A. An example with countably infinite K

We give an example to demonstrate our method when there is a countably infinite number of categories. Consider the N-mixture model,

$$N \sim \text{Poisson}(\lambda) \tag{43}$$
$$y_i \sim \text{Binomial}(N, p) \quad \text{for} \quad i = 1, ..., n, \tag{44}$$

a model used in ecological modeling of species counts (Royle, 2004).

In our experiment, we take p and λ to be known parameters. We want to infer N given data $y_1, ..., y_n$. Since the support of N is the integers greater than or equal to $y_{max} := \max_n \{y_n\}$, we use a negative binomial distribution shifted by y_{max} to approximate the posterior. Let $\hat{r}$ and $\hat{p}$ be the number of failures and the probability of success, respectively, for a negative binomial. We optimize the ELBO,

$$\mathcal{L}(\hat{r}, \hat{p}) = E_{q(N;\hat{r},\hat{p})}[\log p(y|N)p(N) - \log q(N; \hat{r}, \hat{p})] \tag{45}$$

This expectation is taken over N, and is given by an infinite sum. The exact expectation is intractable. However, we have a closed form variational distribution, and for any $\hat{r}$ and $\hat{p}$, it is easy to find the integers N where $q(N; \hat{r}, \hat{p})$ places most of its mass. We therefore can apply our Rao-Blackwellization procedure to compute stochastic gradients of the ELBO.

In our experiment, we take the true $N = 10$ and $p = 0.2$. We drew 1000 data points from Equation (44). We set our Poisson prior with $\lambda = 10$.

We found that the REINFORCE estimator was too high variance to be useful in this example, so we start with REINFORCE$^+$. Figure 7 compares the REINFORCE$^+$ estimator with its Rao-Blackwellization, using either $k = 1$ or $k = 3$ categories summed.

We find that our Rao-Blackwellization improves the convergence rate of the ELBO. This is because our variational distribution eventually concentrates around the true N (Figure 8), and only a few categories have significant mass.

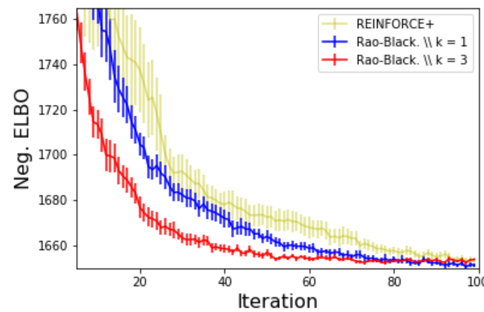

Figure 7. Negative ELBO per iteration in the N-mixture experiment. We compare the REINFORCE$^+$ estimator with its Rao-Blackwellization, using either $k = 1$ or $k = 3$ categories summed. Vertical lines denote standard errors over 10 trials from the same initialization.

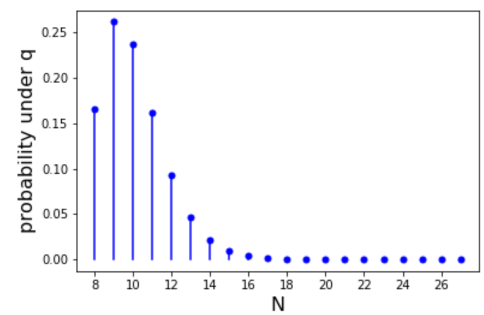

Figure 8. Negative binomial variational distribution q at convergence for the N-mixture experiment.

B. Experimental details

Implementations of all methods in our paper as well as code to reproduce our results can be found in the git repository `https://github.com/Runjing-Liu120/RaoBlackwellizedSGD`.

B.1. Generative semi-supervised classification

In this experiment, our classifier $q_\phi(y|x)$ consists of three fully connected hidden layers, each with 256 nodes and ReLU activations. The inference and generative models, $q_\phi(z|x, y)$ and $p_\theta(x|z, y)$, both have one hidden layer with 128 nodes and ReLU activations, similar to the MLPs used in Kingma et al. (2014). The latent variable z is five dimensional and $q_\phi(z|x)$ is multivariate Gaussian with diagonal covariance.

