Rejection Sampling Variational Inference

Karan Grewal

CSC2547 / STA4273

Overview

Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms

Christian A. Naesseth[†‡] **Francisco J. R. Ruiz**[†§] **Scott W. Linderman**[†] **David M. Blei**[†]
[†]Linköping University [‡]Columbia University [§]University of Cambridge

Abstract

Variational inference using the reparameterization trick has enabled large-scale approximate Bayesian inference in complex probabilistic models, leveraging stochastic optimization to sidestep intractable expectations. The reparameterization trick is applicable when we can simulate a random variable by applying a differentiable deterministic function on an auxiliary random variable whose distribution is fixed. For many distributions of interest (such as the gamma or Dirichlet), simulation of random variables relies on acceptance-rejection sampling. The discontinuity introduced by the accept–reject step means that standard reparameterization tricks are not applicable. We propose a new method that lets us leverage reparameterization gradients even when variables are outputs of a acceptance-rejection sampling algorithm. Our approach enables reparameterization on a larger class of variational distributions.

2014] and text [Hoffman et al., 2013]. By definition, the success of variational approaches depends on our ability to (i) formulate a flexible parametric family of distributions; and (ii) optimize the parameters to find the member of this family that most closely approximates the true posterior. These two criteria are at odds—the more flexible the family, the more challenging the optimization problem. In this paper, we present a novel method that enables more efficient optimization for a large class of variational distributions, namely, for distributions that we can efficiently simulate by acceptance-rejection sampling, or rejection sampling for short.

For complex models, the variational parameters can be optimized by stochastic gradient ascent on the evidence lower bound (ELBO), a lower bound on the marginal likelihood of the data. There are two primary means of estimating the gradient of the ELBO: the score function estimator [Paisley et al., 2012, Ranganath et al., 2014, Mnih and Gregor, 2014] and the reparameterization trick [Kingma and Welling, 2014, Rezende et al., 2014, Price, 1958, Bonnet, 1964], both of which rely on Monte Carlo sampling. While the

Variational Inference

- Interested in computing posterior $p(z|x)$, but it is often intractable

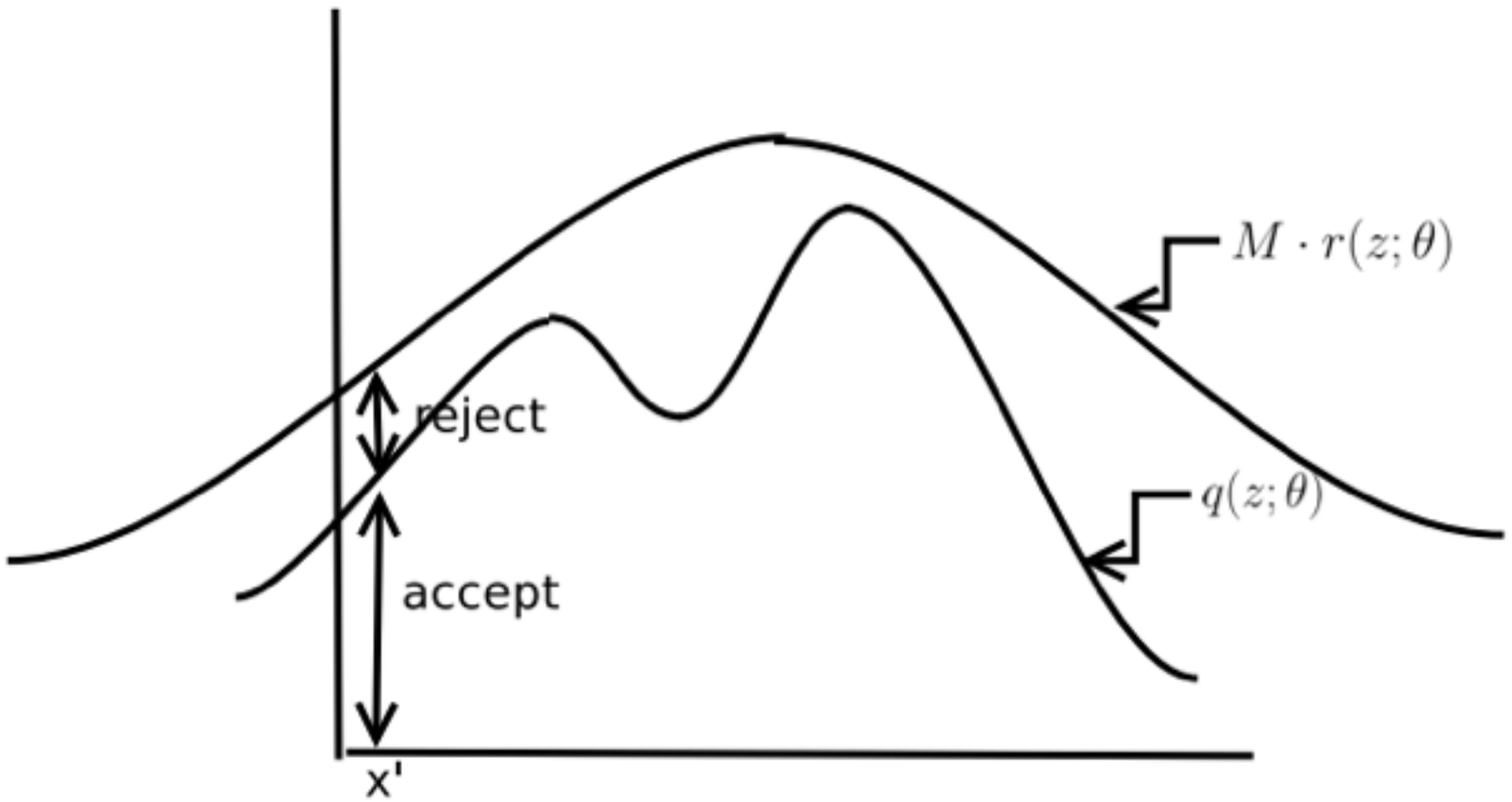
- parametrize a *variational family* of distributions $q(z;\theta)$ to approximate true posterior

