Technical Report on Bayesian Computation for Intractable Posteriors with a Fixed Prior and Varying Data

Ryan Warnick[1]

[1]Independent Work

July 15, 2023

Abstract

Bayesian computation has been a major area of research in Bayesian analysis for the last two decades, and many novel developments have increased reach of the Bayesian approach to previously untouched domains. However, one primary challenge is that many algorithms for inference must be rerun when provided with new observations. The goal of the presented approach is to combine Importance Sampling (IS) with an approach taken from the machine learning community termed Mixture Density Networks (MDNs). This approach, similar in spirit to variational bayesian methods, returns a function which, when provided with data and prior parameter values, returns a distribution over the parameters of interest. This is useful in industrial applications of Bayesian methods, where, once a model has been designed, it might be desired to apply it to various data sets without expensive computational burdens.

1 Introduction

In typical Bayesian inference, a model is constructed incorporating a prior $P(\theta)$ for $\theta \in \Theta$ and a likelihood $P(X|\theta)$, and data is observed X where $X \in \mathcal{X}$. For simple models,

analytic expressions for the posterior distribution $P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} = \frac{P(X|\theta)P(\theta)}{\int_\theta P(X|\theta)P(\theta)}$ may be constructed. However, in modern Bayesian approaches the likelihood and prior specifications have become increasingly complex; leading to sophisticated inference algorithms to "look at" properties of posterior distributions without analytical representations. This is a broad research agenda in Bayesian analysis, and there are many papers offering various approaches or refinements of existing approaches. Two common methods are Variational Inference (VI) and Markov Chain Monte Carlo (MCMC); both with many variants thereof. However, one distinguishing characteristic of all existing approaches towards inference for complex models is that they need to be re-run with new data and/or a new prior specification. This complicates both sensitivity analysis, as well as limiting the utility of the Bayesian paradigm in industrial settings where the goal is a model that has reusability on new sets of data.

What the the work presented here seeks to do is find an approach that can be trained on the "model", instead of on the model in concordance with the data. That is, what we seek is a function defined on $\mathcal{X}$ which outputs a probability distribution $\hat{P}(\theta|X);\ \forall X \in \mathcal{X}$. This serves as an operator mapping the data space $\mathcal{X}$ to the probability distributions on Θ; $\hat{P} : \mathcal{X} \to \mathcal{P}_\Theta$. For all data $X \in \mathcal{X}$, we want to have it such that $\hat{P}(X)(\theta)$ is as close as possible to $P(\theta|X)$. This problem is non-trivial, as the set of probability distributions on Θ is large and the functional form of the posterior distribution $P(\theta|X)$ is unrestricted. Additionally adding to the complexity is that the function must be defined over all $\mathcal{X}$; including extreme values in the tails of the likelihood. This can be accomplished using a combination of MDNs and IS.

2 Background

Here some preliminary background is presented that is needed to understand the approach.

2.1 Mixture Density Networks

In typical neural networks, we have a function $f(\circ;\eta) : \mathcal{T} \to \mathcal{U}$ which is a composition of simple functions parameterized by a set of weights $\eta \in \mathbb{R}^q$ for some $q \in \mathbb{N}$. The composition of the simple functions creates a function with a high amount of flexibility.

From observations $\{(t_i, u_i)\}_{i=1}^n$, we can minimize a loss function $L(\circ, \circ)$ such that:

$$\hat{\eta} = \text{argmin}_\eta L(\{u_i\}_{i=1}^n, \{f(t_i; \eta)\}_{i=1}^n) \tag{1}$$

is the optimal set of weights to fit a suitable function from $\mathcal{T}$ to $\mathcal{U}$.

MDNs expand on this approach, but instead of learning a function from $\mathcal{T}$ to $\mathcal{U}$, they use a neural network to learn a function between $\mathcal{T}$ and the parameters Ξ, $f(\circ; \eta) : \mathcal{T} \to \Xi$, of a known probability distribution $P(U|\xi)$ indexed by a parameter $\xi \in \Xi$. Assuming that the data are generated by $U_i \sim P(U|\xi_i)$; where $\xi_i = f(T_i; \eta)$; where $\forall i T_i \in \mathcal{T} U_i \in \mathcal{U}$. To do this, we use the log-likelihood as a loss. Denoting by $p(U|\xi)$ the density of $U \sim P(U|\xi)$, our modified loss function is given by:

$$L(\{U_i\}_{i=1}^n, \{f(T_i; \eta)\}_{i=1}^n) = -\sum_{i=1}^n \log p(U_i|f(T_i; \eta)) \tag{2}$$

