Reinforcement Learning through the Lenses of Variational Inference

Sergey Bartunov
Research Scientist, DeepMind

Outline

Outline

- (very brief) Introduction to Reinforcement Learning

Outline

- (very brief) Introduction to Reinforcement Learning
- Entropy-regularized RL via variational inference

Outline

- (very brief) Introduction to Reinforcement Learning
- Entropy-regularized RL via variational inference
- Policy gradients

Outline

- (very brief) Introduction to Reinforcement Learning
- Entropy-regularized RL via variational inference
- Policy gradients
- Stable policy gradients

Outline

- (very brief) Introduction to Reinforcement Learning
- Entropy-regularized RL via variational inference
- Policy gradients
- Stable policy gradients
- Hierarchical RL with Options as auxiliary variables

Unsupervised Learning

Prediction

Gregor et al, 2016

$p(x)$

Model

Ground-truth

Objective $\log p(x)$

Supervised Learning

Observation

$p(y|x)$

Model

Prediction

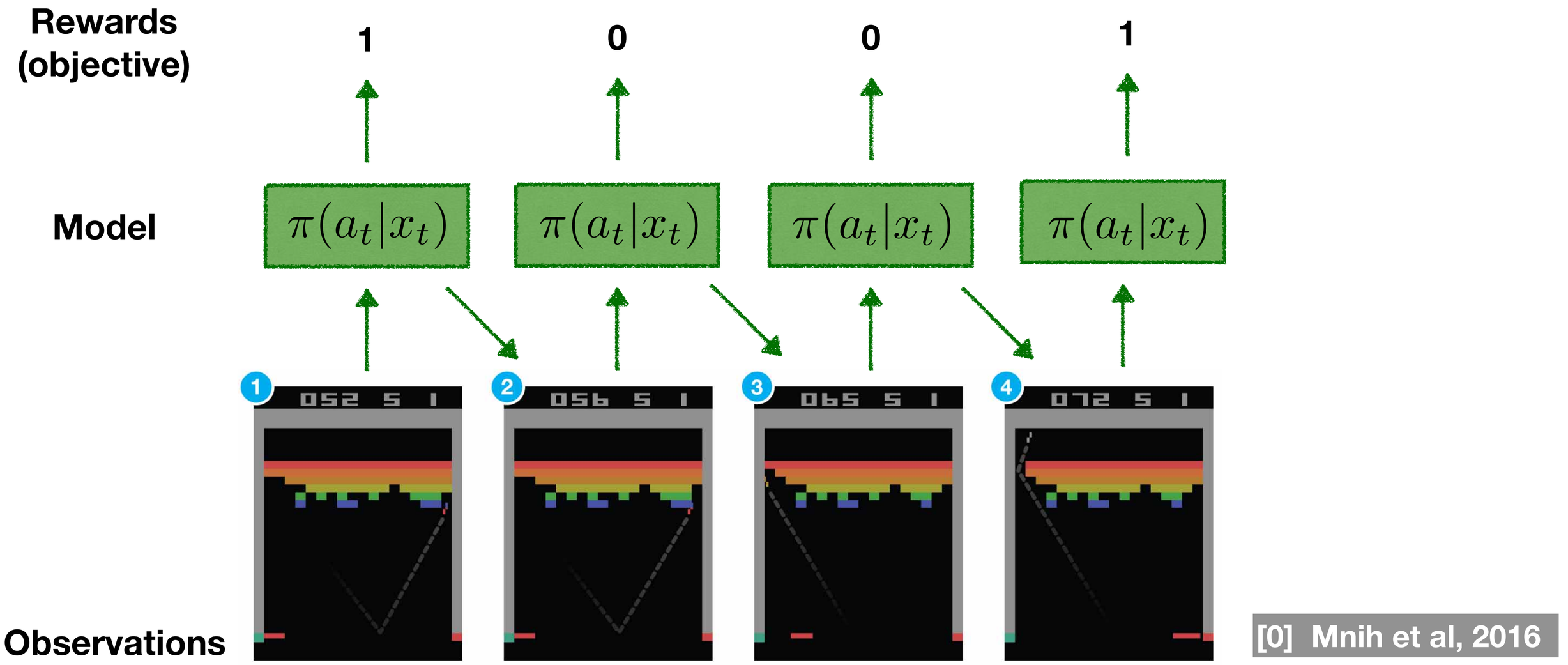
cat

dog

Correct label

Objective $\log p(y|x)$

Reinforcement Learning



[0] Mnih et al, 2016

Reinforcement Learning

Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making

Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- Potentially sparse, noisy and delayed rewards
- Credit assignment problem

Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- Potentially sparse, noisy and delayed rewards
  - Credit assignment problem
- Exploration / exploitation trade-off

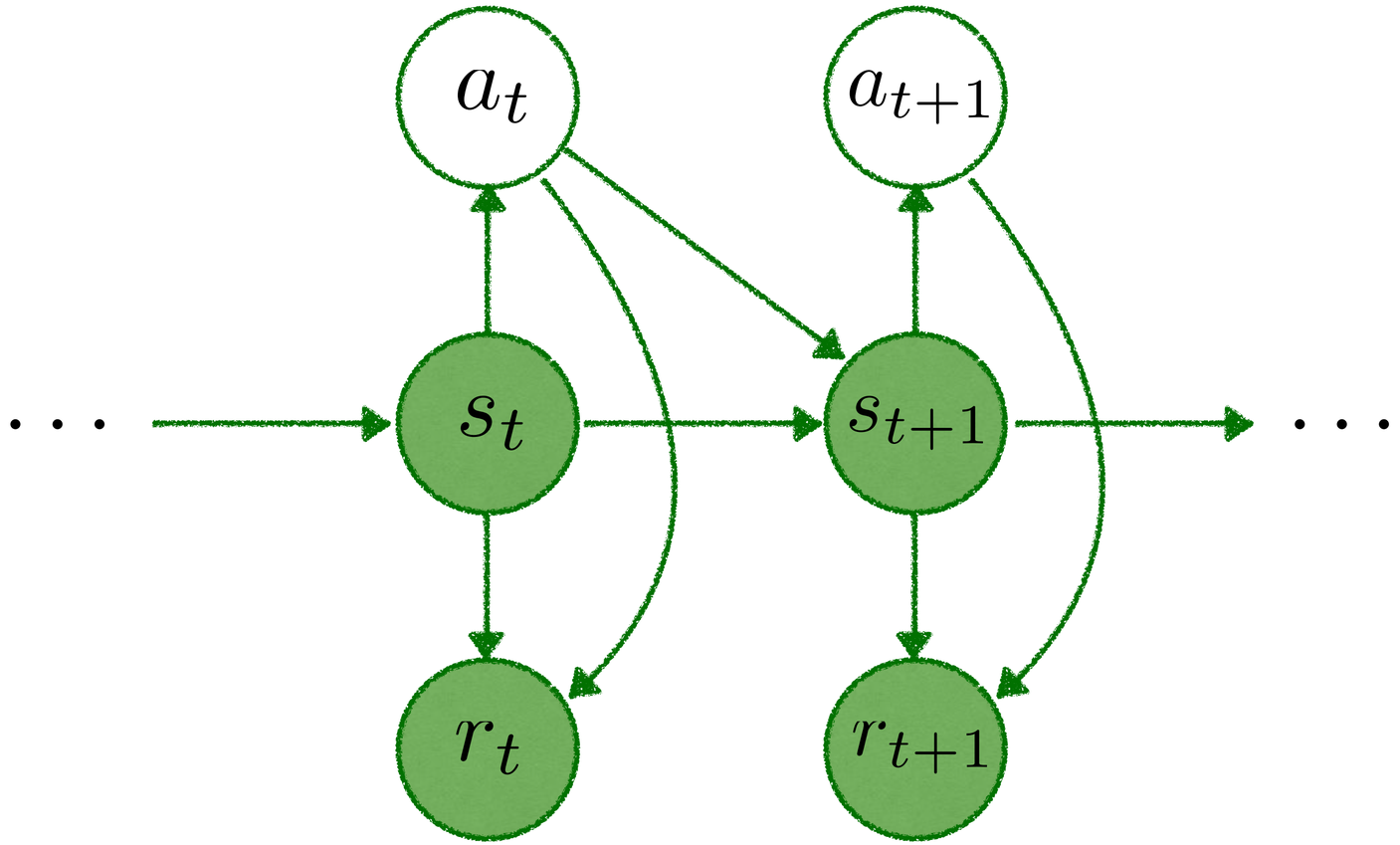
Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- Potentially sparse, noisy and delayed rewards
  - Credit assignment problem
- Exploration / exploitation trade-off
- Partial observability

