Stochastic gradient estimation for discrete latent variables

Rakitin Denis
Higher School of Economics
Bayesian Methods Research Group

Gradient estimation

$$L(\phi) = \mathbb{E}_{q_\phi(z)} f(z)$$

Relaxation-based methods:

- Gumbel-Softmax / Stochastic Softmax Tricks
- REBAR

Score-function methods:

- REINFORCE (+ baselines)
- REBAR/RELAX

AR estimator

Dirichlet reparameterization of $b \sim Cat(\sigma(\theta))$:

$$\pi \sim Dir(1_C) \text{ and } b = \arg\min \pi_i e^{-\theta_i}$$

Augmented functional:

$$\mathbb{E}_{b \sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_{\pi \sim Dir(1_C)} f(\arg\min \pi_i e^{-\theta_i})$$

Augmented true gradient:

$$\nabla_{\theta_l} \mathbb{E}_{b \sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_{\pi} f(b)(1 - C\pi_l) \approx f(b)(1 - C\pi_l)$$

$$\pi \sim Dir(1_C), \quad b = \arg\min \pi_i e^{-\theta_i}$$

https://arxiv.org/abs/1905.01413

AR estimator

Dirichlet reparameterization of $b \sim Cat(\sigma(\theta))$:

$$\pi \sim Dir(1_C) \text{ and } b = \arg\min \pi_i e^{-\theta_i}$$

Augmented functional:

$$\mathbb{E}_{b \sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_{\pi \sim Dir(1_C)} f(\arg\min \pi_i e^{-\theta_i})$$

Augmented true gradient:

$$\nabla_{\theta_l} \mathbb{E}_{b \sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_\pi f(b)(1 - C\pi_l) \approx f(b) \underbrace{\boxed{(1 - C\pi_l)}}_{\text{Score term}}$$

$$\pi \sim Dir(1_C), \quad b = \arg\min \pi_i e^{-\theta_i}$$

AR: what happened

Exponential racing:

$$\tau_1, \ldots, \tau_C \sim Exp(e^{\theta_1}), \ldots Exp(e^{\theta_C})$$

$$b = \arg\min \tau_i \sim Cat(\sigma(\theta))$$

$$b \to \arg\min \tau_i \to \arg\min \varepsilon_i \, e^{-\theta_i}$$

$$\to \arg\min \xi \, \pi_i \, e^{-\theta_i} \to \arg\min \pi_i \, e^{-\theta_i}$$

Exponential reparameterization:

$\varepsilon \sim Exp(1)$ then $(1/\lambda)\,\varepsilon \sim Exp(\lambda)$

Dirichlet vs Gamma:

$(\varepsilon_1, \ldots, \varepsilon_C) \overset{d}{=} \pi \cdot \xi$, where

$\pi \sim Dir(1_C)$ and $\xi \sim \Gamma(C, 1)$

AR: symmetries

Notation:

$$\pi^{j \rightleftharpoons l} = (\pi_1, \ldots, \underset{j}{\pi_l}, \ldots, \underset{l}{\pi_j}, \ldots, \pi_C)$$

$$b = \arg\min_i \pi_i e^{-\theta_i}, \quad b^{j \rightleftharpoons l} = \arg\min_i \pi_i^{j \rightleftharpoons l} e^{-\theta_i}$$

Reference index j
Differentiate w.r.t. l

Dirichlet symmetry:

$\pi \sim Dir(1_C)$ then $\pi^{j \rightleftharpoons l} \sim Dir(1_C)$

Rewrite the gradient:

$$\nabla_{\theta_l} \mathbb{E}_{b \sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_\pi f(b)(1 - C\pi_l) = \mathbb{E}_\pi f(b^{j \rightleftharpoons l})(1 - C\pi_j)$$

AR: deriving a baseline

$$\nabla_{\theta_l}\mathbb{E}_{b\sim Cat(\sigma(\theta))}f(b)=\mathbb{E}_{\pi}f(b)(1-C\pi_l)=\mathbb{E}_{\pi}f(b^{j\rightleftharpoons l})(1-C\pi_j)$$

Let's first write:

$$0=\frac{1}{C}\sum_{m=1}^{C}(1-C\pi_m)=\frac{1}{C}\sum_{m=1}^{C}f(b)(1-C\pi_m)=$$

$$=\frac{1}{C}\mathbb{E}_{\pi}\sum_{m=1}^{C}f(b)(1-C\pi_m)=\frac{1}{C}\mathbb{E}_{\pi}\sum_{m=1}^{C}f(b^{j\rightleftharpoons m})(1-C\pi_j)$$

Reference index j
Differentiate w.r.t. l

ARS estimator follows:

$$\mathbb{E}_{\pi}f(b^{j\rightleftharpoons l})(1-C\pi_j)=\mathbb{E}_{\pi}\left(f(b^{j\rightleftharpoons l})-\frac{1}{C}\sum_{m=1}^{C}f(b^{j\rightleftharpoons m})\right)(1-C\pi_j)$$

ARS and ARSM

ARS = AR + symmetries + baseline

Number of function evaluations

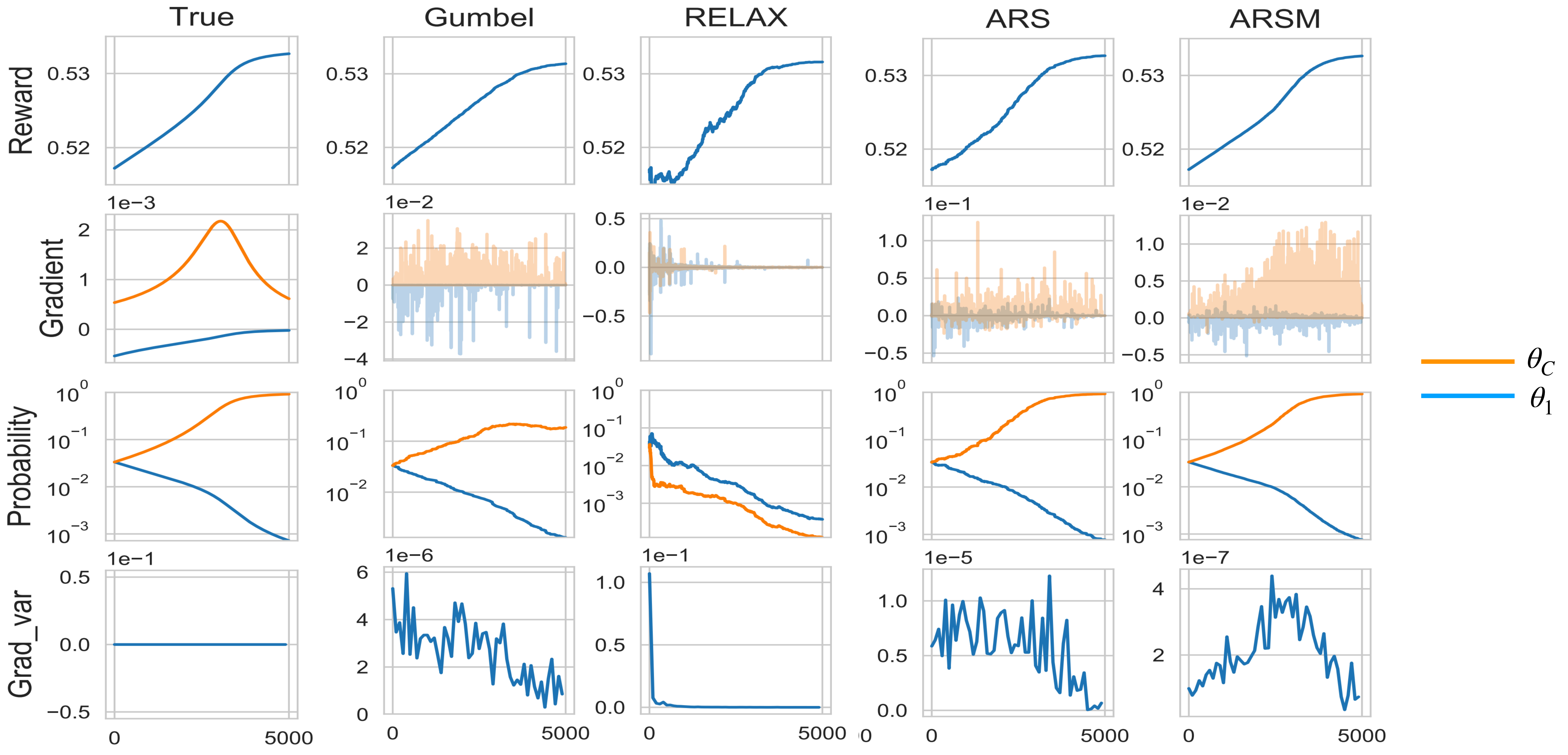
$$\nabla_{\theta_l} L(\theta) = \mathbb{E}_\pi \underbrace{\left(f(b^{j \rightleftharpoons l}) - \frac{1}{C} \sum_{m=1}^{C} f(b^{j \rightleftharpoons m}) \right)(1 - C\pi_j)}_{g_{ARS,l}} \qquad \leq C$$

