Nonparametric Bayesian analysis using Normalizing Flows (better funner title, please)

Ji, Jiang, Tan, Wang, Schwartz

June 6, 2022

1 Introduction

All explanations should be about one to two sentences only, and should primarily rely upon appropriate referencing.

1.1 Bayesian Analysis (Haining/Yichen)

- Bayesian updating, including posteriors as subsequent priors
- General arguments: full uncertainty characterization, Bayesian model averaging [Yichen]
- General criticisms: priors [Yichen]
- MCMC/MH/HMC
- VI
- Importance Sampling

1.2 Bayesian Deep Learning (Yichen/Eric/Haining)

- General criticisms
- BBB
- GP approximation with MC-dropout [Eric]
- SWAG
- criticize VAE "bayesian language" usage (to clarify what of focus is in Bayesian analysis) [Haining]

1.3 Normalizing Flows (Ryan/Yichen)

- Introduction to Generative Models [Yichen]
- Conditioners
- Transformer/Coupling functions
- Alternatives such as stochastic ODEs
- Computation
- Importance sampling under base-to-target as prior-to-posterior (e.g., SNF, Müller)

2 Method (Scott first draft)

The sequential nature of Bayesian learning $p(\theta|x_1, x_2) \propto p(\theta)f(x_1|\theta)f(x_2|\theta) \propto q(\theta)f(x_2|\theta)$ for data partition $x = (x_1, x_2)$ allows a composite analysis based on incorporating x_1 into an intermediate prior $q(\theta) \propto p(\theta)f(x_1|\theta)$. For some intermediate prior approximation $q_\phi(\theta) \approx p(\theta|x_1)$ as the proposal distribution for the target distribution $q(\theta|x_2) = p(\theta|x_1, x_2)$, the target to proposal densities ratio defining (unnormalized) importance sampling weights [**?**] are

$$w(\theta) \propto \frac{q(\theta|x_2)}{q_\phi(\theta)} \propto \frac{f(x_2|\theta)f(x_1|\theta)p(\theta)}{q_\phi(\theta)} \propto \frac{f(x_2|\theta)q(\theta)}{q_\phi(\theta)} \approx f(x_2|\theta)$$

where the approximation is accurate insofar as the cancellation holds. For $\theta \in I\!R^d$ for large d, using a proposal distribution $q_\phi(\theta)$ which approximates the (intermediate prior) partial posterior $p(\theta|x_1)$ facilitates efficient importance sampling by targeting proposals around $E[\theta|x_1] \approx E[\theta|x]$ and balancing importance weights $w(\theta^{(k)}) \approx f(x_2|\theta^{(k)})/\sum_j f(x_2|\theta^{(j)})$ by controlling posterior concentration to bound relative tail ratios $\frac{q(\theta|x_2)}{q_\phi(\theta)} \approx f(x_2|\theta) < c$.

The approximation of the partial posterior $q_\phi(\theta) \approx p(\theta|x_1)$ is produced by SWAG [**?**] for the data model $f(x_1|\theta)$, the x_1 subset of x, and the unupdated prior $p(\theta)$. Samples from this intermediate prior approximation are then (approximately) representative of the full posterior $p(\theta|x_1, x_2) = q(\theta|x_2)$ in proportion to the likelihood approximation of their (unnormalized) importance weights $w(\theta)$ computed from the data model $f(x_2|\theta)$, the x_2 subset of x, and the sample from $q_\phi(\theta)$. The data model itself can be flexibly estimated through a likelihood-defining neural network (NN) such as a normalizing flow (NF) [**?**] defining $f(x|\theta) = f(z = g_{\theta_0}^{-1} \circ \cdots \circ g_{\theta_T}^{-1}(x)) \prod_{t=0}^{T} |\det J_{g_{\theta_t}^{-1}}(x)|$. Thus, $p(\theta|x)$ characterizes posterior uncertainty in the data model $f(x|\theta)$ given data x.