Sutton & Barto, 1998

Markov Decision Process

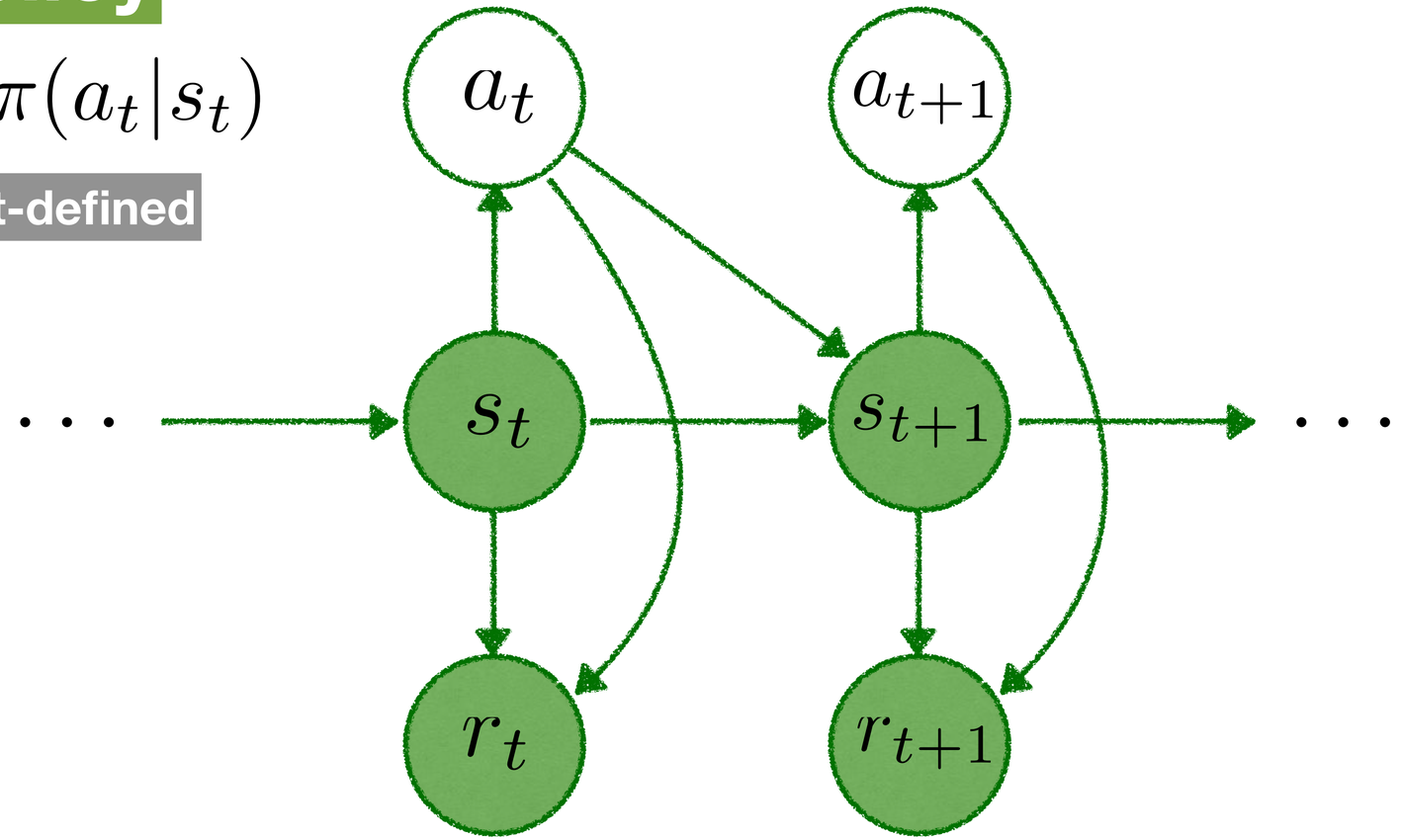


Markov Decision Process

Policy

$$a_t \sim \pi(a_t|s_t)$$

Agent-defined

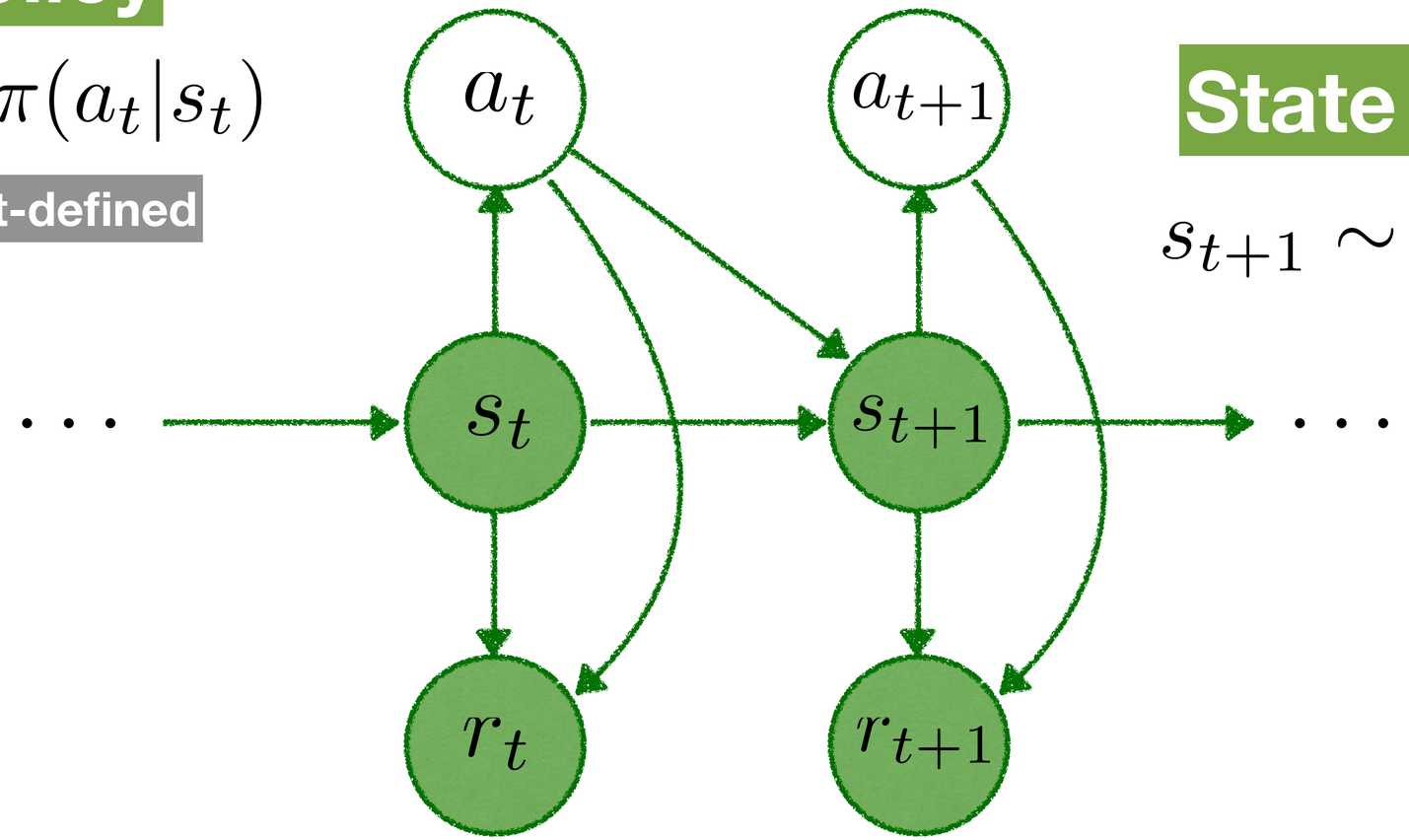


Markov Decision Process

Policy

$$a_t \sim \pi(a_t|s_t)$$

Agent-defined



State transitions

$$s_{t+1} \sim p(s_{t+1}|s_t, a_t)$$

Likely unknown

Markov Decision Process

Policy

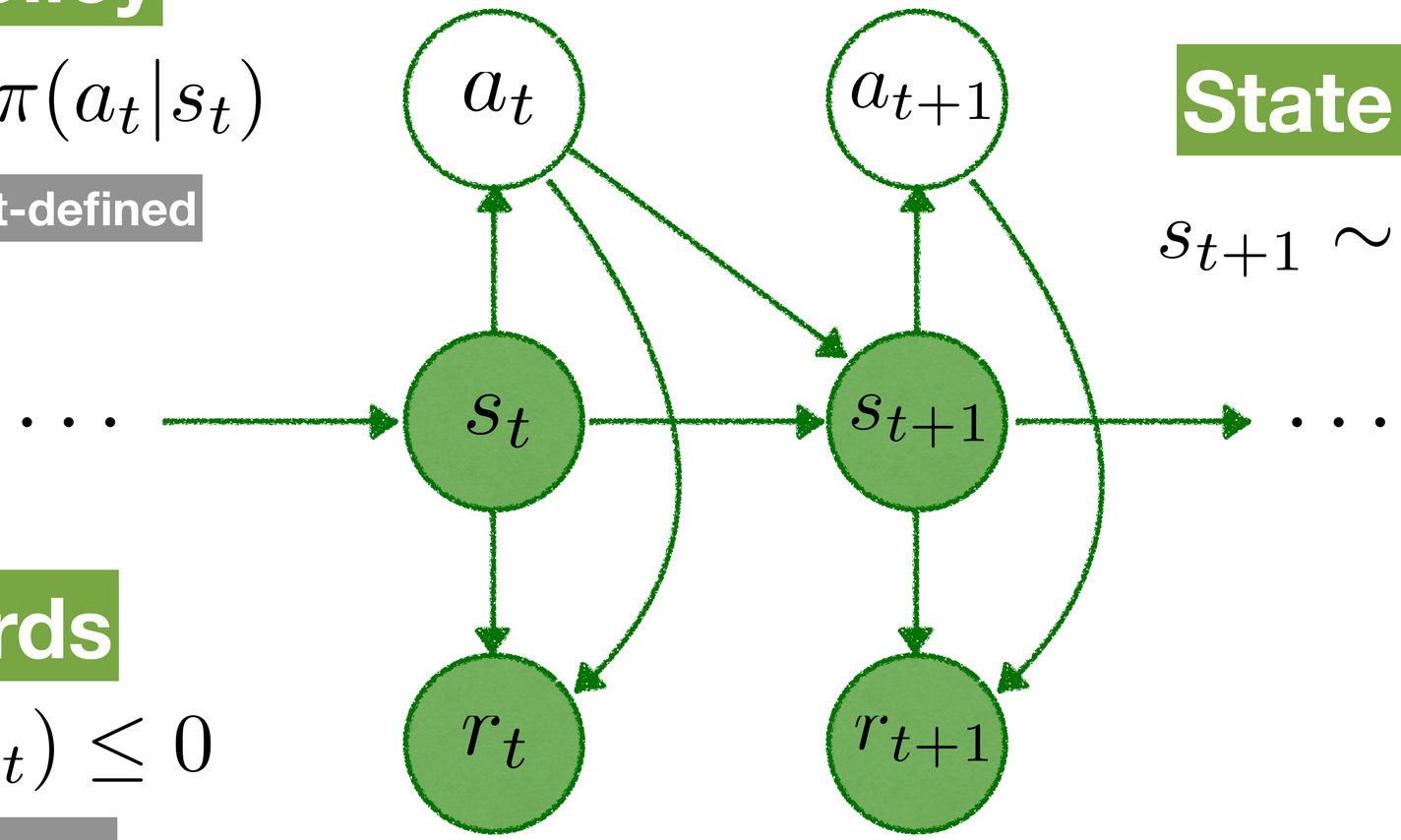
$$a_t \sim \pi(a_t|s_t)$$

Agent-defined

State transitions

$$s_{t+1} \sim p(s_{t+1}|s_t, a_t)$$

Likely unknown



Rewards

$$r_t = r(s_t, a_t) \leq 0$$

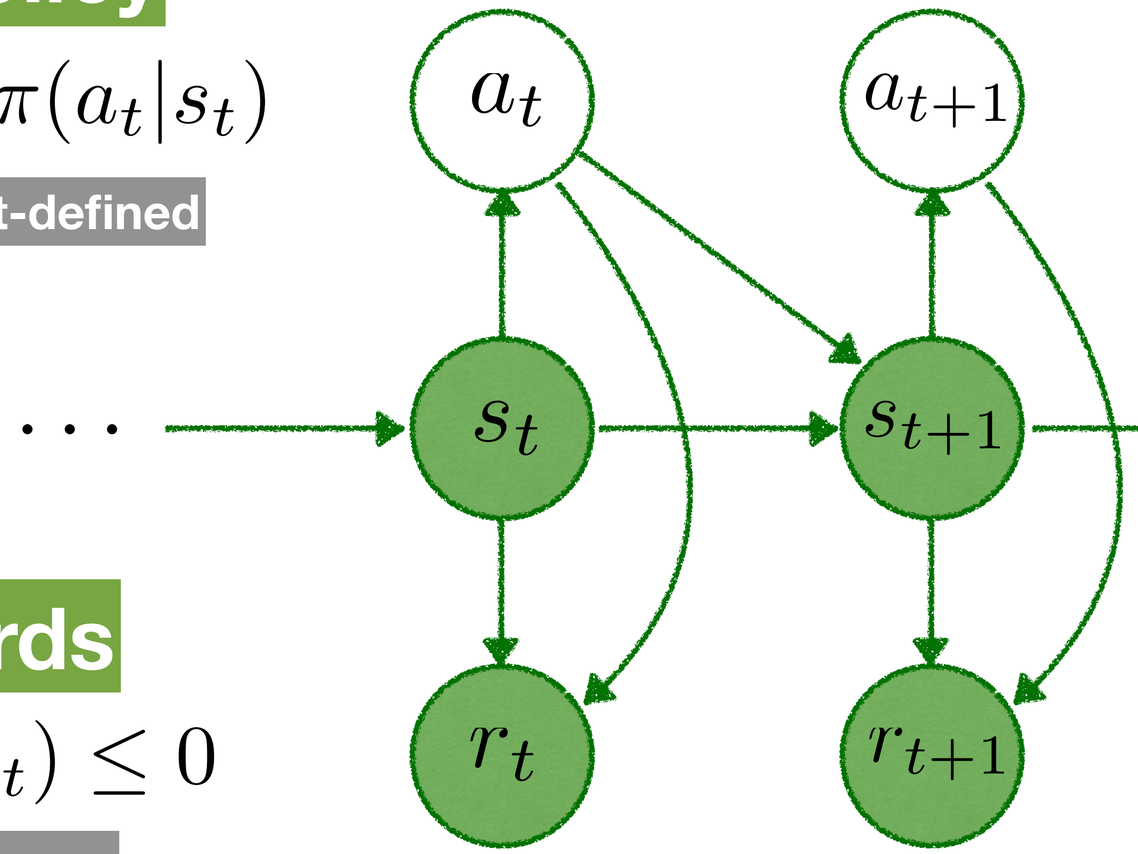
Likely unknown

Markov Decision Process

Policy

$$a_t \sim \pi(a_t|s_t)$$

Agent-defined



State transitions

$$s_{t+1} \sim p(s_{t+1}|s_t, a_t)$$

Likely unknown

Rewards

$$r_t = r(s_t, a_t) \leq 0$$

Likely unknown

Goal

$$\max_{\pi} \mathbb{E}_{\mathbf{s}_{1:T}, \mathbf{a}_{1:T}} \sum_{t=1}^{T} r_t$$

Variational inference

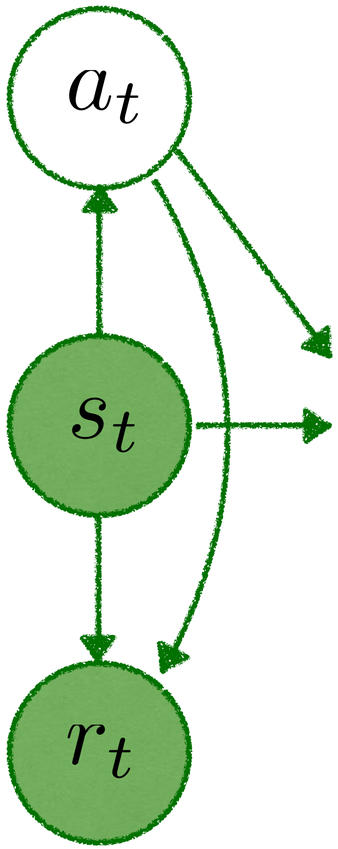
Latent-variable model

- Latent variable $z \sim p(z)$
- Observation $x \sim p(x|z)$
- Marginal likelihood $p(x) = \int p(z)p(x|z)dz$

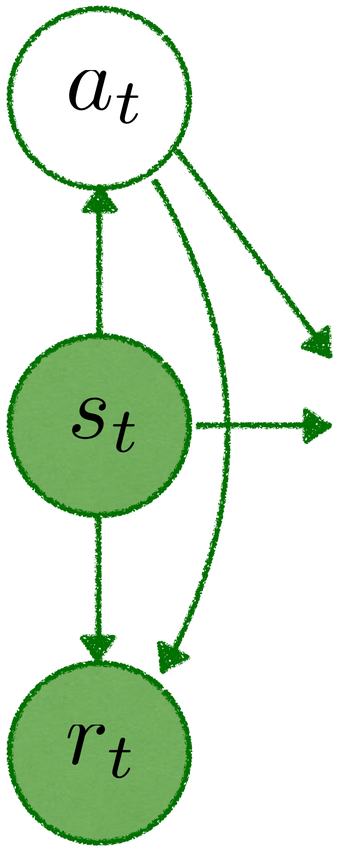
Variational lower bound

$$\begin{aligned}
\log p(x) &= \log \int p(z)p(x|z)dz = \log \int q(z)\frac{p(z)}{q(z)}p(x|z)dz \\
&\geq \mathbb{E}_{q(z)}\left[\log p(x|z)\right] - \mathrm{KL}(q(z)||p(z)) \\
&= \log p(x) - \mathrm{KL}(q(z)||p(z|x)) \\
&= \mathcal{L}(q,p)
\end{aligned}$$