- Maximize *Evidence Lower Bound* (ELBO):

$$\mathcal{L}(\theta) = \mathbb{E}_{q(z;\theta)}\left[f(z)\right] + \mathbb{H}[q(z;\theta)],$$
$$f(z) := \log p(x,z),$$
$$\mathbb{H}[q(z;\theta)] := \mathbb{E}_{q(z;\theta)}[-\log q(z;\theta)].$$

Rejection Sampling

- Want to sample from $q(z;\theta)$, parametrize a *proposal distribution* $r(z;\theta)$ s.t. $q(z;\theta) \leq M \cdot r(z;\theta)$
- Accept sample $z \sim r(z;\theta)$ with probability $\frac{q(z;\theta)}{M \cdot r(z;\theta)}$

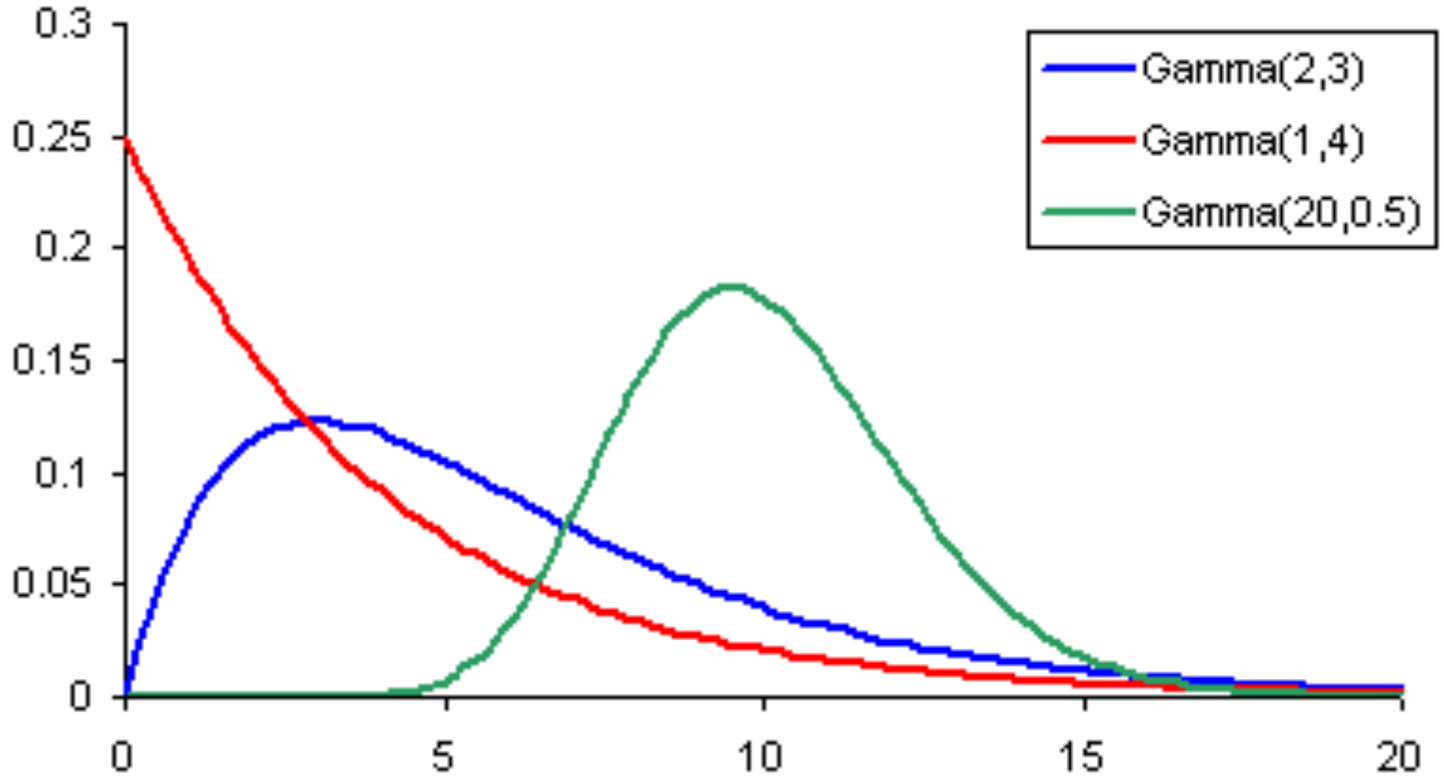


source: https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/lectures/lecture17.pdf

Reparameterized Rejection Sampler

- Problem: what if we want our variational family $q(z;\theta)$ to follow a distribution that requires rejection sampling to approximate?
- Rejection sampling causes discontinuities

Example: Gamma Distribution



source: http://www.epixanalytics.com/

- To sample from $\mathrm{Gamma}(\theta, \beta)$, sample from $\mathrm{Gamma}(\theta, 1)$ and divide by β, the acceptance probability is dependent on θ

