BACKPROPAGATING THROUGH THE VOID: OPTIMIZING DIFFERENTIABLE CONTROL VARIATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient-based optimization is the foundation of deep learning and reinforcement learning. Even when the mechanism being optimized is unknown or not differentiable, optimization using high-variance or biased gradient estimates is still often the best strategy. We introduce a general framework for learning low-variance, unbiased gradient estimators for black-box functions of random variables. These estimators can be jointly trained with model parameters or policies, and are applicable in both discrete and continuous settings. We give unbiased, adaptive analogs of state-of-the-art reinforcement learning methods such as deep deterministic policy gradients and advantage actor-critic. We also demonstrate this framework for training discrete latent-variable models.

1 INTRODUCTION

Gradient-based optimization has been key to most recent advances in machine learning and deep reinforcement learning. The back-propagation algorithm (Rumelhart & Hinton, 1986), also known as reverse-mode automatic differentiation (Speelpenning, 1980; Rall, 1981) computes exact gradients of deterministic, differentiable objective functions. The reparameterization trick (Williams, 1992; Kingma & Welling, 2014; Rezende et al., 2014) allows backpropagation to give unbiased, low-variance estimates of gradients of expectations of continuous random variables. This has allowed effective stochastic optimization of large probabilistic latent-variable models.

Unfortunately, backpropagation cannot be straightforwardly applied to problems involving discrete random variables, or when the function being optimized is a black box (Schulman et al., 2015). This is the case in most reinforcement learning settings, or when fitting probabilistic models with discrete latent variables. Much recent work has been devoted to constructing gradient estimators for these situations. In reinforcement learning, advantage actor-critic methods (Mnih et al., 2016) and deep deterministic policy gradients (Lillicrap et al., 2015) give low-variance but biased(MAY NEED TO CLARIFY THIS, a2s is not biased) gradient estimates by jointly optimizing policy parameters together with baselines. In discrete latent-variable models, low-variance but biased gradient estimates can be given by continuous relaxations of discrete variables (Maddison et al., 2016; Jang et al., 2016).

A recent advance by Tucker et al. (2017) used low-variance but biased gradients from continuous relaxations to construct unbiased, low-variance gradient estimates. Furthermore, Tucker et al. (2017) showed how to tune the free parameters of these relaxations to minimize their variance during training.

We generalize the method of Tucker et al. (2017) to learn a free-form control variate parameterized by a neural network, giving lower-variance, unbiased gradient estimates. Importantly, our method is applicable even when no continuous relaxation is available, as in reinforcement learning. We derive improved variants of popular reinforcement learning methods with unbiased gradients and more stable training dynamics.

2 MOTIVATION

(WE SHOULD BREAK THIS UP INTO MOTIVATION AND BACKGROUND) IE WHY DO WE WANT THIS, RL, DISCRETE RANDOM VARIABLE MODELS, YADDA, YADDA, YADDA

THEN HOW HAS IT BEEN DONE IN THE PAST

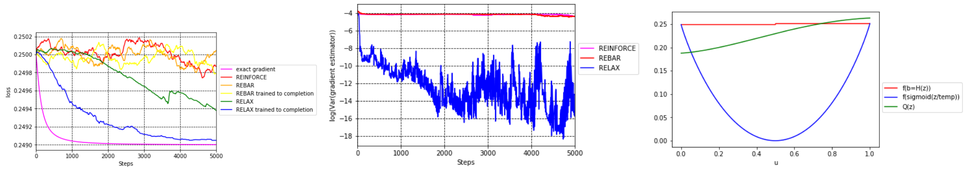

Figure 1: *Left:* Estimator losses. *Centre:* Estimator variance. *Right:*Learned relaxation function.

THEN WHATS WRONG WITH THAT, LIMITATIONS, SHIT LIKE THAT

2.1 OPTIMIZING CONTINUOUS BLACK-BOX FUNCTIONS

Deep deterministic policy gradients (Lillicrap et al., 2015)

Also: (Levine et al., 2016)

3 BACKGROUND: GRADIENT ESTIMATORS

How to choose the parameters of a distribution to maximize an expectation? This problem comes up in reinforcement learning, where we must choose a policy π on actions a to maximize the expected reward $\mathbb{E}_{p(a|\pi)}\left[r(a)\right]$. It also comes up in fitting latent-variable models, when we wish to maximize the marginal probability $p(x|\theta) = \sum p(x|z)p(z|\theta) = \mathbb{E}_{p(z|\theta)}\left[p(x|z)\right]$. In this paper, we'll consider the general problem of optimizing

$$\mathcal{L}(\theta) = \mathbb{E}_{p_\theta(b)}[\, f(b)] \,. \tag{1}$$

Later, we will discuss the case when $f(b)$ depends directly on θ.

