# Marginalized Stochastic Natural Gradients for Black-Box Variational Inference

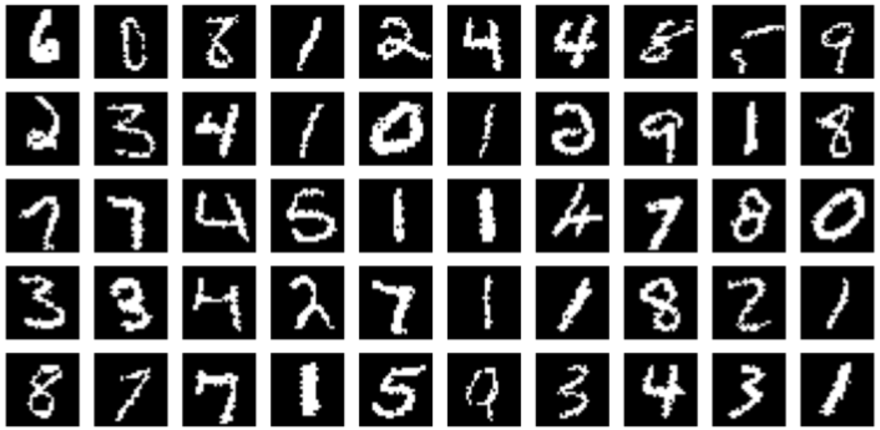
Geng Ji, Debora Sujono, Erik Sudderth

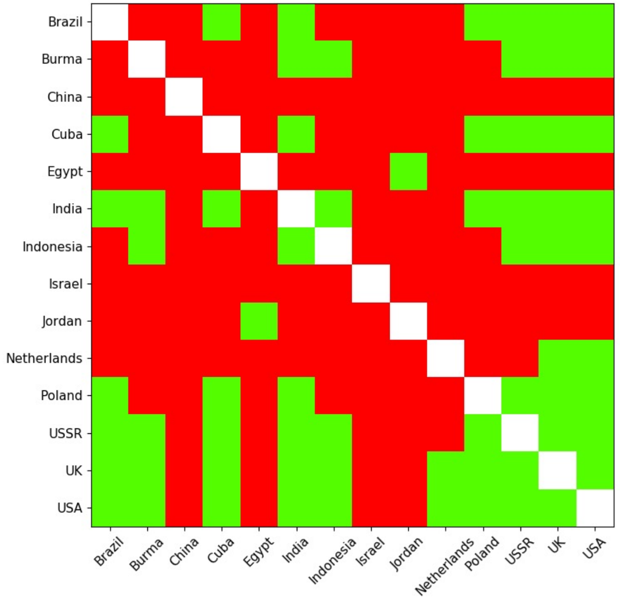
ICML | 2021
Thirty-eighth International Conference on Machine Learning

# Introduction

**We propose a black-box variational inference algorithm for discrete-variable models**

- Black-box: Integration with probabilistic programming languages (PPL), no manual derivations
- Discrete models: Not reparameterizable, existing methods are either biased or have high variance
- Examples: Binary belief networks for image data (e.g. MNIST digits) or text documents, relational models for network data (e.g. country relations, NeurIPS co-authors)





# Variational Inference (VI)

- **Given a model with discrete latent variables $\boldsymbol{z}$ and observations $\boldsymbol{x}$, VI seeks approximate posterior $\boldsymbol{q(z)}$ by maximizing the *evidence lower bound* (ELBO):**

$$\mathcal{L}(x;\, q) = \mathbb{E}_{q(z)}[\log p(z, x) - \log q(z)]$$

Assume mean-field variational distribution: $q(z) = \prod_i q(z_i)$

# Variational Inference (VI)

- **Given a model with discrete latent variables $z$ and observations $x$, VI seeks approximate posterior $q(z)$ by maximizing the *evidence lower bound* (ELBO):**

