Title

Gian Carlo Diluvi
STAT 520C Final Project
April 2023

1 Introduction

The Bayesian statistical framework provides practitioners with a principled means to obtain finite-sample uncertainty quantification guarantees. Specifically, the posterior distribution encodes the uncertainty around the unobserved quantities given the observations and a prior distribution. Posterior credible intervals (CIs) or more general regions allow practitioners to quantify the accuracy of point estimates.

The quality of credible intervals can be measured by the proportion of times they contain the "true" value of the parameter. This is known as the frequentist coverage, and Bayesian credible intervals are asymptotically exact in this sense as a consequence of the Bernstein-von Mises theorem. Another way to assess the quality of an interval is via its *Bayesian* coverage, i.e., the proportion of times the interval contains the parameter value that generated the data and itself is a realization from the prior distribution. The Bayesian coverage of an interval leverages the fact that the prior distribution quantifies the uncertainty around the parameter before observing data.

In class, we showed that credible intervals attain nominal Bayesian coverage (i.e., their coverage is equal to the credibility level) for any sample size and without any regularity conditions. This follows from the definition of Bayesian coverage and credible intervals. However, this result assumes that we have access to *exact* credible intervals. For all but the simplest of models, however, the posterior distribution is intractable and has to be approximated numerically. In this setting, the resulting credible intervals are approximate and their Bayesian coverage need not be exact.

Variational inference (VI) is a scalable framework to learn posterior distributions. Succincly, VI casts inference as an optimization problem by approximating the posterior with an element of a family of candidate approximations that minimizes some divergence to the posterior. The quality of the approximation depends on the flexibility of the family of approximations, the divergence being minimized, and whether the optimization problem can be solved reliably. A simple instantiation of VI consists on finding the best Gaussian approximation to the posterior distribution (a framework called Gaussian VI). In this case, the optimization problem is usually amenable to off-the-shelf stochastic optimization algorithms for many common divergences.

In this work, I study the Bayesian coverage of credible intervals of Gaussian VI when minimizing the Kullback-Leibler (KL) divergence. Specifically, I focus on the impact of minimizing the reverse KL (from approximation to target) versus the forward KL (from target to approximation). The former is known to produce approximations that underestimate the posterior variance and viceversa. Through two simulation studies, I conclude that the forward KL tends to have better Bayesian coverage when

the posterior has heavier-than-Gaussian tails, which is a common occurrence (e.g., in Bayesian logistic regression, see Section 3.2.)

2 Background: Variational Inference

Let $\pi(x)$ be a density that we are interested in approximating. In the Bayesian setting, $\pi(x) = p(x)/Z$ where $p(x) = \text{prior}(x)\mathcal{L}(y \mid x)$ combines the prior and likelihood (given data y) and $Z = \int p(\tilde{x})\mathrm{d}\tilde{x}$ is the (unknown) normalizing constant (or evidence). We assume that data have already been observed and we are just interested in estimating π as a function of x, so we omit y from notation.

Variational inference refers to approximating π with an element $q^\star \in \mathcal{Q} = \{q_\lambda \mid \lambda \in \Lambda\}$, where $\mathcal{Q}$ is a family of probability distributions parametrized by λ. In the remainder of this work, $\mathcal{Q}$ will be the family of Gaussian distributions: $\lambda = (\mu, \Sigma)$ and $\mathcal{Q} = \{q_\lambda = \mathcal{N}(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma \geq 0\}$. The approximation $q^\star$ is chosen to minimize some divergence D from elements in $\mathcal{Q}$ to π:

$$q_\star = q_{\lambda^\star}, \quad \lambda^\star = \arg\min_{\lambda \in \Lambda} \mathrm{D}\left(q_\lambda || \pi\right).$$

Approximate credible intervals can be generated by taking the quantiles of the optimal approximation $q^\star$.