ARSM = ARS + mean over reference indices

$$\nabla_{\theta_l} L(\theta) = \mathbb{E}_\pi \underbrace{\left[\frac{1}{C} \sum_{j=1}^{C} \left(f(b^{j \rightleftharpoons l}) - \frac{1}{C} \sum_{m=1}^{C} f(b^{j \rightleftharpoons m}) \right)(1 - C\pi_j) \right]}_{g_{ARSM,l}} \qquad \leq C(C-1)/2 + 1$$

Toy experiment

$$L(\theta) = \mathbb{E}_{b\sim Cat(\sigma(\theta))} f(b) = \mathbb{E}_{b\sim Cat(\sigma(\theta))} 0.5 + b/(CR)$$



VAE on binarized MNIST

ARM+ and ARSM+

Ideal case - Rao-Blackwellization

$$\nabla_{\theta_l} L(\theta) = \mathbb{E}_\pi \left(f(b^{j \rightleftharpoons l}) - \frac{1}{C} \sum_{m=1}^{C} f(b^{j \rightleftharpoons m}) \right)(1 - C\pi_j) =$$

$$= \mathbb{E}_{b^{j \rightleftharpoons 1}, \ldots, b^{j \rightleftharpoons C}} \mathbb{E}_{\pi | b^{j \rightleftharpoons 1}, \ldots, b^{j \rightleftharpoons C}} \left(f(b^{j \rightleftharpoons l}) - \frac{1}{C} \sum_{m=1}^{C} f(b^{j \rightleftharpoons m}) \right)(1 - C\pi_j)$$

Reference index j
Differentiate w.r.t. l

We only need to compute:

$$\mathbb{E}_{\pi | b^{j \rightleftharpoons 1}, \ldots, b^{j \rightleftharpoons C}}[\pi_j] = \mathbb{E}_{\pi_{-(out,\, j)} | b^{j \rightleftharpoons 1}, \ldots, b^{j \rightleftharpoons C}} \boxed{\mathbb{E}_{\pi_j | \pi_{-(out,j)},\, b^{j \rightleftharpoons 1}, \ldots, b^{j \rightleftharpoons C}}[\pi_j]}$$

Compute analytically

https://openreview.net/pdf?id=lcs9Dg0F0g5

Partial integrating

$$\mathbb{E}_{\pi_j|\pi_{-(out,j)},\, b^{j\rightleftharpoons 1},\ldots,b^{j\rightleftharpoons C}}[\pi_j]$$