When the parameters θ are high-dimensional, gradient-based optimization is a appealing because it provides information about how to adjust each parameter individually. Stochastic optimization is essential for large problems, but is only guaranteed to converge to a fixed point of the original objective (Robbins & Monro, 1951) when the stochastic gradients $\hat{g}$ are unbiased, i.e. $\mathbb{E}\left[\hat{g}\right] = \nabla_\theta \mathbb{E}_{p(b|\theta)}\left[f(b)\right]$.

The score-function gradient estimator One of the most generally-applicable gradient estimators is known as the score-function estimaor, or REINFORCE (Williams, 1992):

$$\hat{g}_{\text{reinforce}} = f\left(b\right) \nabla_\theta \log p(b|\theta), \qquad b \sim p(b|\theta) \tag{2}$$

This estimator is unbiased [todo: give conditions], but in general has high variance (CAN ME MAKE THAT CLAIM?). Intuitively, this estimator is limited by the fact that it doesn't use any information about how f depends on b, only on the final outcome of evaluating $f(b)$.

The reparameterization trick When f is continuous and differentiable, and the latent variables b can be written as a deterministic, differentiable function of a random draw from a fixed distribution, the reparameterization trick (Williams, 1992; Kingma & Welling, 2014; Rezende et al., 2014) creates a low-variance, unbiased gradient estimator by making the dependence of b on θ explicit:

$$\hat{g}_{\text{reparam}} = \nabla_\theta f\left(b(\theta, \epsilon)\right), \qquad \epsilon \sim p(\epsilon) \tag{3}$$

This gradient estimator is used when training high-dimensional, continuous latent-variable models, such as variational autoencoders or GANs. One intuition for why this gradient estimator is preferable to REINFORCE is that it depends on $\partial f / \partial b$, which directly exposes the dependence of f on b.

Concrete relaxation When b is discrete and f is known, one general approach is to differentiate a continuous relaxation of the discrete random variables. Maddison et al. (2016) and Jang et al. (2016) developed a differentiable relaxation of the categorical distribution, called the concrete distribution:

$$\hat{g}_{\text{concrete}} = \nabla_\theta f\left(\sigma_\lambda(\log\theta - \log(-\log \mathbf{u}))\right), \qquad \mathbf{u} \sim \text{uniform}[0, 1] \tag{4}$$

where σ_λ is the softmax function with temperature λ [TODO: Reference this in appendix]. This gradient estimator is fairly effective in practice, but produces biased gradients. Additionally, it is not clear how to set the temperature λ.

Control variates Control variates are a general trick for reducing the variance of a Monte Carlo estimator. Given an estimator $f(b)$, a control variate is a function $\hat{f}(b)$ with a known mean $\mathbb{E}_{p(b)}[\hat{f}]$. Subtracting the control variate from our estimator and adding its mean gives us a new estimator:

$$\hat{f}_{\text{new}}(b) = f(b) - \hat{f}(b) + \mathbb{E}_{p(b)}[\hat{f}(b)] \tag{5}$$

This new estimator has the same expectation as the old one:

$$\mathbb{E}_{p(b)}\left[f_{\text{new}}\right] = \mathbb{E}_{p(b)}\left[f(b) - \hat{f}(b) + \mathbb{E}_{p(b)}\left[\hat{f}(b)\right]\right] = \mathbb{E}_{p(b)}\left[f(b)\right] \tag{6}$$

Importantly, the new estimator has lower variance than $f(b)$ if $\hat{f}(b)$ is positively correlated with $f(b)$.

REBAR ALL THIS STUFF I HAVE COMMENTED OUT SHOULD BE CUT DOWN AND PLACED HERE

4 Scope and Limitations

One major limitation of the REBAR estimator and the concrete relaxation is that they requires the function being optimized, whose input is only defined at discrete inputs, to also accept continuous inputs, and to be differentiable w.r.t. those inputs. This makes REBAR and the concrete relaxation inapplicable for optimizing black-box functions, as in reinforcement learning settings where the environment is unknown.