Therefore, by minimizing the loss with respect to η, we are maximizing the likelihood of the density with respect to the parameters η. $\hat{\eta}$ is the minimum of Equation 2 over η. The next result gives results about the asymptotic behavior of the limiting distribution $p(u|f(t; \eta))$ as $n \to \infty$

Result 1: Denote by $\hat{\eta}$ the minimum over η of the negative log likelihood loss in Equation 2, and suppose $(T_1, U_1), ..., (T_n, U_n) \overset{iid}{\sim} Q$, and that Q has density $q(U, T)$ which factorizes into a marginal and conditional component $q(U, T) = q(U|T)q(T)$. We require the following three assumptions:

1. Suppose that $\text{E}_{Q_{U,T}}[\log q(u|t)]$, exists and is finite.
2. There exists an $m(U, T)$ such that, $\forall \eta$, $|\log p(U|f(T; \eta))| \leq m(U, T)$, and $m(U, T)$ is integrable with respect to Q. An alternative statement would be that for all η $E_{Q_{U,T}}[\log p(U|f(t; \eta))]$ exists and is finite, but the former gives a constructive approach.
3. Denote by Δ_η the distribution on $\mathcal{U} \times \mathcal{T}$ with factorized density $p(U|f(T; \eta))q(T)$, for each η. There exists a unique minimum $\min_\eta \text{KL}(Q||\Delta_\eta)$; where $\text{KL}(\circ||\circ)$ denotes the Kullback-Leibler divergence.

Then we have that $\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_T}[-E_{Q_{U|T}}[\frac{\log p(U|f(T;\eta))}{\log q(U|T)}|T]]$; where the outer expectation is taken with respect to the marginal distribution on T and the inner expectation is the conditional expectation of U for given T.

In other words, the solution convergences to the Kullback-Leibler projection of $Q_{U|T}$ onto the set of distributions with density $p(u|f(t;\eta)))$, for particular T, averaged over the true distribution of the $T \in \mathcal{T}$.

2.2 Importance Sampling

Importance sampling is a powerful method for approximating expectations taken over a random variable with density $p(x)$ using samples from another random variable. The principle is established by a result from basic probability: suppose we are seeking $\mathrm{E}_p[f(X)]$, but calculating this expectation and/or sampling from $p(x)$ is difficult. Then for any random variable with density $q(x)$ such that $q(x) > 0$ whenever $p(x)f(x) > 0$, we have the following:

$$\mathrm{E}_p[f(X)] = \mathrm{E}_q[w(X)f(X)], \qquad w(x) = \frac{p(x)}{q(x)} \tag{3}$$

This allows us to approximate $\mathrm{E}_p[f(X)]$ with $\hat{\mu_f} = \frac{1}{n}\sum_{i=1}^{n} w(x^{(j)})f(x^{(j)})$; where $x^{(j)} \overset{iid}{\sim} Q$ for all $j \in \{1, ..., n\}$. If $p(x)$ or $q(x)$ are known only up to a normalizing constant, as is frequently the case in Bayesian analysis, we can approximate $\mathrm{E}_p[f(X)]$ with $\widetilde{\mu_f} = \frac{\sum_{i=1}^{n} w(x^{(j)})f(x^{(j)})}{\sum_{i=1}^{n} w(x^{(j)})}$. The strong law of large numbers gives us that both $\hat{\mu_f}$ and $\widetilde{\mu_f}$ converge almost surely to $\mathrm{E}_p[f(X)]$ as $n \to \infty$.

3 Using MDNs As an Approximate Function Mapping Data to an Approximate Posterior

We seek an approximation to $P(\theta|X)$ for all X, and this can be accomplished using mixture density networks. If we sample $(X_1, \theta_1), ..., (X_n, \theta_n)$ from a distribution Q on $\mathcal{X} \times \Theta$ specified by $P(X, \theta) = P(X|\theta)P(\theta)$, optimize the previous log likelihood loss function for to find an MDN $p_\eta(\theta|f(X;\eta))$ approximating $P(\theta|X)$.

Assuming the assumptions of Result 1 are met, we have that $\hat{\eta} \overset{a.s.}{\to} \mathrm{argmin}_\eta \mathrm{E}_X[-E_{\theta|X}[\log \frac{p_\eta(\theta|f(X;\eta))}{P(\theta|X)}|X]]$. This gives us that the MDN converges asymptotically as $n \to \infty$ to the minimizer in the MDN family of the Kullback-Leibler divergence from the posterior, averaged over all the X.