For all methods, we used performance on a validation set to choose between the possible step-sizes, {5e-5, 1e-4, 5e-4, 1e-3, 5e-3}. For Gumbel-softmax, we also chose the annealing rate among {1e-5, 5e-5, 1e-4, 5e-4}. For RELAX, the relaxation temperature was chosen adaptively using gradients, while the scaling parameter was set at 1.0.

The step-size for REINFORCE was chosen to be 1e-4 and the step-size for RELAX was chosen to be 5e-4. The step-size for the remaining methods were chosen to be 1e-3. The annealing rate for Gumbel-softmax was chosen to be 5e-4.

Optimization was done with Adam (Kingma & Ba, 2015), with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. An initialization for $q_\phi(z|x, y)$ and $p_\theta(x|z, y)$ was obtained by first optimizing $\mathcal{L}^L(x, y)$ on the labeled data only. We also initialized $q_\phi(y|x)$ on the labeled data using cross-entropy loss. The results in the paper show the optimization of the semi-supervised ELBO starting from this initialization.

B.1.1. Conditional generation results

Figure 9 displays the conditional generation of MNIST digits obtained after 100 epochs of running our Rao-Blackwellized gradient method.

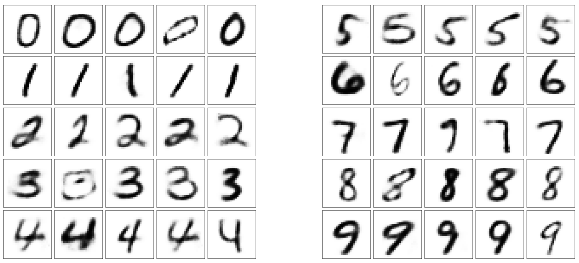

Figure 9. The conditional generation of MNIST digits. Each row displays five draws from the learned generative model $z \sim \mathcal{N}(0, I)$, $x \sim p_\theta(x|y, z)$, for a different digit y in each row.

B.2. Moving MNIST

For the decoder $p(x|l, z)$ we use one fully connected hidden layer with 256 nodes and tanh activations, similar to the architecture described in Kingma and Welling (2014). Our z is 5 dimensional.

The attention mechanism $q(l|x)$ contains four convolutional layers, each with 7 output channels and ReLU activations; the final layer is a fully connected layer with a softmax. The encoder network $q(z|x)$ has one fully connected hidden layer with 256 nodes and tanh activations, mirroring the decoder network.

We again used performance on the validation set to choose between the possible step-sizes and model parameters as described in the section above. The learning rate and annealing rate for Gumbel-sofmax was chosen to be 5e-5 and 5e-4, respectively. For RELAX, the learning rate was 5e-4. The step-sizes for the remaining procedures were chosen to be 1e-3. We again use Adam (Kingma & Ba, 2015) for optimization, and we set $\beta_1 = 0.9$, $\beta_2 = 0.999$.

B.2.1. VAE reconstruction

Figure 10 displays (1) the original non-centered MNIST digit; (2) the reconstruction of the MNIST digits after passing through our attention mechanism and VAE; and (3) the learned pixel locations.

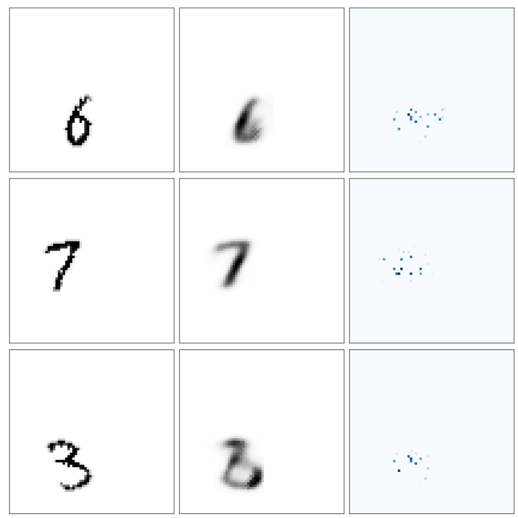

Figure 10. (Left column) The original MNIST digit. (Center column) The reconstructed MNIST digit. (Right column) The learned probability distribution over the grid of pixels. Brighter spots indicate higher probabilities.