Automated Posterior Interval Evaluation for Inference in Probabilistic Programming

Edward Kao, Michael Yee | MIT Lincoln Laboratory

ABSTRACT

In probabilistic inference, credible intervals constructed from posterior samples provide ranges of likely values for continuous parameters of interest. Intuitively, an inference procedure is optimal if it produces the most precise posterior intervals that cover the true parameter value with the expected frequency in repeated experiments. We present theories and methods for automating posterior interval evaluation of inference performance in probabilistic programming using two metrics: 1.) truth coverage, and 2.) ratio of the empirical over the ideal interval widths. Demonstrating with inference on popular regression and state-space models, we show how the metrics provide effective comparisons between different inference procedures, and capture the effects of collinearity and model misspecification. Overall, we claim such automated interval evaluation can accelerate the robust design and comparison of probabilistic inference programs by directly diagnosing how accurately and precisely they can estimate parameters of interest.

THEORY

Based on the statistical principle of evaluating Bayesian inference with frequentist properties (1,2), we compute two metrics for inference output in repeated experiments: 1.) posterior credible interval coverage of the true parameter value (90% intervals should cover the truth 90% of the time), and 2.) ratio of the empirical over the ideal interval widths (ratio of 1 indicates precise inference). The ideal interval width can be computed based on the asymptotic theorem:

The Bernstein-von Mises Theorem (3). For regular models, the posterior distribution of continuous parameters in finite dimensions converges asymptotically, with increasing data, in distribution to Normal with mean at the true value $\boldsymbol{\theta}^*$ and covariance equal to the inverse of the Fisher information matrix $\boldsymbol{\mathcal{I}}$ evaluated at $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta} \xrightarrow{D} \text{Normal}\left(\boldsymbol{\theta}^*,\ \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}\right) \qquad [1]$$

The diagonal terms of the asymptotic covariance $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}^*)^{-1}$ provide the ideal interval width for each parameter in the model. In the non-asymptotic regime, the ideal interval width can be computed using the Laplace approximation where $q''(\boldsymbol{\theta}^*)$ is the Hessian of the log posterior distribution evaluated at $\boldsymbol{\theta}^*$:

$$\boldsymbol{\theta} \ \approx\ \text{Normal}\left(\boldsymbol{\theta}^*,\ -q''(\boldsymbol{\theta}^*)^{-1}\right) \qquad [2]$$

AUTOMATION

Computing the proposed metrics based on posterior intervals can be automated for any probabilistic programming systems (4) that ***simulate*** data $\mathcal{D}$ and parameters $\boldsymbol{\theta}$ based on statistical models M and priors, and ***infer*** the posterior distribution such that the likelihood function and the unnormalized posterior distribution are accessible. The Fisher information matrix can be computed via the hessian function on the log likelihood, through auto-differentiation. For non-asymptotic cases using the Laplace approximation, simply replace the likelihood with the unnormalized posterior distribution. We implement and demonstrate this automated evaluation in Gen (5).

function `simulate` (M)
- $\boldsymbol{\theta} \sim p_M(\cdot)$
- $\mathcal{D}|\boldsymbol{\theta} \sim p_M(\cdot|\boldsymbol{\theta})$
- **return** $(\boldsymbol{\theta}, \mathcal{D})$

end

function `ll-hessian` $(M, \boldsymbol{\theta}, \mathcal{D})$
- **return** $\left[\dfrac{\partial \log p_M(\mathcal{D}|\boldsymbol{\theta})}{\partial\theta_u \theta_v}\right]_{u,v \in \boldsymbol{\theta}\times\boldsymbol{\theta}}$

end

Algorithm for Computing the Truth Coverage Rate

Input: Model M, inference program `infer`, interval probability ci, #simulations S, #inference samples T
Output: Coverage rates for S simulated datasets

```
for s ← 1 to S do
    (θ*, D) ← simulate(M)
    outcomes ← [ ]
    θ^{1:T} ← infer(M, D, T)
    for u ∈ θ do
        (θ_lo, θ_hi) ← empirical-interval(θ_u^{1:T}, ci)
        outcomes.append(θ_lo ≤ θ*_u ≤ θ_hi)
    end
end
return mean(outcomes)
```

Algorithm for Computing the Interval Width Ratios

Input: Model M, inference program `infer`, interval probability ci, #simulations S, #inference samples T
Output: Empirical to ideal width ratios
ratios ← [ ]

```
for s ← 1 to S do
    (θ*, D) ← simulate(M)
    I(θ*) ← -ll-hessian(M, θ*, D)
    σ* ← sqrt(diag(I(θ*)^{-1}))
    θ^{1:T} ← infer(M, D, T)
    for u ∈ θ do
        (θ_lo, θ_hi) ← empirical-interval(θ_u^{1:T}, ci)
        (θ*_lo, θ*_hi) ← ideal-interval(θ*_u, σ*_u, ci)
        ratios.append((θ_hi − θ_lo)/(θ*_hi − θ*_lo))
    end
end
return ratios
```