The choice of divergence can influence the geometry of the optimization problem as well as the characteristics of the optimal approximation $q^\star$. Arguably, the most popular choice of divergence is the Kullback-Leibler divergence from q to π, known as the *reverse* KL:

$$\mathrm{D}_{\mathrm{KL}}^{\mathrm{rev}}\left(q||\pi\right) = \int q(x) \log \frac{q(x)}{\pi(x)} \,\mathrm{d}x = \int q(x) \log \frac{q(x)}{p(x)} \,\mathrm{d}x + Z := -\mathcal{E}(q,p) + Z.$$

In the last equation, I factorized the evidence Z out of the integral and defined $\mathcal{E}(q,p)$ as the (negative) "unnormalized" KL. $\mathcal{E}(q,p)$ is known as the Evidence Lower BOund (ELBO). Since Z does not depend on λ, minimizing the KL divergence is equivalent to maximizing the ELBO.

We can solve this optimization problem through the use of gradient-based optimization algorithms. Specifically, the gradient of the ELBO is an expectation under the variational approximation q_λ:

$$\nabla_\lambda \mathcal{E}(q,p) = -\int q(x) \log \frac{q(x)}{p(x)} \nabla_\lambda q_\lambda(x) \,\mathrm{d}x.$$

We can approximate the ELBO gradient via Monte Carlo samples from q within a stochastics gradient ascent routine (black box vi cite).

The minimizer of the reverse KL, $q_{\mathrm{rev}}(x)$, is known to underestimate the variance of $\pi(x)$ (cite ML book). One way to address this is to minimize the *forward* KL divergence instead, i.e., the KL divergence from the posterior π to the approximation:

$$\mathrm{D}_{\mathrm{KL}}^{\mathrm{fwd}}\left(\pi||q\right) = \int \pi(x) \log \frac{\pi(x)}{q(x)}.$$

Common folk wisdom in the machine learning community suggests that $\mathrm{D}_{\mathrm{KL}}^{\mathrm{fwd}}\left(\pi||q\right)$ produces better approximations, $q_{\mathrm{fwd}}(x)$, but it is considerably more difficult to optimize. For example, its gradient can be expressed as an expectation but under the intractable π:

$$\nabla_\lambda \mathrm{D}_{\mathrm{KL}}^{\mathrm{fwd}}\left(q||\pi\right) = -\int \pi(x) \nabla_\lambda q(x) \,\mathrm{d}x.$$

Importance sampling can be numerically unstable for high-dimensional π, so recent work has attempted to estimate this gradient by running an MCMC chain in parallel to produce approximate samples from π (cite Markovian score climbing).

3 Experiments

In this section, we carry out a simulation study to compare the Bayesian coverage of credible intervals from Gaussian approximations to two different targets. The Gaussian distributions were parametrized by mean and log standard deviation, which is more numerically stable. In each simulation study, we minimized both the reverse and the forward KL divergence using 10,000 iterations of stochastic gradient descent with learning rate $\propto 1/\sqrt{t}$; the proportionality constant was fine-tuned for every experiment. The code to reproduce our experiments is available at .

3.1 Cauchy

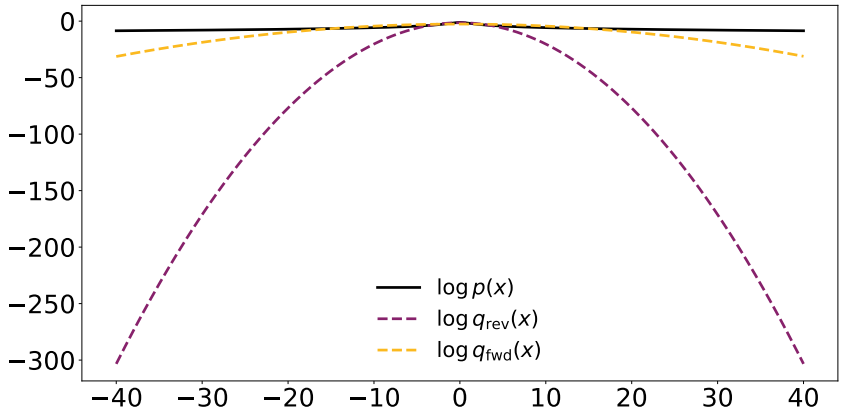
First we consider approximating a Cauchy distribution $\pi(x) = (\pi(1 + x^2))^{-1}$. This example was chosen because it gives us access to exact credible intervals and also because of its heavy tails, which should encourage the reverse KL-optimal approximation to severly underestimate the variance and viceversa with the forward KL-optimal approximation. This is seen in Fig. 1a, where the tails of q_{fwd} are a better approximation to those of π. Figs. 1b and 1c show the limits and resulting coverage of each approximation. As expected, the q_{fwd} tends to have better coverage, especially for large credibility values. Specifically, for $\alpha = 0.05$ the coverage is nearly exact. On the other hand, q_{rev} produces intervals with consistently low coverage.