Following Tucker et al. (2017), the following overview focuses on a single discrete Bernoulli random variable. However, the generalization to categorial variables is straightforward. (LETS ADD THE REST OF OUR CRITICISMS GRIEVANCES HERE)

5 A General Family of Gradient Estimators

In this work, we seek to synthesize and generalize the approaches introduced above. Mainly, we desire a simple and generic gradient estimator that can be applied to expectations of known or black-box functions of discrete or random variables. We would like these approaches to be simple to implement and add minimal computational overhead.

The foundation of our approach is the score function gradient estimator with a control variate who's expectation can be estimated with low variance using the reparameterization trick. This control variate is neural network which is trained directly to minimize the variance of the estimated gradients. The central result of this paper is that learning the function in the control variate leads to even better convergence properties and lower variance gradient estimates.

We first outline the algorithm when we are optimizing $E_{p_\theta(b)}[f(b)]$ where b is a continuous random variable and then when b is discrete.

5.1 Continuous Variables

We wish to compute $\frac{\partial}{\partial\theta}\mathbb{E}_{p_\theta(b)}[f(b)]$ where $p_\theta(b)$ is a distribution with continuous support that admits a reparameterization T such that $T(\theta, \epsilon) = b \sim p_\theta(b)$ where $\epsilon \sim p(\epsilon)$. We assume that f is a black-box or involves operations such that $\frac{\partial}{\partial\theta}f = 0$ almost everywhere or is not computable. Given the reparameerization, we can re-write this as an expectation over ϵ as $\frac{\partial}{\partial\theta}\mathbb{E}_{p(\epsilon)}[f(T(\epsilon,\theta)]$. We introduce a new, differentiable function r_ϕ and define our estimator below.

$$\frac{\partial}{\partial\theta}\mathbb{E}_{p_\theta(b)}[f(b)] \approx g_\phi(\epsilon) = \Big[f(b) - r_\phi(b)\Big] \cdot \frac{\partial}{\partial\theta}\log p(b) + \frac{\partial}{\partial\theta}r(b) \tag{7}$$

$$= \Big[f(T(\epsilon,\theta)) - r_\phi(T(\epsilon,\theta))\Big] \cdot \frac{\partial}{\partial\theta}\log p(b) + \frac{\partial}{\partial\theta}r(T(\epsilon,\theta)) \tag{8}$$

We note that our estimator is unbiased for all choices of r_ϕ, i.e $\mathbb{E}_\epsilon[g_\phi(\epsilon)] = \frac{\partial}{\partial\theta}\mathbb{E}_{p_\theta(b)}[f(b)]$. A proof can be found in the appendix.

5.2 Discrete Variables

In the case where $p_\theta(b)$ is a distribution over discrete variables, we introduce some notation. We assume there exist a continuous distribution $p_\theta(z)$ and a determanistic mapping H such that $H(z) = b \sim p_\theta(b)$. If $p(b)$ is Bernoulli then $p_\theta(z) = \text{Logistic}(z|\theta)$ and $H(z) = \mathbb{I}(z > 0)$. If b is Categorical, then $p_\theta(z) = \text{Softmax}(z|\theta)$. We also assume that $p_\theta(z)$ is reparameterizable. Moreover, we assume the distribution $p_\theta(z|b)$ is also reparameterizable meaning there exists $\hat{T}$ such that $\hat{T}(\hat{\epsilon}, b, \theta) = \hat{z} \sim p_\theta(\hat{z}|b)$ (see Appendix for details).

The estimator is based on a 2-step monte-carlo procedure. First we sample $\epsilon \sim p(\epsilon)$ from which we produce $z = T(\theta, \epsilon)$ and $b = H(z)$. We then sample $\hat{\epsilon} \sim p(\hat{\epsilon})$ and produce $\hat{z} = \hat{T}(\hat{\epsilon}, b, \theta)$. We use these values to produce our estimator below.

$$\frac{\partial}{\partial\theta}\mathbb{E}_{p_\theta(b)}[f(b)] \approx g_\phi(\epsilon, \hat{\epsilon}) = \Big[f(b) - r_\phi(\hat{z})\Big] \cdot \frac{\partial}{\partial\theta}\log p(b) + \frac{\partial}{\partial\theta} r(z) - \frac{\partial}{\partial\theta} r(\hat{z}) \tag{9}$$

$$= \Big[f(H(T(\epsilon,\theta))) - r_\phi(\hat{T}(\epsilon, H(T(\epsilon,\theta)),\theta))\Big] \cdot \frac{\partial}{\partial\theta}\log p(b) \tag{10}$$

$$+ \frac{\partial}{\partial\theta} r(T(\epsilon,\theta)) - \frac{\partial}{\partial\theta} r_\phi(\hat{T}(\epsilon, H(T(\epsilon,\theta)),\theta)) \tag{11}$$

This estimator is also unbiased i.e $E_{\epsilon,\hat{\epsilon}}[g_\phi(\epsilon,\hat{\epsilon})] = \frac{\partial}{\partial\theta}\mathbb{E}_{p_\theta(b)}[f(b)]$.

