Coupled Gradient Estimators for Discrete Latent Variables

Zhe Dong ZHEDONG@GOOGLE.COM
Google Research, Brain Team

Andriy Mnih AMNIH@GOOGLE.COM
DeepMind

George Tucker GJT@GOOGLE.COM
Google Research, Brain Team

1. Introduction

Training models with discrete latent variables is challenging due to the difficulty of estimating the gradients. The gradient can be expressed as an expectation, however, in all but the simplest settings, it is analytically intractable and we must resort to estimating the gradient using Monte Carlo sampling. This problem is encountered, for example, in modern variational inference, where we would like to maximize a variational lower bound with respect to the parameters of the variational posterior. The *pathwise gradient estimator*, also known as the reparameterization trick, has low variance and has been instrumental to the success of variational autoencoders (Kingma and Welling, 2014; Rezende et al., 2014). Unfortunately, it can only be used with continuous random variables, and finding a similarly effective estimator for discrete random variables remains an important open problem.

Score-function estimators (Glynn, 1990; Fu, 2006), also known as REINFORCE (Williams, 1992), have historically been the estimators of choice for models with discrete random variables due to their unbiasedness and few requirements. As they usually exhibit high variance, previous work has augmented them with variance reduction methods to improve their practicality (Williams, 1992; Ranganath et al., 2014; Mnih and Gregor, 2014). Motivated by the efficiency of the pathwise estimator, recent progress in gradient estimators for discrete variables has primarily been driven by leveraging gradient information. While the model may only be defined for discrete inputs and hence gradients w.r.t. the random variables may not be defined, if we can construct a continuous relaxation of the system, then we can compute gradients of the continuous system and use them in an estimator (Gu et al., 2016; Jang et al., 2017; Maddison et al., 2017; Tucker et al., 2017; Grathwohl et al., 2018).

While such relaxation techniques are appealing because they result in low variance estimators, they do so by taking advantage of gradient information from an appropriate continuous relaxation. In some cases, constructing a natural continuous relaxation is nontrivial. In other cases, the computational cost of evaluating the function at the relaxed variable values will be prohibitive, e.g. in conditional computation (Bengio et al., 2013), where discrete variables specify which parts of a large model should be evaluated and using a relaxation would require evaluating the entire model every time.

Recently, Yin et al. (2019) introduced a promising alternative to relaxation-based estimators for discrete latent variables, the Augment-REINFORCE-Swap (ARS) and Augment-REINFORCE-Swap-Merge (ARSM) estimators. Instead of relaxing the variables, ARS and ARSM reparameterize them as deterministic transformations of underlying continuous variables. The estimators leverage coupled samples and a careful construction relying on symmetries of the Dirichlet distribution and exponential racing. We observe however that the continuous augmentation, which is the first step in ARS and ARSM, increases the variance of the REINFORCE estimator. Inspired by recent work (Dong et al., 2020), we improve both estimators by analytically integrating out unnecessary randomness introduced by the augmentation and reducing the variance of the estimator substantially. We show that the resulting estimators consistently outperform ARS and ARSM. However, we find that REINFORCE with a leave-one-out-baseline (Kool et al., 2019) greatly outperforms ARS and ARSM in all cases and is competitive or outperforms our improved estimators. As it is a simpler estimator to implement, we recommend it in practice.

2. Background

We consider the problem of optimizing

$$\mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right], \tag{1}$$

with respect to the parameters θ of a factorial discrete distribution $q_\theta(z) = \prod_k \mathrm{Cat}(z_k; \alpha_{\theta,k})$ where k indexes dimension and $\alpha_{\theta,k}$ are the logits of the discrete distribution with C choices[1]. This situation covers many problems with discrete latent variables, for example, in variational inference $f_\theta(z)$ could be the instantaneous ELBO (Jordan et al., 1999) and $q_\theta(z)$ the variational posterior.

The gradient with respect to θ is

$$\nabla_\theta \mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\right] = \mathbb{E}_{q_\theta(z)}\left[f_\theta(z)\nabla_\theta \log q_\theta(z) + \nabla_\theta f_\theta(z)\right]. \tag{2}$$

It typically suffices to estimate the second term with a single Monte Carlo sample, so for notational clarity, we omit the dependence of f on θ in the following sections. Monte Carlo estimates of the first term can have large variance. Low-variance, unbiased estimators of the first term will be our focus.

