BACKPROPAGATION THROUGH THE VOID

Optimizing Control Variates for Black-Box Gradient Estimation

27 Nov 2017, University of Cambridge

Speaker: Geoffrey Roeder, University of Toronto

OPTIMIZING EXPECTATIONS

$$\mathcal{L}(\theta) = \mathbb{E}_{p(b|\theta)} f(b)$$

- Variational inference: Evidence Lower Bound
- Reinforcement learning: Expected Reward Function
- Hard attention mechanism
- How to choose the parameters θ to maximize this expectation?

GRADIENT-BASED OPTIMIZATION

- Reverse-mode automatic differentiation (backpropagation) computes exact gradients of deterministic, differentiable objectives
- Reparameterization trick (Williams, 1992; Kingma & Welling 2014; Rezende 2014): using backprop, gives unbiased, low-variance estimates of gradients of expectations
- This has allows effective stochastic optimization of large probabilistic *continuous* latent-variable models

GRADIENT-BASED OPTIMIZATION: LIMITATIONS

- There many relevant objective functions in ML to which backpropagation **cannot** be applied
- In RL, in fact, the reward function is unknown: a black box from the perspective of an agent
- Discrete latent variable models: discrete sampling creates discontinuities, giving the objective function zero gradient w.r.t. its parameters

GRADIENT-BASED OPTIMIZATION: LIMITATIONS

- But, gradients are appealing: in high dimensions, provides information on how to adjust each parameter individually
- Moreover, stochastic optimization is essential for scalability
- However, are only guaranteed to converge to a fixed point of an objective if a gradient estimator is unbiased

How can we build unbiased stochastic estimators of $\frac{\partial}{\partial \theta}\mathcal{L}(\theta)$?

SCORE-FUNCTION ESTIMATOR (“REINFORCE”, WILLIAMS 1992)

SCORE-FUNCTION ESTIMATOR (“REINFORCE”, WILLIAMS 1992)

$$\frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}f(b) = \int \frac{\partial}{\partial\theta}p(b|\theta)f(b)d\theta$$

SCORE-FUNCTION ESTIMATOR ("REINFORCE", WILLIAMS 1992)

$$\frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}f(b) = \int \boxed{\frac{\partial}{\partial\theta}p(b|\theta)} f(b)d\theta$$

$$= \mathbb{E}_{p(b|\theta)}\left[f(b)\boxed{\frac{\partial}{\partial\theta}\log p(b|\theta)}\right]$$

- **Log-derivative trick** allows us to rewrite gradient of expectation as expectation of gradient (under weak regularity conditions)

SCORE-FUNCTION ESTIMATOR ("REINFORCE", WILLIAMS 1992)

$$\frac{\partial}{\partial\theta}\mathbb{E}_{p(b|\theta)}f(b) = \int \frac{\partial}{\partial\theta}p(b|\theta)f(b)d\theta$$
$$= \mathbb{E}_{p(b|\theta)}\left[f(b)\frac{\partial}{\partial\theta}\log p(b|\theta)\right]$$

$$\hat{g}_{SF} = f(b)\frac{\partial}{\partial\theta}\log p(b|\theta)$$

- **Log-derivative trick** allows us to rewrite gradient of expectation as expectation of gradient (under weak regularity conditions)
- Yields unbiased, but high variance estimator

REPARAMETERIZATION TRICK

$$g_{REP}\left[f(b)\right] = \frac{\partial}{\partial \theta} f(b) = \frac{\partial f}{\partial \mathcal{T}} \frac{\partial \mathcal{T}}{\partial \theta}, b = \mathcal{T}(\theta, \epsilon), \epsilon \sim p(\epsilon)$$

- Requires function to be known and differentiable
- Requires distribution $p(b|\theta)$ to be reparameterizable through a transformation $\mathcal{T}(\theta, \epsilon)$
- Unbiased; lower variance empirically

CONCRETE REPARAMETERIZATION (MADDISON ET AL. 2016)

$$g_{CON}\left[f(b)\right] = \frac{\partial}{\partial\theta} f(b) = \frac{\partial f}{\partial \sigma_\lambda(z)} \frac{\partial \sigma_\lambda(z)}{\partial \theta}, z = \mathcal{T}(\theta, \epsilon), \epsilon \sim p(\epsilon)$$

- Works well with careful hyper parameter choices
- Lower variance than score-function estimator due to reparameterization

- Biased estimator
- Temperature parameter λ
- Requires f to be known and differentiable
- Requires $p(b|\theta)$ to be reparamaterizable

REBAR (TUCKER ET AL. 2017)

- Improves over concrete distribution (*rebar* is stronger than *concrete*)
- Uses continuous relaxation of discrete random variables (concrete) to build unbiased, lower-variance gradient estimator
- Using the reparameterization from the Concrete distribution, construct a control variate for the score-function estimator
- Show how tune additional parameters of the estimator (e.g., temperature λ) online