MDP as a probabilistic model



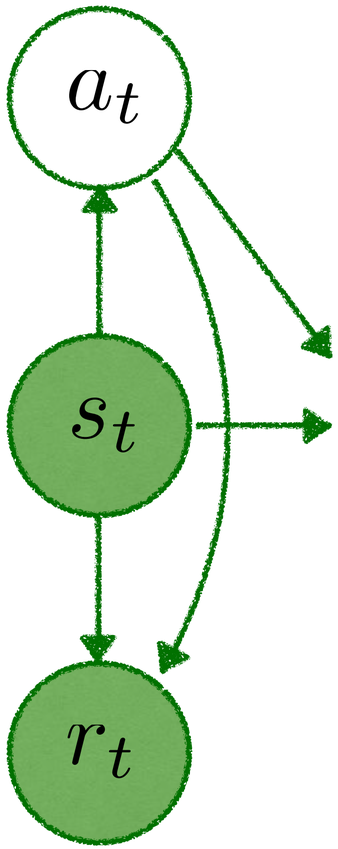
MDP as a probabilistic model



Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T)$$

MDP as a probabilistic model



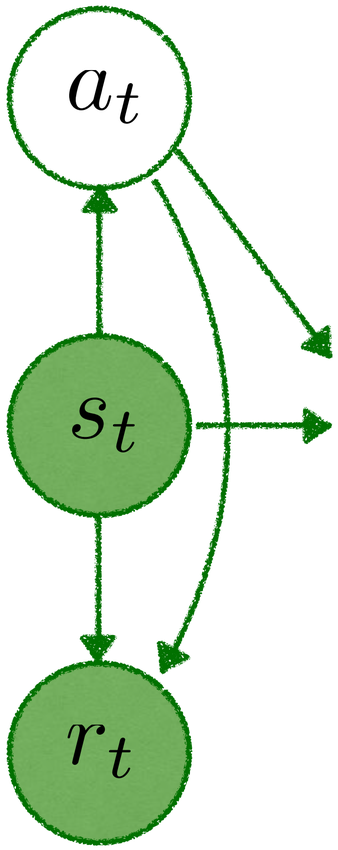
Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T)$$

Likelihood

$$p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \prod_{t=1}^{T} p(\hat{R}_t = 1|s_t, a_t) = \prod_{t=1}^{T} \exp(\alpha \cdot r_t)$$

MDP as a probabilistic model



Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T)$$

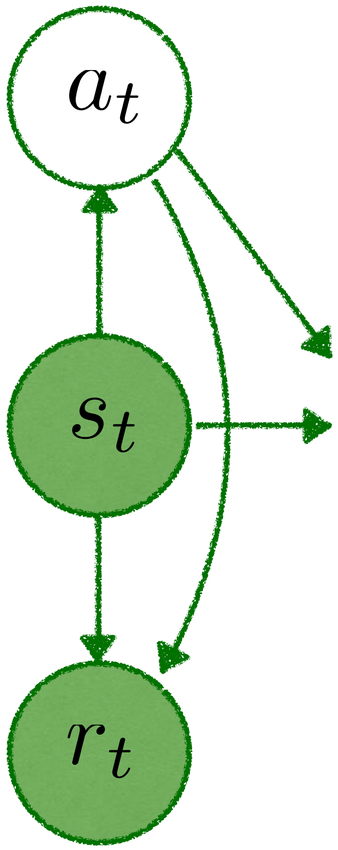
Likelihood

$$p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \prod_{t=1}^{T} p(\hat{R}_t = 1|s_t, a_t) = \prod_{t=1}^{T} \exp(\alpha \cdot r_t)$$

Approximate posterior (w.r.t. some other policy)

$$q_{\pi}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi(a_T|s_T)$$

MDP as a probabilistic model



Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)] \, \pi_0(a_T|s_T)$$

Likelihood

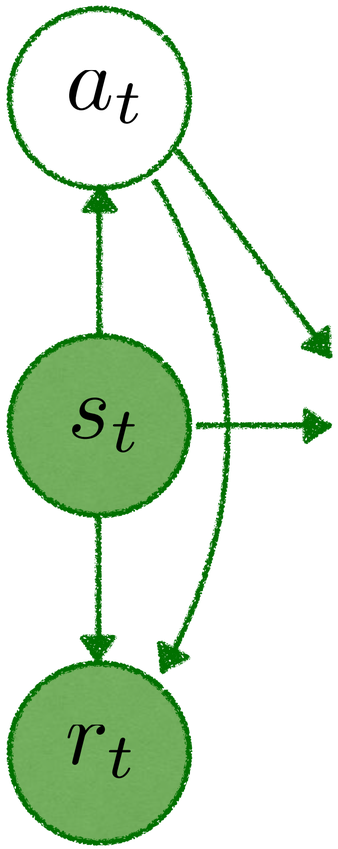
$$p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \prod_{t=1}^{T} p(\hat{R}_t = 1|s_t, a_t) = \prod_{t=1}^{T} \exp(\alpha \cdot r_t)$$

Approximate posterior (w.r.t. some other policy)

$$q_{\pi}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\pi(a_t|s_t) p(s_{t+1}|s_t, a_t)] \, \pi(a_T|s_T)$$

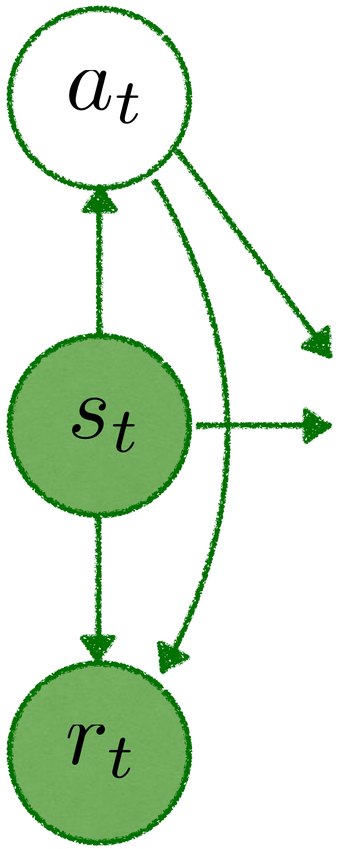
Different Same

Solving MDP via approximate inference



[1] Rawlik et al, 2012

Solving MDP via approximate inference

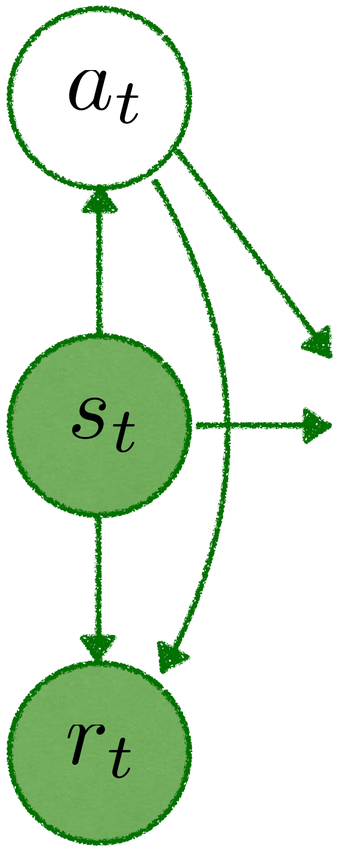


Marginal likelihood

$$\log p(\hat{\mathbf{R}}_{1:T}) = \log \mathbb{E}_{p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} p(\hat{\mathbf{R}}_{1:T} | \mathbf{s}_{1:T}, \mathbf{a}_{1:T})$$

[1] Rawlik et al, 2012

Solving MDP via approximate inference



Marginal likelihood

$$\log p(\hat{\mathbf{R}}_{1:T}) = \log \mathbb{E}_{p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} p(\hat{\mathbf{R}}_{1:T} | \mathbf{s}_{1:T}, \mathbf{a}_{1:T})$$

Variational lower bound

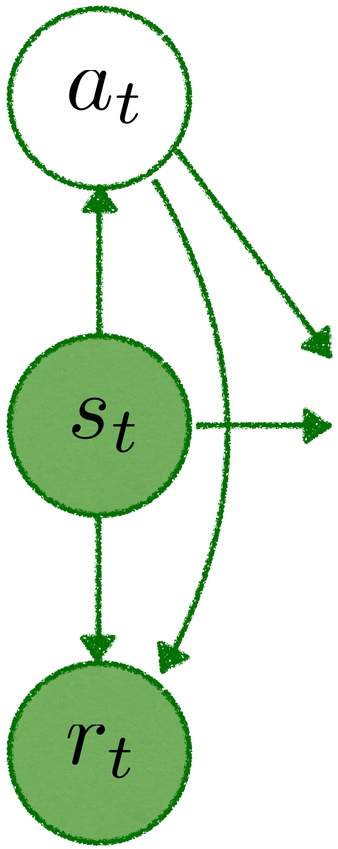
$$\log p(\hat{\mathbf{R}}_{1:T}) = \log \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\frac{p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} p(\hat{\mathbf{R}}_{1:T} | \mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \right]$$

$$\geq \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\alpha \sum_{t=1}^{T} r_t \right] - \mathbb{E}_{q_\pi(\mathbf{s}_{1:T})} \sum_{t=1}^{T} \underbrace{\mathrm{KL}(\pi(\cdot|s_t) || \pi_0(\cdot|s_t))}_{\mathrm{KL}_t}$$

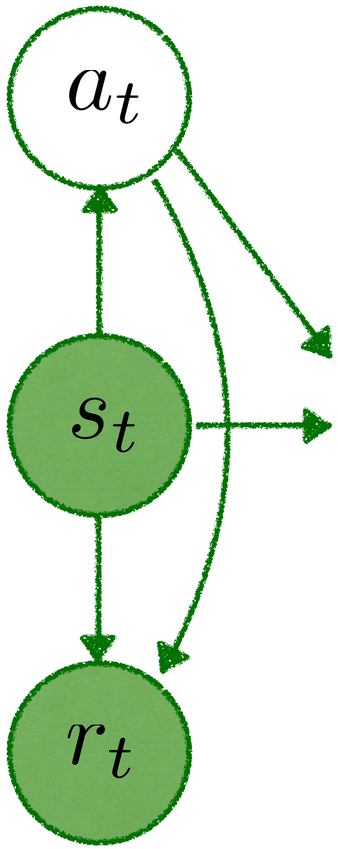
$$= \mathcal{L}(q_\pi, p_{\pi_0})$$

[1] Rawlik et al, 2012

Entropy-regularized reinforcement learning



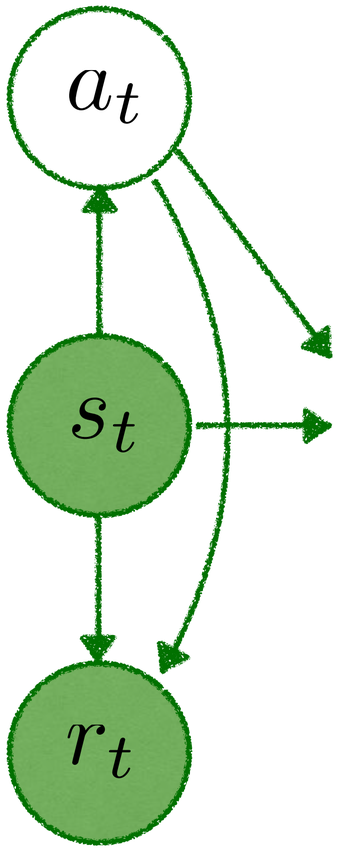
Entropy-regularized reinforcement learning



Variational lower bound

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \underbrace{\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))}_{\mathrm{KL}_t}\Big]$$

Entropy-regularized reinforcement learning

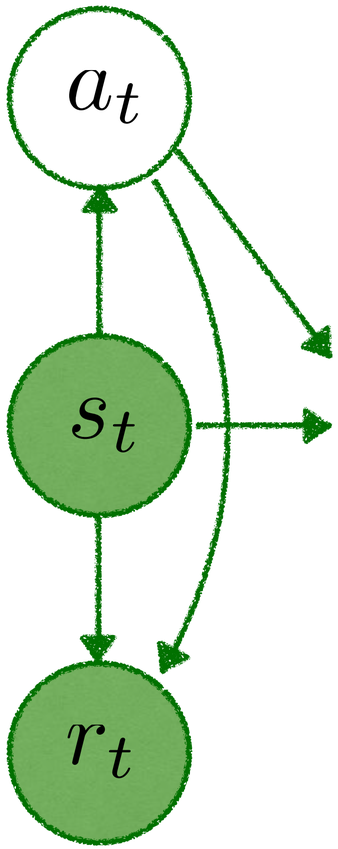


Variational lower bound

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \underbrace{\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))}_{\mathrm{KL}_t}\right]$$

Entropy regularization as a special case

Entropy-regularized reinforcement learning



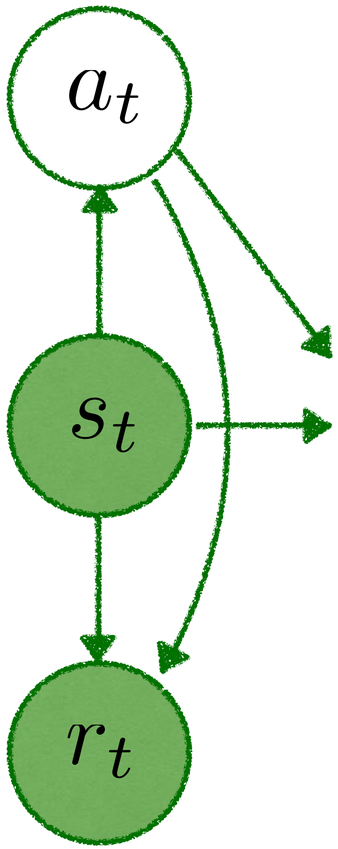
Variational lower bound

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \underbrace{\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))}_{\mathrm{KL}_t}\Big]$$

Entropy regularization as a special case

- $\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) = -\mathcal{H}(\pi(\cdot|s_t)) - \mathbb{E}_{\pi(a_t|s_t)} \log \pi_0(a_t|s_t)$