5.3 Optimizing The Control Variate

For both estimators, we would like to optimize the parameters ϕ to produce the lowest variance. Following Tucker et al. (2017), we note that for any unbiased estimator g_ϕ we have

$$\frac{\partial}{\partial\phi}\text{Var}(g_\phi) = \frac{\partial}{\partial\phi}\mathbb{E}[g_\phi^2] - \frac{\partial}{\partial\phi}\mathbb{E}[g_\phi]^2 = \frac{\partial}{\partial\phi}\mathbb{E}[g_\phi^2] - \frac{\partial}{\partial\phi}\mathbb{E}_{p_\theta(b)}[f(b)] = \frac{\partial}{\partial\phi}\mathbb{E}[g_\phi^2] = \mathbb{E}\left[\frac{\partial}{\partial\phi} g_\phi^2\right]. \tag{12}$$

Therefore, we can minimize the variance of the estimator by training to minimize the square of the gradients it produces. [TODO: talk about how this is probably better than minimizing the squared prediction error]

6 Applications

(I THINK WE SHOULD RENAME THIS APPLICATIONS. THIS ALLOWS ME TO GO INTO MORE OF THE THEORY WHEN I GET TO THE RL STUFF) We demonstrate the effectiveness of our estimator on a number of challenging optimization problems. Following Tucker et al. (2017) we begin with a simple toy example to illuminate the potential of our method and then continue to the more relevant problems of optimize binary VAE's and reinforcement learning.

6.1 Toy Experiment

We seek to minimize $\mathbb{E}_{b\sim p(b|\theta)}[(b-t)^2]$ as a function of the parameter θ where $p(b|\theta) = \texttt{Bern}(b|\theta)$. Tucker et al. (2017) set the target $t = .45$. We focus on the more challenging case where $t = .499$. With this setting of the target, REBAR and competing methods suffer from high variance and are unable to discover the optimal solution $\theta = 0$.

Figure **??** plots the learned relaxations for a fixed value of θ. It can be seen that RELAX learns a relaxation whose derivative points in the direction of decreased loss for all values of reparameterization noise u, whereas REBAR's fixed relaxation only does so for values of $u > t$.

6.2 Discrete Variational Autoencoder

As in (Tucker et al., 2017), we benchmark the RELAX estimator on the task of training a variational autoencoder (Kingma & Welling, 2014; Rezende et al., 2014) where all random variables are Bernoulli taking values in $\{-1, 1\}$. As in Tucker et al. (2017), we compare training the variational lower-bound across the MNIST and Omniglot (Lake et al., 2015) datasets. As in Tucker et al. (2017) we test models with 1 and 2 of Bernoulli random variables with linear mappings between them and a model with 1 layer of Bernoulli random variables with non-linear mappings between layers.

We found that due to the complicated structure of the loss function, the RELAX estimator performed worse than REBAR. Instead we add a learned relaxation to REBAR's control variate which we denote relaxed-REBAR. Our estimator takes the form of (**??**) with

$$\bar{r}(z) = r(z) + f(\sigma_\lambda(z))$$

where $r(z)$ is a learned neural network and $f(\sigma_\lambda(z))$ is the Concrete relaxation of REBAR with temperature parameter λ. In all experiments, adding the learned $r(z)$ reduced the variance of the gradients and improved the final results.

	NVIL	MuProp	REBAR	REBAR	RELAX
MNIST			Tucker et al. (2017)	ours	
Nonlinear	-102.2	-101.5	-101.1	-83.02	**-79.49**
Linear 1 layer	-112.5	-111.7	-111.6	-111.66	**-111.22**
Linear 2 Layer	-99.6	-99.07	-98.8	-98.23	**-98.04**
Omniglot					
Nonlinear	-110.4	-109.58	-108.72	-62.28	**-58.55**
Linear 1 layer	-117.44	-117.09	-116.83	-116.75	**-116.62**
Linear 2 Layer	-109.98	-109.55	-108.99	-108.74	**-108.59**

Table 1: Training variational lower bound after training.

