Reinforced Variational Inference

Theophane Weber
Google DeepMind

(joint work with: Nicolas Heess, Ali Eslami, John Schulman, David Wingate, David Silver)

This talk

- Variational Inference: powerful method for leveraging optimization techniques in inference problems
- Reinforcement Learning: powerful framework for sequential decision making under uncertainty
- We formalize mapping from variational inference to reinforcement learning
  - Unifies many concepts in variational inference from a graphical standpoint
  - Derive new methods by leveraging known RL ideas
  - Derive intuition about when variational inference is hard

Previous work

Control as inference: a rich field

- Dayan and Hinton, *Using Expectation-Maximization for Reinforcement Learning* (1997)
- Furmston and Barber, *Variational Methods for Reinforcement Learning* (2010)
- Botvinick and Toussaint, *Planning as probabilistic inference* (2012)
- Rawlik et al. *On Stochastic Optimal Control and Reinforcement Learning by Approx. Inference* (2012)

Inference as RL is more recent, and less developed:

- Wingate *A Reinforcement Learning approach to Variational Inference* (2012)
- Mnih and Gregor, *Neural Variational Inference* (2014)
- Bachman, Precup, *Data Generation as Sequential Decision Making* (2015)
- Schulman, Heess, W., Abbeel, *Gradient estimation using Stochastic Computation Graph* (NIPS15)

Modern Variational Inference

Variational inference was recently revolutionized by two key ideas:

- Turnkey: ‘Automated’ / ‘black-box’ inference by general purpose Monte Carlo estimates of the cost function gradient.

- Faster and scalable: Amortized inference (data-conditional fast inference schemes), minibatches in VI (‘SVI’)

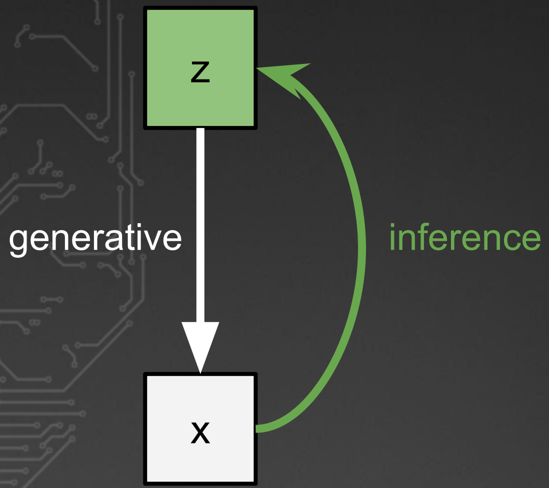
Variational Inference

Objective function

$$\mathcal{L}(\theta) = \int_z q_\theta(z|x) \log p(x|z) - KL(q_\theta(z|x), p(z))$$

Stochastic gradient estimate
(score function method)

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}\left[\nabla_\theta \log q_\theta(z|x) \log\left(\frac{p(z,x)}{q_\theta(z|x)}\right)\right]$$



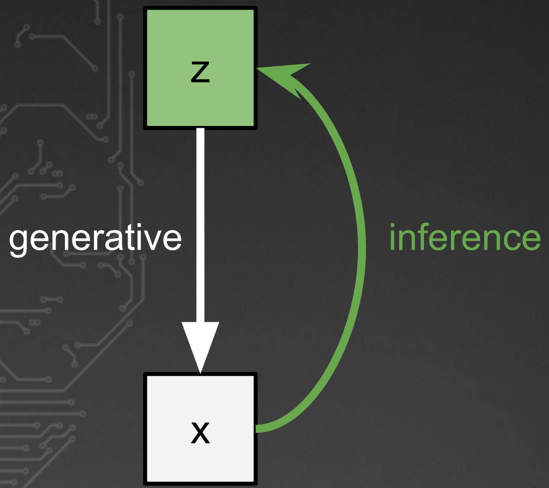
Variational Inference

Objective function

$$\mathcal{L}(\theta) = \int_z q_\theta(z|x) \log p(x|z) - KL(q_\theta(z|x), p(z))$$