Reparameterized Rejection Sampler

1. Reparameterize z: $z = h(\varepsilon, \theta), \varepsilon \sim s(\varepsilon)$

2. Find marginal distribution of accepted sample ε:

$$
\begin{aligned}
\pi(\varepsilon\,;\theta) &= \int \pi(\varepsilon, u\,;\theta)\mathrm{d}u \\
&= \int M{\cdot}\, s(\varepsilon)\mathbb{1}\left[0 < u < \frac{q\left(h(\varepsilon,\theta)\,;\theta\right)}{M{\cdot}\, r\left(h(\varepsilon,\theta)\,;\theta\right)}\right]\mathrm{d}u \\
&= s(\varepsilon)\frac{q\left(h(\varepsilon,\theta)\,;\theta\right)}{r\left(h(\varepsilon,\theta)\,;\theta\right)},
\end{aligned}
$$

Reparameterized Rejection Sampler

3. Rewrite ELBO:

$$\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)]$$
$$= \nabla_\theta \mathbb{E}_{\pi(\varepsilon;\theta)}[f(h(\varepsilon,\theta))]$$
$$= \mathbb{E}_{\pi(\varepsilon;\theta)}[\nabla_\theta f(h(\varepsilon,\theta))] + \mathbb{E}_{\pi(\varepsilon;\theta)}\left[f(h(\varepsilon,\theta))\nabla_\theta \log \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)}\right]$$