3.2 Logistic regression

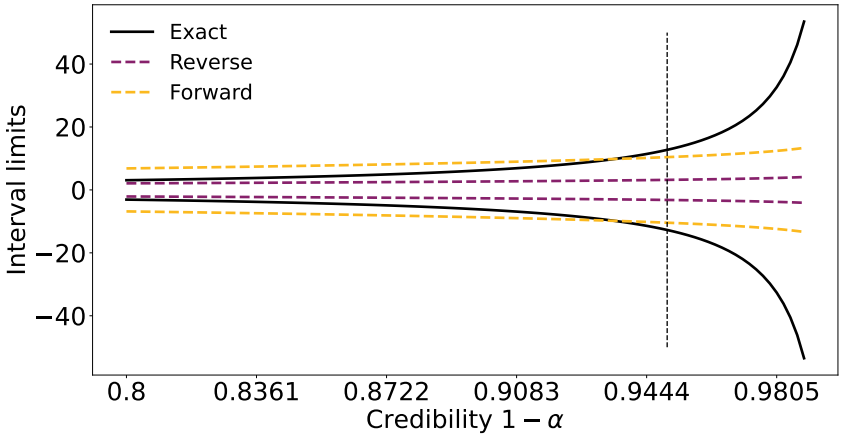
Now we consider a logistic regression example where we observe $N = 100$ observations from the model

$$Y_n \sim \mathsf{Bern}(p_n), \quad p_n = \frac{1}{1 + \exp\{-\beta_0 - \beta_1 x_1\}}, \quad x_n \sim \mathcal{N}(-1, 1.5^2).$$

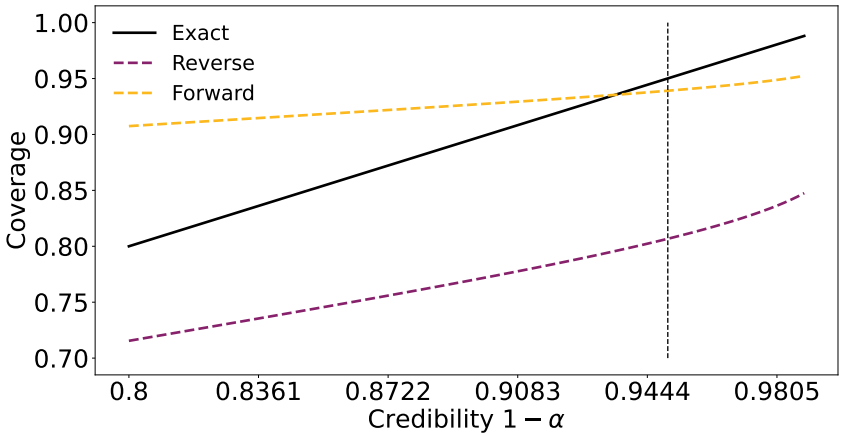
We set $\beta = (2, 3)$ and assumed that $\beta_0 = 2$ is known, so we only approximate the posterior distribution of β_1. We estimated the normalizing constant Z using a simple numerical approximation to the integral. The posterior distribution is unimodal and symmetri, and thus should be well-approximated by a Gaussian, but it has a slightly heavy right tail. We also ran Hamiltonian Monte Carlo for 10,000 iterations using Stan to assess the fidelity of q_{rev} and q_{fwd}. Fig. 2a shows the resulting densities. The forward KL-optimal approximation seems to be the better fit, and it also better captures the right-tail of the target as seen in Fig. 2b (although neither it nor q_{bwd} do an amazing job). In terms of the interval limits, the light left tail is well approximated by both VI densities and by the MCMC samples, but the heavier right tail results in credible intervals with smaller lower limits, as seen in Fig. 2c. All methodologies result in intervals with smaller-than-nominal coverage, although q_{fwd} does have the coverage closest to nominal for credibility values larger than 0.05; see Fig. 2d. This is potentially due to the slight over correction in the lower limit of the interval in Fig. 2c.