Entropy-regularized reinforcement learning



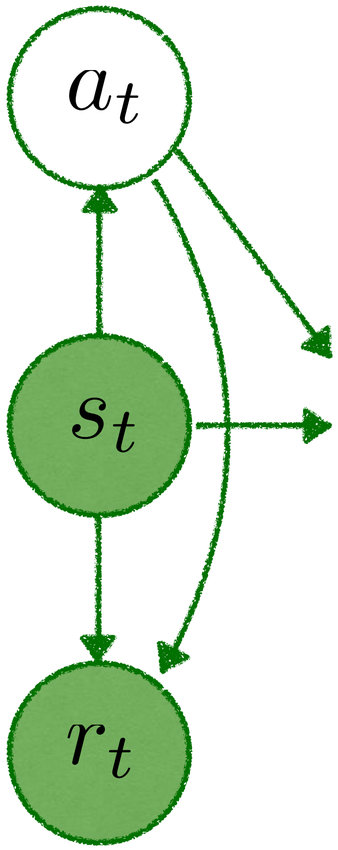
Variational lower bound

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \underbrace{\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))}_{\mathrm{KL}_t}\Big]$$

Entropy regularization as a special case

- $\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) = -\mathcal{H}(\pi(\cdot|s_t)) - \mathbb{E}_{\pi(a_t|s_t)} \log \pi_0(a_t|s_t)$
- Consider $\pi_0(a_t|s_t) = \mathrm{Uniform}(a_t)$

Entropy-regularized reinforcement learning



Variational lower bound

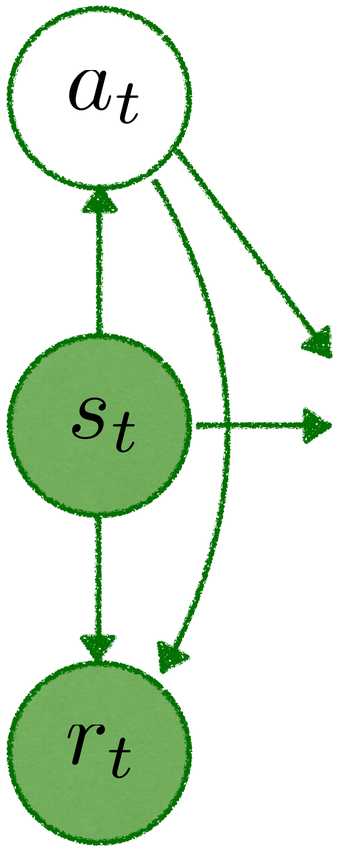
$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \underbrace{\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))}_{\mathrm{KL}_t}\right]$$

Entropy regularization as a special case

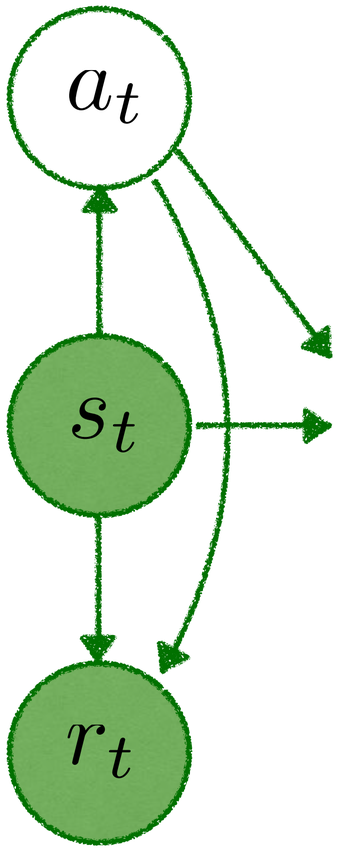
- $\mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) = -\mathcal{H}(\pi(\cdot|s_t)) - \mathbb{E}_{\pi(a_t|s_t)} \log \pi_0(a_t|s_t)$
- Consider $\pi_0(a_t|s_t) = \mathrm{Uniform}(a_t)$
- Classic objective: $\mathbb{E}_{\mathbf{s},\mathbf{a}}\left[\sum_{t=1}^{T} \alpha \cdot r_t + \mathcal{H}(\pi(\cdot|s_t))\right] \to \max_\pi$

[16] Neu et al, [17] Haarnoja et al, [18] Schulman et al

Value functions



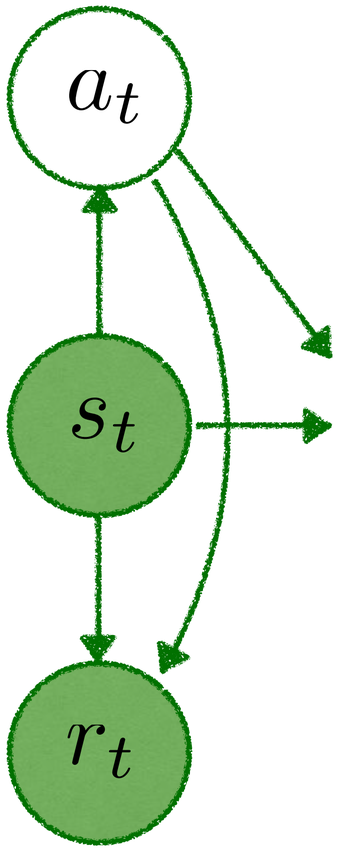
Value functions



State value

$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k\geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right] \qquad V_\pi(s_1) = \mathcal{L}(q_\pi, p_{\pi_0})$$

Value functions



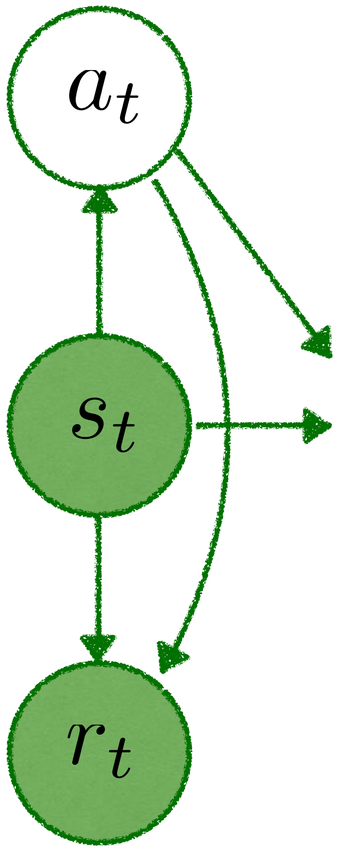
State value

$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k \geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right] \qquad V_\pi(s_1) = \mathcal{L}(q_\pi, p_{\pi_0})$$

State-action value

$$Q_\pi(s_t, a_t) = \alpha \cdot r_t + \mathbb{E}_{q_\pi}\left[\sum_{k > t}^{T} (\alpha \cdot r_k - \mathrm{KL}_k)\right]$$

Value functions



State value

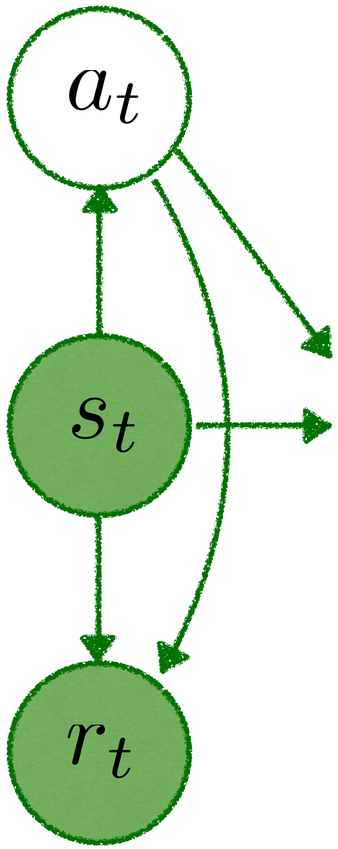
$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k\geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right] \qquad V_\pi(s_1) = \mathcal{L}(q_\pi, p_{\pi_0})$$

State-action value

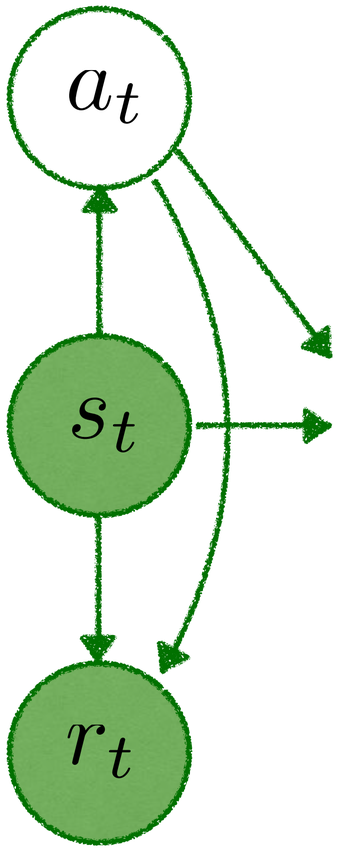
$$Q_\pi(s_t, a_t) = \alpha \cdot r_t + \mathbb{E}_{q_\pi}\left[\sum_{k>t}^{T} (\alpha \cdot r_k - \mathrm{KL}_k)\right]$$

$$V_\pi(s_t) = \mathbb{E}_{\pi(a_t|s_t)} Q_\pi(s_t, a_t) - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))$$

Backup equations



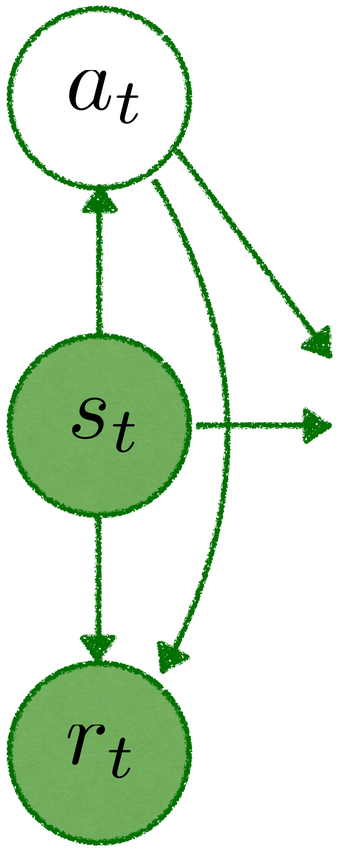
Backup equations



$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k\geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right]$$

$$= \mathbb{E}_{a_t}\left[\alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V(s_{t+1})\right] - \mathrm{KL}_t$$

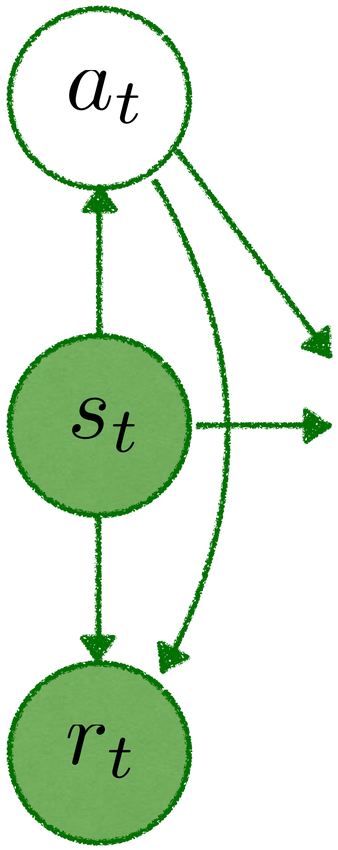
Backup equations



$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k \geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right]$$

$$= \mathbb{E}_{a_t}\left[\alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V(s_{t+1})\right] - \mathrm{KL}_t$$

Backup equations



$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k\geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right]$$

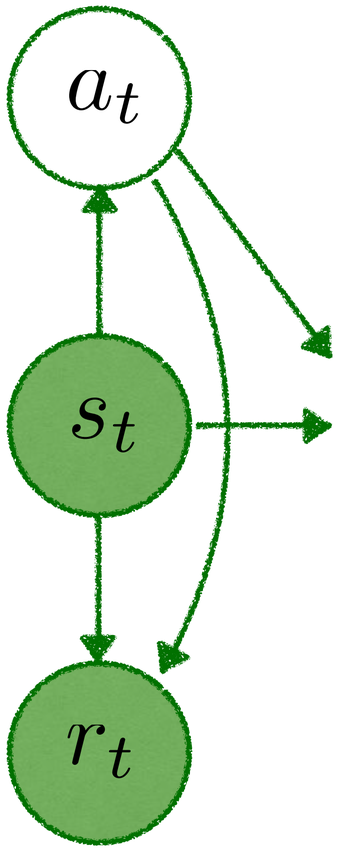
$$= \mathbb{E}_{a_t}\left[\alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V(s_{t+1})\right] - \mathrm{KL}_t$$

$$Q_\pi(s_t, a_t) = \alpha \cdot r_t + \mathbb{E}_{q_\pi}\left[\sum_{k>t}^{T} (\alpha \cdot r_k - \mathrm{KL}_k)\right]$$

$$= \alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V_\pi(s_{t+1})$$

$$= \alpha \cdot r_t + \mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1}) - \mathrm{KL}_{t+1}\right]$$

Backup equations



$$V_\pi(s_t) = \mathbb{E}_{q_\pi}\left[\sum_{k \geq t}^{T} \alpha \cdot r_k - \mathrm{KL}_k\right]$$