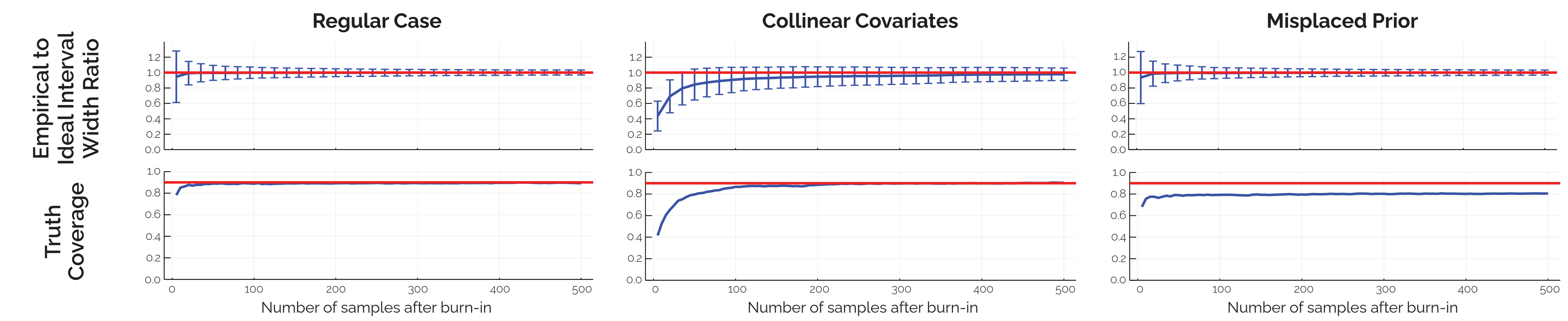

1. Rubin, D.B. Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics (1984).
2. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. Bayesian data analysis. Taylor & Francis (2014).
3. Van der Vaart, A.W. Asymptotic statistics, 10.2 Bernstein–von Mises Theorem. Cambridge University Press (1998).
4. Gordon, A.D., Henzinger, T.A., Nori, A.V., and Rajamani, S.K. Probabilistic programming. ACM (2014).
5. Cusumano-Towner, M.F., Saad, F.A., Lew, A.K., and Mansinghka V.K. Gen: a general-purpose probabilistic programming system with programmable inference. ACM (2019).

DEMONSTRATION ON BAYESIAN LINEAR REGRESSION

$$\boldsymbol{w}|\alpha \sim \text{Normal}(\boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I})$$
$$\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w},\beta \sim \text{Normal}(\boldsymbol{X}\boldsymbol{w}, \beta^{-1}\boldsymbol{I})$$

where $\boldsymbol{w}$ and $\boldsymbol{m}_0 \in \mathbb{R}^d$, $\boldsymbol{y} \in \mathbb{R}^N$, and $\boldsymbol{X}$ is an $N{\times}d$ covariate matrix. Here, d = 2, N = 30 and the covariates $\boldsymbol{X}$ are generated independently, or with collinearity of 0.9 correlation

For purpose of demonstration, we infer $\boldsymbol{w}$ using Gibbs sampling, even though the posterior has a closed-form expression. Evaluation cases: 1.) Regular Gibbs sampling, 2.) Covariates generated with collinearity, and 3.) Prior location is misplaced



Posterior interval evaluation identifies the expected effects of collinearity and misplaced prior on inference