Digression: control variates for Monte Carlo estimators

CONTROL VARIATES: DIGRESSION

$$\hat{g}_{new}(b) = \hat{g}(b) + \eta\left(c(b) - \mathbb{E}_{p(b)}[c(b)]\right)$$

$$\eta^\star = -\frac{\mathrm{Cov}[\hat{\mathrm{g}}, \mathrm{c}]}{\mathrm{Var}[\hat{g}]}$$

- New estimator is equal in expectation to old estimator (bias is unchanged)
- Variance is reduced when |corr(c, g)| > 0
- We exploit the difference between the function c and its known mean during optimization to "correct" the value of the estimator

CONTROL VARIATES: FREE-FORM

$$\hat{g}_{new}(b) = \hat{g}(b) - c_\phi(b) + \mathbb{E}_{p(b)}\left[c_\phi(b)\right]$$

- If we choose a neural network as our parameterized differentiable function, then the above formulation can be simplified to the above
- The scaling constant will be absorbed into the weights of the network, and optimality is determined by training
- How should we update the weights of the free-form control variate?

What is essential for a control variate?

LEARNING FREE-FORM CONTROL VARIATE: LOSS FUNCTION

$$\begin{aligned}\frac{\partial}{\partial \phi} \mathrm{Var}[\hat{g}] &= \frac{\partial}{\partial \phi}\mathbb{E}[\hat{g}^2] - \frac{\partial}{\partial \phi}\mathbb{E}[\hat{g}]^2 \\ &= \frac{\partial}{\partial \phi}\mathbb{E}[\hat{g}^2] = \mathbb{E}[2\hat{g}\frac{\partial \hat{g}}{\partial \phi}]\end{aligned}$$

- For unbiased estimator, we can form a Monte-Carlo estimate for the variance of the estimator overall
- We use this as the training signal for the parameters of the control variate, adapting the parameters during training

GENERALIZING REBAR

$$\hat{g}_{LAX} = g_{SF}[f] - g_{SF}[c_\phi] + g_{REP}[c_\phi]$$
$$= [f(b) - c_\phi(b)] \frac{\partial}{\partial \theta} \log p(b|\theta) + \frac{\partial}{\partial \theta} c_\phi(b),$$
$$b = \mathcal{T}(\theta, \epsilon), \epsilon \sim p(\epsilon)$$

- Start with score function (SF) estimator of gradient of f
- Introduce a parametrized differentiable function c_ϕ
- Use SF estimator of c_ϕ as a control variate, subtracting its mean estimated through the lower-variance reparameterization estimator
- This generalizes Tucker et al. 2017 to free-form control variates: no longer require continuous relaxations

RELAX: EXTENSION TO DISCRETE RANDOM VARIABLES

$$\hat{g}_{RELAX} = [f(b) - c_\phi(\tilde{z})] \frac{\partial}{\partial \theta} \log p(b|\theta) + \frac{\partial}{\partial \theta} c_\phi(z) - \frac{\partial}{\partial \theta} c_\phi(\tilde{z}),$$

$$z \sim p(z|\theta), b = H(z), \tilde{z} \sim p(z|b, \theta)$$

- When b is discrete, we introduce a related distribution and a function H where $H(z) = b \sim p(b|\theta)$
- We use a conditional reparameterization scheme developed by Tucker et al. 2017 for REBAR
- This estimator is unbiased for all choices of c_ϕ

RELAX: EXTENSION TO DISCRETE RANDOM VARIABLES

$$\hat{g}_{RELAX} = [f(b) - c_\phi(\tilde{z})] \frac{\partial}{\partial\theta} \log p(b|\theta) + \frac{\partial}{\partial\theta} c_\phi(z) - \frac{\partial}{\partial\theta} c_\phi(\tilde{z}),$$

$$z \sim p(z|\theta), b = H(z), \tilde{z} \sim p(z|b, \theta)$$

- When b is discrete, we introduce a related distribution and a function H where $H(z) = b \sim p(b|\theta)$
- We use a conditional reparameterization scheme developed by Tucker et al. 2017 for REBAR
- This estimator is unbiased for all choices of c_ϕ