Related Work

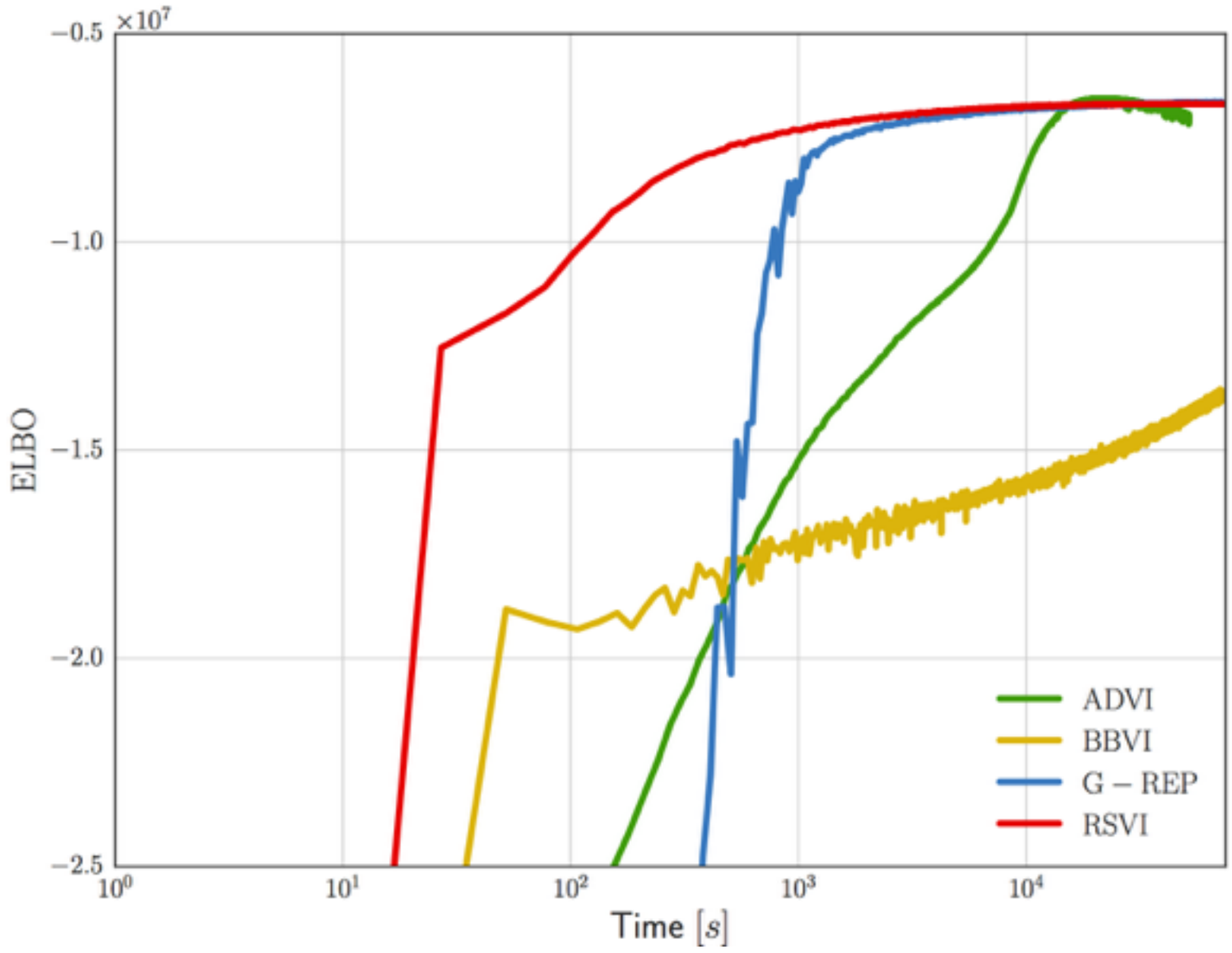
- Automatic Differentiation Variational Inference **(ADVI)**
  - fit z with Gaussian posterior; cannot learn a Gamma or Dirichlet posterior

- Black-Box Variational Inference **(BBVI)**
  - sample from $q(z;\theta)$ to approximate gradient

- Generalized Reparameterization Gradient **(G-REP)**
  - find a distribution $s(\varepsilon)$ that makes ε dependent on choice of variational family

Results

- model = sparse gamma Deep Exponential Family



source: https://arxiv.org/pdf/1610.05683.pdf

Future Work

- Combining Rejection Sampling Variational Inference with Metropolis-Hastings
- Metropolis-Hastings: Acquire a sequence of samples from a distribution that is difficult to sample from directly; use rejection sampling

Supplementary: Gradient Derivation

$$\nabla_\theta \mathbb{E}_{q(z;\theta)}[f(z)] = \nabla_\theta \mathbb{E}_{\pi(\varepsilon;\theta)}[f(h(\varepsilon,\theta))] = \int s(\varepsilon)\nabla_\theta \left(f(h(\varepsilon,\theta)) \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} \right) d\varepsilon$$

$$= \int s(\varepsilon) \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} \nabla_\theta f(h(\varepsilon,\theta)) d\varepsilon$$

$$+ \int s(\varepsilon) f(h(\varepsilon,\theta)) \nabla_\theta \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} d\varepsilon$$

$$= \int s(\varepsilon) \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} \nabla_\theta f(h(\varepsilon,\theta)) d\varepsilon$$

$$+ \int s(\varepsilon) \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} f(h(\varepsilon,\theta)) \nabla_\theta \log \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} d\varepsilon$$

$$= \mathbb{E}_{\pi(\varepsilon;\theta)}[\nabla_\theta f(h(\varepsilon,\theta))] + \mathbb{E}_{\pi(\varepsilon;\theta)} \left[f(h(\varepsilon,\theta)) \nabla_\theta \log \frac{q(h(\varepsilon,\theta);\theta)}{r(h(\varepsilon,\theta);\theta)} \right]$$