Rejection Sampling Variational Inference

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 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 rejection sampling algorithm. Our approach enables reparameterization on a larger class of variational distributions. In several studies of real and synthetic data, we show that the variance of the estimator of the gradient is significantly lower than other state-of-the-art methods. This leads to faster convergence of stochastic optimization variational inference.

Let $p(x, z)$ be a probabilistic model, i.e., a joint probability distribution of *data* x and *latent* (unobserved) variables z. In Bayesian inference, we are interested in the posterior distribution $p(z|x) = \frac{p(x,z)}{p(x)}$. For most models, the posterior distribution is analytically intractable and we have to use an approximation, such as Monte Carlo methods or variational inference. In this paper, we focus on variational inference.

In variational inference, we approximate the posterior with a *variational family* of distributions $q(z\,;\theta)$, parameterized by θ. Typically, we choose the *variational parameters* θ that minimize the Kullback-Leibler (KL) divergence between $q(z\,;\theta)$ and $p(z|x)$. This minimization is equivalent to maximizing the evidence lower bound (ELBO) [Jordan et al., 1999],

$$\begin{aligned} &\mathcal{L}(\theta) = \mathbb{E}_{q(z\,;\theta)}\left[f(z)\right] + \mathbb{H}[q(z\,;\theta)], \\ &f(z) := \log p(x, z), \quad \mathbb{H}[q(z\,;\theta)] := \mathbb{E}_{q(z\,;\theta)}[-\log q(z\,;\theta)]. \end{aligned} \tag{1}$$

When the model and variational family satisfy conjugacy requirements, we can use coordinate ascent to find a local optimum of the ELBO [Ghahramani and Beal, 2001, Blei et al., 2016]. If the conjugacy requirements are not satisfied, a common approach is to build a Monte Carlo estimator of the gradient of the ELBO [Paisley et al., 2012, Ranganath et al., 2014, Salimans and Knowles, 2013]. Empirically, the *reparameterization trick* has been shown to be beneficial over direct Monte Carlo estimation of the gradient using the score fuction estimator [Rezende et al., 2014, Kingma and Welling, 2014, Titsias and Lázaro-Gredilla, 2014, Fan et al., 2015]. However, it is not generally applicable, it requires that: (i) the latent variables z are continuous; and (ii) we can simulate from $q(z\,;\theta)$ as follows,

$$z = h(\varepsilon, \theta), \qquad \text{with } \varepsilon \sim s(\varepsilon). \tag{2}$$

Here, $s(\varepsilon)$ is a distribution that does not depend on the variational parameters; it is typically a standard normal or a standard uniform. Further, $h(\varepsilon, \theta)$ is differentiable with respect to θ. Using (2), we can move the derivative inside the expectation and rewrite the gradient of the ELBO as

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{s(\varepsilon)}\left[\nabla_z f(h(\varepsilon, \theta))\nabla_\theta h(\varepsilon, \theta)\right] + \nabla_\theta \mathbb{H}[q(z\,;\theta)].$$

*Corresponding author: christian.a.naesseth@liu.se

Advances in Approximate Bayesian Inference (NIPS 2016 Workshop), Barcelona, Spain.

Algorithm 1 Reparameterized Rejection Sampling

Input: target $q(z\,;\theta)$, proposal $r(z\,;\theta)$, and constant M_θ, with $q(z\,;\theta) \le M_\theta r(z\,;\theta)$
Output: ε such that $h(\varepsilon,\theta) \sim q(z\,;\theta)$

$i \leftarrow 0$
repeat
  $i \leftarrow i + 1$
  Propose $\varepsilon_i \sim s(\varepsilon)$
  Simulate $u_i \sim \mathcal{U}[0,1]$
until $u_i < \frac{q(h(\varepsilon_i,\theta)\,;\theta)}{M_\theta r(h(\varepsilon_i,\theta)\,;\theta)}$
return ε_i

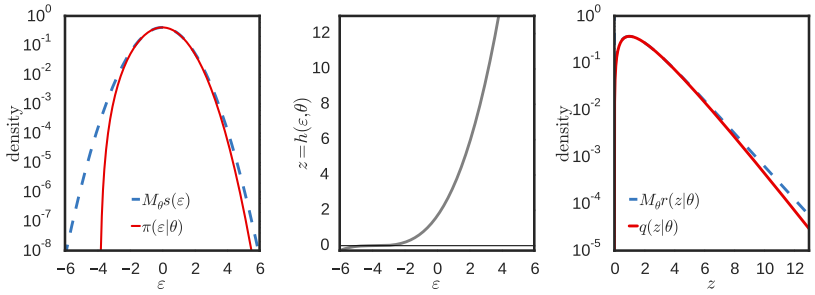


Figure 1: Example of a reparameterized rejection sampler for $q(z\,;\theta) = \mathrm{Gamma}(\theta, 1)$, shown here with $\theta = 2$. We use the rejection sampling algorithm of Marsaglia and Tsang [2000], which is based on a nonlinear transformation $h(\varepsilon,\theta)$ of a standard normal $\varepsilon \sim \mathcal{N}(0,1)$, and has acceptance probability of 0.98 for $\theta = 2$. The marginal density of the accepted value of ε (integrating out the acceptance variables, $u_{1:i}$) is given by $\pi(\varepsilon\,;\theta)$. We compute unbiased estimates of the gradient of the ELBO (6) via Monte Carlo, using Algorithm 1 to rejection sample $\varepsilon \sim \pi(\varepsilon\,;\theta)$. By reparameterizing in terms of ε, we obtain a low-variance estimator of the gradient for challenging variational distributions.

We next show that taking a novel view of the rejection sampler lets us perform exact reparameterization for variational families where it was previously not possible.

1 Reparameterizing the Rejection Sampler

The basic idea behind reparameterization is to rewrite simulation from a complex distribution as a deterministic mapping of its parameters and a set of simpler random variables. We can view the rejection sampler as a (complicated) deterministic mapping of a (random) number of simple random variables such as uniforms and normals. This makes it tempting to take the standard reparameterization approach when we consider random variables generated by rejection samplers. However, this mapping is in general *not continuous*, and thus moving the derivative inside the expectation and using direct automatic differentiation would not give the correct answer.

Our insight is that we can overcome this problem by instead considering only the marginal over the accepted sample, analytically integrating out the accept–reject variable. Thus, the mapping comes from the proposal step. This is continuous under mild assumptions, enabling us to greatly extend the class of variational families amenable to reparameterization.

We first review rejection sampling and present the reparameterized rejection sampler. Next we show how to use it to calculate low-variance gradients of the ELBO. Finally, we present the complete stochastic optimzation for variational inference, rejection sampling variational inference (RSVI).

1.1 Reparameterized Rejection Sampling

Rejection sampling is a powerful way of simulating random variables from complex distributions whose inverse cumulative distribution functions are not available or are too expensive to evaluate [Devroye, 1986, Robert and Casella, 2004]. We consider an alternative view of rejection sampling

in which we explicitly make use of the reparameterization trick. This view of the rejection sampler enables our variational inference algorithm in Section 1.2.

To generate samples from a distribution $q(z\,;\theta)$ using rejection sampling, we first sample from a *proposal distribution* $r(z\,;\theta)$ such that $q(z\,;\theta) \leq M_\theta r(z\,;\theta)$ for some $M_\theta < \infty$. In our version of the rejection sampler, we assume that the proposal distribution is reparameterizable, i.e., that generating $z \sim r(z\,;\theta)$ is equivalent to generating $\varepsilon \sim s(\varepsilon)$ (where $s(\varepsilon)$ does not depend on θ) and then setting $z = h(\varepsilon, \theta)$ for a differentiable function $h(\varepsilon, \theta)$. We then accept the sample with probability $\min\left\{1, \frac{q(h(\varepsilon,\theta)\,;\theta)}{M_\theta r(h(\varepsilon,\theta)\,;\theta)}\right\}$; otherwise, we reject the sample and repeat the process. We illustrate this in Figure 1 and provide a summary of the method in Algorithm 1, where we consider the output to be the (accepted) variable ε, instead of z.

The ability to simulate from $r(z\,;\theta)$ by a reparameterization through a differentiable $h(\varepsilon, \theta)$ is not needed for the rejection sampler to be valid. However, this is indeed the case for the rejection sampler of many common distributions.

1.2 The Reparameterized Rejection Sampler in Variational Inference

We now use reparameterized rejection sampling to develop a novel Monte Carlo estimator of the gradient of the ELBO. We first rewrite the ELBO in (1) as an expectation in terms of the transformed variable ε,

$$\mathcal{L}(\theta) = \mathbb{E}_{q(z\,;\theta)}\left[f(z)\right] + \mathbb{H}[q(z\,;\theta)] = \mathbb{E}_{\pi(\varepsilon\,;\theta)}\left[f\left(h(\varepsilon,\theta)\right)\right] + \mathbb{H}[q(z\,;\theta)]. \tag{3}$$

In this expectation, $\pi(\varepsilon\,;\theta)$ is the distribution of the *accepted sample* ε in Algorithm 1. We construct it by marginalizing over the auxiliary uniform variable u,

$$\pi(\varepsilon\,;\theta) = \int \pi(\varepsilon, u\,;\theta)\mathrm{d}u = \int M_\theta s(\varepsilon)\mathbb{1}\left[0 < u < \frac{q\left(h(\varepsilon,\theta)\,;\theta\right)}{M_\theta r\left(h(\varepsilon,\theta)\,;\theta\right)}\right]\mathrm{d}u = s(\varepsilon)\frac{q\left(h(\varepsilon,\theta)\,;\theta\right)}{r\left(h(\varepsilon,\theta)\,;\theta\right)}, \tag{4}$$

where $\mathbb{1}[x \in A]$ is the indicator function, and recall that M_θ is a constant used in the rejection sampler. This can be seen by the algorithmic definition of the rejection sampler, where we propose values $\varepsilon \sim s(\varepsilon)$ and $u \sim \mathcal{U}[0,1]$ until acceptance, i.e., until $u < \frac{q(h(\varepsilon,\theta)\,;\theta)}{M_\theta r(h(\varepsilon,\theta)\,;\theta)}$.

We can now compute the gradient of $\mathbb{E}_{q(z\,;\theta)}[f(z)]$ based on Eq. 3,

$$\begin{aligned}\nabla_\theta \mathbb{E}_{q(z\,;\theta)}[f(z)] &= \nabla_\theta \mathbb{E}_{\pi(\varepsilon\,;\theta)}[f(h(\varepsilon,\theta))] \\ &= \underbrace{\mathbb{E}_{\pi(\varepsilon\,;\theta)}[\nabla_\theta f(h(\varepsilon,\theta))]}_{=:g_{\text{rep}}} + \underbrace{\mathbb{E}_{\pi(\varepsilon\,;\theta)}\left[f(h(\varepsilon,\theta))\nabla_\theta \log \frac{q(h(\varepsilon,\theta)\,;\theta)}{r(h(\varepsilon,\theta)\,;\theta)}\right]}_{=:g_{\text{cor}}},\end{aligned} \tag{5}$$

where we have used the log-derivative trick and rewritten the integrals as expectations with respect to $\pi(\varepsilon\,;\theta)$ (see Naesseth et al. [2016, Section 3] for all details.) We define g_{rep} as the reparameterization term, which takes advantage of gradients with respect to the model and its latent variables; we define g_{cor} as a correction term that accounts for *not* using $r(z\,;\theta) \equiv q(z\,;\theta)$.

Using (5), the gradient of the ELBO in (1) can be written as

$$\nabla_\theta \mathcal{L}(\theta) = g_{\text{rep}} + g_{\text{cor}} + \nabla_\theta \mathbb{H}[q(z\,;\theta)], \tag{6}$$

and thus we can build an unbiased one-sample Monte Carlo estimator $\hat{g} \approx \nabla_\theta \mathcal{L}(\theta)$ as

$$\begin{aligned}&\hat{g} := \hat{g}_{\text{rep}} + \hat{g}_{\text{cor}} + \nabla_\theta \mathbb{H}[q(z\,;\theta)], \\ &\hat{g}_{\text{rep}} = \nabla_z f(z)\big|_{z=h(\varepsilon,\theta)} \nabla_\theta h(\varepsilon,\theta), \quad \hat{g}_{\text{cor}} = f(h(\varepsilon,\theta))\nabla_\theta \log \frac{q(h(\varepsilon,\theta)\,;\theta)}{r(h(\varepsilon,\theta)\,;\theta)},\end{aligned} \tag{7}$$

where ε is a sample generated using Algorithm 1. (We can also generate more samples of ε and average; however, one sample is enough in practice to ensure a reasonable gradient estimate.)

We now describe the full variational algorithm based on reparameterizing the rejection sampler. We make use of Eq. 6 to obtain a Monte Carlo estimator of the gradient of the ELBO. We use this estimate to take stochastic gradient steps. We use the step-size sequence ρ^n proposed by Kucukelbir et al. [2016] (also used by Ruiz et al. [2016]), which combines RMSPROP [Tieleman and Hinton, 2012] and Adagrad [Duchi et al., 2011]. We summarize the full method in Algorithm 2. We refer to our method as RSVI.