2.1. Augment-REINFORCE-Swap (ARS)

Yin et al. (2019) show that the discrete distribution can be reparameterized by an underlying continuous augmentation: if $\pi \sim \prod_k \mathrm{Dirichlet}(1_C)$ and $z_k := \arg\min_i \pi_{k,i} e^{-\alpha_{k,i}}$, then $z_k \sim \mathrm{Cat}(\alpha_k)$ and

$$\nabla_{\alpha_{k,c}} \mathbb{E}_{q_\theta(z)}\left[f(z)\right] = \mathbb{E}_\pi\left[f(z)(1 - C\pi_{k,c})\right].$$

Furthermore, they define a swapped probability matrix

$$\pi_{k,c}^{i \rightleftharpoons j} := \begin{cases} \pi_{k,i} & c = j \\ \pi_{k,j} & c = i \\ \pi_{k,c} & \text{o.w.} \end{cases},$$

[1] To simplify notation, we omit the subscripted θ on α in the following sections.

where $\pi_k^{i\leftrightharpoons j}$ has the entries at indices i and j swapped, and $z_k^{i\leftrightharpoons j} := \arg\min_c \pi_{k,c}^{i\leftrightharpoons j} e^{-\alpha_{k,c}}$. Using these constructions, they show an important identity

$$\nabla_{\alpha_{k,c}} \mathbb{E}_{q_\theta(z)}\left[f(z)\right] = \mathbb{E}_\pi \left[\underbrace{\left[f(z^{c\leftrightharpoons j}) - \frac{1}{C}\sum_{m=1}^{C} f(z^{m\leftrightharpoons j}) \right] (1 - C\pi_{k,j})}_{g_{\mathrm{ARS}k,c}} \right],$$

which shows that $g_{\mathrm{ARS}k,c}$ is an unbiased estimator. Importantly, if we fix a reference j and compute $f(z^{c\leftrightharpoons j})$ for all c, we can compute $g_{\mathrm{ARS}k,c}$ for all k and c with at most C potentially expensive function evaluations.

2.2. Augment-REINFORCE-Swap-Merge (ARSM)

To further improve the estimator, Yin et al. (2019) suggest averaging over the choice of the reference j resulting in the ARSM estimator

$$g_{\mathrm{ARSM}k,c} := \frac{1}{C}\sum_{j=1}^{C} \left[f(z^{c\leftrightharpoons j}) - \frac{1}{C}\sum_{m=1}^{C} f(z^{m\leftrightharpoons j}) \right] (1 - C\pi_{k,j}).$$

We can compute $g_{\mathrm{ARSM}k,c}$ for all k and c with at most $C(C-1)/2+1$ function evaluations. When $C = 2$, g_{ARSM} reduces to the ARM estimator for binary variables (Yin and Zhou, 2019).

Notably, both ARS and ARSM only evaluate f at discrete values, so do not require a continuous relaxation. Both are unbiased and we expect them to have low variance because the learning signal is a difference of evaluations of f. Yin et al. (2019) empirically show that it performs comparably or outperforms previous methods.

3. Methods

ARS and ARSM heavily rely on a continuous reparameterization of the problem, yet the original problem only depends on the discrete values. Inspired by the ideas in (Dong et al., 2020), we can derive an improved estimator by integrating out the extra randomness. Starting with ARS, ideally, we would like to compute

$$\begin{aligned}
\mathbb{E}_{\pi|z^{1\leftrightharpoons j},\ldots,z^{C\leftrightharpoons j}}\left[g_{\mathrm{ARS}k,c}\right] &= \left[f(z^{c\leftrightharpoons j}) - \frac{1}{C}\sum_{m=1}^{C} f(z^{m\leftrightharpoons j}) \right] (1 - C\mathbb{E}_{\pi|z^{1\leftrightharpoons j},\ldots,z^{C\leftrightharpoons j}}\left[\pi_{k,j}\right]) \\
&= \left[f(z^{c\leftrightharpoons j}) - \frac{1}{C}\sum_{m=1}^{C} f(z^{m\leftrightharpoons j}) \right] (1 - C\mathbb{E}_{\pi_{k,j}|z_k^{1\leftrightharpoons j},\ldots,z_k^{C\leftrightharpoons j}}\left[\pi_{k,j}\right]),
\end{aligned}$$

taking advantage of independence between dimensions (indexed by k). To reduce notational clutter, we omit the dimension index in the following derivation.

