PRACTICAL BOUNDS TO THE BAYESIAN MARGINAL LIKELIHOOD USING DEEP LEARNING

A PREPRINT

Gregory A. Ross
Schrödinger Inc,
New York, New York, 10036

July 23, 2022

ABSTRACT

Model selection in Bayesian analysis is critically important but remains immensely challenging despite the increasing ease of sampling methods. To aid formal model comparison, simple bounds to the marginal likelihood are introduced as well as straightforward ways to compute them using deep learning techniques.

Keywords First keyword · Second keyword · More

1 Introduction

When attempting to create quantitative and predictive models of data, one seldom has a single model in mind. At some point, one inevitably has to select aspects of the model like the functional form and the number of parameters. The formal approach to this choice is known as model selection, which can be an immensely complex and difficult task to do rigorously. The singular interest of this work is model selection in a Bayesian context, whose framework naturally admits a quantity - the marginal likelihood or marginal evidence - through which all models can be judged and compared. Although there remains some debate as to the usefulness of marginal likelihoods, model selection based on marginal likelihoods is popular because of its consistency (in the formal sense), its interpretability, and its connection with other model selection methods such as cross validation - a technique that is important within both frequentist and Bayesian frameworks.

In the majority of "real life" Bayesian modelling exercises, one uses computational techniques to sample parameters of interest from the posterior distribution. While there are various types of sampling methods, the most common use Markov chain Monte Carlo, owing to its adaptability and ease of use. There exist a number of probabilistic programming languages, such as BUGS, pyMC, and Edward, that greatly facilitate the design and implementation of Bayesian models. In these languages, sampling from the posterior can essentially be done automatically, usually with a robust MCMC method after the model is specified. Nowadays, sampling methods tap into the increased speed and available of graphical processing units, making sampling less and less of a computational burden. The increasing ease and speed of Bayesian model building and sampling has not corresponded to an ease in the computation of marginal likelihoods, whose calculation remains a rarefied domain of expertise.

The deviance information criterion (DIC) is a quantity that can be calculated using samples from the posterior alone and was designed to aid model comparison. In this framework, the best model is one that has the lowest DIC. The ease of use of the DIC has made it a very popular model comparison tool, and since its introduction, there now exist and number of variants and improvements. However, unlike the marginal likelihood, its use as a model selection tool is of somewhat questionable validity, and, unlike the marginal likelihood, a model with the lowest DIC does not necessarily have the highest prediction accuracy on new data.

This works aims to bridge the gap between the utility of the DIC and the rigor of marginal likelihoods for comparing Bayesian models. Given samples from the prior or posterior, the methods presented here work by framing the computation of the marginal likelihood as an optimization problem, in a manner that is not wholly dissimilar to the framework of variational Bayesian analysis. Upper and lower bounds to the marginal likelihood are derived that, with the aid of deep learning techniques, can be computed automatically and in a straightforward manner. The greater the expressibility of the neural network, the tighter the bounds. In particular, it is shown that the upper bound, which uses samples from the posterior, is tighter than the lower bound and of more practical use. These upper and lower bounds are applied to classical Bayesian problems and are shown to be competitive or superior to alternative fast model comparison tools.

2 Background theory

In Bayesian modelling, one is interested in estimating the parameters of a model, denoted $\theta = \{\theta_1, \theta_2, ...,\theta_k\}$, given some data, $x = \{x_1, x_2, ..., x_N\}$. We will assume that $x_i \in \mathbb{R}^D$, with the actual value of D being unimportant. All inferences of θ are summarized by the posterior distribution, $p(\theta|x)$, which is proportional to the product of the likelihood, $p(x|\theta)$, and prior, $p(\theta)$. For a single model m out of a total of M models, the relationship between these quantities is given by Bayes Theorem:

$$p_m(\theta|x) = \frac{p_m(x|\theta)p_m(\theta)}{\int p_m(x|\theta)p_m(\theta)\, d\theta}. \tag{1}$$

The normalising factor on the right-hand side is the marginal likelihood of model m and whose estimation is of primary interest in this work. As will hopefully be clear below, it is advantages to consider the logarithm of the marginal likelihood

$$\mathcal{L}_m(x) = \ln \int p_m(x|\theta)p_m(\theta)\, d\theta. \tag{2}$$

When one has a total of M models to choose from, the most faithfully Bayesian approach is to use all models and weight each prediction from m proportional to $e^{\mathcal{L}_m(x)}$. However, this approach can be computationally very expensive when M is large. In that case, a pragmatic approach is to select the model with the highest $\mathcal{L}_m(x)$ for prospective use.

2.1 Information criteria in model selection

2.1.1 Bayesian information criterion

Although $\mathcal{L}_m(x)$ is very difficult to calculate for most models, it greatly simplifies in the asymptotic data limit. In 1978, Schwarz showed that, under certain assumptions, as $N \to \infty$, $\mathcal{L}_m(x)$ has the approximate limit

$$BIC_m(x, \theta^*) = \ln(\, p_m(x|\theta^*_m)\,) - \frac{1}{2} k_m \ln(N) \tag{3}$$

where θ^*_m is the vector of parameters that maximizes the likelihood. k_m is the dimensionality of θ_m in the model m (i.e. the number of parameters). This limiting form of $\mathcal{L}_m(x)$ is known as the Bayesian information criterion (BIC); its simplicity and ease of use has made it a popular model comparison score. Similiar to the Akaike information criterion (AIC) that came before it, model selection using the BIC can be seen as compromise between the goodness of fit (via the maximum likelihood term) and the model complexity (via the the $k_m \ln(N)$ term).

2.1.2 Deviance information criterion

From a purely Bayesian perspective, the lack of dependence of the BIC on the prior may not be desirable. Priors can reflect important, pre-existing information and help regularize models by reducing the variance of future predictions. Priors can also serve to reduce the effective dimensionality of model, such that the BIC may penalize the model complexity too severely in cases when N is not large.

To aid Bayesian model comparison, Spiegelhalter et al. introduced the deviance information criterion (DIC). It is not an approximation to the marginal likelihood, but instead was designed as a practical way to compare Bayesian models using samples from the posterior distribution. To motivate the DIC in manner differently from Spiegelhalter et al. that will help clarify the results later in this manuscript, consider the Taylor series of the log-likelihood when it is expanded about the posterior mean of the parameters $\hat{\theta} \equiv \mathbb{E}_{\theta|x}[\theta]$. Let all terms except the first term be grouped into the remainder $R(\theta, \hat{\theta})$

$$\begin{gathered} \ln(p(x|\theta)) = \ln(p(x|\hat{\theta})) + R(\theta, \hat{\theta}) \quad \text{such that} \\ \mathbb{E}_{\theta|x}\left[\ln(p(x|\theta))\right] = \ln(p(x|\hat{\theta})) + \mathbb{E}_{\theta|x}\left[R(\theta, \hat{\theta})\right] \end{gathered} \tag{4}$$

If the posterior mean, $\hat{\theta}$, and the maximum likelihood estimate, θ^*, are equal, the above expansion has the same form as the BIC, with $\mathbb{E}_{\theta|x}[R(\theta, \hat{\theta})]$ acting the regularizing term - a negative number - that penalizes model complexity. This suggests, at least with Gaussian like posterior distributions, that the posterior mean of the log-likelihood, i.e. $\mathbb{E}_{\theta|x}[\ln(p(x|\theta))]$, may serve as a score for model comparison as it trades-off goodness of fit with model complexity. The DIC exploits the regularizing effects of the mean log-likelihood for model comparison. In the notation of this manuscript, the DIC is defined as

$$DIC = -2\,\mathbb{E}_{\theta|x}\left[\ln(p(x|\theta))\right] - 2\,\mathbb{E}_{\theta|x}\left[R(\theta, \hat{\theta})\right] \tag{5}$$

$$= -2\ln(p(x|\hat{\theta})) - 4\,\mathbb{E}_{\theta|x}\left[R(\theta, \hat{\theta})\right]. \tag{6}$$

Unlike the marginal likelihood, models are preferred that have a lower DIC. As $\mathbb{E}_{\theta|x}[R(\theta, \hat{\theta})]$ arises from the Taylor expansion of $\mathbb{E}_{\theta|x}[\ln(p(x|\theta))]$ (equation 4), by adding it to the mean log-likelihood, the DIC can be thought of as doubling the complexity penalty that is inherent in the mean log-likelihood. Spiegelhalter et al. defined the effective number of parameters in a model as $-2\,\mathbb{E}_{\theta|x}[R(\theta, \hat{\theta})]$, and they showed this produces intuitive results in a number of Bayesian models.

The DIC can be computed from samples from the posterior distribution, which allows practitioners to straightforwardly perform inference and model comparison with a single set of samples. However, as recognized by Spiegelhalter et al., the DIC is not invariant with respect to reparameterization, making its use in some instances unreliable. Being heuristic in nature, the domain where the DIC is appropriate for model selection is less well defined than the marginal likelihood.

3 Bounding marginal likelihoods

3.1 Upper bounds

Starting with the Bayes theorem, we can express the marginal likelihood, $\mathcal{L}(x)$, as

$$\mathcal{L}(x) = \ln p(x|\theta) - \ln \frac{p(\theta|x)}{p(\theta)} \tag{7}$$

$$= \int p(\theta|x) \ln p(x|\theta)\, d\theta - \int p(\theta|x) \ln \frac{p(\theta|x)}{p(\theta)}\, d\theta$$

$$= \mathbb{E}_{\theta|x}[\ln p(x|\theta)] - D\big(p(\theta|x) \parallel p(\theta)\big), \tag{8}$$

where $\mathbb{E}_{\theta|x}[\ln p(x|\theta)]$ is the expectation of the likelihood using samples from the posterior, and $D\big(p(\theta|x) \parallel p(\theta)\big)$ is the Kullback-Leibler (KL) divergence from the prior to the posterior. The second line uses the fact that $\int \mathcal{L}(x)\, p(\theta|x)\, d\theta = \mathcal{L}(x)$. Pertinently for this work, the KL divergence is strictly non-negative. This leads to a trivial upper bound to $\mathcal{L}(x)$

$$\mathcal{L}(x) \leq \mathbb{E}_{\theta|x}[\ln p(x|\theta)], \tag{9}$$

which can be easily estimated as long as one has samples from the posterior distribution. One can have a tighter upper bound to $\mathcal{L}(x)$ if one can compute a lower bound to the KL divergence, $D_{lb}\big(p(\theta|x) \parallel p(\theta)\big)$:

$$\mathcal{L}(x) \leq \mathbb{E}_{\theta|x}[\ln p(x|\theta)] - D_{lb}\big(p(\theta|x) \parallel p(\theta)\big). \tag{10}$$

These simple upper bounds to $\mathcal{L}(x)$ are the primary theoretical result of this work. While there exist a number of lower bounds to the KL divergence [refs?], amongst the most useful for our purpose is that of Nguyen et al. [Nguyen2009Estimating] which was used by Nowozin et al. [Nowozin2016fGAN] in their development of generative adversarial networks. Let $V_\omega(\theta) : \Theta \rightarrow \mathbb{R}$ be some function (here an neural network) parameterized by ω, then

$$D_{lb}\big(a(\theta) \parallel b(\theta)\big) = \mathbb{E}_{\theta\sim a}[V_\omega(\theta)] - \mathbb{E}_{\theta\sim b}[\exp(V_\omega(\theta) - 1)] \tag{11}$$

Thus, to find a lower bound to the KL divergence, we must maximise the right-hand side, which can be achieved using stochastic gradient decent.