EXPERIMENTAL RESULTS

SIMPLE EXAMPLE

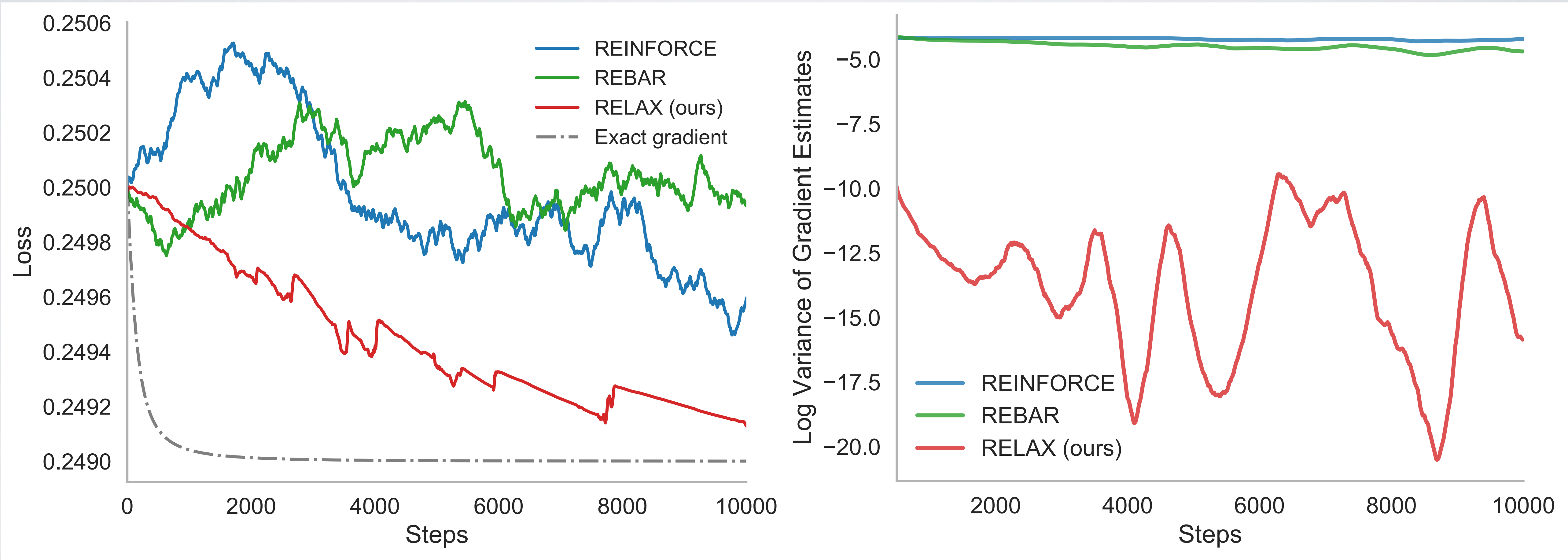
$$\mathbb{E}_{p(b|\theta)}\left[(t-b)^2\right]$$

$$b \sim \mathrm{Ber}(\theta)$$

- Validated idea with simple function above
- Used to validate REBAR estimator, fixing t=0.45
- We chose t = 0.499

SIMPLE EXAMPLE

$$\mathbb{E}_{p(b|\theta)}\left[(t-b)^2\right]$$



- (Right) RELAX finds a reasonable solution, REINFORCE and REBAR oscillate
- (Left) Variance is *considerably* reduced in our estimator