$$\mathcal{L}(x;\ q) = \mathbb{E}_{q(z)}[\log p(z,x) - \log q(z)]$$

Assume mean-field variational distribution: $q(z) = \prod_i q(z_i)$

- **Suppose $z_i$ is binary and $q(z_i)$ is Bernoulli parameterized by logit:**

$$\tau_i = \log \frac{q(z_i = 1)}{q(z_i = 0)}$$

Generalization to categorical variables in paper

# Limitations of Existing VI Algorithms

- **Coordinate ascent variational inference (CAVI)**

$$\tau_i = \mathbb{E}_{q(z_{-i})}\left[\log \frac{p(z_i = 1 | z_{-i}, x)}{p(z_i = 0 | z_{-i}, x)}\right]$$

  - Intractable for complex discrete models
  - Guaranteed to converge only when updated **sequentially**
  - Auxiliary-variable methods: looser bounds (local optima), handcrafted derivations (not black-box)

# Limitations of Existing VI Algorithms

- **Coordinate ascent variational inference (CAVI)**

$$\tau_i = \mathbb{E}_{q(z_{-i})}\left[\log \frac{p(z_i = 1 | z_{-i}, x)}{p(z_i = 0 | z_{-i}, x)}\right]$$

  - Intractable for complex discrete models
  - Guaranteed to converge only when updated **sequentially**
  - Auxiliary-variable methods: looser bounds (local optima), handcrafted derivations (not black-box)

- **Score-function estimator (REINFORCE)**

$$\frac{\partial \mathcal{L}}{\partial \tau_i} \approx \frac{1}{M} \sum_{m=1}^{M} \left. \frac{\partial \log q(z_i)}{\partial \tau_i} \right|_{z_i^{(m)}} \cdot \left( \log p\left(z_i^{(m)} \middle| z_{-i}^{(m)}, x\right) - \log q(z_i^{(m)}) \right)$$

  - Estimates have **high variance**

# Limitations of Existing VI Algorithms

- **Coordinate ascent variational inference (CAVI)**

$$\tau_i = \mathbb{E}_{q(z_{-i})}\left[\log \frac{p(z_i = 1|z_{-i}, x)}{p(z_i = 0|z_{-i}, x)}\right]$$

  - Intractable for complex discrete models
  - Guaranteed to converge only when updated **sequentially**
  - Auxiliary-variable methods: looser bounds (local optima), handcrafted derivations (not black-box)

- **Score-function estimator (REINFORCE)**

$$\frac{\partial \mathcal{L}}{\partial \tau_i} \approx \frac{1}{M}\sum_{m=1}^{M} \left.\frac{\partial \log q(z_i)}{\partial \tau_i}\right|_{z_i^{(m)}} \cdot \left(\log p\left(z_i^{(m)}\middle|z_{-i}^{(m)}, x\right) - \log q(z_i^{(m)})\right)$$

  - Estimates have **high variance**

- **Gumbel-softmax relaxations (CONCRETE)**
  - Estimates are **biased**
  - Require gradients of model log-probability

# Marginalized Stochastic Natural Gradients (MSNG)

REINFORCE

$$\frac{\partial \mathcal{L}}{\partial \tau_i} = \mathbb{E}_{q(z)}\left[\frac{\partial \log q(z_i)}{\partial \tau_i}(\log p(z_i|z_{-i}, x) - \log q(z_i))\right]$$

✓ Unbiased ✓ Black-box ✓ No gradients of model log-probability

# Marginalized Stochastic Natural Gradients (MSNG)

REINFORCE → Natural Gradient

$$F^{-1}(\tau_i)\frac{\partial \mathcal{L}}{\partial \tau_i} = \cancel{F^{-1}(\tau_i)\frac{\partial \mu_i}{\partial \tau_i}}\frac{\partial \mathcal{L}}{\partial \mu_i} = \mathbb{E}_{q(z)}\left[\frac{\partial \log q(z_i)}{\partial \mu_i}(\log p(z_i|z_{-i},x) - \log q(z_i))\right]$$