Observe that $b^{j\rightleftharpoons m}$ imply bounds on π_j :

$$b^{j\rightleftharpoons m} = k \text{ means } \arg\min_i \pi_i^{j\rightleftharpoons m} e^{-\theta_i} = k$$

$$\pi_i^{j\rightleftharpoons m} e^{-\theta_i} \geq \pi_k^{j\rightleftharpoons m} e^{-\theta_k} \iff \pi_i^{j\rightleftharpoons m} e^{\theta_k - \theta_i} \geq \pi_k^{j\rightleftharpoons m}$$

Dirichlet is uniform on the simplex, then conditional of π_j is also uniform

Leveraging symmetry

Multidimensional case:

$$f(b^{j\rightleftharpoons m}) = f(b_1^{j_1\rightleftharpoons m}, \ldots, b_K^{j_1\rightleftharpoons m})$$

Introduce

$$\delta_k = I\{b_k^{j\rightleftharpoons 1} = \ldots = b_k^{j\rightleftharpoons C}\}$$

$$\mathbb{E}_{\pi|\delta_k=1}\left(f(b^{j\rightleftharpoons l}) - \frac{1}{C}\sum_{m=1}^{C} f(b^{j\rightleftharpoons m})\right)(1 - C\pi_{k,j}) =$$

$$= \mathbb{E}_{\pi_k|\delta_k=1}\left(\mathbb{E}_{\pi_{-k}} f(b^{j\rightleftharpoons l}) - \frac{1}{C}\sum_{m=1}^{C} \mathbb{E}_{\pi_{-k}} f(b^{j\rightleftharpoons m})\right) = 0$$

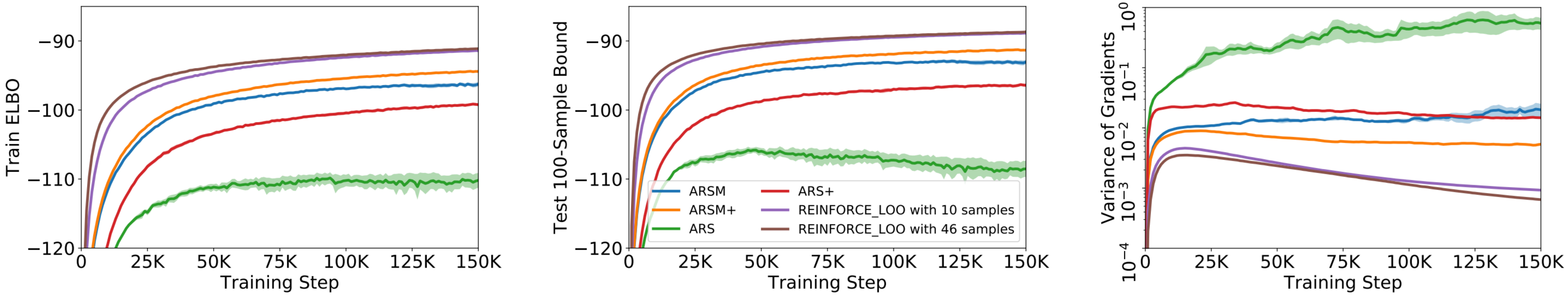
Final estimators

$$g_{ARS+,k,l} = \underbrace{\mathbb{E}_{\pi_{k,j}|\pi_{k,-(1,j)},\, b_k^{1\rightleftharpoons j},\ldots,b_k^{C\rightleftharpoons j}}}_{\text{partial integration}}\, g_{ARS,k,l}\underbrace{(1-\delta_k)}_{\text{symmetry}}$$