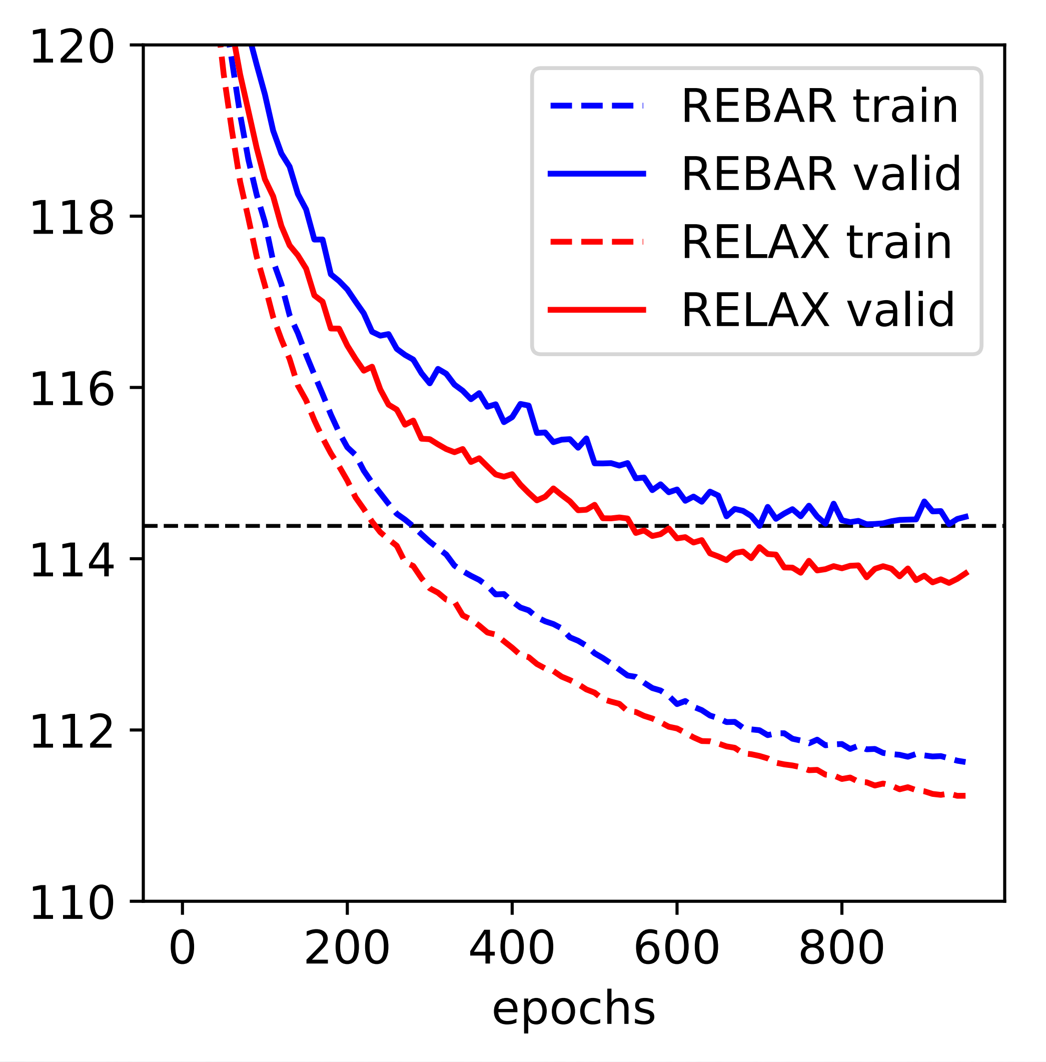
A MORE INTERESTING APPLICATION

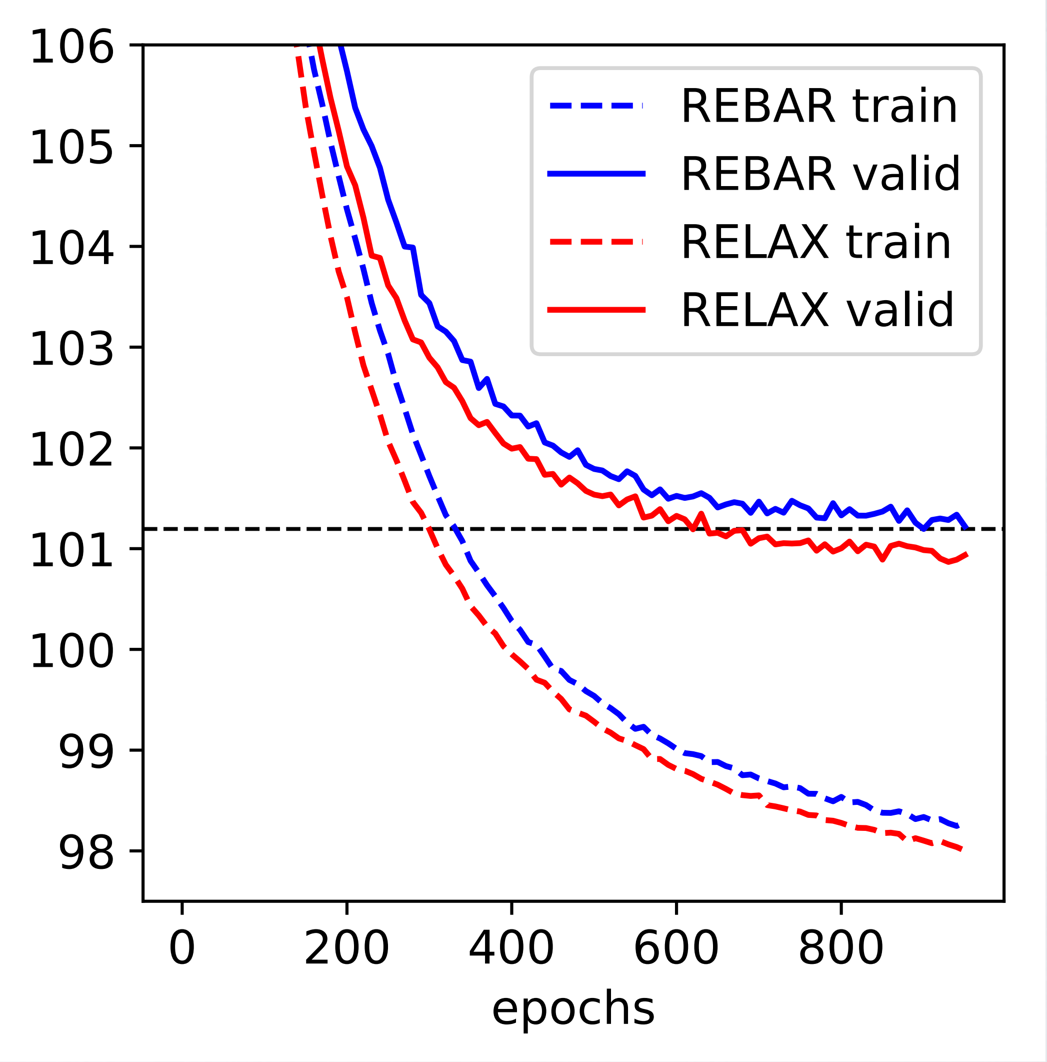
$$\log p(x) \geq \mathcal{L}(\theta) = \mathbb{E}_{q(b|x)}\left[\log p(x|b) + \log p(b) - \log q(b|x)\right]$$