We have reduced the problem to computing

$$\mathbb{E}_{\pi_j|z^{1\leftrightharpoons j},\ldots,z^{C\leftrightharpoons j}}\left[\pi_j\right],$$

however, we were unable to compute the expectation analytically. Instead, we analytically integrate out some dimensions of π and use Monte Carlo sampling to deal with the rest. First, we know that $\sum_i \pi_i = 1$, so one variable is redundant (denote this choice by l). Next, we show how to integrate the reference index $j \neq l$ (i.e., compute $\mathbb{E}_{\pi_j | \pi_{-j,l}, z^{1 \leftrightharpoons j}, ..., z^{C \leftrightharpoons j}} \left[\pi_j\right]$, where $\pi_{-j,l}$ denotes π excluding its j-th and l-th elements.).

The known values of $\pi_{-j,l}, z^{1 \leftrightharpoons j}, ..., z^{C \leftrightharpoons j}$ imply lower and upper bounds on π_j. First, because $1 - \sum_{i \neq l} \pi_i = \pi_l \geq 0$, we conclude that $\pi_j \leq 1 - \sum_{i \neq j,l} \pi_i$. To determine the implications of the configurations $z^{1 \leftrightharpoons j}, ..., z^{C \leftrightharpoons j}$, it is helpful to define additional notation. Let $s^{c \leftrightharpoons j} := \pi^{c \leftrightharpoons j} e^{-\alpha}$. Let's look at what the value of $z^{m \leftrightharpoons j} := \arg\min_i s_i^{m \leftrightharpoons j}$ tells us about π_j. We need to consider two cases:

- $z^{m \leftrightharpoons j} = m$: This means that $s_m^{m \leftrightharpoons j} = \pi^j e^{-\alpha_m}$ is the smallest entry in $s^{m \leftrightharpoons j}$: $\pi^j e^{-\alpha_m} \leq \min_{i \neq m} s_i^{m \leftrightharpoons j}$ which implies that $\pi^j \leq \min_{i \neq m} e^{\alpha_m} s_i^{m \leftrightharpoons j}$.

  $e^{\alpha_m} s_i^{m \leftrightharpoons j}$ contains π_l when $m = l$ and $i = j$ or $m \neq l$ and $i = l$. When $m = l$ and $i = j$, we have that $e^{\alpha_l} s_j^{m \leftrightharpoons j} = e^{\alpha_l} \pi_l e^{-\alpha_j} = (1 - \sum_{n \neq j,l} \pi_n - \pi_j) e^{\alpha_l - \alpha_j}$. Therefore,

  $$\pi_j \leq \frac{(1 - \sum_{n \neq j,l} \pi_n) e^{-\alpha_j}}{e^{-\alpha_j} + e^{-\alpha_l}}.$$

  A similar computation is required for the case $m \neq l$ and $i = l$.

- $z^{m \leftrightharpoons j} \neq m$: This means that $\pi^j e^{-\alpha_m}$ is larger than the smallest entry in $s^{m \leftrightharpoons j}$: $\pi^j e^{-\alpha_m} \geq \min_i s_i^{m \leftrightharpoons j}$ which implies that $\pi^j \geq \min_i e^{\alpha_m} s_i^{m \leftrightharpoons j}$. As above, we can eliminate π_l from the bounds.

Finally, we aggregate the inequalities to compute the lower and upper bounds. Because $\pi \sim \text{Dirichlet}(1_C)$ is a uniform distribution over the simplex, $\pi_j | \pi_{-j,l}, z^{1 \leftrightharpoons j}, ..., z^{C \leftrightharpoons j}$ will be uniformly distributed over an interval, which means that it suffices to compute the lower and upper bounds to compute the expectation.

In principle, we should be able to apply similar ideas to ARSM. We plan to implement this in the full version of the paper.