Algorithm 2 Rejection Sampling Variational Inference

Input: Data x, model $p(x, z)$, variational family $q(z\,;\theta)$
Output: Variational parameters θ^*

1: **repeat**
2: Run Algorithm 1 for θ^n to obtain a sample ε
3: Estimate the gradient $\hat{g}^n$ at $\theta = \theta^n$ (Eq. 7)
4: Calculate the stepsize ρ^n
5: Update $\theta^{n+1} = \theta^n + \rho^n \hat{g}^n$
6: **until convergence**

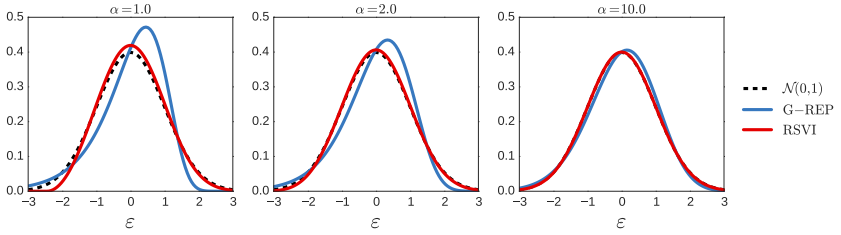
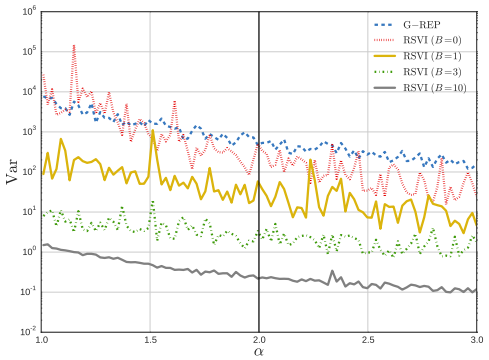


Figure 2: In the distribution on the transformed space ε for a gamma distribution we can see that the rejection sampling-inspired transformation converges faster to a standard normal. Therefore it is less dependent on the parameter α, which implies a smaller correction term. We compare the transformation of RSVI (this paper) with the standardization procedure suggested in Ruiz et al. [2016] (G-REP), for shape parameters $\alpha = \{1, 2, 10\}$.

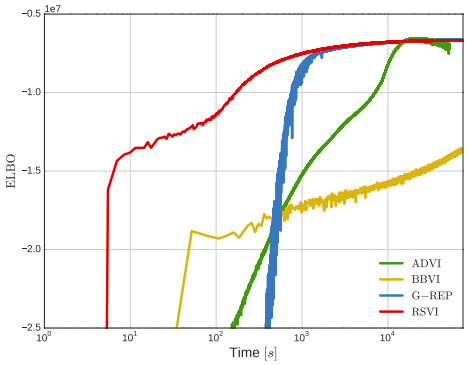
2 Experiments

In Figure 2 we show that the distribution $\pi(\varepsilon\,;\theta)$ for the gamma rejection sampler converges to $s(\varepsilon)$ (a standard normal) as the shape parameter α increases. For large α, $\pi(\varepsilon\,;\theta) \approx s(\varepsilon)$ and the acceptance probability of the rejection sampler approaches 1, which makes the correction term negligible.

In Figure 3a we study the variance of the stochastic gradient for a synthethic example for which we can calculate the true gradient. RSVI achieves significantly lower variance compared to generalized reparameterization (G-REP) [Ruiz et al., 2016]. We also show in Figure 3b that RSVI outperforms other state of the art methods in terms of convergence speed, due to the reduced variance.



(a) RSVI (this paper) achieves lower variance compared to G-REP [Ruiz et al., 2016]. The estimated variance is for a component of Dirichlet approximation to a multinomial likelihood with uniform Dirichlet prior. Optimal concentration (parameter value) is $\alpha = 2$, and B denotes shape augmentation steps (see Naesseth et al. [2016, Section 5] for details).



(b) RSVI (this paper) presents a significantly faster initial improvement of the ELBO as a function of wall-clock time. The model is a sparse gamma DEF [Ranganath et al., 2015], applied to the Olivetti faces dataset, and we compare with ADVI [Kucukelbir et al., 2016], BBVI [Ranganath et al., 2014], and G-REP.

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv:1601.00670*, 2016.

L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, jul 2011.

K. Fan, Z. Wang, J. Beck, J. Kwok, and K. A. Heller. Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, 2015.

Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems*, 2001.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov. 1999.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *arXiv:1603.00788*, 2016.

G. Marsaglia and W. W. Tsang. A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, 26(3):363–372, Sept. 2000.

C. A. Naesseth, F. J. R. Ruiz, S. W. Linderman, and D. M. Blei. Rejection Sampling Variational Inference. *arXiv:1610.05683*, Oct. 2016.

J. W. Paisley, D. M. Blei, and M. I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

R. Ranganath, L. Tang, L. Charlin, and D. M. Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2004.

F. J. R. Ruiz, M. K. Titsias, and D. M. Blei. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, 2016.

T. Salimans and D. A. Knowles. Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):837–882, 2013.

T. Tieleman and G. Hinton. Lecture 6.5-RMSPROP: Divide the gradient by a running average of its recent magnitude. Coursera: Neural Networks for Machine Learning, 4, 2012.

M. K. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.