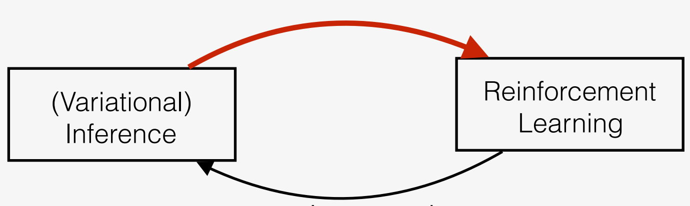
Reinforced Variational Inference

Théophane Weber[1], Nicolas Heess[1], Ali Eslami[1], John Schulman[2], David Wingate[3], David Silver[1]

[1]Google DeepMind [2] University of California, Berkeley [3] Brigham Young University

Overview

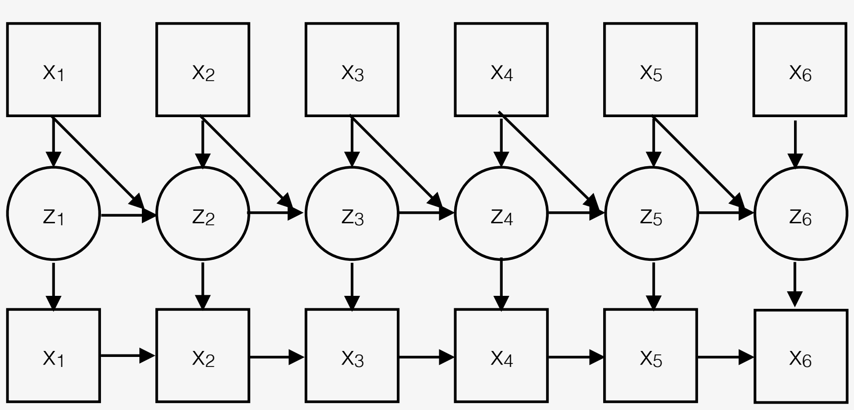
- Variational Inference: Powerful method that leverages optimization technique for inference problems
- Reinforcement Learning: Powerful framework for sequential decision making under uncertainty



⇒ Unifies many concepts of VI from an RL standpoint.

⇒ Suggests new algorithms and methods for approximate inference.

Example: time series with inference network



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

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

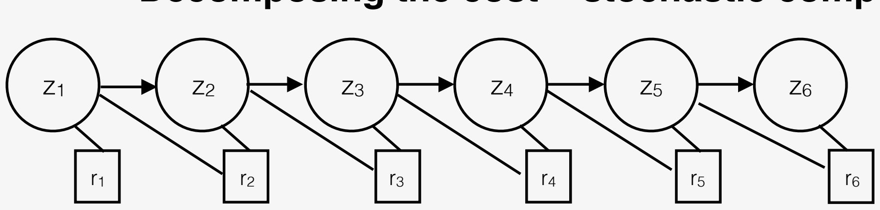
Cost function

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

Stochastic gradient (score function method)

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

Decomposing the cost - stochastic computation graph



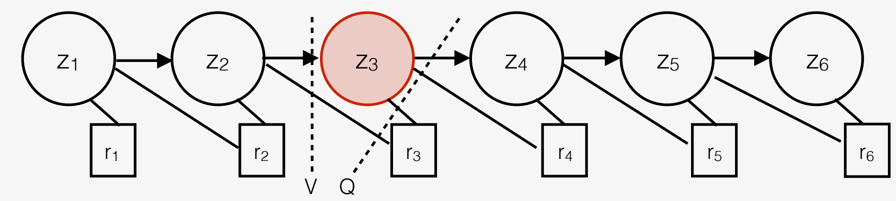
Factored prior and posterior ⇒ cost can be distributed across time steps

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

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

Problem takes the form of sequential decision making

Policy gradient, Value functions and Critics



Classically, REINFORCE gradient: $\nabla_\theta \log q_\theta(z_3)(R_3 - b)$ $\quad R_3 = \sum_{t\geq 3} r_t$

Novel Advantage estimate for stochastic gradients:

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

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

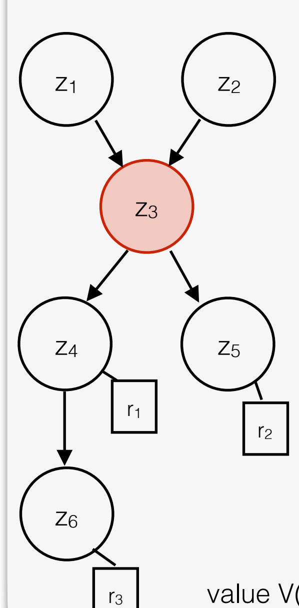
$Q(z_2,z_3) = \mathbb{E}[R_3|z_2,z_3]$ Critic

Similar to ACTOR-CRITIC algorithms in RL. Trades variance for possible bias in a flexible fashion.

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	$\log\left(\frac{p(x,z)}{q_\theta(z\|x)}\right)$

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



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
???	...

References

- [NVIL] Mnih & Gregor. *Neural variational inference and learning in belief networks* (2014).
- [SCG] Schulman, Heess, W., Abbeel, Gradient estimation using Stochastic Computation Graph (2015)
- [EMRL] Dayan & Hinton, *Using Expectation-Maximization for Reinforcement Learning* (1997)
- [VMRL] Furmston & Barber, *Variational Methods for Reinforcement Learning* (2010)
- [PLANINF] Botvinick & Toussaint, Planning as probabilistic inference (2012)
- [RLINF] Rawlik et al. On Stoc. Optimal Control and Reinforcement Learning by Approx. Inference (2012)
- [DATASEQ] Bachman & Precup, Data Generation as Sequential Decision Making (2015)
- [REINFORCE] Williams *Simple statistical gradient-following algorithms for connectionist reinforcement learning* (1992)