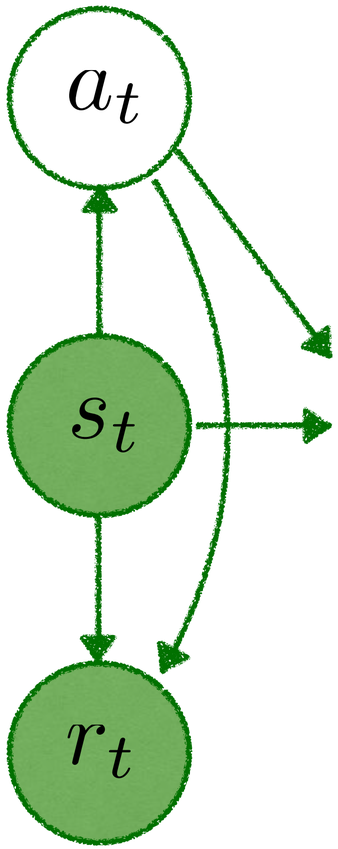
$$= \mathbb{E}_{a_t}\left[\alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V(s_{t+1})\right] - \mathrm{KL}_t$$

$$Q_\pi(s_t, a_t) = \alpha \cdot r_t + \mathbb{E}_{q_\pi}\left[\sum_{k > t}^{T} (\alpha \cdot r_k - \mathrm{KL}_k)\right]$$

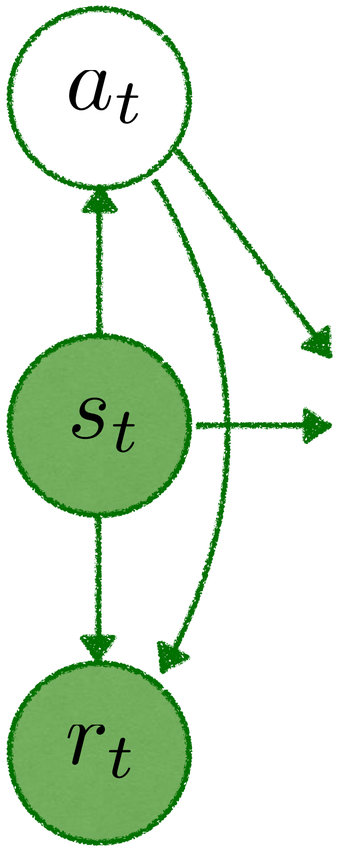
$$= \alpha \cdot r_t + \mathbb{E}_{s_{t+1}} V_\pi(s_{t+1})$$

$$= \alpha \cdot r_t + \mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1}) - \mathrm{KL}_{t+1}\right]$$

Value iteration



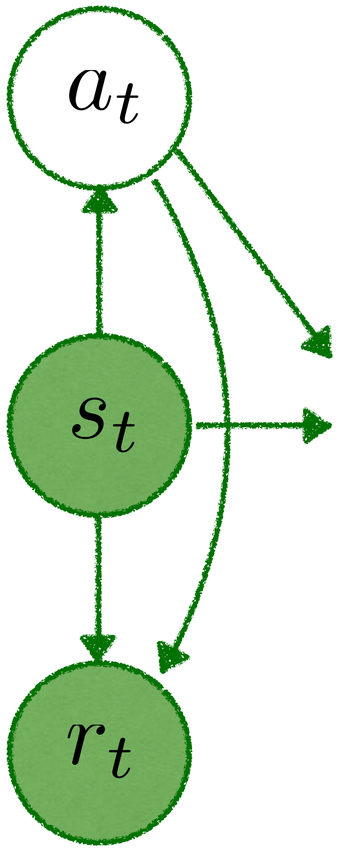
Value iteration



Locally-optimal policy

$$\mathbb{E}_{\pi(a_t|s_t)}Q(s_t, a_t) - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) \rightarrow \max_{\pi(\cdot|s_t)}$$

Value iteration



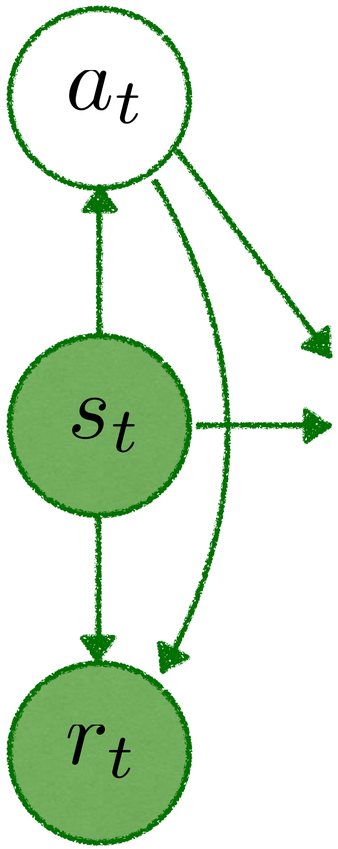
Locally-optimal policy

$$\mathbb{E}_{\pi(a_t|s_t)} Q(s_t, a_t) - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) \to \max_{\pi(\cdot|s_t)}$$

$$\pi_Q(a_t|s_t) = \pi_0(a_t|s_t) \exp\{Q(s_t, a_t) \underbrace{-V_Q(s_t)}_{-\text{log-normalizer}}\}$$

$$V_Q(s_t) = \log \mathbb{E}_{a \sim p_0} \exp\{Q(s_t, a_t)\}$$

Value iteration



Locally-optimal policy

$$\mathbb{E}_{\pi(a_t|s_t)} Q(s_t, a_t) - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) \to \max_{\pi(\cdot|s_t)}$$

$$\pi_Q(a_t|s_t) = \pi_0(a_t|s_t) \exp\{Q(s_t, a_t) \underbrace{-V_Q(s_t)}_{-\text{log-normalizer}}\}$$

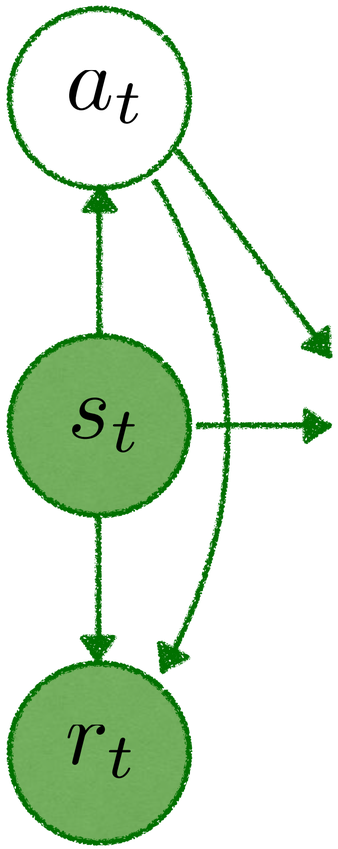
$$V_Q(s_t) = \log \mathbb{E}_{a \sim p_0} \exp\{Q(s_t, a_t)\}$$

Bellman equation

$$Q(s_t, a_t) \leftarrow r_t + \mathbb{E}_{s_{t+1}}\left[\mathbb{E}_{a_{t+1} \sim \pi_Q} Q(s_{t+1}, a_{t+1}) - \mathrm{KL}_t\right]$$

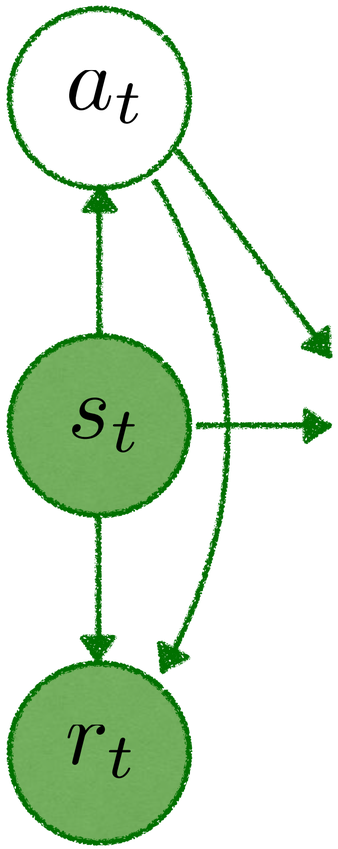
$$= r_t + \mathbb{E}_{s_{t+1}}\left[\log \mathbb{E}_{a_{t+1} \sim \pi_0} \exp\{Q(s_{t+1}, a_{t+1})\}\right]$$

Policy gradients as inference in MDP



$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\right]$$

Policy gradients as inference in MDP

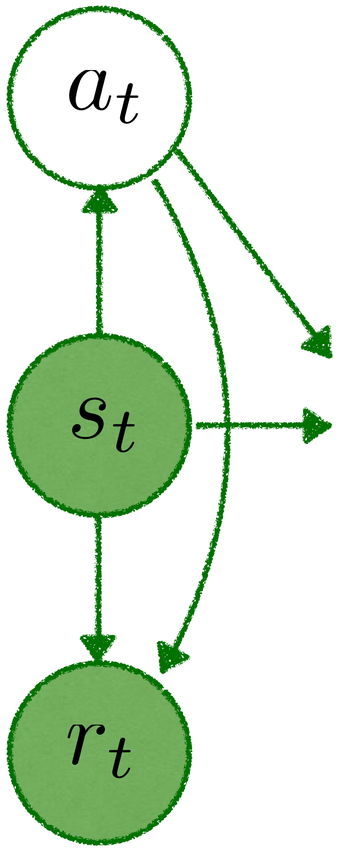


$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\right]$$

Stochastic gradient inference

$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \left(\sum_{k \geq t} \alpha r_k + \mathrm{KL}_k\right) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathrm{KL}_t\right]$$

Policy gradients as inference in MDP



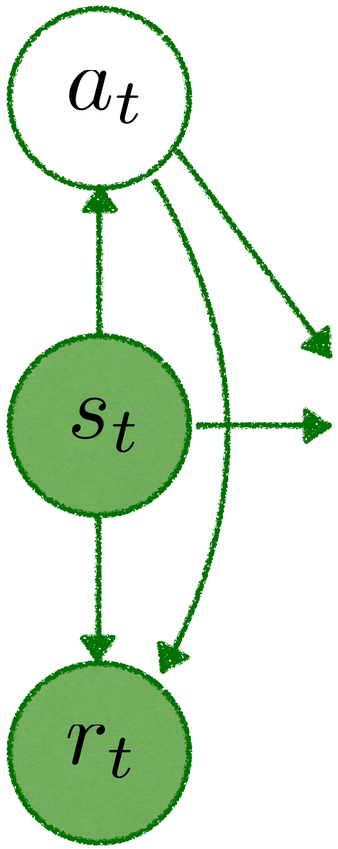
$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\right]$$

Stochastic gradient inference

$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \left(\sum_{k\geq t} \alpha r_k + \mathrm{KL}_k\right) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathrm{KL}_t\right]$$

- Sample a trajectory $\hat{\mathbf{s}}_{1:T}, \hat{\mathbf{a}_{1:T}} \sim q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$

Policy gradients as inference in MDP



$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\right]$$

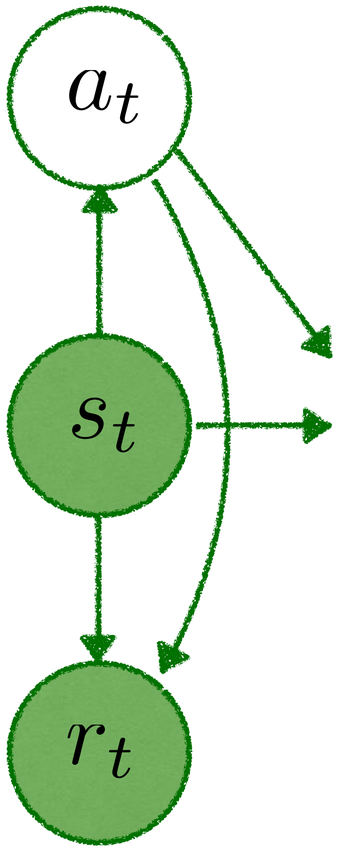
Stochastic gradient inference

$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\sum_{t=1}^{T} \left(\sum_{k\geq t} \alpha r_k + \mathrm{KL}_k\right) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathrm{KL}_t\right]$$

- Sample a trajectory $\hat{\mathbf{s}}_{1:T}, \hat{\mathbf{a}_{1:T}} \sim q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$
- Estimate gradient w.r.t. policy parameters:

$$\hat{\nabla}_\pi \mathcal{L} = \sum_{t=1}^{T} \left(\sum_{k\geq t} \alpha r_k + \mathrm{KL}_k\right) \nabla_\pi \log \pi(\hat{a}_t|\hat{s}_t) + \nabla_\pi \mathrm{KL}_t$$

Policy gradients as inference in MDP



$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\Big]$$

Stochastic gradient inference

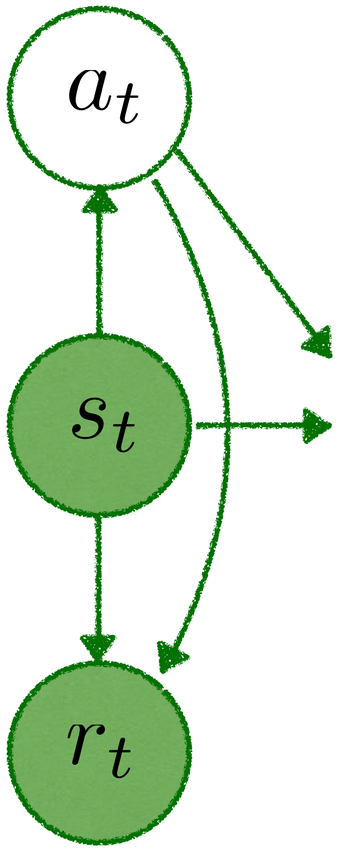
$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \big(\sum_{k\geq t} \alpha r_k + \mathrm{KL}_k\big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathrm{KL}_t\Big]$$

- Sample a trajectory $\hat{\mathbf{s}}_{1:T}, \hat{\mathbf{a}_{1:T}} \sim q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$
- Estimate gradient w.r.t. policy parameters:

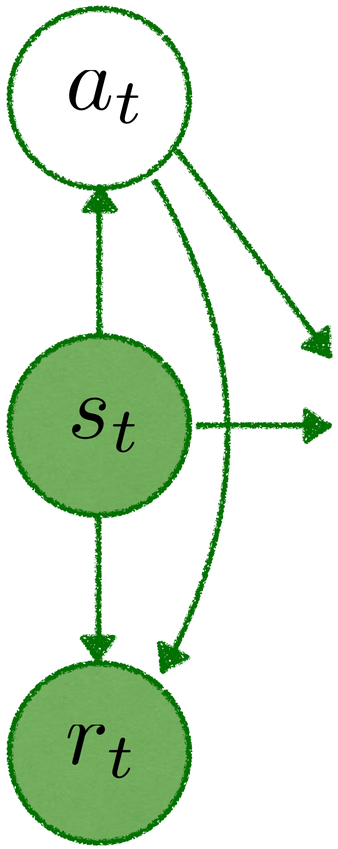
$$\hat{\nabla}_\pi \mathcal{L} = \sum_{t=1}^{T} \big(\sum_{k\geq t} \alpha r_k + \mathrm{KL}_k\big)\nabla_\pi \log \pi(\hat{a}_t|\hat{s}_t) + \nabla_\pi \mathrm{KL}_t$$

- Similar to REINFORCE (but not equal!)