This limiting solution means that the MDN approximation to the posterior will be more accurate for X more likely under the prior predictive distribution. To see this, assume that $P(X) = \int_\Theta P(X|\theta)P(\theta)d\theta = (1-\epsilon)h(X)+\epsilon\delta_{X^*}(X)$ for some density $h(X)$ and some $X^* \in \mathcal{X}$; where $\delta_{X^*}(\circ)$ denotes the dirac delta at X^* and $\epsilon \in (0,1)$. This is an ϵ-contaminated class of prior predictive distributions. We have then that:

$$\hat{\eta} = \text{argmin}_\eta \ \ (1-\epsilon)\text{E}_h[-E_{\theta|X}[\log \frac{p_\eta(\theta|f(X;\eta))}{P(\theta|X)}|X]] + \epsilon[-E_{\theta|X^*}[\log \frac{p_\eta(\theta|f(X^*;\eta))}{P(\theta|X^*)}]] \tag{4}$$

So we see as $\epsilon \to 1$ we have that the minimum becomes weighted towards the more likely data under the prior predictive.

A common approach in Bayesian analysis is hierarchical modeling. That is, priors are placed on top of each other in successive fashion, forming a hierarchy of distributions. In the Bayesian paradigm this allows the model to incorporate structural information about the system under observation into the prior, while also diminishing the needs of an accurate prior; as the prior becomes more diffuse as the layers of the hierarchy increase. For some model formulations this diffuse prior makes the prior predictive more diffuse as well. The classical example is in a normal means model with known variance: $X|\mu,\sigma \sim \text{N}(\mu,\sigma^2)$ and $\mu \sim \text{N}(\mu_0,\sigma_0^2)$, where σ^2 is known.

Elaborating on the normal means model as an exemplary case, we see that $X = \mu+\epsilon$ where $\mu \sim \text{N}(\mu_0,\sigma_0^2)$ and $\epsilon \sim \text{N}(0,\sigma^2)$. Thus $X \sim \text{N}(\mu_0,\sigma^2+\sigma_0^2)$. If we chain normals on top of normals in a hierarchy of means, for K levels in the hierarchy, placing $X \sim \text{N}(\mu_0,\sigma^2)$, and a hierarchy of priors $\{\mu_{k-1} \sim \text{N}(\mu_k,\sigma_k)\}_{k=1}^K$, then we have that $X \sim \text{N}(0,\sigma^2+\sum_{k=0}^K \sigma_k^2)$. This follows by induction from the previous proof for one level in the hierarchy. Thus we see that increasing the hierarchy increases the prior predictive variance; making the distribution more diffuse.

As the distribution gets more diffuse, the expectation over P_X does not lend as much weight to any particular X value. This causes the approximation to $P(\theta|X)$ by $p_\eta(\theta|f(X;\eta))$ to be more accurate across a broader range of X values.

3.1 Accounting for Outlier Data Under the Marginal Likelihood With Importance Reweighting

Suppose we want the model to be fit to a higher diversity of X values, but don't want to give up the potential of using an informative prior elicited from some secondary information or subject matter expertise. This informative prior could bias the prior predictive towards particular values, limiting the viability of the approximation to X values observed in the future. In the setting that we have the marginal likelihood up to a normalizing constant $P(X) = cg(X)$ with $c > 0$ this is possible using an importance reweighting.

If we have a distribution Q_X on $\mathcal{X}$ with density $Q(X)$ which is informative in regions of the possible X values we are interested in, or potentially uninformative to allow the model to fit well for a wider variety of $X \in \mathcal{X}$, such that $Q(X) = 0$ whenever $P(X) = 0$ almost surely, we can modify the loss function in Equation 2 to:

$$L(\{\theta_i\}_{i=1}^n, \{f(X_i;\eta)\}_{i=1}^n) = \frac{-\sum_{i=1}^n \frac{Q(X_i)}{g(X_i)} \log p(\theta_i | f(X_i;\eta))}{\sum_{i=1}^n \frac{Q(X_i)}{g(X_i)}} \tag{5}$$

Then with $\hat{\eta} = \text{argmin}_\eta L(\{\theta_i\}_{i=1}^n, \{f(X_i;\eta)\}_{i=1}^n)$, we get the following result:

Result 2:

Suppose $E_{P_{\theta,X}}[\frac{Q(X)}{P(X)} \log P(\theta|X)]$ exists and is finite. Furthermore, suppose that $E_{P_{\theta,X}}[\frac{Q(X)}{P(X)} \log p(\theta|f(X;\eta))]$ exists and is finite for all η, and that $P(X) = cg(X)$ for some $c > 0$ and known function $g(X)$. The last point specifies that $P(X)$ is known up to a normalizing constant.

Then with $(X_1, \theta_1), ..., (X_n, \theta_n) \overset{iid}{\sim} P(X, \theta)$ we have, as $n \to \infty$, with the the loss specified in Equation 5:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta E_{Q_X}[-E_{P_{\theta|X}}[\log \frac{p(\theta|f(X;\eta))}{P(\theta|X)}|X]] \tag{6}$$

This allows us to improve performance in regions of $\mathcal{X}$ where the prior predictive would not give a high probability to the X values by re-weighting the loss to have an outer expectation of Q_X in the limit.

4 Example: Normal Likelihood Normal-Inverse Gamma Prior

For the sake of comparison, we examine a simple model for which a closed form solution for $P(\theta|X)$ already exists. Suppose we have m observed X values, $(X_1, ..., X_m)$ where the $X_j \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$. If we place a normal-inverse gamma prior on μ, σ^2, $\mathrm{NIG}(\mu_0, \nu, \alpha, \beta)$, where the normal-inverse gamma is specified as $\mu|\sigma \sim \mathrm{N}(\mu_0, \frac{\sigma^2}{\nu})$ and $\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$ with $\mathrm{IG}(\alpha, \beta)$ denoting the Inverse-Gamma distribution with shape parameter α and scale parameter β. This prior specification is conjugate and has the distribution $\mu, \sigma^2|\{X_j\}_{j=1}^m \sim \mathrm{NIG}(\mu_m, \nu_m, \alpha_m, \beta_m)$, with the parameters provided below:

$$\begin{aligned}
\mu_m &= \frac{\nu\mu_0 + m\bar{X}}{\nu + m} \\
\nu_m &= \nu + m \\
\alpha_m &= \alpha + \frac{m}{2} \\
\beta_m &= \beta + \frac{1}{2}\sum_{j=1}^{m}(X_j - \bar{X})^2 + \frac{m\nu}{m+\nu}\frac{(\bar{X} - \mu_0)^2}{2}
\end{aligned}$$

Where $\bar{X}$ corresponds to the mean of $(X_1, ..., X_m)$, $\bar{X} = \frac{1}{m}\sum_{j=1}^m X_j$.

This conjugacy allows us to study the approximation specified previously, and design an $p(\mu, \sigma^2|f(X;\eta))$ that might accurately approximate the posterior over a range of $\{X_j\}_{j=1}^m \subseteq \mathcal{X}$.

Suppose that we choose $p(\mu, \sigma^2|f(X;\eta))$ such that $p(\circ|\circ)$ is the density of a $\mathrm{NIG}(f(X;\eta)_{\mu^*}, f(X;\eta)_{\nu^*}, f(X;\eta)_{\alpha^*}, f(X;\eta)_{\beta^*})$ distribution. We then need to try and come up with the appropriate functions $f_\circ$ mapping $\mathcal{X}$ the parameters of p.

We have that $X_j \in \mathcal{X} = \mathbb{R}$, so this gives us that $(X_1, ..., X_m) \in \mathcal{X}^m = \mathbb{R}^m$. A linear activation function is defined by $h(x;\eta) = h(x;A,b) = Ax + b$, and note that this is sufficiently robust to model μ_m, ν_m, and α_m. Denote by A_{θ^*} and b_{θ^*} the weights of for $f(X;\eta)_{\theta^*} = h(X;\eta)$ for $\theta \in \{\mu, \nu, \alpha\}$. We have that $\hat{\eta}$ has a closed form and is given by:

$$A_{\mu^*} = \frac{1}{\nu + m}\underline{1}^T, \quad b_{\mu^*} = \frac{\nu\mu_0}{\nu + m}$$
$$A_{\nu^*} = \underline{0}^T, \quad b_{\nu^*} = \nu + m$$
$$A_{\alpha^*} = \underline{0}^T, \quad b_{\alpha^*} = \alpha + \frac{m}{2}$$

However, β_m has a more complicated expression, and needs to be specified as a function $f(X;\eta)_{\beta^*}$ of $\{X_j\}_{j=1}^m$ with a more nuanced structure. If we let denote by $c(X;\eta) = c(X;A,b) = (AX+b)^2$ the quadratic activation function, we get that an expression for $\hat{\beta}$ as a function of η in Equation 7:

$$f(X;\eta)_{\beta^*} = h_1([c_1(X - h_2(X;A_{h_2},b_{h_2})\underline{1};A_{c_1},b_{c_1}),c_2(X;A_{c_2,c_2})];A_{h_1},b_{h_1}) \quad (7)$$

With the following minimizing values for the parameters:

$$A_{h_1} = [\frac{1}{2}, \frac{m\nu}{2(m+\nu)}], \quad b_{h_1} = \beta$$
$$A_{h_2} = \frac{1}{n}\underline{1}^T, \quad b_{h_2} = 0$$
$$A_{c_1} = \underline{1}^T, \quad b_{c_1} = 0$$
$$A_{c_2} = \frac{1}{n}\underline{1}^T, \quad b_{c_2} = \mu_0$$

So we see that the set of all A and b parameters are our weight space η, and it has a unique minimum that can be expressed in terms of a series of simple functions.

5 Appendix

5.1 Proof of Result 1:

The proof follows along with Theorem 2.1 and 2.2 of a 1982 paper by Halbert White titled Maximum Likelihood Estimation in Misspecified Models in Econometrica. From Assumption 2 we have that we can apply the strong law of large numbers, giving us that:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_{U,T}}[-\log p(U|f(T;\eta)] \tag{8}$$

Using the assumption that $E_{Q_{U,T}}[\log q(U|T)]$ exists and is finite, and noting that it does not depend on η, we can add it to the function being minimized to obtain:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_{U,T}}[-\log p(U|f(T;\eta)] + E_{Q_{U,T}}[\log q(U|T)] \tag{9}$$

Consolidating the expectations and rearranging the logarithms together we get:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_{U,T}}[-\log \frac{p(U|f(T;\eta)}{q(U|T)}] \tag{10}$$

Using properties of conditional expectations we can expand the expectation into a marginal and conditional component:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_T}[E_{Q_{U|T}}[-\log \frac{p(U|f(T;\eta)}{q(U|T)}|T]] \tag{11}$$

5.2 Proof of Result 2:

Using the strong law of large numbers result outlined in the section on importance sampling, we have that, because the expectation exists and is finite for all η:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{P_{\theta,X}}[-\frac{Q(X)}{P(X)} \log p(\theta|f(X;\eta)] \tag{12}$$

Then using our assumption that $E_{P_{\theta,X}}[\frac{Q(X)}{P(X)} \log P(\theta|X)]$ exists, and noting that it does not depend on η, we can add it to the function being minimized:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \quad \text{E}_{P_{\theta,X}}[-\frac{Q(X)}{P(X)} \log p(\theta|f(X;\eta)] + \text{E}_{P_{\theta,X}}[\frac{Q(X)}{P(X)} \log P(\theta|X)] \tag{13}$$

Consolidating the expectations we get:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{P_{\theta,X}}[-\frac{Q(X)}{P(X)} \log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}] \tag{14}$$

Using properties of conditional expectations, we get:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{P_X}[-\text{E}_{P_{\theta|X}}[\frac{Q(X)}{P(X)} \log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}|X]] \tag{15}$$

Noting that the $\frac{Q(X)}{P(X)}$ factor inside the expectations does not depend on θ, we can take it out of the inner expectation being done with respect to $P_{\theta|X}$:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{P_X}[-\frac{Q(X)}{P(X)} E_{P_{\theta|X}}[\log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}|X]] \tag{16}$$

Evaluating the expectation as an integral, either with respect to Lebesgue or counting measure, and canceling the density/probability mass function of P_X, gives us:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta - \int_{\mathcal{X}} P(X) \frac{Q(X)}{P(X)} \text{E}_{P_{\theta|X}}[\log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}|X] dX \tag{17}$$

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta - \int_{\mathcal{X}} Q(X) \text{E}_{P_{\theta|X}}[\log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}|X] dX \tag{18}$$

Which, rewriting as an expectation with respect to Q_X, gives us:

$$\hat{\eta} \overset{a.s.}{\to} \text{argmin}_\eta \text{E}_{Q_X}[-\text{E}_{P_{\theta|X}}[\log \frac{p(\theta|f(X;\eta)}{P(\theta|X)}|X]] \tag{19}$$

Which completes the proof.