3.1. Leveraging symmetry

We can take advantage of symmetry to further integrate out extraneous randomness. Let $\delta_k = \mathbb{1}_{z_k^{1 \leftrightharpoons j} = \cdots = z_k^{C \leftrightharpoons j}}$. Then, we have

$$\begin{aligned} \mathbb{E}_{\pi | \delta_k = 1} \left[g_{\text{ARS}k,c}\right] &= \mathbb{E}_{\pi | \delta_k} \left[\left[f(z^{c \leftrightharpoons j}) - \frac{1}{C} \sum_{m=1}^{C} f(z^{m \leftrightharpoons j}) \right] (1 - C\pi_{k,j}) \right] \\ &= \mathbb{E}_{\pi_k | \delta_k = 1} \left[\left(\mathbb{E}_{\pi_{-k}} \left[f(z^{c \leftrightharpoons j})\right] - \frac{1}{C} \sum_{m=1}^{C} \mathbb{E}_{\pi_{-k}} \left[f(z^{m \leftrightharpoons j})\right] \right) (1 - C\pi_{k,j}) \right]. \end{aligned}$$

Now, we claim that inside the expectation $\mathbb{E}_{\pi_{-k}} \left[f(z^{m \leftrightharpoons j})\right]$ is constant with respect to m. First, we know that $z_k^{1 \leftrightharpoons j} = \cdots = z_k^{C \leftrightharpoons j}$ inside the expectation and that the dimensions indexed by k are independent. Because $\pi \sim \prod_k \text{Dirichlet}(1_C)$, $\text{Dirichlet}(1_C)$ is symmetric, and we are taking an unconditional expectation over the remaining dimensions, the value is invariant to the swapping operation. As a result, the entire expression vanishes. Thus, we conclude that

$$g_{\mathrm{ARS}k,c}(1-\delta_k)$$

is still an unbiased estimator. A similar argument holds for g_{ARSM}. This is complementary to the approach in the previous subsection and can done in combination

$$g_{\mathrm{ARS+}k,c} \coloneqq \mathbb{E}_{\pi_{k,j}|\pi_{k,-jl}, z_k^{1\leftrightharpoons j},\ldots,z_k^{C\leftrightharpoons j}}\left[g_{\mathrm{ARS}k,c}(1-\delta_k)\right],$$

where we choose $l \neq j$ uniformly randomly. This is the estimator we use in our experiments.