- Discrete Variational Autoencoder
- Latent state: 2 layers of 200 Bernoulli variables
- Discrete sampling renders reparameterization estimator unusable

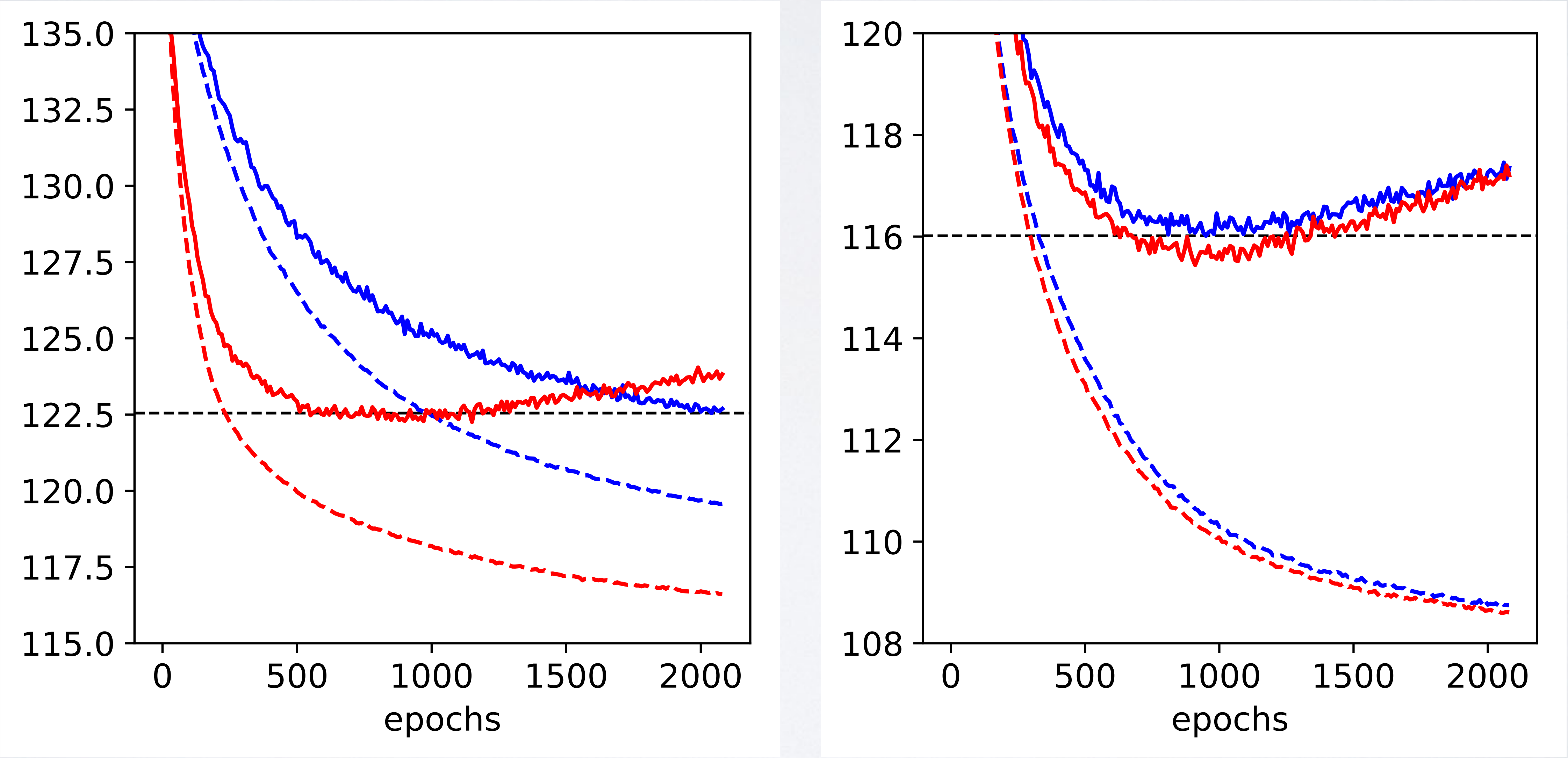
$$c_\phi(z) = f(\sigma(z)) + r_\rho(z)$$