(a) Log density

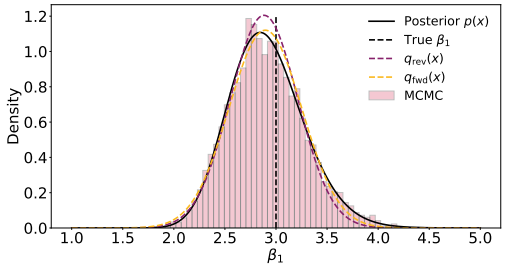


(b) CI limits

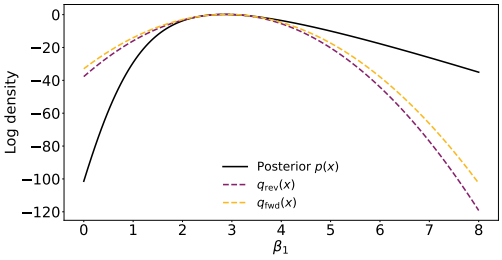


(c) CI coverage

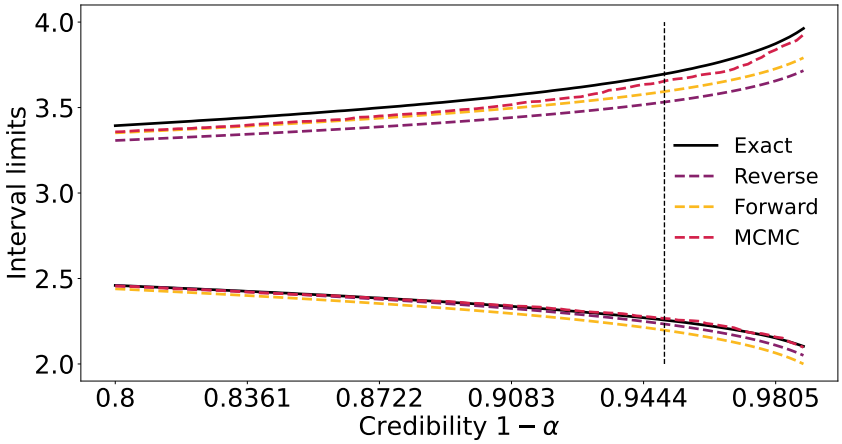
Figure 1: Results on the Cauchy distribution.



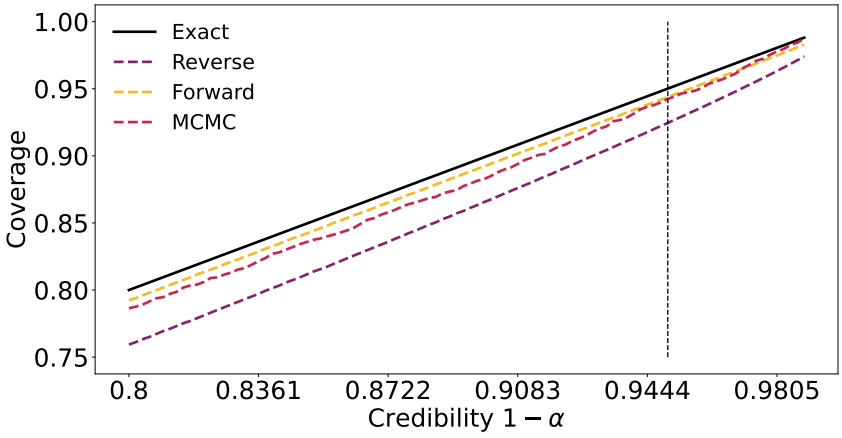
(a) Density



(b) Log density



(c) CI limits



(d) CI coverage

Figure 2: Results on the logistic regression example.

References