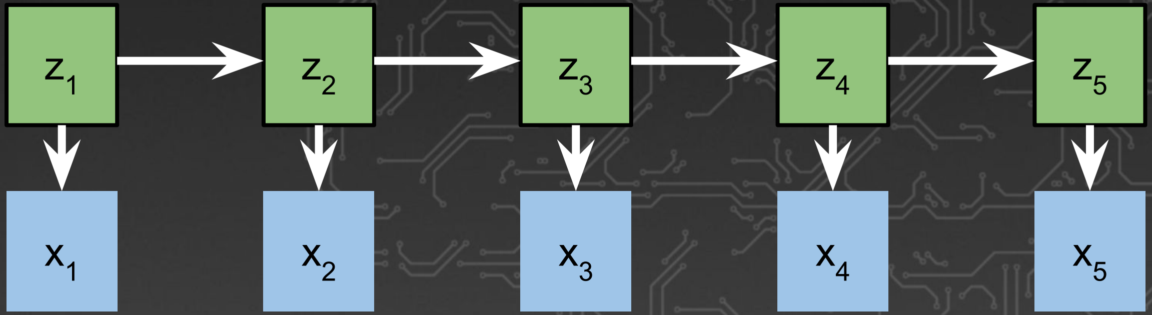
Stochastic gradient estimate
(score function method)

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}\left[\nabla_\theta \log q_\theta(z|x) \log\left(\frac{p(z,x)}{q_\theta(z|x)}\right)\right]$$



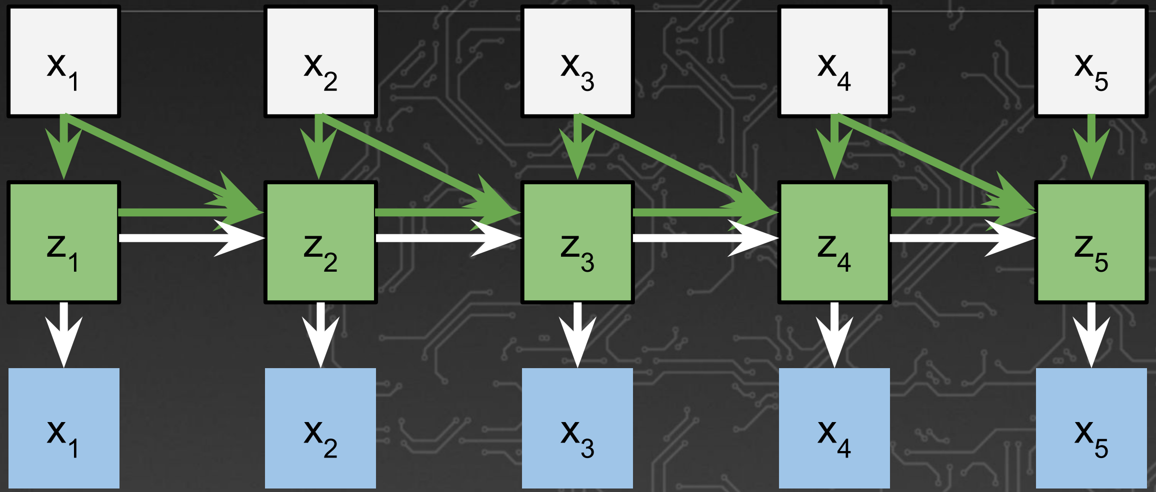
$\Rightarrow$ Issues with: variance of estimate, credit assignment

An example: time series with inference network



$$p(z, x) = \prod_t p(z_t | z_{t-1}) p(x_t | z_t)$$

An example: time series with inference network

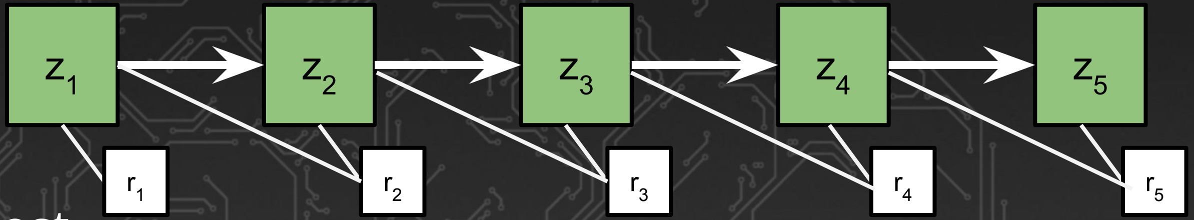


$$q(z|x) = \prod_t q(z_t|z_{t-1}, x_t, x_{t-1})$$

$$p(z, x) = \prod_t p(z_t|z_{t-1})p(x_t|z_t)$$

$$\mathcal{L} = \mathbb{E}\left[\sum_t \log p(z_t|z_{t-1}) + \log p(x_t|z_t) - \log q(z_t|x_t, x_{t-1}, z_{t-1})\right]$$

