions for parameter sets in econometric models." *Econometrica*, 75: 1243–1284. MR2347346. doi: https://doi.org/10.1111/j.1468-0262.2007.00794.x. 239

Drton, M. and Plummer, M. (2017). "A Bayesian information criterion for singular models." *Journal of Royal Statistical Society: Series B*, 79: 323–380. MR3611750. doi: https://doi.org/10.1111/rssb.12187. 255

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96: 1348–1360. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 235, 241, 242, 255

Grünwald, P. and van Ommen, T. (2014). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." ArXiv:1412.3730. MR3724979. doi: https://doi.org/10.1214/17-BA1085. 255

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models: Exploring the Limits of Limited Data*. CRC Press, New York. MR3642458. 250, 255

Hong, H. and Preston, B. (2012). "Bayesian averaging, prediction and nonnested model selection." *Journal of Econometrics*, 167: 358–369. MR2892081. doi: https://doi.org/10.1016/j.jeconom.2011.09.021. 240, 254

Ishwaran, H. and Rao, J. S. (2011). "Consistency of spike and slab regression." *Statistics and Probability Letters*, 81: 1920–1928. MR2845909. doi: https://doi.org/10.1016/j.spl.2011.08.005. 235, 237, 240, 254

Jiang, W. and Li, C. (2015). "A note on Bayesian oracle properties." ArXiv: 1507.05723v1. 239, 252

Jiang, W. and Li, C. (2019). "Supplement to "On Bayesian Oracle Properties"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1097SUPP. 239

Johnson, V. E. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 107: 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 237, 241, 244

Jun, S. J., Pinkse, J., and Wan, Y. (2011). "Root-n consistent robust integration-based estimation." *Journal of Multivariate Analysis*, 102: 828–846. MR2772339. doi: https://doi.org/10.1016/j.jmva.2011.01.003. 243

Jun, S. J., Pinkse, J., and Wan, Y. (2015). "Classical Laplace estimation for cube-root-n consistent estimators: Improved convergence rates and rate-adaptive inference." *Journal of Econometrics*, 187: 201–216. MR3347303. doi: https://doi.org/10.1016/j.jeconom.2015.01.005. 238, 242, 247, 248, 249, 250, 254

Kim, J. and Pollard, D. (1990). "Cubic root asymptotics." *The Annals of Statistics*, 18: 191–219. MR1041391. doi: https://doi.org/10.1214/aos/1176347498. 247, 250

Kline, B. and Tamer, E. (2016). "Bayesian inference in a class of partially identified models." *Quantitative Economics*, 7: 329–366. MR3536642. doi: https://doi.org/10.3982/QE399. 250, 255

Li, C. and Jiang, W. (2016). "On oracle property and asymptotic validity of Bayesian generalized method of moments." *Journal of Multivariate Analysis*, 145: 132–147. MR3459943. doi: https://doi.org/10.1016/j.jmva.2015.12.009. 235, 237, 240, 254

Li, C., Jiang, W., and Tanner, M. A. (2014). "General inequalities for Gibbs posterior with nonadditive empirical risk." *Econometric Theory*, 30: 1247–1271. MR3278163. doi: https://doi.org/10.1017/S0266466614000152. 247

Liang, F., Song, Q., and Yu, K. (2013). "Bayesian subset modeling for high-dimensional generalized linear models." *Journal of the American Statistical Association*, 108: 589–606. MR3174644. doi: https://doi.org/10.1080/01621459.2012.761942. 237, 241, 244

Manski, C. (1975). "Maximum score estimation of the stochastic utility model of choice." *Journal of Econometrics*, 3: 205–228. MR0436905. doi: https://doi.org/10.1016/0304-4076(75)90032-9. 238, 247

Moon, H. R. and Schorfheide, F. (2012). "Bayesian and frequentist inference in partially identified models." *Econometrica*, 80: 755–782. MR2951948. doi: https://doi.org/10.3982/ECTA8360. 239, 242, 250, 252, 253, 254, 255

Newey, W. K. (1991). "Uniform convergence in probability and stochastic equicontinuity." *Econometrica*, 59: 1161–1167. MR1113551. doi: https://doi.org/10.2307/2938179. 253

Poirier, D. J. (1998). "Revising beliefs in nonidentified models." *Econometric Theory*, 14: 483–509. MR1650041. doi: https://doi.org/10.1017/S0266466698144043. 250, 255

Schwartz, G. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6: 461–464. MR0468014. 241, 244, 245

Shen, X. (2002). "Asymptotic normality of semiparametric and nonparametric posterior distributions." *Journal of the American Statistical Association*, 97: 222–235. MR1947282. doi: https://doi.org/10.1198/016214502753479365. 244

Syring, N. and Martin, R. (2017). "Gibbs posterior inference on the minimum clinically important difference." *Journal of Statistical Planning and Inference*, 187: 67–77. MR3638043. doi: https://doi.org/10.1016/j.jspi.2017.03.001. 237

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 244

Wan, Y. (2013). "An integration-based approach to moment inequality models." Manuscript. University of Toronto. http://individual.utoronto.ca/yuanyuanwan/Integration_Wan.pdf. 239, 250, 255

Wasserman, L. (2000). "Bayesian model selection and model averaging." *Journal of Mathematical Psychology*, 44: 92–107. MR1770003. doi: https://doi.org/10.1006/jmps.1999.1278. 244

Acknowledgments

We thank the Editor, the Associate Editor, and two referees for providing additional references and constructive comments that have improved this paper. The first author is partially supported by a special fund from the Taishan Scholar Construction Project. The second author is supported by the National University of Singapore start-up grant (R155000172133).