4. Experiments

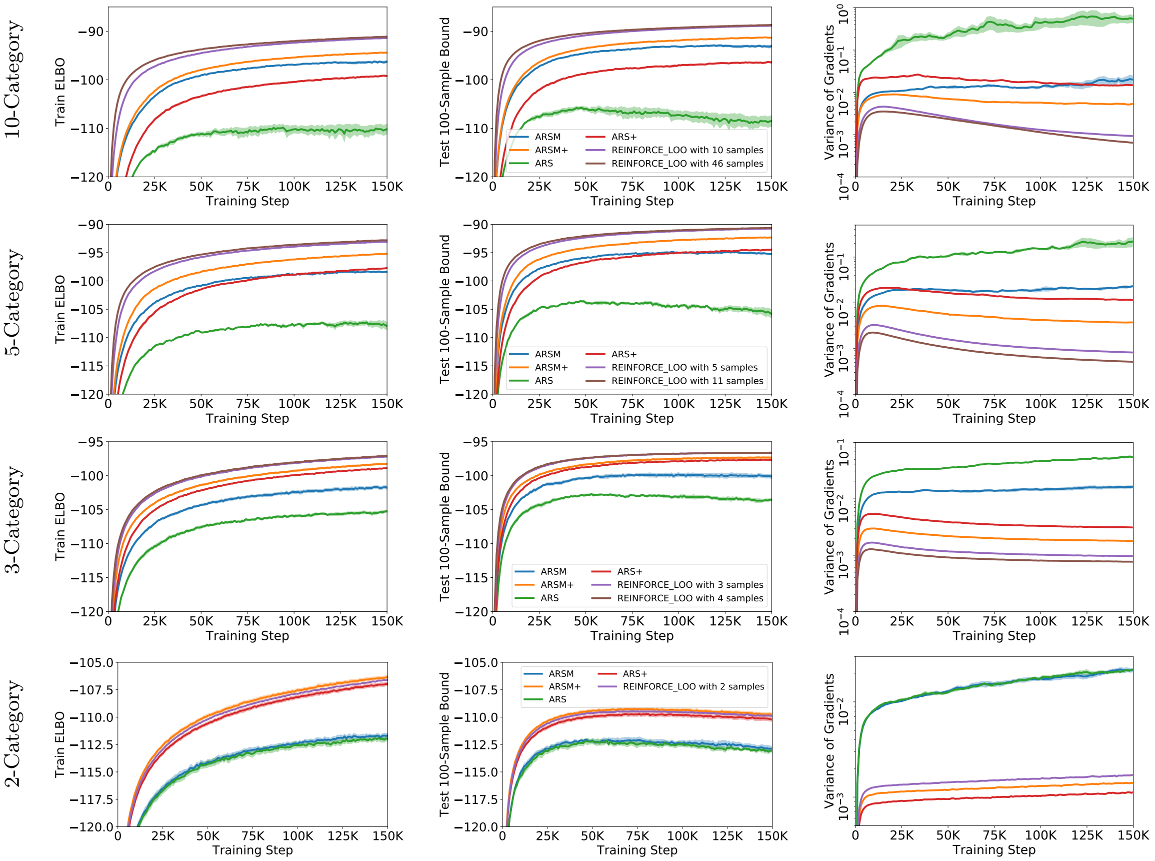

Figure 1: Training a non-linear Categorical VAE with stochastic hidden units of 10/5/3/2 categories on dynamically binarized MNIST dataset by maximizing the ELBO. We plot the train ELBO (left column), test 100-sample bound (middle column), and the variance of gradient estimator (right column). We plot the mean and one standard error based on 5 runs from different random initializations

We benchmark the proposed gradient estimators by training Variational Auto-Encoders with categorical latent variables. To facilitate comparison with ARS and ARSM, we use the same architecture as in (Yin et al., 2019). Briefly, the model has a single layer of 20 categorical latent variables which are mapped to Bernoulli logits with an MLP with two hidden layers of 256 and 512 of LeakyReLU units (Xu et al., 2015) with 0.2 negative slope. The encoder mirrors the structure with two hidden layers of 512 and 256 LeakyReLU units.

For ARSM, we reduce the variance only by leveraging the symmetry (Section 3.1) and call the resulting estimator ARSM+. We train models with 10/5/3/2-dimensional categorical latent variables on dynamically binarized MNIST. For comparison, we train models with ARM, ARSM, and an n-sample REINFORCE estimator with a leave-one-out baseline (REINFORCE LOO) (Kool et al., 2019). To match computation, REINFORCE LOO uses C samples for comparing against ARS/ARS+, and uses $C(C-1)/2+1$ samples for ARSM/ARSM+, where C is the number of categories.

As shown in Figure 1, the proposed estimators, ARS+/ARSM+, significantly outperform ARS/ARSM. Surprisingly, we find that both ARS and ARSM underperform the simpler REINFORCE LOO baseline in all cases. For $C = 2$, ARS+[2] and ARSM+ outperform REINFORCE LOO; however, for $C > 2$, REINFORCE LOO is superior and the gap increases as C does. This suggests that partially integrating out the randomness is insufficient to account for the variance introduced by the continuous augmentation.

5. Conclusion

While we achieved our goal of improving ARS and ARSM, along the way we discovered that the methods that we built upon were inferior to a simpler baseline. In practice, we suggest using REINFORCE LOO as it achieves competitive or superior performance in all of the cases we evaluated. However, in spite of this negative result, we still believe the techniques and careful benchmarking in this paper will be valuable to the community.

References

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Zhe Dong, Andriy Mnih, and George Tucker. DisARM: An antithetic gradient estimator for binary latent variables. *arXiv preprint arXiv:2006.10680*, 2020.

Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

[2] In this case, ARS+ reduces to DisARM (Dong et al., 2020).

Shixiang Gu, Sergey Levine, Ilya Sutskever, and Andriy Mnih. MuProp: Unbiased backpropagation for stochastic neural networks. In *International Conference on Learning Representations*, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

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

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! In *Deep RL Meets Structured Prediction ICLR Workshop*, 2019.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1791–1799, 2014.

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286, 2014.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems 30*, 2017.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Mingzhang Yin and Mingyuan Zhou. ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*, 2019.

Mingzhang Yin, Yuguang Yue, and Mingyuan Zhou. ARSM: Augment-REINFORCE-swap-merge estimator for gradient backpropagation through categorical variables. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.