✓ Unbiased ✓ Black-box ✓ No gradients of model log-probability

# Marginalized Stochastic Natural Gradients (MSNG)

REINFORCE → Natural Gradient → Marginalization

$$F^{-1}(\tau_i)\frac{\partial \mathcal{L}}{\partial \tau_i} = \cancel{F^{-1}(\tau_i)}\frac{\partial \cancel{\mu_i}}{\cancel{\partial}\tau_i}\frac{\partial \mathcal{L}}{\partial \mu_i} = \mathbb{E}_{q(z)}\left[\frac{\partial \log q(z_i)}{\partial \mu_i}(\log p(z_i|z_{-i},x) - \log q(z_i))\right]$$

$$= \sum_{z_i} q(z_i)\frac{\partial \log q(z_i)}{\partial \mu_i}\mathbb{E}_{q(z_{-i})}[\log p(z_i|z_{-i},x) - \log q(z_i)]$$

✓ Unbiased ✓ Black-box ✓ No gradients of model log-probability

✓ Low variance

# Marginalized Stochastic Natural Gradients (MSNG)

REINFORCE → Natural Gradient → Marginalization

Weighted average

$$\tau_i^{\text{new}} = \alpha \frac{1}{M} \sum_{m=1}^{M} \log \frac{p(z_i = 1 | z_{-i}^{(m)}, x)}{p(z_i = 0 | z_{-i}^{(m)}, x)} + (1 - \alpha)\tau_i$$

Monte Carlo approximation of CAVI update

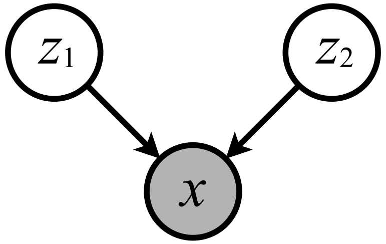
✓ Unbiased ✓ Black-box ✓ No gradients of model log-probability