MNIST RESULTS





OMNIGLOT RESULTS

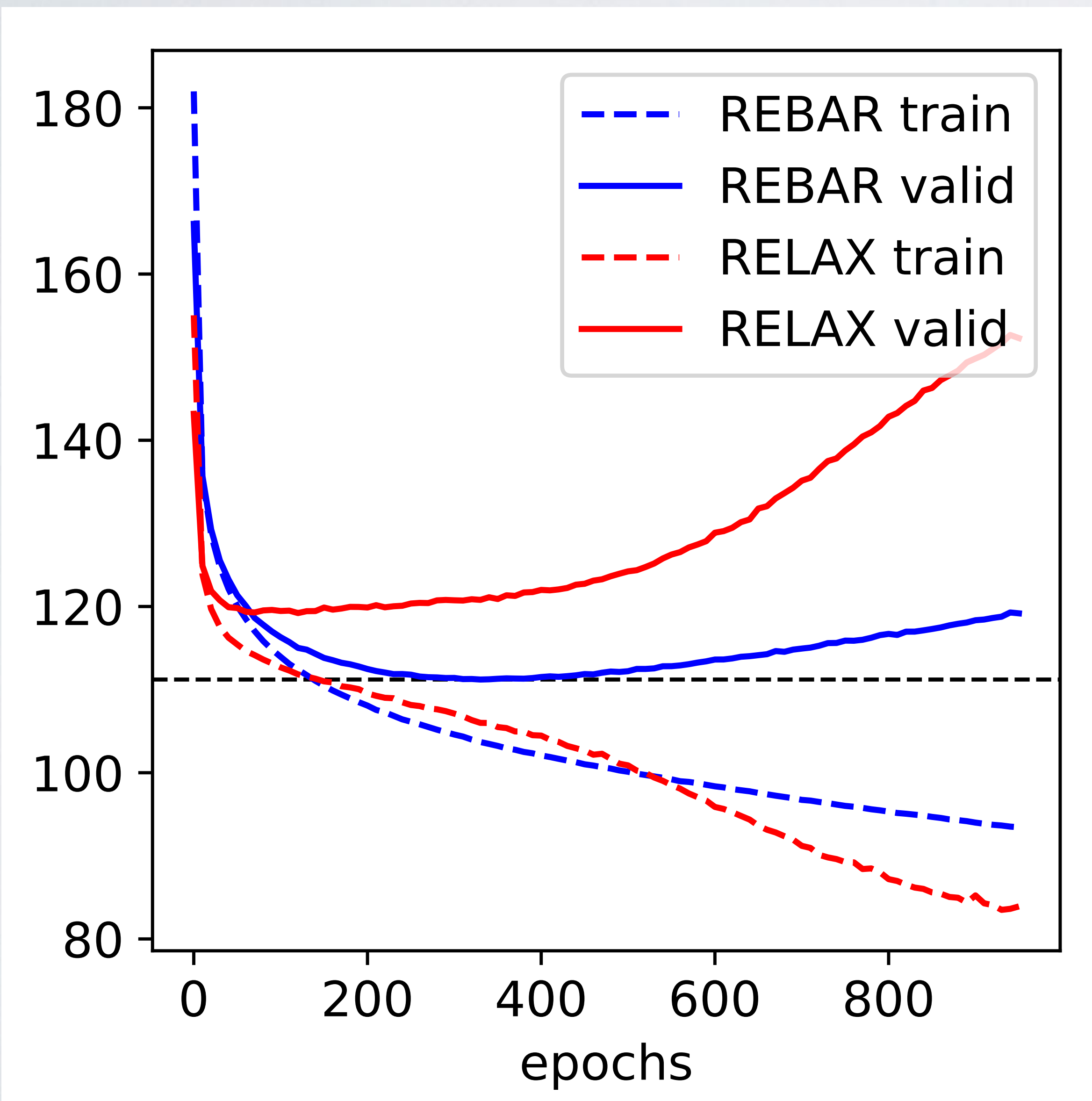


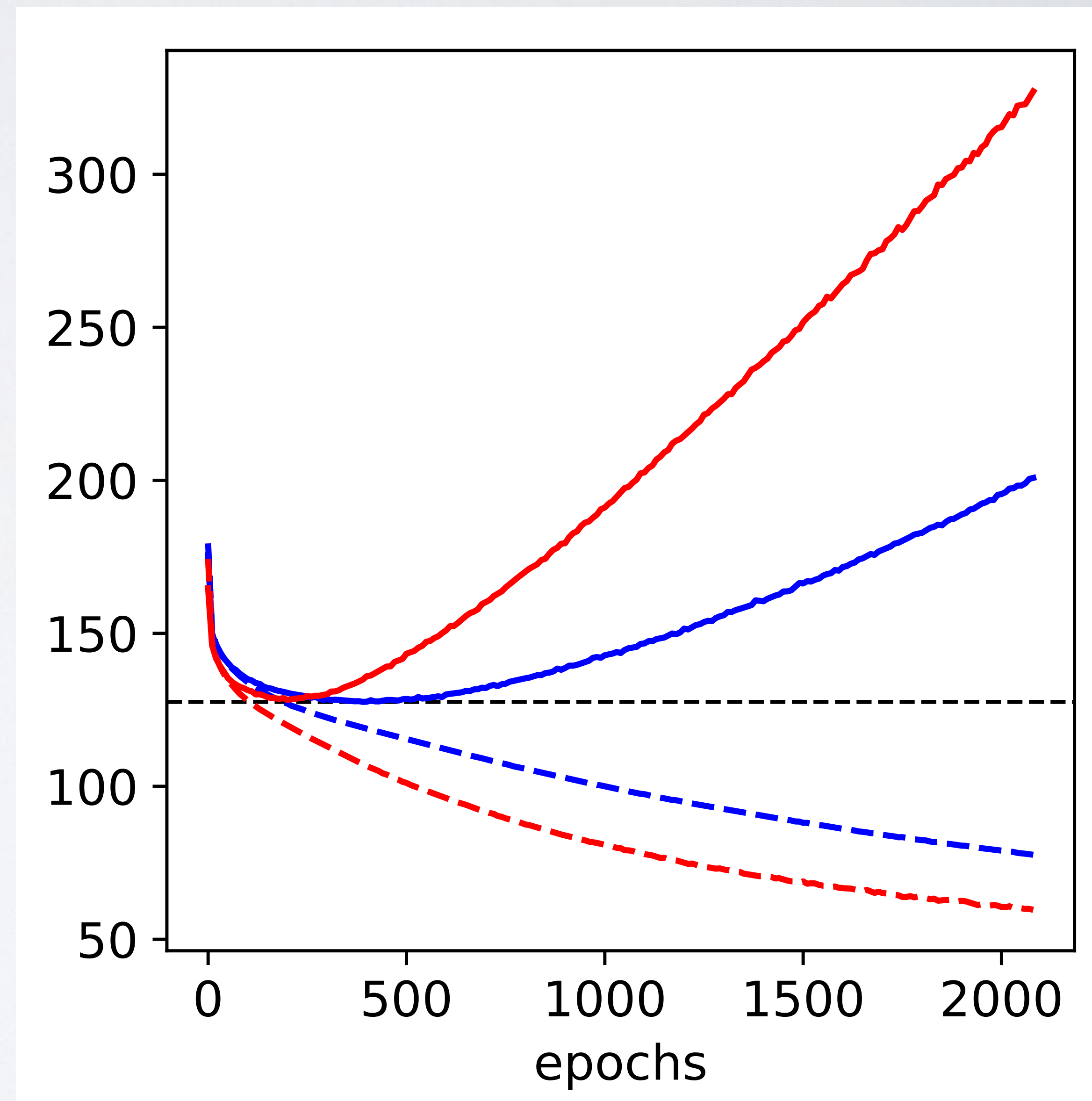
QUANTITATIVE RESULTS

Dataset	Model	Concrete	NVIL	MuProp	REBAR	RELAX
MNIST	Nonlinear	-102.2	-101.5	-101.1	-81.01	**-78.13**
	linear 1 layer	-111.3	-112.5	-111.7	-111.6	**-111.20**
	linear 2 layer	-99.62	-99.6	-99.07	-98.22	**-98.00**
Omniglot	Nonlinear	-110.4	-109.58	-108.72	-56.76	**-56.12**
	linear 1 layer	-117.23	-117.44	-117.09	-116.63	**-116.57**
	linear 2 layer	-109.95	-109.98	-109.55	-108.71	**-108.54**

Table 1: Best obtained training objective for discrete variational autoencoders.

OVERFITTING 1 LAYER: MNIST (LEFT), OMNIGLOT (RIGHT)





REINFORCEMENT LEARNING

- Policy gradient methods effective for finding policy parameters (A2C, A3C, ACKTR)
- Goal: $\operatorname{argmax}_\theta \mathbb{E}_{\tau \sim \pi(\tau|\theta)} [R(\tau)]$
- Need estimate of $\frac{\partial}{\partial \theta} \mathbb{E}_{\tau \sim \pi(\tau|\theta)} [R(\tau)]$
- True reward function unknown (black-box, from environment)

ADVANTAGE ACTOR CRITIC (SUTTON, 2000)

$$\hat{g}_{A2C} = \sum_{t=1}^{\infty} \frac{\partial}{\partial \theta} \log \pi(a_t|s_t, \theta) \left[\sum_{t'=t}^{\infty} r_{t'} - c_\phi(s_t) \right], a_t \sim \pi(a_t|s_t, \theta)$$

- c_ϕ is an estimate of the value function
- This is exactly the REINFORCE estimator using an estimate of the value function as a control variate
- Why not use action in control variate?
- Dependence on action would add bias