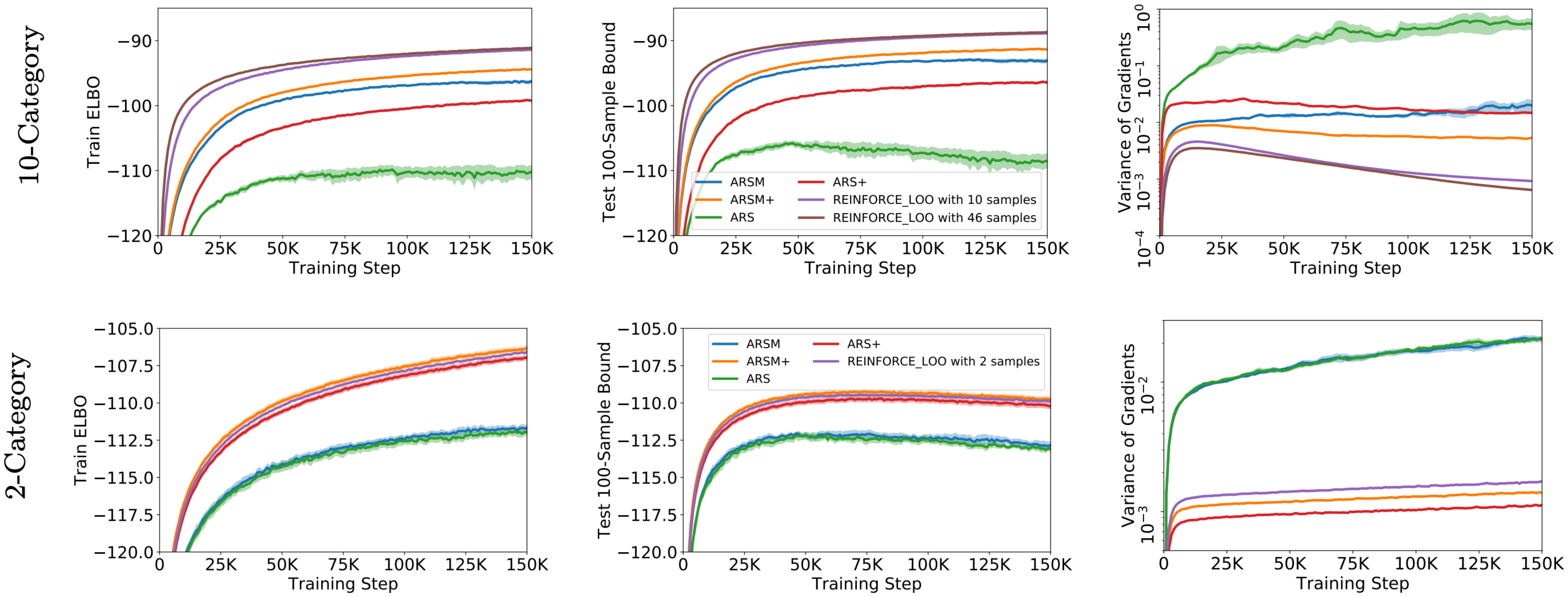
ARSM is modified only by symmetry argument

Experiments

10-Category



Experiments



REINFORCE+ (LOO)

$$L(\phi) = \mathbb{E}_{q_\phi(z)} f(z)$$

$$g_{RF} = f(z_i)\,\nabla_\phi \log q_\phi(z_i)$$

$$g_{RF+} = \frac{1}{S}\left(\sum_{i=1}^{S}\left(f(z_i) - \underbrace{\sum_{j\neq i} f(z_j)}_{\text{loo baseline}}\right)\nabla_\phi \log q_\phi(z_i)\right) = \frac{1}{S-1}\left(\sum_{i=1}^{S}\left(f(z_i) - \bar{f}\right)\nabla_\phi \log q_\phi(z_i)\right)$$

https://openreview.net/pdf?id=r1lgTGL5DE

Log-variance loss

$$L(\phi) = ELBO(\phi) = \mathbb{E}_{q_\phi(z)} \log \frac{p(x,z)}{q_\phi(z)}$$

Log-variance loss:

$$\mathcal{L}_r(q_\phi(z) || p(z|x)) = \frac{1}{2} Var_r \left(\log \frac{q_\phi(z)}{p(z|x)} \right)$$