Policy gradients as inference in MDP



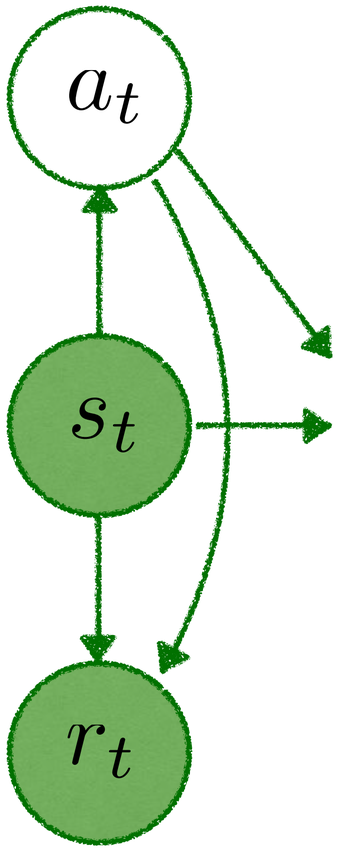
Policy gradients as inference in MDP



Stochastic gradient inference

$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\sum_{k\geq t}\alpha r_k + \mathcal{H}(\pi(\cdot|s_k))\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

Policy gradients as inference in MDP



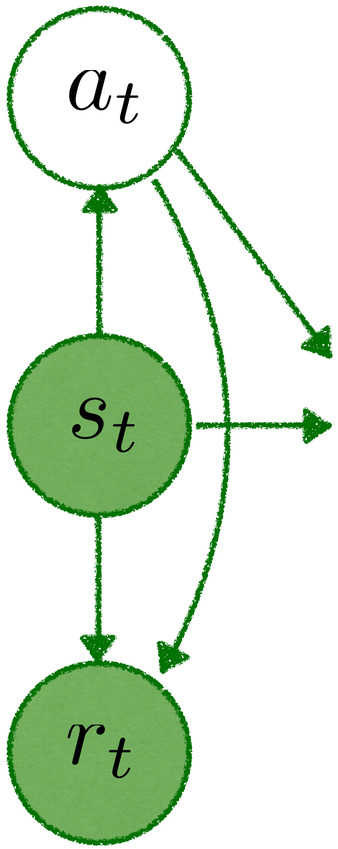
Stochastic gradient inference

$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\sum_{k\geq t}\alpha r_k + \mathcal{H}(\pi(\cdot|s_k))\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

Stochastic policy gradient

$$\nabla_\pi R = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\alpha\sum_{k\geq t} r_k\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

Policy gradients as inference in MDP



Stochastic gradient inference

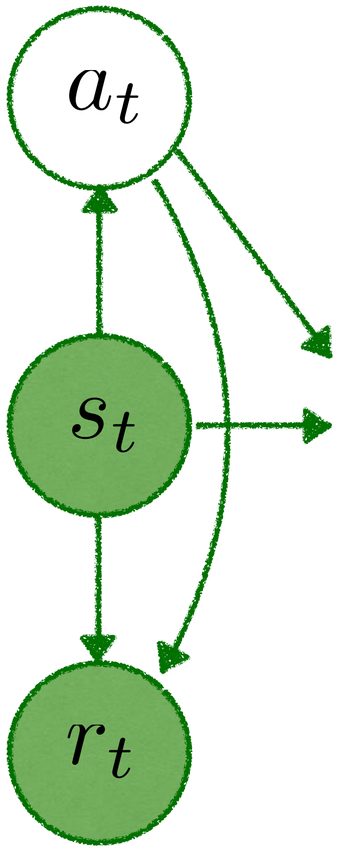
$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\sum_{k\geq t}\alpha r_k + \mathcal{H}(\pi(\cdot|s_k))\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

Stochastic policy gradient

$$\nabla_\pi R = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\alpha\sum_{k\geq t} r_k\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

- Model-free learning algorithm

Policy gradients as inference in MDP



Stochastic gradient inference

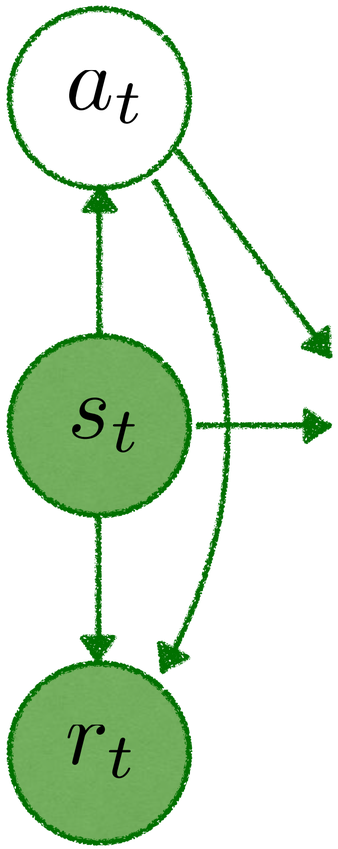
$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\sum_{k\geq t}\alpha r_k + \mathcal{H}(\pi(\cdot|s_k))\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

Stochastic policy gradient

$$\nabla_\pi R = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\alpha\sum_{k\geq t} r_k\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|s_t))\Big]$$

- Model-free learning algorithm
- Known as REINFORCE [2] Williams, 1992

Policy gradients as inference in MDP



Stochastic gradient inference

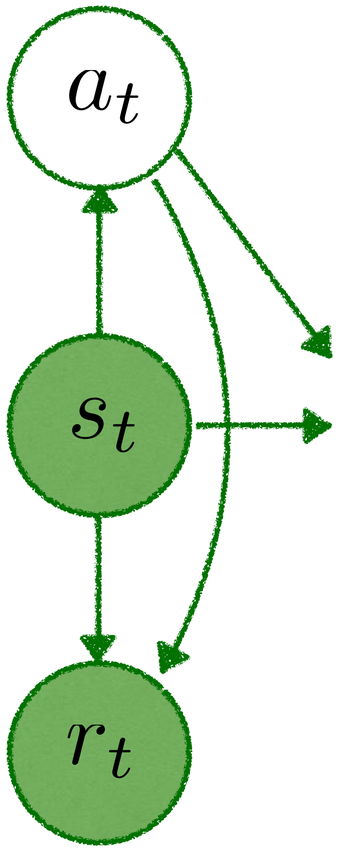
$$\nabla_\pi \mathcal{L} = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\sum_{k\geq t}\alpha r_k + \mathcal{H}(\pi(\cdot|s_k))\Big)\nabla_\pi \log\pi(a_t|s_t) + \nabla_\pi\mathcal{H}(\pi(\cdot|s_t))\Big]$$

Stochastic policy gradient

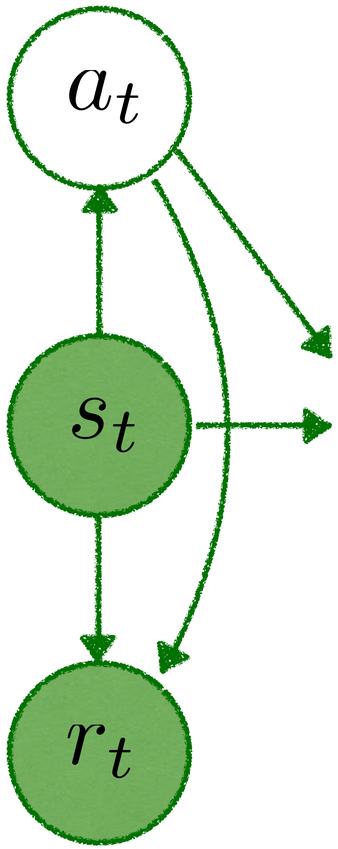
$$\nabla_\pi R = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T}\Big(\alpha\sum_{k\geq t} r_k\Big)\nabla_\pi \log\pi(a_t|s_t) + \nabla_\pi\mathcal{H}(\pi(\cdot|s_t))\Big]$$

- Model-free learning algorithm
- Known as REINFORCE [2] Williams, 1992
- Instantaneous vs long-term regularization

Improving policy optimization

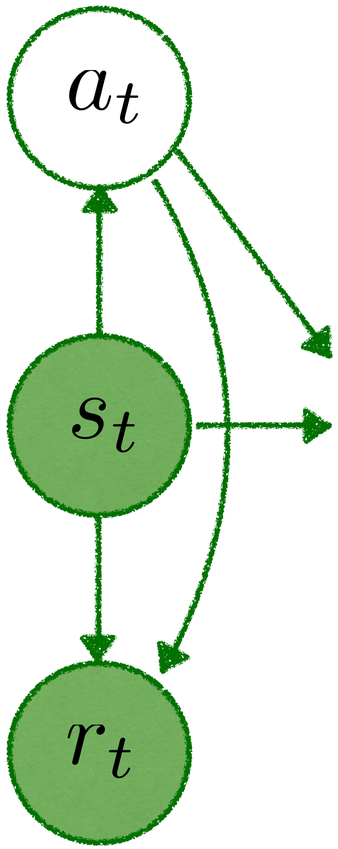


Improving policy optimization

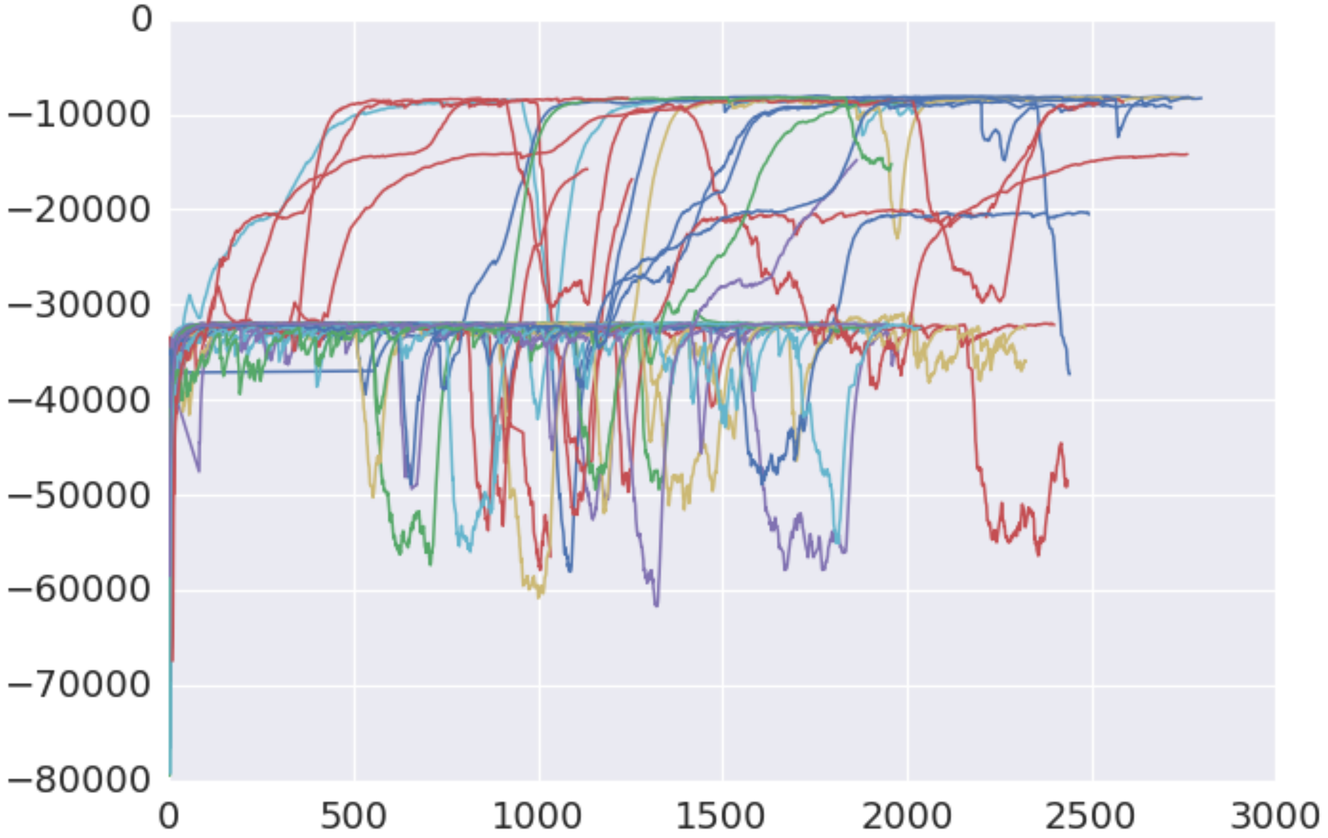


- Reward optimization can be a **very** hard optimization problem

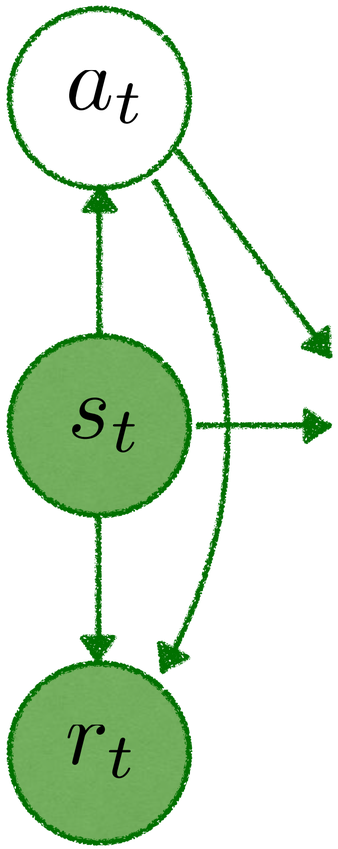
Improving policy optimization



- Reward optimization can be a **very** hard optimization problem
- Typical rewards vs time plot:

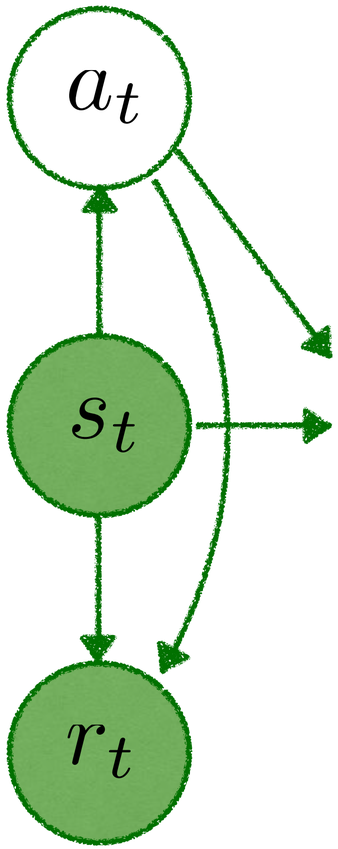


Improving policy optimization

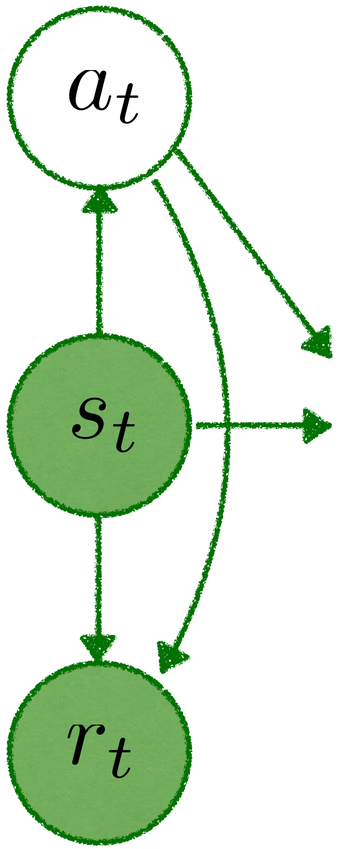


Improving policy optimization

- Policy can change too fast

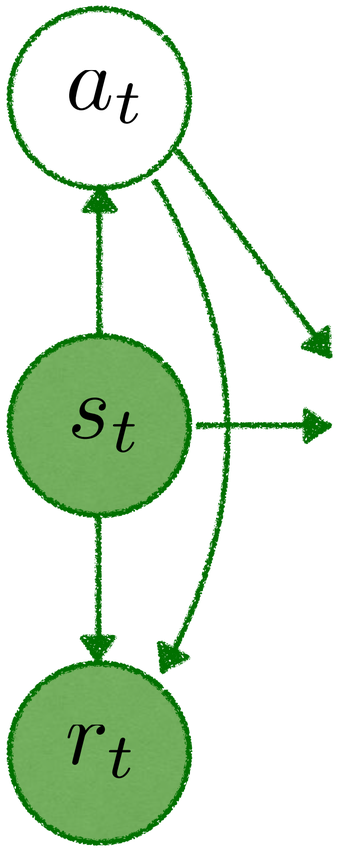


Improving policy optimization



- Policy can change too fast
- We want to make small, reliable improvements

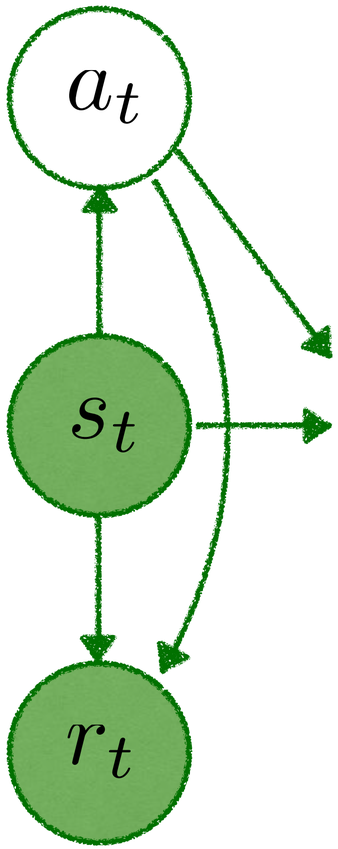
Improving policy optimization



- Policy can change too fast
- We want to make small, reliable improvements
- Our lower bound has already all necessary ingredients

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\Big]$$

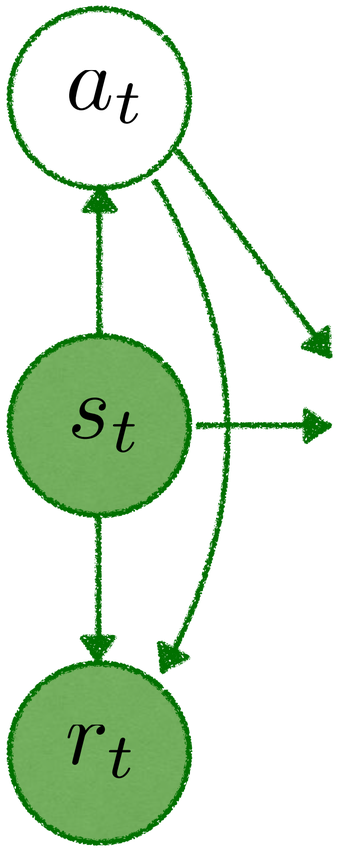
Improving policy optimization



- Policy can change too fast
- We want to make small, reliable improvements
- Our lower bound has already all necessary ingredients

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\Big]$$

Improving policy optimization

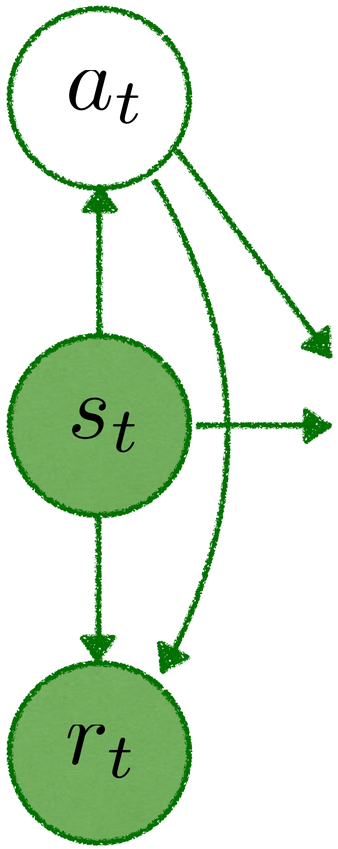


- Policy can change too fast
- We want to make small, reliable improvements
- Our lower bound has already all necessary ingredients

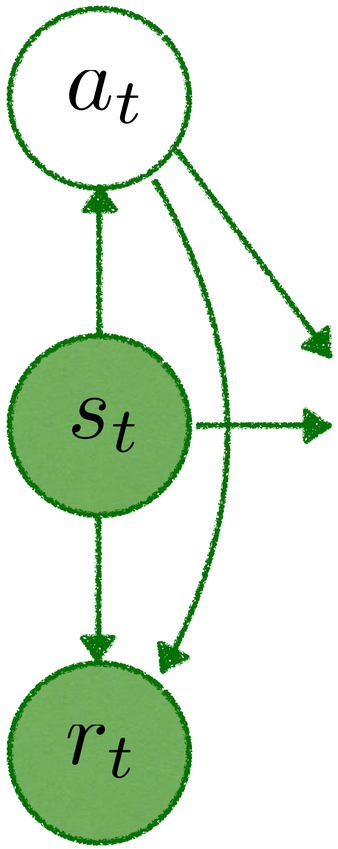
$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\Big]$$

- Use this constraint to prevent too rapid changes in the policy

Stable EM-algorithm

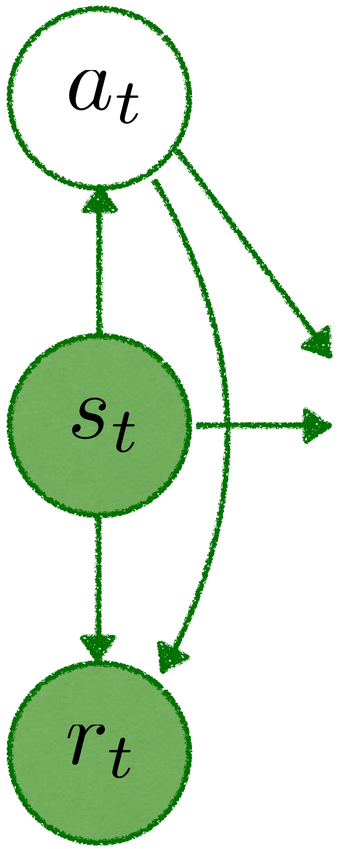


Stable EM-algorithm



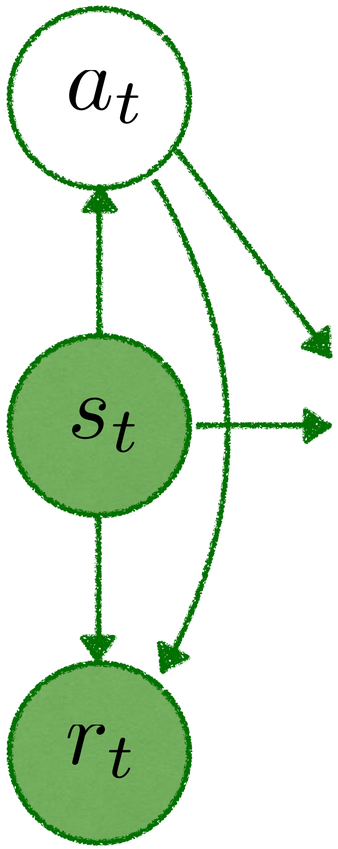
- Initialize π_0

Stable EM-algorithm



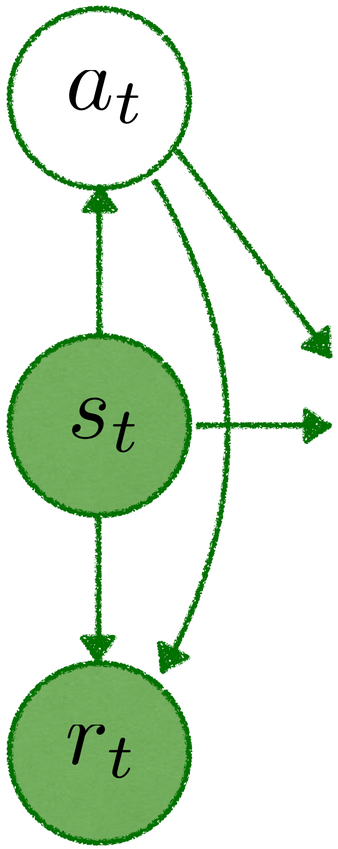
- Initialize π_0
- For each j = 1, 2, …:

Stable EM-algorithm



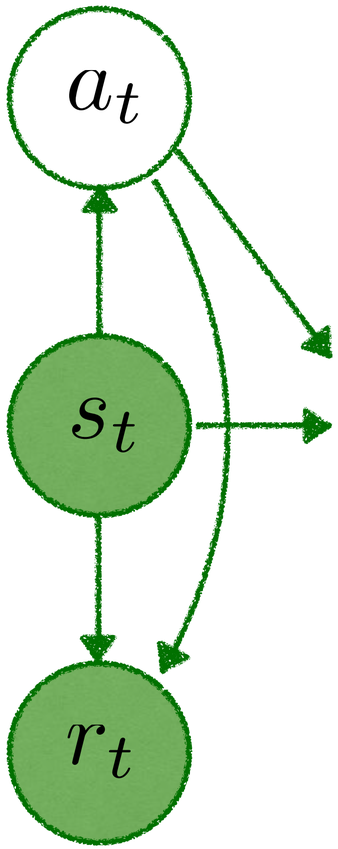
- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$

Stable EM-algorithm



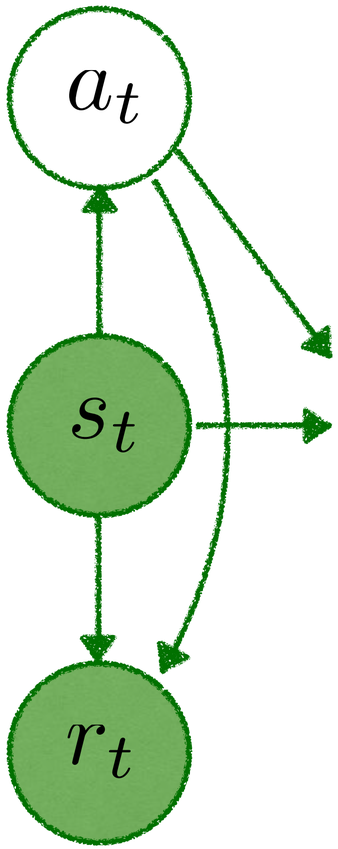
- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$
  - Set prior policy to π_j