✓ Low variance ✓ Parallelizable

# Experiments: Toy Data

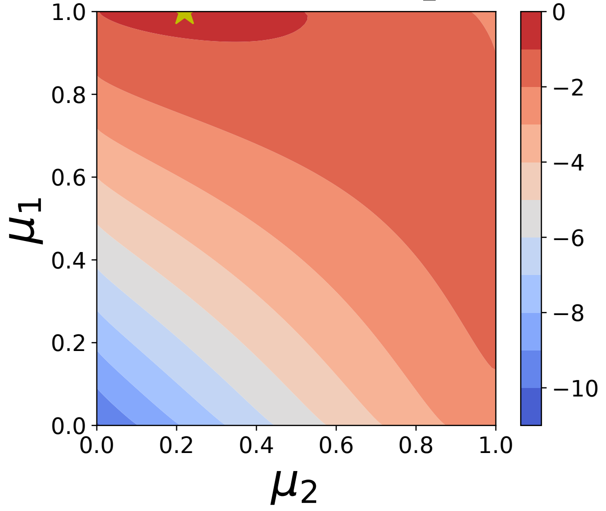
Noisy-OR model

$p(z_1) = 0.5 \quad p(z_2) = 0.2$




$$p(x = 1 \mid z_1, z_2)$$
$$= 1 - 0.9999 \cdot 0.1^{z_1} \cdot 0.1^{z_2}$$

ELBO contour plot

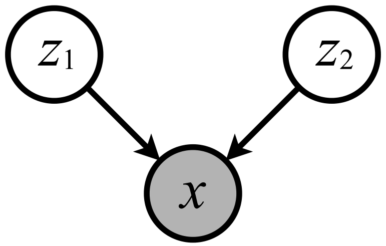



- Yellow star indicates global maximum

# Experiments: Toy Data

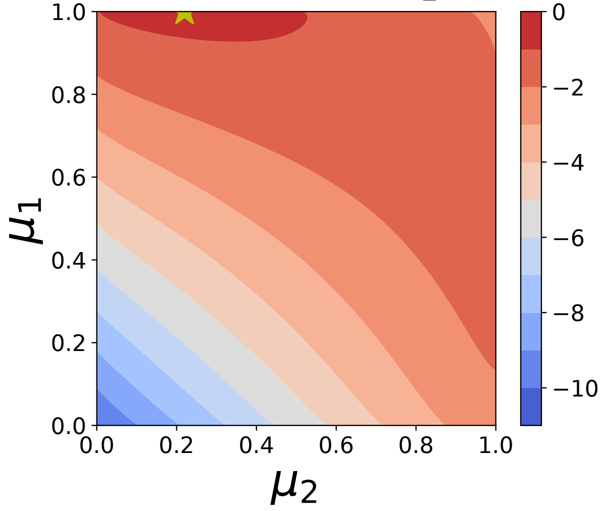
Noisy-OR model

$p(z_1) = 0.5 \quad p(z_2) = 0.2$

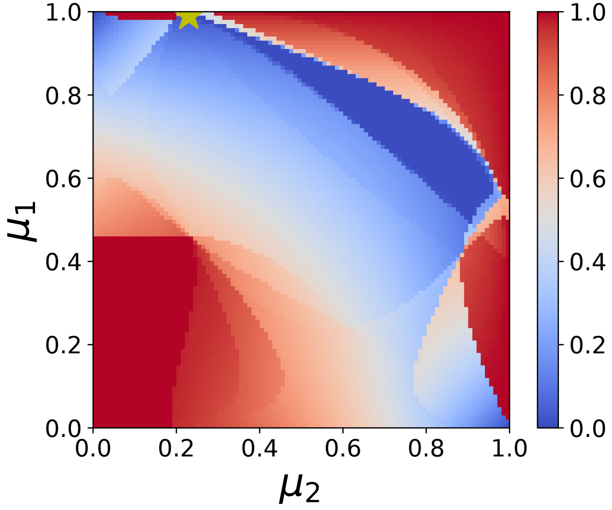



$$p(x = 1 \mid z_1, z_2) = 1 - 0.9999 \cdot 0.1^{z_1} \cdot 0.1^{z_2}$$

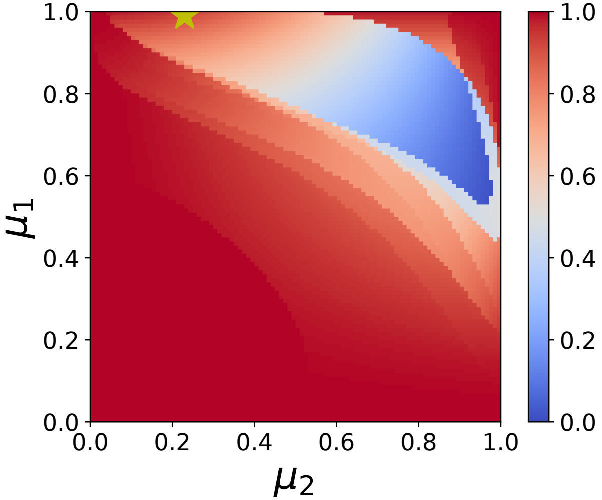
ELBO contour plot




REINFORCE

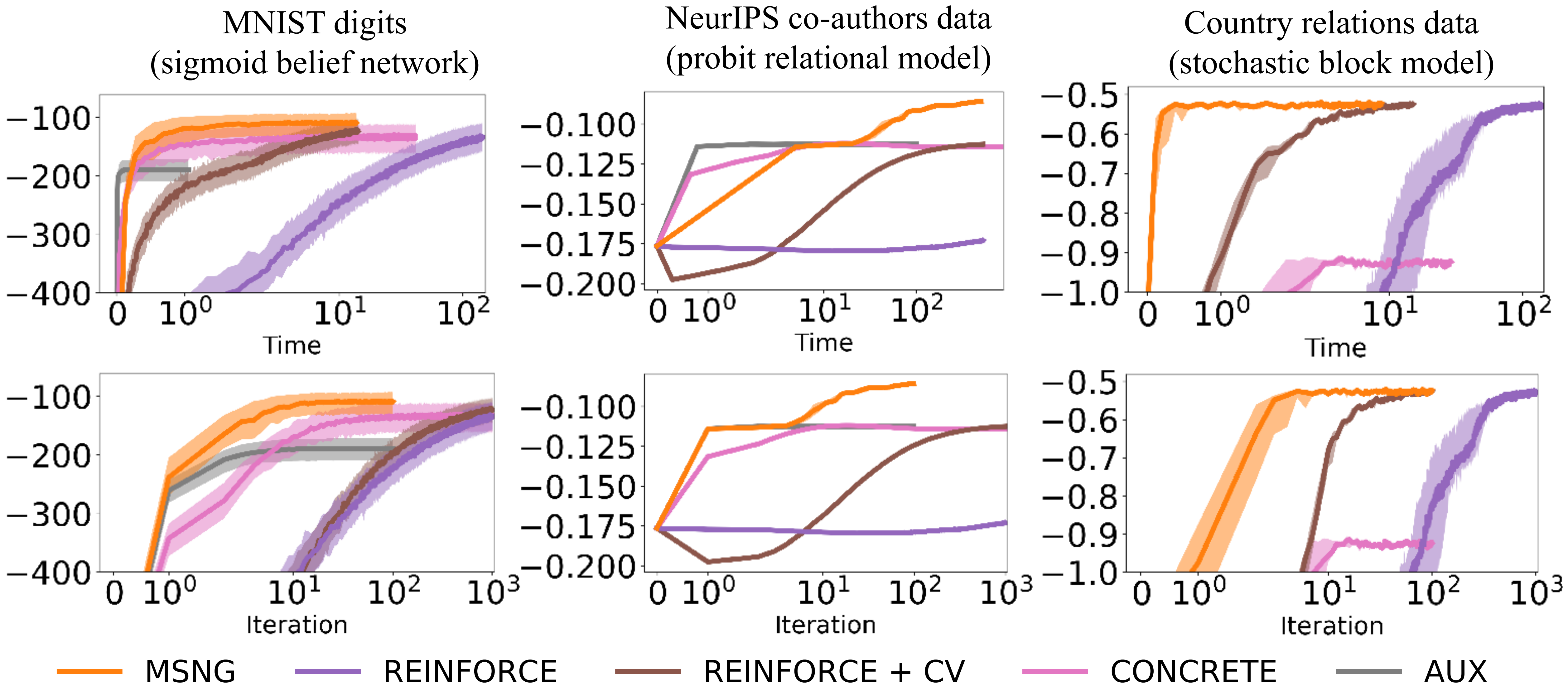



MSNG




- Yellow star indicates global maximum
- Red area: ELBO likely to increase
  Blue area: ELBO likely to decrease
- REINFORCE is more likely to *decrease* ELBO in blue region near maximum, and shows sensitivity to step-size
- MSNG has promising red area near the yellow star to "attract" variational parameters toward global maximum

# Experiments: Image and Network Data



# Summary

Classic CAVI updates are not tractable for general discrete models.
Our **Marginalized Stochastic Natural Gradients (MSNG)** have attractive properties:

| | CAVI (AUX) | REINFORCE | CONCRETE | MSNG |
|---|---|---|---|---|
| Unbiased optimization of discrete ELBO | ✗ | ✓ | ✗ | ✓ |
| Black-box variational inference | ✗ | ✓ | ✓ | ✓ |
| No model log-likelihood derivatives | ✓ | ✓ | ✗ | ✓ |
| Low variance | ✓ | ✗ | ✓ | ✓ |
| Parallelizable | ✗ | ✓ | ✓ | ✓ |

**See our ICML 2021 paper:** Categorical variable updates, integration with Pyro PPL, additional stochastic VI baselines, applications to models of text topics & crowd-sourcing