An example: time series with inference network



Decomposing the cost

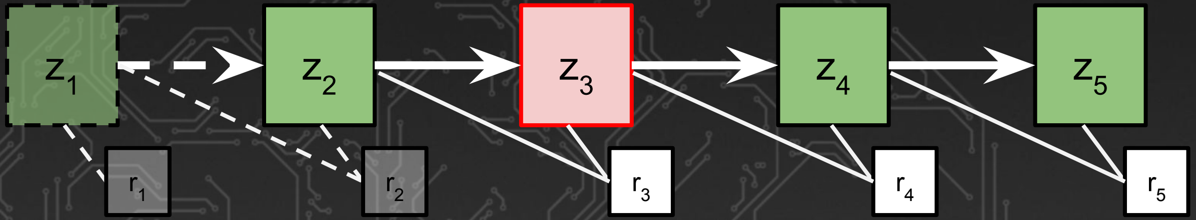
$$\mathcal{L} = \mathbb{E}\left[\sum_t \log p(z_t|z_{t-1}) + \log p(x_t|z_t) - \log q(z_t|x_t, x_{t-1}, z_{t-1})\right]$$

$$\mathcal{L} = \mathbb{E}\left[\sum_t r_t\right]$$

$$r_t = \log p(z_t|z_{t-1}) + \log p(x_t|z_t) - \log q(z_t|x_t, x_{t-1}, z_{t-1})$$

An example: time series with inference network

At time t=3⁻



State: z_2

Stochastic action: $z_3 \sim q_\theta(z_3)$

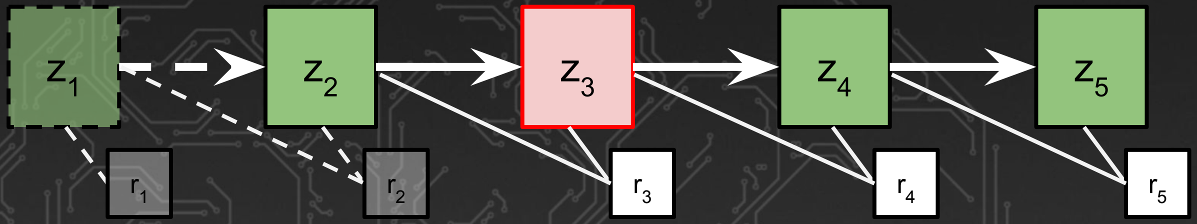
Stochastic gradient:

$$\nabla_\theta \mathbb{E}[R] = \mathbb{E}[\nabla_\theta \log q_\theta(z_3) R]$$

$$\approx \nabla_\theta \log q_\theta(z_3) R, z_3 \sim q_\theta(z_3)$$

An example: time series with inference network

At time t=3$^-$



State: z_2

Stochastic action: $z_3 \sim q_\theta(z_3)$

Sequential decision

Stochastic gradient:

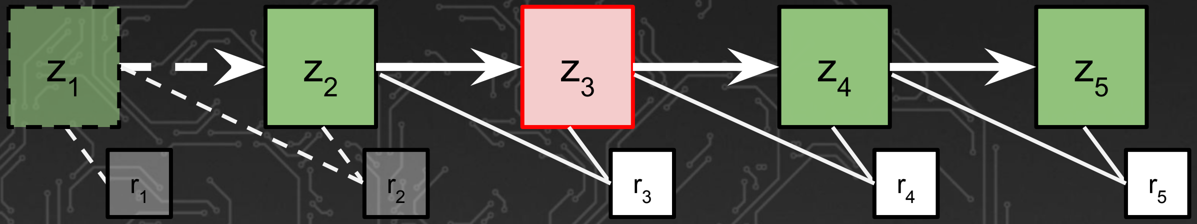
$$\nabla_\theta \mathbb{E}[R] = \mathbb{E}[\nabla_\theta \log q_\theta(z_3) R]$$

$$\approx \nabla_\theta \log q_\theta(z_3) R, z_3 \sim q_\theta(z_3)$$

Noise in gradient hinders learning

An example: time series with inference network

At time t=3⁻



State: z_2

Stochastic action: $z_3 \sim q_\theta(z_3)$

Sequential decision

Stochastic gradient:

$$\nabla_\theta \mathbb{E}[R] = \mathbb{E}[\nabla_\theta \log q_\theta(z_3) R]$$

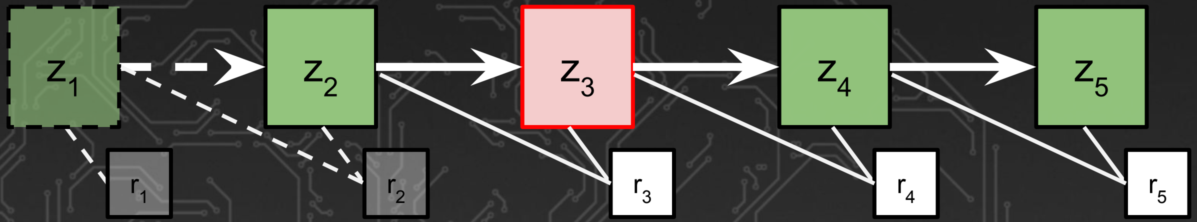
$$\approx \nabla_\theta \log q_\theta(z_3) R, z_3 \sim q_\theta(z_3)$$

Noise in gradient hinders learning

RL deals with sequential decision making, and has developed techniques in variance reduction

A first idea: Value functions

At time $t=3^-$



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t) R$$

$$\approx \sum_t \nabla_\theta \log q_\theta(z_t) R_t \quad \text{with} \quad R_t = r_t + r_{t+1} + .. + r_T$$

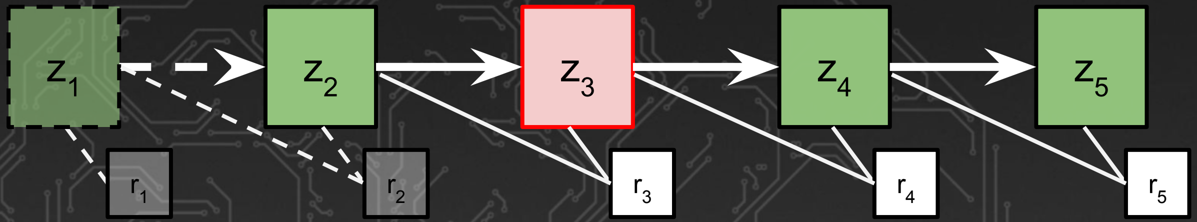
$$\approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - b) \quad \text{since} \quad \mathbb{E}[\nabla_\theta \log q_\theta(z_t)] = 0$$

Appropriate value of b reduces variance - what value to use?

$$b^* = \frac{\mathbb{E}[\nabla \log q_\theta^2 R_t]}{\mathbb{E}[\nabla \log q_\theta^2]}$$

A first idea: Value functions

At time t=3⁻



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t) R$$

$$\approx \sum_t \nabla_\theta \log q_\theta(z_t) R_t \quad \text{with} \quad R_t = r_t + r_{t+1} + .. + r_T$$

$$\approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - b) \quad \text{since} \quad \mathbb{E}[\nabla_\theta \log q_\theta(z_t)] = 0$$

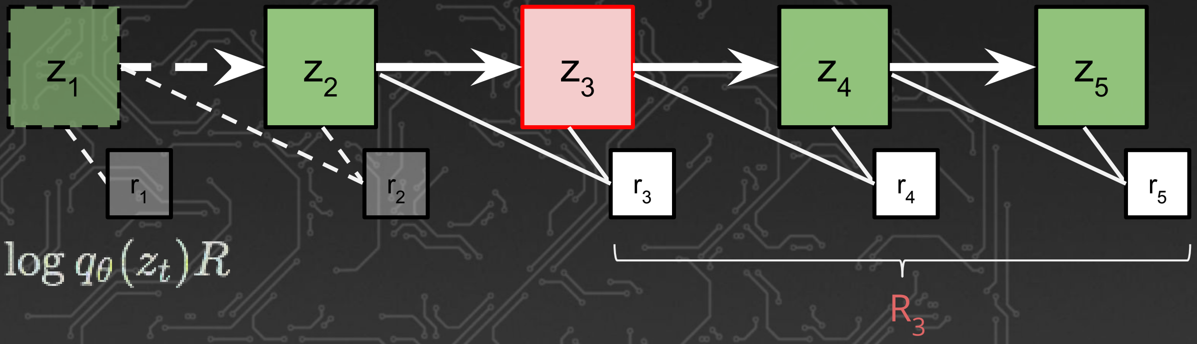
Appropriate value of b reduces variance - what value to use?

$$b^* = \frac{\mathbb{E}[\nabla \log q_\theta^2 R_t]}{\mathbb{E}[\nabla \log q_\theta^2]}$$

perhaps not ideal
> multidimensional
> low intuition

A first idea: Value functions

At time t=3⁻



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t) R$$

$$\approx \sum_t \nabla_\theta \log q_\theta(z_t) R_t \quad \text{with} \quad R_t = r_t + r_{t+1} + .. + r_T$$

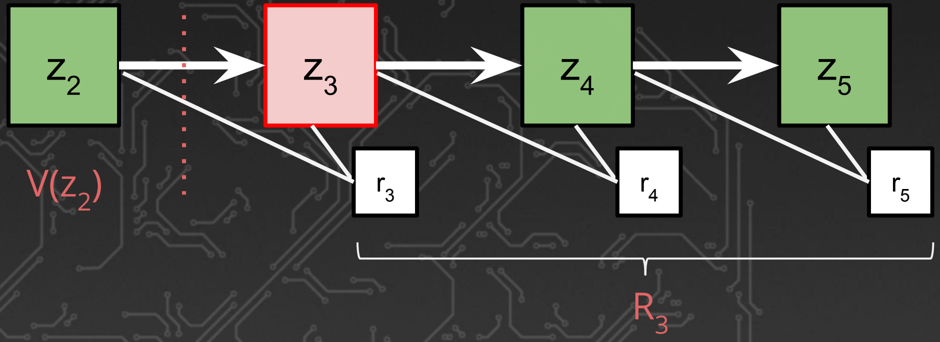
$$\approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - b) \quad \text{since} \quad \mathbb{E}[\nabla_\theta \log q_\theta(z_t)] = 0$$

Appropriate value of b reduces variance - what value to use?

$$b = \mathbb{E}[R_t]$$

A first idea: Value functions

At time t=3⁻



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t) R$$

$$\approx \sum_t \nabla_\theta \log q_\theta(z_t) R_t \quad \text{with} \quad R_t = r_t + r_{t+1} + .. + r_T$$

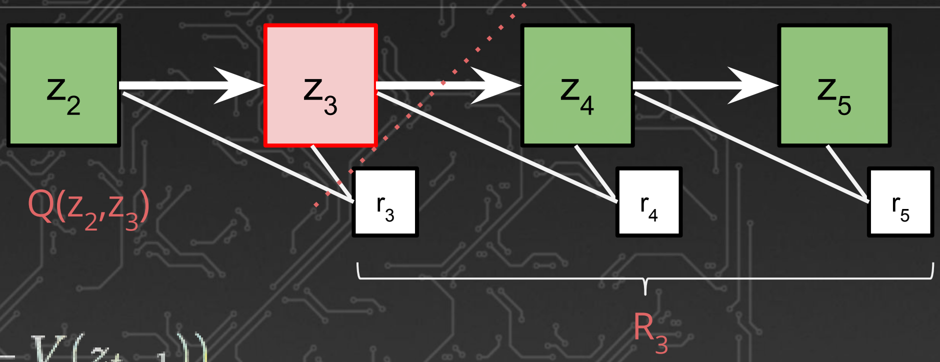
$$\approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - b) \quad \text{since} \quad \mathbb{E}[\nabla_\theta \log q_\theta(z_t)] = 0$$

Appropriate value of b reduces variance - what value to use?
Can use state-conditional value function!

$$b = V(z_{t-1}) = \mathbb{E}[R_t | z_{t-1}]$$

A second idea: critics

At time t=3-



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - V(z_{t-1}))$$

Can further reduce variance by replacing return by its expectation over future choices

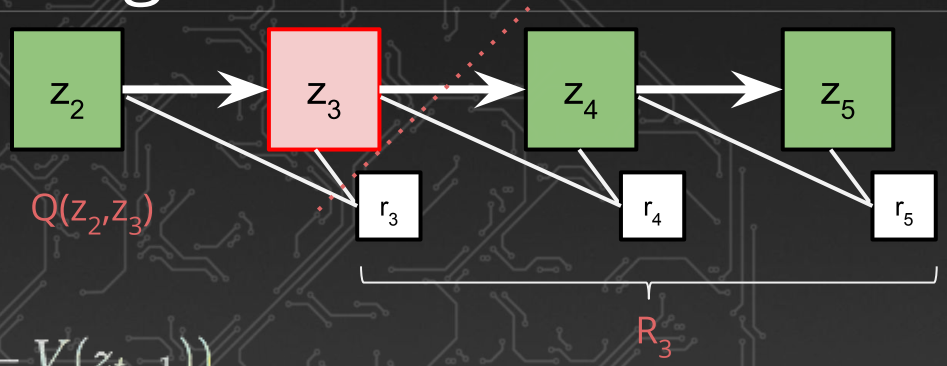
$$\mathbb{E}[\nabla_\theta \log q_\theta(z_t) R_t] = \mathbb{E}_{z_t}[\nabla_\theta \log q_\theta(z_t) \mathbb{E}_{z_{>t}}[R_t]]$$

define: $Q(z_{t-1}, z_t) = \mathbb{E}_{z_{>t}}[R_t]$ — critic

$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t)(Q(z_{t-1}, z_t) - V(z_{t-1}))$$

A third idea: advantage functions

At time t=3⁻



$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t)(R_t - V(z_{t-1}))$$

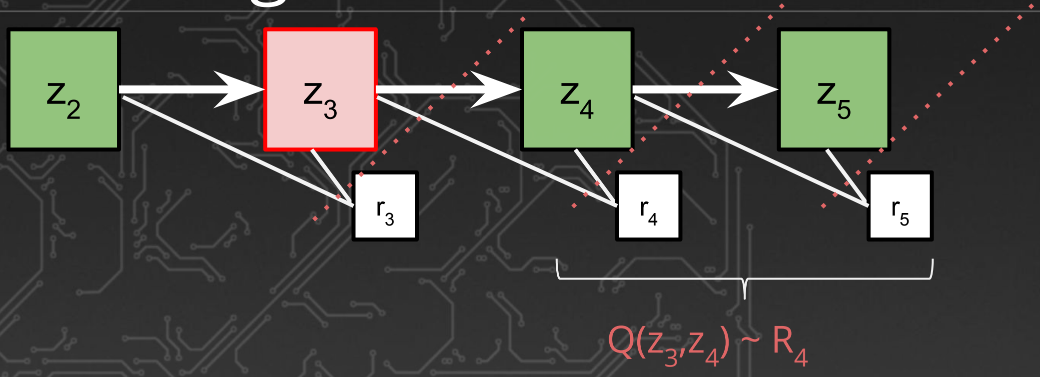
Can further reduce variance by replacing return by its expectation over future choices

$$\mathbb{E}[\nabla_\theta \log q_\theta(z_t) R_t] = \mathbb{E}_{z_t}[\nabla_\theta \log q_\theta(z_t) \mathbb{E}_{z_{>t}}[R_t]]$$

$$\nabla_\theta \mathbb{E}[R] \approx \sum_t \nabla_\theta \log q_\theta(z_t) A(z_{t-1}, z_t)$$

A fourth idea: TD learning

At time t=3⁻

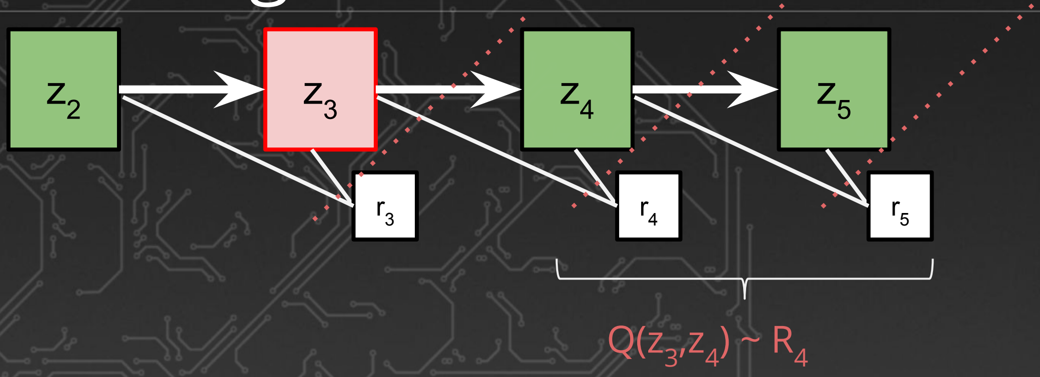


Advantage estimates:

$$Q(z_2, z_3) - V(z_2)$$

A fourth idea: TD learning

At time t=3⁻



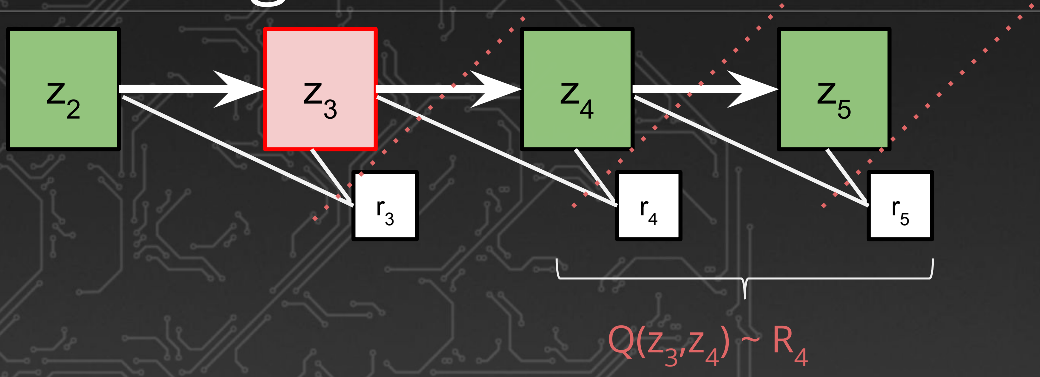
Advantage estimates:

$$Q(z_2, z_3) - V(z_2)$$

$$r_3 + Q(z_3, z_4) - V(z_2)$$

A fourth idea: TD learning

At time t=3⁻



Advantage estimates:

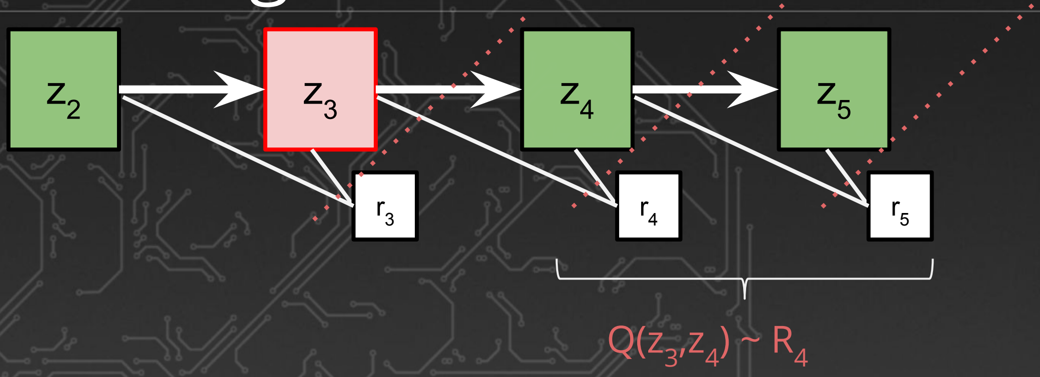
$$Q(z_2, z_3) - V(z_2)$$

$$r_3 + Q(z_3, z_4) - V(z_2)$$

$$r_3 + r_4 + Q(z_4, z_5) - V(z_2)$$

A fourth idea: TD learning

At time t=3⁻



Advantage estimates:

$$Q(z_2, z_3) - V(z_2)$$

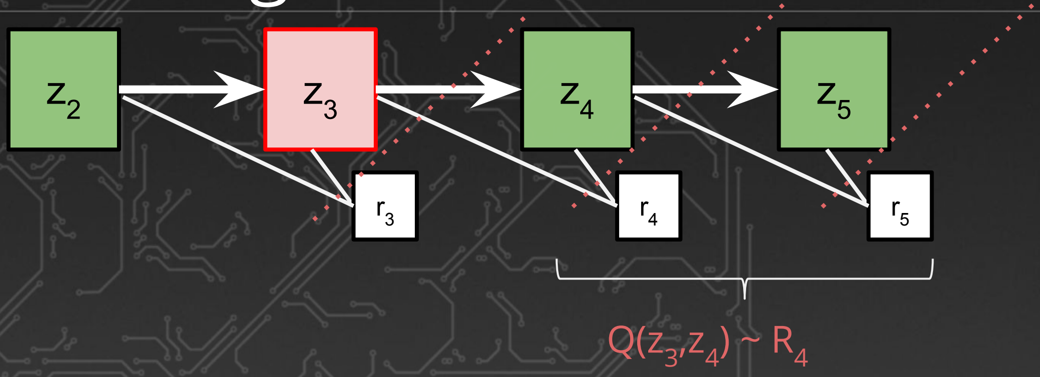
$$r_3 + Q(z_3, z_4) - V(z_2)$$

$$r_3 + r_4 + Q(z_4, z_5) - V(z_2)$$

$$r_3 + r_4 + r_5 - V(z_2)$$

A fourth idea: TD learning

At time t=3⁻

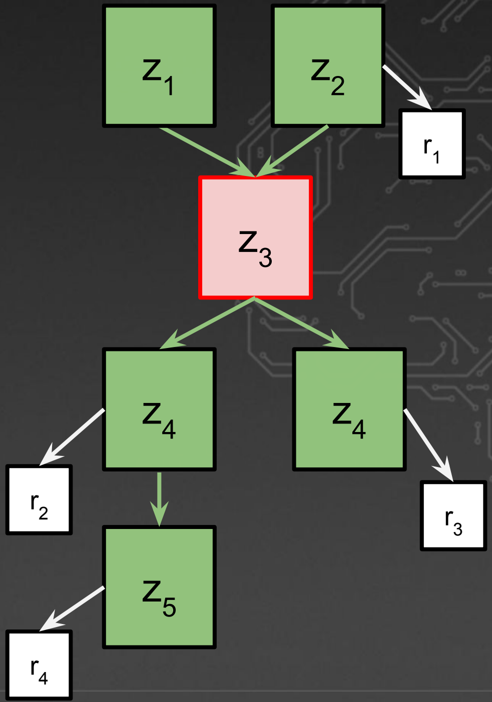


Advantage estimates:

$$Q(z_2, z_3) - V(z_2)$$
$$r_3 + Q(z_3, z_4) - V(z_2)$$
$$r_3 + r_4 + Q(z_4, z_5) - V(z_2)$$
$$r_3 + r_4 + r_5 - V(z_2)$$

can be combined

Arbitrary graphs (stochastic computation graph)



State: (z_1, z_2)

Downstream costs: $R_3 = r_2 + r_3 + r_4$

Value function: $V(z_1, z_2) = \mathbb{E}[R_3 | z_1, z_2]$

Critic: $Q(z_1, z_2, z_3) = \mathbb{E}[R_3 | z_1, z_2, z_3]$

Stochastic gradient estimate:

$$\nabla_\theta \log q_\theta(z_3)(Q(z_1, z_2, z_3) - V(z_1, z_2))$$

The high level general mapping

Generic expectation		RL		VI	
Optimization var.	θ	Policy param.	θ	Variational param.	θ
Integration var.	y	Trajectory	τ	Latent trace	z
Distribution	$p_\theta(y)$	Trajectory dist.	$p_\theta(\tau)$	Posterior dist.	$q_\theta(z\|x)$
Integrand	$f(y)$	Total return	$R(\tau)$	Free energy	$\log\left(\frac{p(x,z)}{q_\theta(z\|x)}\right)$

Conceptual mapping

Variational Inference:	Reinforcement learning
Log partition function	Expected total cost
Free-energies	Rewards
Rao-blackwellized free energies	Returns
Mean-field posterior	Open-loop control
Structured posterior	Closed-loop control
Per data point inference	Trajectory optimization
Amortized inference	Context-based control
Baselines	Value function
???	Critics
???	TD-learning
???	Exploration
???	Experience replay
???	Your favorite RL technique
???	...

Conceptual mapping

Variational Inference:	Reinforcement learning
Log partition function	Expected total cost
Free-energies	Rewards
Rao-blackwellized free energies	Returns
Mean-field posterior	Open-loop control
Structured posterior	Closed-loop control
Per data point inference	Trajectory optimization
Amortized inference	Context-based control
Baselines	Value function
???	Critics
???	TD-learning
???	Exploration
???	Experience replay
???	Your favorite RL technique
???	...

so much to explore!

Sequential mapping

	RL	VIMDP
Context	—	x
Dynamic state	s_t	z_{k-1}
State	s_t	(z_{k-1}, x)
Action	a_t	$z_k \sim q_\theta(z_k \mid z_{k-1}, x)$
Transition	$(s_t, a_t) \rightarrow s_{t+1} \sim P(s \mid s_t, a_t)$	$((z_{k-1}, x), z_k) \rightarrow (z_k, x)$
Instant reward	r_t	$\log\left(\frac{p(z_k \mid z_{k-1}, x)}{q_\theta(z_k \mid z_{k-1}, x)}\right)$
Final reward	0	$\log p(x \mid z_K)$