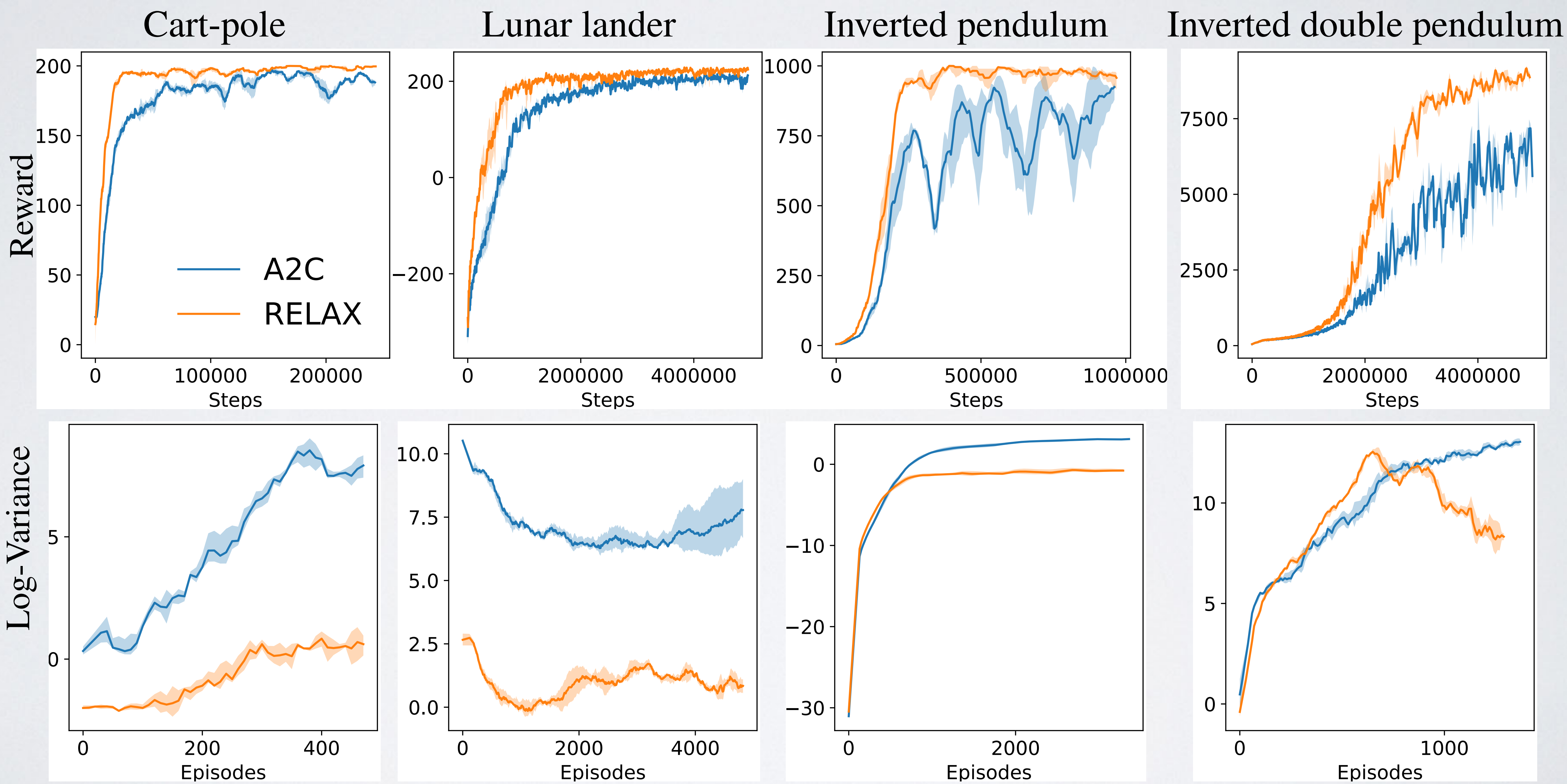
EXTENDING LAX TO RL

$$g_{LAX}^{RL} = \sum_{t=1}^{\infty} \frac{\partial}{\partial \theta} \log \pi(a_t|s_t;\theta) \left[\sum_{t'=t}^{\infty} r_{t'} - c_\phi(a_t, s_t) \right] + \frac{\partial}{\partial \theta} c_\phi(a_t, s_t)$$

$$a_t = a_t(\epsilon_t, s_t, \theta), \epsilon_t \sim p(\epsilon_t)$$

- Allows for action-dependence in control variate
- Remains unbiased estimator
- Similar extension possible for discrete action spaces, see paper Appendix C.2

RL BENCHMARK RESULTS



BERNOULLI REPARAM

Bernoulli When $p(b|\theta)$ is Bernoulli distribution we let $H(z) = \mathbb{I}(z > 0)$ and we sample from $p(z|\theta)$ with

$$z = \log \frac{\theta}{1-\theta} + \log \frac{u}{1-u}, \qquad u \sim \text{uniform}[0,1].$$

We can sample from $p(z|b,\theta)$ with

$$v' = \begin{cases} v \cdot (1-\theta) & b = 0 \\ v \cdot \theta + (1-\theta) & b = 1 \end{cases}$$

$$\tilde{z} = \log \frac{\theta}{1-\theta} + \log \frac{v'}{1-v'}, \qquad v \sim \text{uniform}[0,1].$$