In (Tucker et al., 2017), a separate REBAR estimator was used to estimate the gradients of each model parameter (each weight matrix and bias vector). To apply our estimator to this formulation, we would need to learn a separate relaxation for each model parameter. To get around this, we use our gradient estimator for approximate $g_\phi \approx \frac{\partial}{\partial\theta}\mathbb{E}_{q(b|\theta)}[f(b)]$ where $x \cdot W = \theta$ is the parameters of the Bernouli latent variables, W is our layer's weight matrix. We then obtain an estimate of $\frac{\partial}{\partial W}\mathbb{E}_{q(b|\theta)}[f(b)] = g_\phi \cdot \frac{\partial\theta}{\partial W}$. We note this gives us unbiased gradients because

$$\mathbb{E}_\epsilon[g_\phi(\epsilon) \cdot \frac{\partial\theta}{\partial W}] = \mathbb{E}_\epsilon[g_\phi(\epsilon)] \cdot \frac{\partial\theta}{\partial W} = \frac{\partial}{\partial\theta}\mathbb{E}_{q(b|\theta)}[f(b)] \cdot \frac{\partial\theta}{\partial W} = \frac{\partial}{\partial W}\mathbb{E}_{q(b|\theta)}[f(b)] \tag{13}$$

To provide a fair comparison, we re-implemented REBAR in this way (denoted REBAR-ours in table 6.2). We believe this explains the large difference in performance between our implementation and that of (Tucker et al., 2017) for the nonlinear models since there are 3 layers of parameters that all share the same gradient estimator. In the linear models, each layer has its own gradient estimator making our implementation closer to that of (Tucker et al., 2017).

6.3 Reinforcement Learning

Reinforcement learning is strong motivating problem for methods similar to ours. In reinforcement learning we seek to optimize the paremters of a policy distribution $\pi(a|s;\phi)$ to maximize the (often discounted) sum of future rewards given that policy $\mathbb{E}_{\pi(\phi)}[\sum_{t=1}^{\infty} r_t]$. We can view the sum of future rewards as a black-box function of our policy's actions $f(a)$. Thus, as before we have reduced the problem to that of estimating $\frac{\partial\mathbb{E}_{\pi(a|\phi)}[f(a)]}{\partial\phi}$ which is the standard policy gradient algorithm **?**.

We test our approach on simple reinforcement learning environments with discrete actions. We use the RELAX estimator and compare with the advantage actor critic algorithm (A2C **?**) as a baseline. In all experiments we utilized the same learning rate for the policy network for RELAX and A2C to so differences in performance depended solely on the control variate used.

To compare these approaches in the most illustrative setting possible we run these algorithms on one environment at a time, running each episode to completion. After completion, we generate the discounted reward for each timestep, treat the episode as a single batch of data, and perform one step of gradient decent. We test our alrgorithm on the Cart-Pole and Lunar-Lander environments from the OpenAI Gym **?**. We run the Cart-Pole and Lunar-Lander environments for 250 and 1000 episodes, respectively and plot reward and the log-variance of the policy gradients in figure X.

6.4 RL INTRODUCTION

We seek to compute

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial \theta} = \mathbb{E}\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} \sum_{t'=t}^{T} r_t\Big]$$

but the estimator on the right hand side can have potentially high variance. Instead, we typically compute

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial \theta} = \mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} [(\sum_{t'=t}^{T} r_{t'}) - b(s_t)]\Big]$$

This is unbiased because

$$(14)$$

$$E_{a_{1:T},s_{1:T}}\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} \cdot b(s_t)\Big] = E_{a_{1:t-1},s_{1:t}}\Big[E_{a_{t:T},s_{t+1:T}}\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} \cdot b(s_t)\Big]\Big] \quad (15)$$

$$= E_{a_{1:t-1},s_{1:t}}\Big[b(s_t) \cdot E_{a_{t:T},s_{t+1:T}}\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial \theta}\Big]\Big] \quad (16)$$

$$= E_{a_{1:t-1},s_{1:t}}\Big[b(s_t) \cdot \frac{\partial}{\partial \theta} E_{a_{t:T},s_{t+1:T}}[1]\Big] \quad (17)$$

$$= 0 \quad (18)$$

Where $b(s_t)$ is trained to minimize $(b(s_t) - \sum_{i=t}^{T} r_t)^2$. This is unbiased for any choice of b since b does not depend on a_t.

6.4.1 VARIANCE REDUCTION WITH RELAX

We begin the in the case where actions are continuous and any expectation over an action is replaced with an expectation over reparameterization variables ϵ

$$\mathbb{E}_a[f(a)] = \mathbb{E}_\epsilon[f(a(\epsilon))]$$

. We are interested in replacing $b(s_t)$ with a function $m(a_t, s_t)$ which is a function of both actions and states.

This changes the first equation to

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial \theta} = \mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} [(\sum_{t'=t}^{T} r_{t'}) - m(a_t, s_t)]\Big]$$

but this makes the estimator biased as $\mathbb{E}_\tau[\frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} m(a_t, s_t)] \neq 0$. Because of the dependence on a_t we cannot pull this out of the expectation as we could in line 2 of the above equation.

Instead we subtract out a term who's expectation should be the same as the score function version of the control variate changing our estimator to

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial \theta} = \mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} [(\sum_{t'=t}^{T} r_{t'}) - m(a_t, s_t)] + \frac{\partial m(a_t, s_t)}{\partial \theta}\Big].$$

For this estimator to be unbiased, we must have

$$\mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial \theta} m(a_t, s_t) - \frac{\partial m(a_t, s_t)}{\partial \theta}\Big] = 0$$

and we claim that $\forall t$, we have

$$\mathbb{E}_\tau\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial\theta} m(a_t,s_t) - \frac{\partial m(a_t,s_t)}{\partial\theta}\Big] = 0$$

We begin with

$$\mathbb{E}_\tau\Big[\frac{\partial m(a_t,s_t)}{\partial\theta}\Big] = \mathbb{E}_{a_{1:t-1},s_{1:t}}\Big[E_{a_{t:T},s_{t+1:T}}\Big[\frac{\partial m(a_t,s_t)}{\partial\theta}\Big]\Big] \tag{19}$$

$$= \mathbb{E}_{a_{1:t-1},s_{1:t}}\Big[E_{a_t}\Big[\frac{\partial m(a_t,s_t)}{\partial\theta}\Big]\Big] \tag{20}$$

$$= \mathbb{E}_{a_{1:t-1},s_{1:t}}\Big[\frac{\partial E_{a_t}[m(a_t,s_t)]}{\partial\theta}\Big] \tag{21}$$

$$= \mathbb{E}_{a_{1:t-1},s_{1:t}}\Big[E_{a_t}\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial\theta} m(a_t,s_t)\Big]\Big] \tag{22}$$

$$= E_\tau\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial\theta} m(a_t,s_t)\Big] \tag{23}$$

which completes our proof.

6.4.2 DISCRETE CASE

In the discrete action setting our policy parameterizes a soft-max distribution which we use to sample actions. In the discrete case we define $z_t = f(\pi, u) = \sigma(\log \pi - \log(-\log(u)))$ where $u \sim \text{Unif}[0,1]$, $a_t = \text{argmax}(z_t)$, σ is the soft-max function.

We also define $\tilde{z}_t \sim p(z_t|a_t)$ so if the z_t are gumbel softmax samples, then $\tilde{z}_t$ are gumbel softmax samples constrainted so that the largest value is that of the index of a_t. See appendix for how to efficiently generate samples $\tilde{z}_t$.

In the discrete case, we use $m(\tilde{z}_t, s_t) \cdot \frac{\log \pi(a_t|s_t,\theta)}{\partial\theta}$ as our control variate giving us

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial\theta} = \mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial\theta}[(\sum_{t'=t}^{T} r_{t'}) - m(\tilde{z}_t,s_t)]\Big]$$

which is biased, so we must add a term to remove this bias which will be $\frac{\partial m(z_t,s_t)}{\partial\theta} - \frac{\partial m(\tilde{z}_t,s_t)}{\partial\theta}$ making the full estimator

$$\frac{\partial \mathbb{E}_\tau[R]}{\partial\theta} = \mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial\theta}[(\sum_{t'=t}^{T} r_{t'}) - m(\tilde{z}_t,s_t)] + \frac{\partial m(z_t,s_t)}{\partial\theta} - \frac{\partial m(\tilde{z}_t,s_t)}{\partial\theta}\Big].$$

For this estimator to be unbiased, we must have

$$\mathbb{E}_\tau\Big[\sum_{t=1}^{T} \frac{\log \pi(a_t|s_t,\theta)}{\partial\theta} m(\tilde{z}_t,s_t) - \Big(\frac{\partial m(z_t,s_t)}{\partial\theta} - \frac{\partial m(\tilde{z}_t,s_t)}{\partial\theta}\Big)\Big] = 0$$

and we claim that $\forall t$

$$\mathbb{E}_\tau\Big[\frac{\log \pi(a_t|s_t,\theta)}{\partial\theta} m(\tilde{z}_t,s_t) - \Big(\frac{\partial m(z_t,s_t)}{\partial\theta} - \frac{\partial m(\tilde{z}_t,s_t)}{\partial\theta}\Big)\Big] = 0$$

We introduce some notation that $a(z) = \text{argmax}(z)$

$$0 = \frac{\partial}{\partial\theta}\mathbb{E}_\tau\Big[m(z_t,s_t) - m(z_t,s_t)\Big] = \mathbb{E}_{a_{<t},s_{\leq t}}\Big[\frac{\partial}{\partial\theta}\mathbb{E}_{z_t|s_t}\Big[m(z_t,s_t)\Big] - \frac{\partial}{\partial\theta}\mathbb{E}_{a_t|s_t}\Big[\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\Big]\Big]\Big] \tag{24}$$

We drop the outer expectation for brevity and continue

$$\frac{\partial}{\partial\theta}\mathbb{E}_{z_t|s_t}\Big[m(z_t,s_t)\Big] - \frac{\partial}{\partial\theta}\mathbb{E}_{a_t|s_t}\Big[\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\Big]\Big] \tag{25}$$

$$= \frac{\partial}{\partial\theta}\mathbb{E}_{z_t|s_t}\Big[m(z_t,s_t)\Big] - \mathbb{E}_{a_t|s_t}\Big[\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\frac{\log\pi(a_t|s_t,\theta)}{\partial\theta} - \frac{\partial}{\partial\theta}\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\Big]\Big] \tag{26}$$

$$= \mathbb{E}_{z_t|s_t}\Big[\frac{\partial}{\partial\theta}m(z_t,s_t)\Big] - \mathbb{E}_{a_t|s_t}\Big[\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\frac{\log\pi(a_t|s_t,\theta)}{\partial\theta} - \mathbb{E}_{z_t|a_t,s_t}\Big[\frac{\partial}{\partial\theta}m(z_t,s_t)\Big]\Big] \tag{27}$$

Thus due to the markov property of the MDP, then we can wrap all of these expectations into the expectation over trajectories giving us

$$\mathbb{E}_{a_{<t},s_{\leq t}}\Big[\mathbb{E}_{z_t|s_t}\Big[\frac{\partial}{\partial\theta}m(z_t,s_t)\Big] - \mathbb{E}_{a_t|s_t}\Big[\mathbb{E}_{z_t|a_t,s_t}\Big[m(z_t,s_t)\frac{\log\pi(a_t|s_t,\theta)}{\partial\theta} - \mathbb{E}_{z_t|a_t,s_t}\Big[\frac{\partial}{\partial\theta}m(z_t,s_t)\Big]\Big]\Big] \tag{28}$$

$$= \mathbb{E}_{\tau}\Big[\frac{\log\pi(a_t|s_t,\theta)}{\partial\theta}m(\tilde{z}_t,s_t) - \Big(\frac{\partial m(z_t,s_t)}{\partial\theta} - \frac{\partial m(\tilde{z}_t,s_t)}{\partial\theta}\Big)\Big] = 0 \tag{29}$$

Mnih et al. (2015)

[Do we use ADAM (Kingma & Ba, 2015) for optimization?]

7 Related Work

Miller et al. (2017) further reduce the variance of reparameterization gradients in an orthogonal way.

As gradient estimators become more complex, checking their unbiasedness numerically becomes difficult. The automatic theorem-proving-based unbiasedness checker developed by Selsam et al. (2017) may become relevant to this line of research.

NVIL (Mnih & Gregor, 2014), VIMCO (Mnih & Rezende, 2016)

Staines & Barber (2012) address the general problem of developing gradient estimators for deterministic black-box functions or discrete optimization. They introduce a sampling distribution, and optimize an objective similar to ours.

Wierstra et al. (2014) also introduce a sampling distribution to build a gradient estimator, and consider optimizing the sampling distribution.

Asadi et al. (2017) reduce the variance of actor-critic gradient estimates by simply summing over all possible actions.

Salimans et al. (2017) estimate gradients using a form of finite differences, evaluating hundreds of different parameter values in pararallel to construct a gradient estimator. In contrast, our method is a simple-sample estimator.

Generalized Reparameterization Gradients REBAR and the generalization in this paper uses a mixture of score function and reparameterization gradients. A recent paper by Ruiz et al. (2016) unifies these two gradient estimators as the generalized reparameterization gradient (GRG). This framework can help disentangle the various components of generalized REBAR.

REBAR innovation as further decomposition the correction term into secondary reparameterization components note this is a recursive application of the principles of GRG observe that the GRG suggests this recursive application to components of an estimator propose that other estimators could be similarly recursively decomposed?

8 Conclusions and Future Work

Other possible applications:

GANs (Goodfellow et al., 2014) that generate text or other discrete objects.

Learning to parse (Kusner et al., 2017)

VAEs with continuous latent variables but non-differentiable likelihood functions.

References

Kavosh Asadi, Cameron Allen, Melrose Roderick, Abdel-rahman Mohamed, George Konidaris, and Michael Littman. Mean actor critic. *arXiv preprint arXiv:1709.00503*, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

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

Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. *arXiv preprint arXiv:1703.01925*, 2017.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Andrew C Miller, Nicholas J Foti, Alexander D'Amour, and Ryan P Adams. Reducing reparameterization gradient variance. *arXiv preprint arXiv:1705.07880*, 2017.

Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1791–1799, 2014.

Andriy Mnih and Danilo Rezende. Variational inference for monte carlo objectives. In *International Conference on Machine Learning*, pp. 2188–2196, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 02 2015. URL http://dx.doi.org/10.1038/nature14236.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pp. 1928–1937, 2016.

Louis B Rall. Automatic differentiation: Techniques and applications. 1981.

Danilo J 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*, pp. 1278–1286, 2014.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, pp. 460–468, 2016.

David E Rumelhart and Geoffrey E Hinton. Learning representations by back-propagating errors. *Nature*, 323:9, 1986.

Tim Salimans, Jonathan Ho, Xi Chen, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, pp. 3528–3536, 2015.

Daniel Selsam, Percy Liang, and David L Dill. Developing bug-free machine learning systems with formal mathematics. *International Conference on Machine Learning*, 2017.

Bert Speelpenning. *Compiling Fast Partial Derivatives of Functions Given by Algorithms*. PhD thesis, University of Illinois at Urbana-Champaign, 1980.

Joe Staines and David Barber. Variational optimization. *arXiv preprint arXiv:1212.4507*, 2012.

George Tucker, Andriy Mnih, Chris J Maddison, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. *arXiv preprint arXiv:1703.07370*, 2017.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *Journal of Machine Learning Research*, 15(1):949–980, 2014.

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

9 Appendix A: Control Variates

Generalizing the reparameterization trick

Write sample from distribution $s(\epsilon)$ as $\epsilon = \mathcal{T}^{-1}(\mathbf{z}; \nu)$ for some invertible transform $\mathcal{T}$ with variational parameters ν. write out transformed density example: normal with standard normal s example: inverse CDF of Gaussian with uniform s write out expected gradient under transformation show decomposition of expected gradient into reparameterization and correction terms

10 Appendix B: Categorical variables

Let $G_{1:k} = -\log - \log(U_{i:k})$ be samples from the Gumbel distribution, and learnable parameters $(\alpha_1, \ldots, \alpha_k)$ be interpreted as some unnormalized parameterization of the discrete distribution under consideration. Then, consider the following sampling procedure: for each k, find the k that maximizes $\log \alpha_k - G_k$, and then set $D_k = 1$ and $D_{i \neq k} = 0$. The Gumbel-Max trick states that sampling from the discrete distribution is equivalent to taking this argmax, that is, $p(D_k = 1) = \alpha_k / \sum_{i=1}^{n} \alpha_i$.

Since taking an argmax is still a discontinuous operation, Maddison et al. (2016) and Jang et al. (2016) proposed further relaxing the argmax operator through the softmax function with an additional temperature parameter λ:

$$x_k = \frac{\exp\{(\log \alpha_k + G_k)/\lambda\}}{\sum_{i=1}^{n} \exp\{(\log \alpha_i + G_i)/\lambda\}} \tag{30}$$

This relaxation allows values within the simplex, but in the low temperature limit, it becomes exactly the discrete argmax. One limitation of the concrete distribution is that it is a biased estimator except in limiting temperature. In other words, a small amount of bias is present for a non-zero temperature.