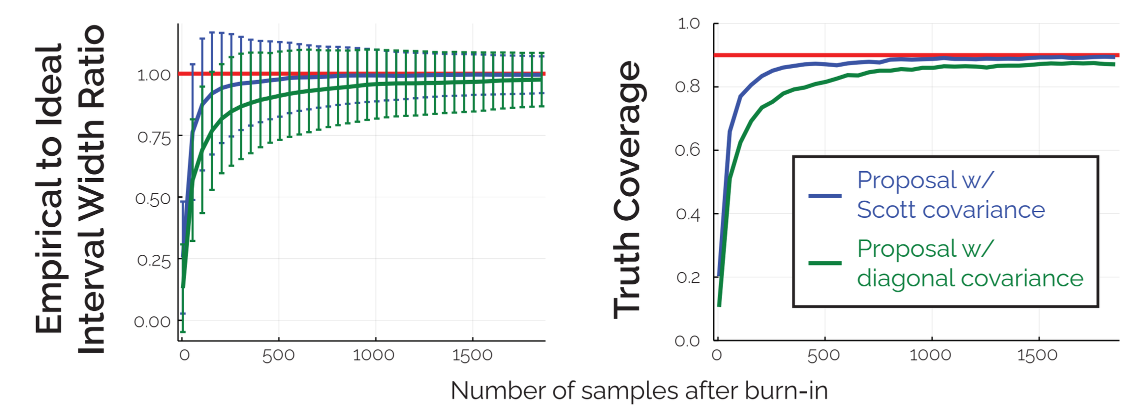
DEMONSTRATION ON BAYESIAN LOGISTIC REGRESSION

$$\boldsymbol{w}|\alpha \sim \text{Normal}(\boldsymbol{m}_0, \alpha^{-1}\boldsymbol{I})$$
$$\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w} \sim \text{Bernoulli}(\text{sigmoid}(\boldsymbol{X}\boldsymbol{w}))$$

where $\boldsymbol{w}$ and $\boldsymbol{m}_0 \in \mathbb{R}^d$, $\boldsymbol{y} \in \{0,1\}^N$, and $\boldsymbol{X}$ is an $N{\times}d$ covariate matrix. Here, d = 10, N = 100 and the covariates $\boldsymbol{X}$ are generated with collinearity of 0.95 correlation

We infer $\boldsymbol{w}$ using Random Walk Metropolis-Hastings with two multivariate normal proposals: $\boldsymbol{w}' \sim \text{Normal}(\boldsymbol{w}, \boldsymbol{\Sigma})$:

1. Scott proposal (approximation to asymptotically optimal proposal): $\boldsymbol{\Sigma}_{Scott} = \left(\boldsymbol{V}_0^{-1} + \frac{6}{\pi^2}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$, where $\boldsymbol{V}_0$ is the covariance of multivariate normal prior on $\boldsymbol{w}$
2. Naïve proposal (diagonal covariance matrix): $\boldsymbol{\Sigma} = 0.2\ \boldsymbol{I}$



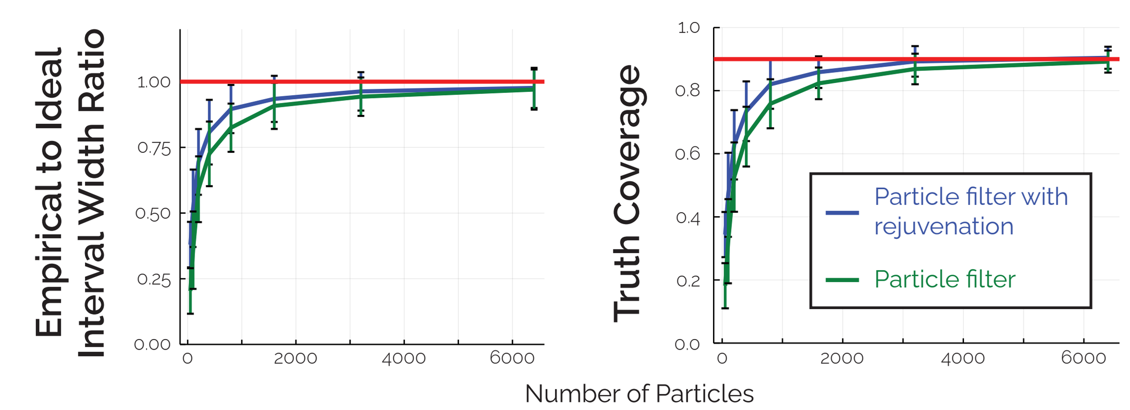
Evaluation quantifies how much faster the Scott proposal converges than the naïve proposal, under collinearity

DEMONSTRATION ON NONLINEAR STATE-SPACE MODEL

$$x_t = \frac{x_{t-1}}{2} + 25\frac{x_{t-1}}{1+x_{t-1}^2} + 8\ \cos(0.1t) + \delta_{t-1}$$
$$y_t = x_t + \epsilon_t \qquad x_1 \sim \text{Normal}(\mu, \nu^2)$$
$$\delta_{t-1} \sim \text{Normal}(0, \omega^2)$$
$$\epsilon_t \sim \text{Normal}(0, \sigma^2)$$

A popular nonlinear state-space model in the literature. Here we generate 100 steps in time to capture both the periodic motion and nonlinear drift.

Infer the states x given the observations y using particle filters. Compare standard particle filters against one with rejuvenation moves on past states.



Evaluation shows with the addition of rejuvenation, same performance is reached with far fewer particles

FUTURE WORK

- Apply the proposed evaluation to real-world scenarios with a single data realization and unknown truth.
- Extend approach to general models with a mixture of regular and irregular parameters through conditioning and exploring generalizations of the Bernstein-von Mises Theorem.
- Explore simple and automated indicators for the adequacy of the normal assumption on the true posterior.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.
This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY