Rao-Blackwellized Stochastic Gradients for Discrete Distributions

Runjing (Bryan) Liu

June 11, 2019

University of California, Berkeley

Objective

- We fit a **discrete latent variable model**.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}} \; \mathbb{E}_{q_\eta(z)}\left[f_\eta(z)\right]$$

where z is a discrete random variable with K categories.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}}\ \mathbb{E}_{q_\eta(z)}\left[f_\eta(z)\right]$$

where z is a discrete random variable with K categories.

- Two common approaches are :

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}}\ \mathbb{E}_{q_\eta(z)}\left[f_\eta(z)\right]$$

  where z is a discrete random variable with K categories.
- Two common approaches are :
  1. Analytically integrate out z.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}} \ \mathbb{E}_{q_\eta(z)} \left[f_\eta(z) \right]$$

  where z is a discrete random variable with K categories.
- Two common approaches are :
  1. Analytically integrate out z.
     Problem: K might be large.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}} \; \mathbb{E}_{q_\eta(z)} \left[f_\eta(z) \right]$$

  where z is a discrete random variable with K categories.
- Two common approaches are :
  1. Analytically integrate out z.
     Problem: K might be large.
  2. Sample $z \sim q_\eta(z)$, and estimate the gradient with $g(z)$.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\text{argmin}}\ \mathbb{E}_{q_\eta(z)}\left[f_\eta(z)\right]$$

  where z is a discrete random variable with K categories.
- Two common approaches are :
  1. Analytically integrate out z.
     Problem: K might be large.
  2. Sample $z \sim q_\eta(z)$, and estimate the gradient with $g(z)$.
     Problem: $g(z)$ might have high variance.

Objective

- We fit a **discrete latent variable model**.
- Fitting such a model involves finding

$$\underset{\eta}{\operatorname{argmin}} \; \mathbb{E}_{q_\eta(z)}\left[f_\eta(z)\right]$$

  where z is a discrete random variable with K categories.
- Two common approaches are :
  1. Analytically integrate out z.
     Problem: K might be large.
  2. Sample $z \sim q_\eta(z)$, and estimate the gradient with $g(z)$.
     Problem: $g(z)$ might have high variance.

We propose a method that uses a combination of these two approaches to reduce the variance of any gradient estimator $g(z)$.

Our method

Suppose g is an unbiased estimate of the gradient, so

$$\nabla_\eta \mathcal{L}(\eta) = \mathbb{E}_{q_\eta(z)}[g(z)] = \sum_{k=1}^{K} q_\eta(k) g(k)$$

Our method

Suppose g is an unbiased estimate of the gradient, so

$$\nabla_\eta \mathcal{L}(\eta) = \mathbb{E}_{q_\eta(z)}[g(z)] = \sum_{k=1}^{K} q_\eta(k) g(k)$$

Key observation: In many applications (e.g. variational Bayes), $q_\eta(z)$ is concentrated on only a few categories.

Our method

Suppose g is an unbiased estimate of the gradient, so

$$\nabla_\eta \mathcal{L}(\eta) = \mathbb{E}_{q_\eta(z)}[g(z)] = \sum_{k=1}^{K} q_\eta(k) g(k)$$

Key observation: In many applications (e.g. variational Bayes), $q_\eta(z)$ is concentrated on only a few categories.

Our idea: Let us analytically sum categories where $q_\eta(z)$ has high probability, and sample the remaining terms.

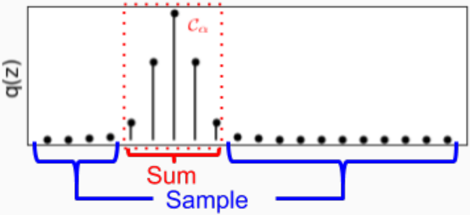
Our method

Suppose g is an unbiased estimate of the gradient, so

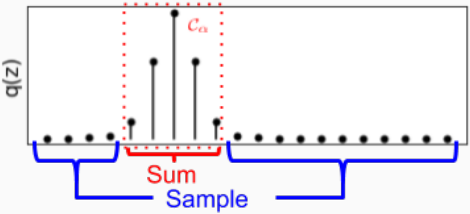
$$\nabla_\eta \mathcal{L}(\eta) = \mathbb{E}_{q_\eta(z)}[g(z)] = \sum_{k=1}^{K} q_\eta(k) g(k)$$

Key observation: In many applications (e.g. variational Bayes), $q_\eta(z)$ is concentrated on only a few categories.

Our idea: Let us analytically sum categories where $q_\eta(z)$ has high probability, and sample the remaining terms.



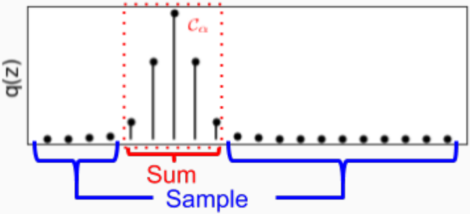
Our method



In math,

$$\sum_{k=1}^{K} q_\eta(k)g(k) = \underbrace{\sum_{z \in \mathcal{C}_\alpha} q_\eta(z)g(z)}_{\text{analytically sum}} + \underbrace{(1 - q_\eta(\mathcal{C}_\alpha))}_{\text{small}} \underbrace{\mathbb{E}_{q_\eta(z)}[g(z) | z \notin \mathcal{C}_\alpha]}_{\text{estimate by sampling}}$$

Our method



In math,

$$\sum_{k=1}^{K} q_\eta(k) g(k) = \underbrace{\sum_{z \in \mathcal{C}_\alpha} q_\eta(z) g(z)}_{\text{analytically sum}} + \underbrace{(1 - q_\eta(\mathcal{C}_\alpha))}_{\text{small}} \underbrace{\mathbb{E}_{q_\eta(z)}[g(z) | z \notin \mathcal{C}_\alpha]}_{\text{estimate by sampling}}$$

The variance reduction is guaranteed by representing our estimator as an instance of **Rao-Blackwellization**.

Results: Generative semi-supervised classification

We train a **classifier** to classify the class label of MNIST digits and learn a **generative model** for MNIST digits conditional on the class label.

Results: Generative semi-supervised classification

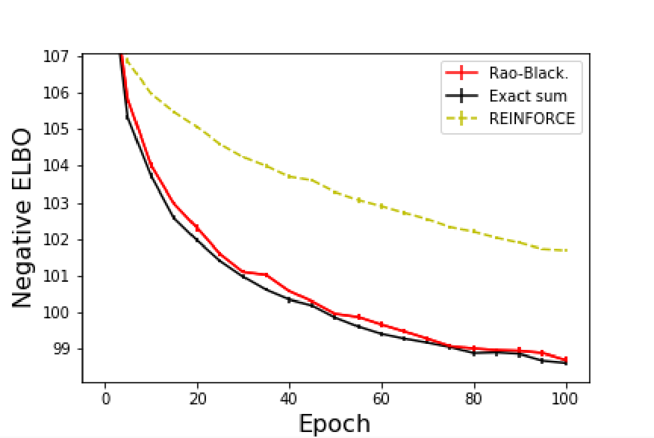
We train a **classifier** to classify the class label of MNIST digits and learn a **generative model** for MNIST digits conditional on the class label.

Our objective is to maximize the evidence lower bound (ELBO),

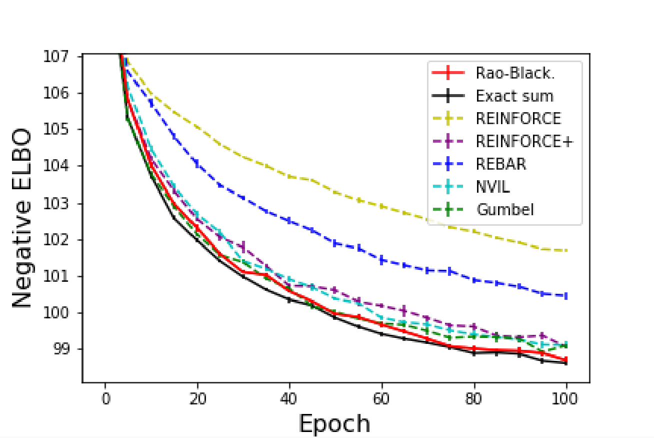
$$p_\eta(x) \geq \mathbb{E}_{q_\eta(z)}[\log p_\eta(x, z) - \log q_\eta(z)]$$

In this problem, the class label z has ten discrete categories.

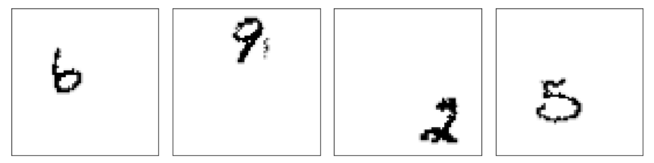
Results: Generative semi-supervised classification



Results: Generative semi-supervised classification

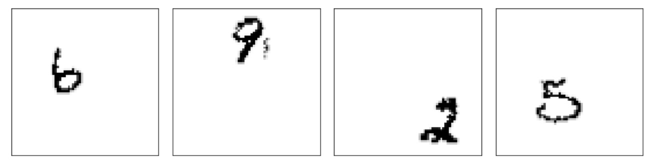


Results: moving MNIST



We train a generative model for non-centered MNIST digits.

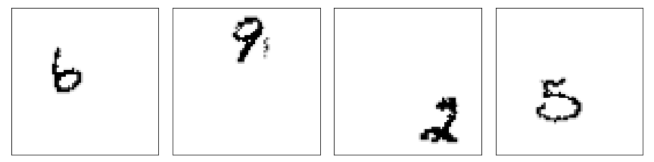
Results: moving MNIST



We train a generative model for non-centered MNIST digits.

To do so, we must first learn the location of the MNIST digit. There are 68×68 discrete categories.

Results: moving MNIST

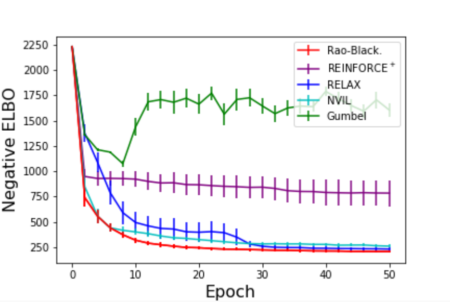


We train a generative model for non-centered MNIST digits.

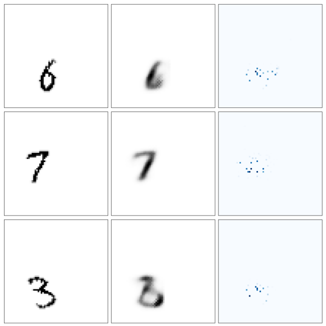
To do so, we must first learn the location of the MNIST digit. There are 68×68 discrete categories.

Thus, computing the exact sum is intractable!

Results: moving MNIST



Trajectory of the negative ELBO



Reconstruction of MNIST digits

Our paper:

Rao-Blackwellized Stochastic Gradients for Discrete Distributions

https://arxiv.org/abs/1810.04777

Our code:

https://github.com/Runjing-Liu120/RaoBlackwellizedSGD

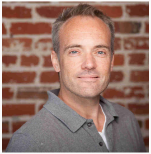
The collaboration:

Bryan Liu

Jeffrey Regier

Nilesh Tripuraneni

Michael I. Jordan



Jon McAuliffe