Importantly, as illustrated in Figure **??**, the NF defines the likelihood $f(x|\theta)$ upon which the SWAG approximation of a (intermediate prior) partial posterior is based, and also defines the approximation of importance sampling weights $w(\theta)$ for posterior importance

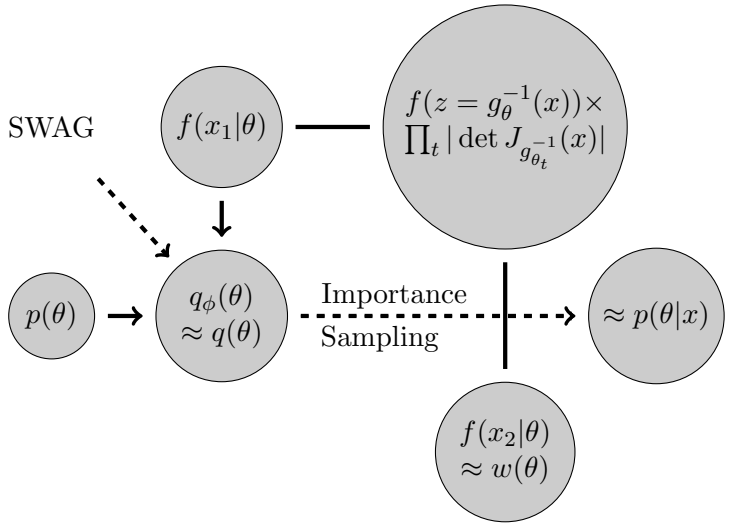

Figure 1: A visual representation of the components of the methodology. The dashed arrows indicated Bayesian posterior sampling methodologies. SWAG approximates the usual posterior derivation indicated by the solid arrows, and importance sampling reweights samples from the (intermediate) prior distribution by their likelihood values to produce a representation of the posterior. The x_1 and x_2 indicate that this occurs sequentially, and the solid lines indicate the use of the (same) data model $f(x|\theta)$ in both posterior derivations. If the SWAG approximation of $q(\theta)$ is solely viewed as a proposal distribution, then the importance weight $w(\theta)$ can be computed exactly, and the posterior approximation is exact. This characterizes posterior uncertainty over the parameters θ of the data model $f(x|\theta)$.

sampling based on $q_\phi(\theta)$. This is different than interpreting a NF as transforming a 'prior' base distribution into a posterior target distribution in order to transform samples from the 'prior' into (potentially importance sampling reweighted) samples from the posterior **??**. While this latter computation is implicitly Bayesian based on its assumption of an externally derived posterior target, the approach presented here is explicitly Bayesian in defining a likelihood and providing Bayesian updates on the parameters θ of the likelihood given the available data.

The approximation $q_\phi(\theta) \approx q(\theta)$ can be pragmatically viewed as a prior specification. This entails some "misuse" of the x_1 posterior update relative to $p(\theta)$, but can nonetheless be viewed as a practical "empirical Bayes" **??** method for prior elicitation **??**. If taken as a prior, the importance sampling weights $w(\theta) = f(x_2|\theta)$ are exact relative to $q_\phi(\theta)$ for the altered posterior $q_\phi(\theta|x_2) \approx p(\theta|x)$, where the approximation improves the larger the x_2 subset of x is. Alternatively, the importance sampling weights $w(\theta) = \frac{f(x|\theta)p(\theta)}{q_\phi(\theta)}$ are exact relative to $q(\theta)$ for the original posterior $q(\theta|x_2) = p(\theta|x)$ as will remain relatively balanced so long as $p(\theta)$ is relatively heavy-tailed.

While focus may be on $p(\theta|x)$ explicitly as in "Bayes by backprop" **??**, more generally interest is in $p\left(h(\theta)|x\right)$. For example, "MC-dropout" **??** approximates a Gaussian process posterior and so is interested in $p\left(f_\theta|x\right)$. Similarly, the NF likelihood $f_{x_0}(\theta) = f(x_0|\theta)$ evaluated at x_0 has the posterior distribution $p\left(f_{x_0}(\theta)|x\right)$ which propagates and aggregates the uncertainty in $p(\theta|x)$. So just as the posterior CDF $F_{\theta|x}$ can be estimated using importance samples representing $p(\theta|x)$, so too can the CDF $F_{f(x_0|\theta)|x}$; namely, for importance samples θ_i and corresponding weights $w(\theta_i)$, $F_{f(x_0|\theta)|x}(\theta) = \sum_{i=1}^{n} w(\theta_i) 1_{[F_{f(x_0|\theta)|x}(\theta) \leq F_{f(x_0|\theta_i)|x}(\theta_i)]}$.

2.1 nonparametric NF likelihood, SWAG prior, importance sampling posterior

2.2 computation: core sets, online covariance estimation, sampling

3 Examples

3.1 mean variance normal posterior

3.2 repeat SWAG analyses

3.3 regression

3.4 mixed effects models

4 Discussion