Gradients property:

$$\nabla_\phi \mathcal{L}_r(q_\phi(z) || p(z|x)) \Big|_{r=q_\phi} = \nabla_\phi KL(q_\phi(z) || p(z|x))$$

https://arxiv.org/abs/2010.10436

Derivation of VarGrad

$$\mathcal{L}_r(q_\phi(z)||p(z|x)) = \frac{1}{2} Var_r\left(\log \frac{q_\phi(z)}{p(z|x)}\right)$$

Sample variance estimation:

$$\mathcal{L}_r(q_\phi(z) \,||\, p(z\,|\,x)) \approx \frac{1}{2(S-1)} \sum_{s=1}^{S} \left(f_\phi(z^{(s)}) - \bar{f}_\phi\right)^2, \quad z^{(s)} \overset{\text{i.i.d.}}{\sim} r(z)$$

differentiate + grad property

$$\widehat{g}_{\text{VarGrad}}(\phi) = \frac{1}{S-1}\left(\sum_{s=1}^{S} f_\phi(z^{(s)}) \nabla_\phi \log q_\phi(z^{(s)}) - \bar{f}_\phi \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)})\right)$$

Unbiased estimate of KL => ELBO

Properties

Another way to introduce RF+:

$$\widehat{g}_{\text{CV}}(\phi) = \widehat{g}_{\text{Reinforce}}(\phi) - a \odot \left(\frac{1}{S} \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)}) \right) \qquad a = \bar{f}_\phi \mathbf{1}$$

Variance minimizing parameter:

$$a_i^* = \frac{\operatorname{Cov}_{q_\phi}\left(f_\phi \partial_{\phi_i} \log q_\phi, \partial_{\phi_i} \log q_\phi\right)}{\operatorname{Var}_{q_\phi}\left(\partial_{\phi_i} \log q_\phi\right)}$$

Lemma about mean of parameter in VarGrad:

$$a^* = \mathbb{E}_{q_\phi}[a^{VarGrad}] + \delta^{CV} = -\text{ELBO}(\phi) + \delta^{CV} \qquad \delta_i^{CV} = \frac{\operatorname{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_i} \log q_\phi)^2\right)}{\operatorname{Var}_{q_\phi}\left(\partial_{\phi_i} \log q_\phi\right)}$$

Properties

Main lemma: correction term is small

Suppose $\sup_z \frac{q_\phi(z)}{p(z\,|\,x)} < C$ and define $\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi] = \frac{\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^4]}{(\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^2])^2}$

Then

$$\left| \frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} \right| \leq \frac{2\sqrt{C\,\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi]}}{\left| \sqrt{\mathrm{KL}(q_\phi(z) \mid\mid p(z\,|\,x))} - \frac{\log p(x)}{\sqrt{\mathrm{KL}(q_\phi(z) \mid\mid p(z\,|\,x))}} \right|}$$

KL is large: $\left| \frac{\delta_i^{\mathrm{CV}}}{\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]} \right| \lessapprox \mathcal{O}\left(\mathrm{KL}(q_\phi(z) \mid\mid p(z\,|\,x))^{-1/2} \right)$

KL is small: $\left| \frac{\delta_i^{\mathrm{CV}}}{\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]} \right| \lessapprox \mathcal{O}\left(\mathrm{KL}(q_\phi(z) \mid\mid p(z\,|\,x))^{1/2} \right)$

Properties

VarGrad is better than REINFORCE:

Suppose $-\frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} = \frac{\delta_i^{CV}}{\mathrm{ELBO}(\phi)} < \frac{1}{2}$, then there exists S_0 such that

$\mathrm{Var}\left(\widehat{g}_{VarGrad,i}(\phi)\right) \leq \mathrm{Var}\left(\widehat{g}_{Reinforce,i}(\phi)\right)$ for all $S > S_0$

Corollary:

If KL grows w.r.t. latent dimension, then:

There exist $S_0, D_0 \in \mathbb{N}$ for which

$\mathrm{Var}\left(\widehat{g}_{VarGrad,i}(\phi)\right) \leq \mathrm{Var}\left(\widehat{g}_{Reinforce,i}(\phi)\right)$ for all $S > S_0$