3.2 Lower bounds

A lower bound to the log marginal likelihood can be obtained if express equation 7 as an expectation over the prior rather than the posterior:

$$\mathcal{L}(x) = \mathbb{E}_{\theta}[\ln p(x|\theta)] + D\big(p(\theta) \parallel p(\theta|x)\big), \tag{12}$$

where $\mathbb{E}_\theta[\ln p(x|\theta)]$ is the expectation of the likelihood over the prior and $D\big(p(\theta) \parallel p(\theta|x)\big)$ is the KL divergence from the posterior to the prior. The mean likelihood term is easy to estimate because in the vast majority of cases, the prior is easy to sample from or one already has samples from the prior (e.g. from Bayesian updating). The non-negativity of KL divergence immediately implies that

$$\mathcal{L}(x) \geq \mathbb{E}_\theta[\ln p(x|\theta)]. \tag{13}$$

Curiously, this suggests that one could perform model comparison using only samples from the prior. However, as we will illustrate, this bound is not as tight as equation 9. Similiarly as with the upper bound, we can construct a tighter lower bound to $\mathcal{L}(x)$ with a lower bound to the KL divergence:

$$\mathcal{L}(x) \geq \mathbb{E}_{\theta|x}[\ln p(x|\theta)] + D_{lb}\big(p(\theta|x) \parallel p(\theta)\big) \tag{14}$$

The utility of these bounds will be explored with mathematical analysis and numerical examples.

Still to include...

- Asymptotic analysis on the bounds, and how the KL divergence is a measure of the number of free parameters in a model.

3.3 An information theoretic bound in the large N limit

As the primary focus of this work is to demonstrate and validate new bounds to the marginal likelihood, it is of interest to establish an absolute upper bound that will serve as a reference in our numerical examples.

Consider the log-likelihood $\ln p(x|\theta)$. When x_i are independent and identically distributed, we have

$$\begin{aligned} \ln p(x|\theta) &= \ln p(x_1, x_2, ..., x_N|\theta) \\ &= \ln\left(p(x_1|\theta)p(x_2|\theta)...p(x_N|\theta)\right) \\ &= \sum_{i=1}^{N} \ln p(x_i|\theta) \end{aligned}$$

Let us now assume that each x_i has been drawn from some "true" but unknown distribution $p_t(x)$ and let $N \to \infty$. Then we have the limit

$$\begin{aligned} \lim_{N\to\infty} \frac{1}{N} \ln p(x|\theta) &= \int p_t(x) \ln p(x|\theta) dx \\ &= \int p_t(x) \ln\left(p(x|\theta)\frac{p_t(x)}{p_t(x)}\right) dx \\ &= \int p_t(x) \ln p_t(x)\, dx + N \int p_t(x) \ln \frac{p(x|\theta)}{p_t(x)}\, dx \\ &= -H(x) - D_{lb}\big(p_t(x) \parallel p(x|\theta)\big) \end{aligned} \tag{15}$$

The KL divergence term in the penultimate line is the divergence between the true data generating distribution and the model, and it features in model selection methods such as the AIC and minimum description length [Akaike1974, Hansen2001Model]. Combining equations 8 and 15 in the asymptotic data limit we have

$$\frac{1}{N}\mathcal{L}(x) = -H(x) - \mathbb{E}_{\theta|x}\left[D_{lb}\big(p_t(x) \parallel p(x|\theta)\big)\right] - \frac{1}{N}D\big(p(\theta|x) \parallel p(\theta)\big) \tag{16}$$

and since when N is large $D\big(p(\theta|x) \parallel p(\theta)\big) \propto \ln N$ we have

$$\lim_{N\to\infty} \frac{1}{N}\mathcal{L}(x) = -H(x) - \mathbb{E}_{\theta|x}\left[D_{lb}\big(p_t(x) \parallel p(x|\theta)\big)\right] \tag{17}$$

$$\leq -H(x) \tag{18}$$

That is, the information entropy is an upper bound to the marginal likelihood, which is tight when the Bayesian model can be close to the true data generating distribution. Ideally, when validating our bounds for the marginal likelihood, we will compare the bounds to the actual marginal likelihood. However, we will encounter cases where then marginal likelihood is intractable. The above bound will allow use to use the information entropy of the data generating distribution as a substitute for $\mathcal{L}(x)$ when it cannot be calculated.