Stable EM-algorithm



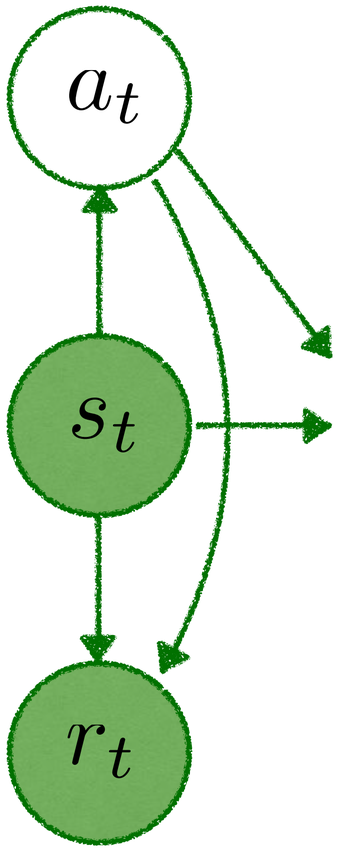
- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$
  - Set prior policy to π_j
  - Repeat

Stable EM-algorithm



- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$
  - Set prior policy to π_j
  - Repeat
- Monotonically improves the (expected) return

Stable EM-algorithm

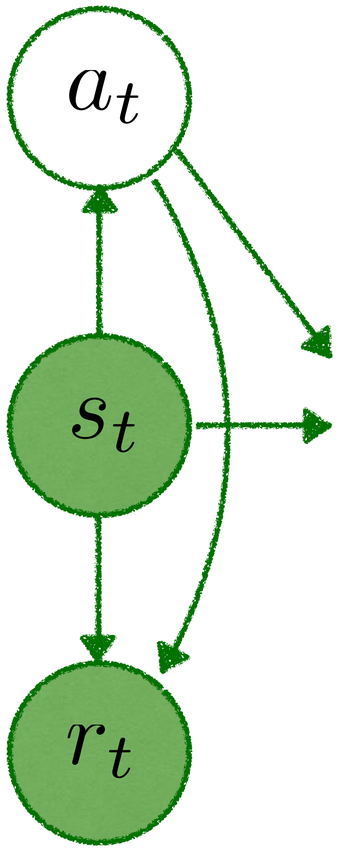


- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$
  - Set prior policy to π_j
  - Repeat
- Monotonically improves the (expected) return
- Closely related to trust-region and proximal policy optimization

[4] Schulman et al, 2015

[8] Schulman et al, 2017

MDP as a probabilistic model



Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)] \, \pi_0(a_T|s_T)$$

Likelihood

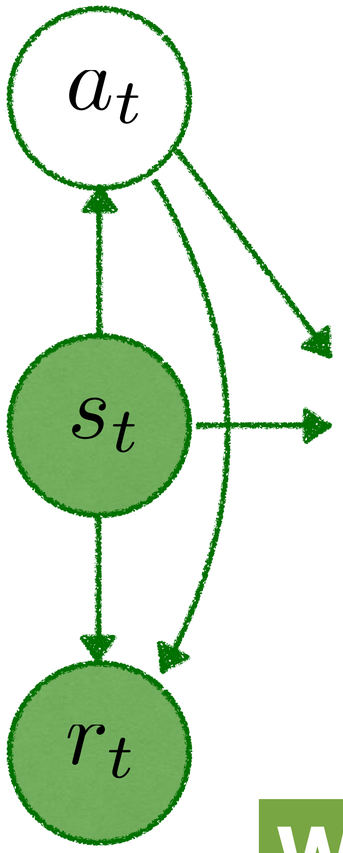
$$p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \prod_{t=1}^{T} p(\hat{R}_t = 1|s_t, a_t) = \prod_{t=1}^{T} \exp(\alpha \cdot r_t)$$

Approximate posterior (w.r.t. some other policy)

$$q_{\pi}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\pi(a_t|s_t) p(s_{t+1}|s_t, a_t)] \, \pi(a_T|s_T)$$

Different Same

MDP as a probabilistic model



Prior (w.r.t. some policy)

$$p_{\pi_0}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)] \, \pi_0(a_T|s_T)$$

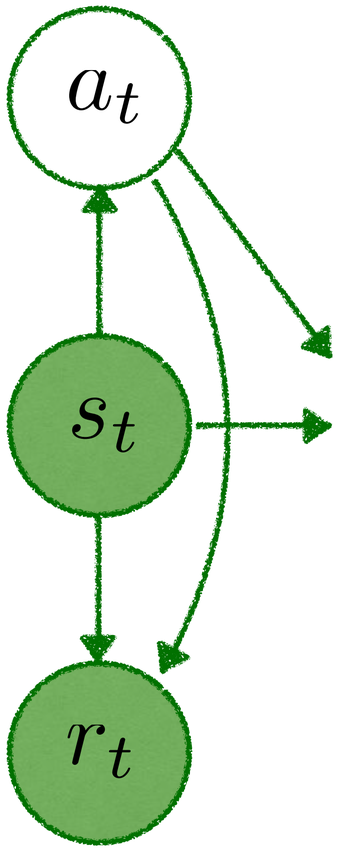
Likelihood

$$p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \prod_{t=1}^{T} p(\hat{R}_t = 1|s_t, a_t) = \prod_{t=1}^{T} \exp(\alpha \cdot r_t)$$

What happens if we use an arbitrary approximation?

$$q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$$

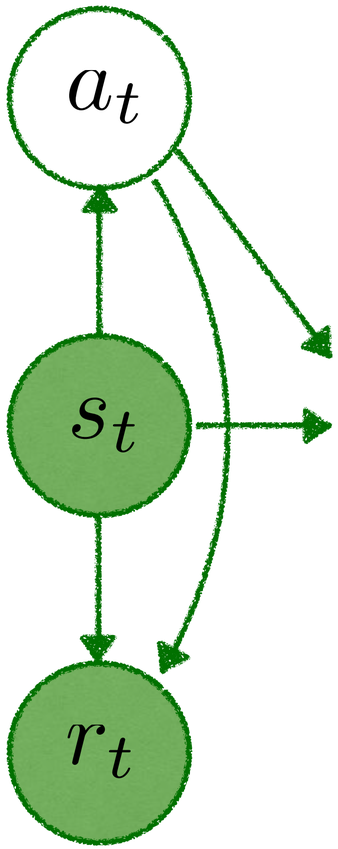
Optimistic MDP



Optimistic lower bound

$$\mathcal{L}(q, p_{\pi_0}) = \mathbb{E}_{q(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t + \log p(s_{t+1}|s_t, a_t) + \log \pi_0(a_t|s_t) + \mathcal{H}(q)\Big]$$

Optimistic MDP

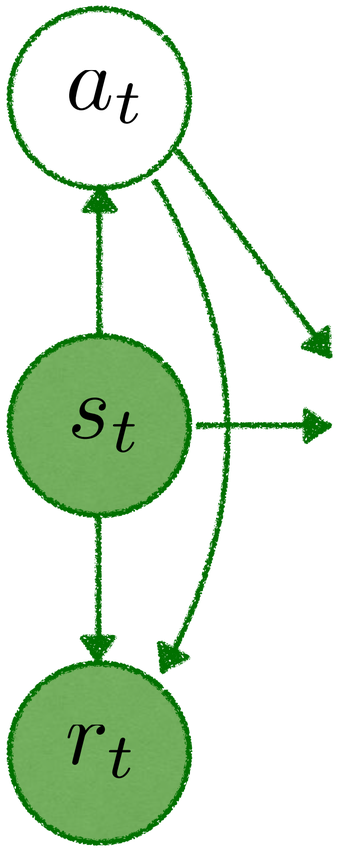


Optimistic lower bound

$$\mathcal{L}(q, p_{\pi_0}) = \mathbb{E}_{q(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t + \log p(s_{t+1}|s_t, a_t) + \log \pi_0(a_t|s_t) + \mathcal{H}(q)\Big]$$

- Difficult to learn (transition dynamics is unknown)

Optimistic MDP

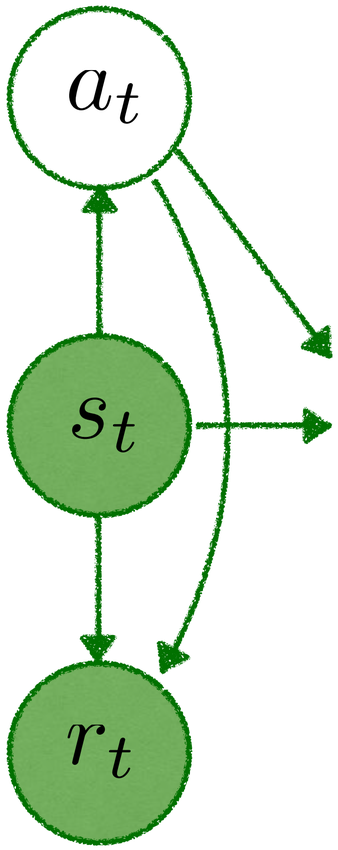


Optimistic lower bound

$$\mathcal{L}(q, p_{\pi_0}) = \mathbb{E}_{q(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t + \log p(s_{t+1}|s_t, a_t) + \log \pi_0(a_t|s_t) + \mathcal{H}(q)\Big]$$

- Difficult to learn (transition dynamics is unknown)
- May achieve unrealistically high rewards

Optimistic MDP



Optimistic lower bound

$$\mathcal{L}(q, p_{\pi_0}) = \mathbb{E}_{q(\mathbf{s},\mathbf{a})}\Big[\sum_{t=1}^{T} \alpha \cdot r_t + \log p(s_{t+1}|s_t, a_t) + \log \pi_0(a_t|s_t) + \mathcal{H}(q)\Big]$$

- Difficult to learn (transition dynamics is unknown)
- May achieve unrealistically high rewards
- At maximum optimizes a different utility function

$$\log p(\hat{\mathbf{R}}_{1:T}) = \log\left(\mathbb{E}_{\mathbf{s}_{1:T},\mathbf{a}_{1:T}}\left[\exp\{\sum_{t=1}^{T} \alpha \cdot r_t\}\right]\right) \neq \mathbb{E}_{\mathbf{s}_{1:T},\mathbf{a}_{1:T}}\left[\sum_{t=1}^{T} \alpha \cdot r_t\right]$$

[19] Levine, 2018; [20] Burda et al, 2015

Hierarchical Reinforcement Learning

Hierarchical Reinforcement Learning

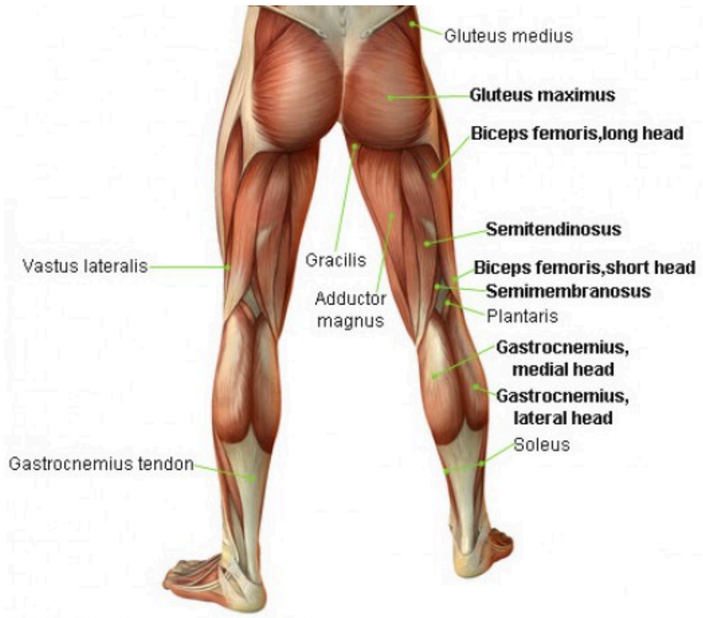
https://en.wikipedia.org/wiki/Roller_skating

Task: roller skating

Hierarchical Reinforcement Learning

https://en.wikipedia.org/wiki/Roller_skating

Task: roller skating

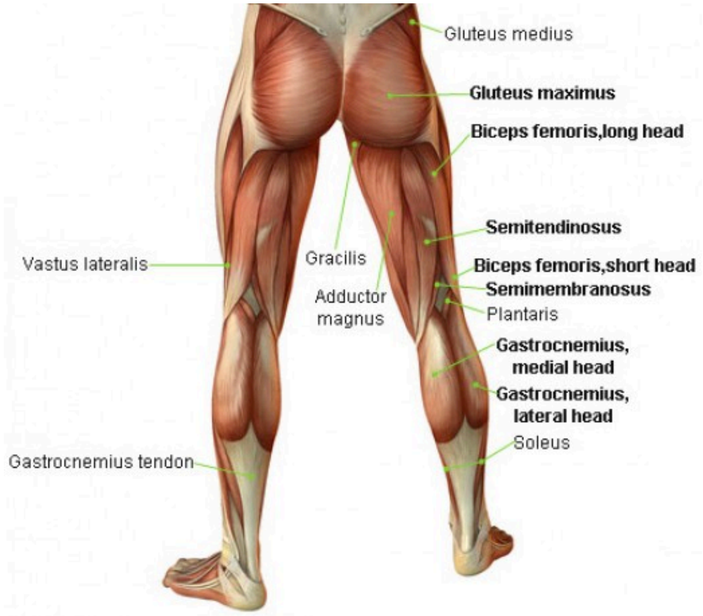


Low-level actions muscles control

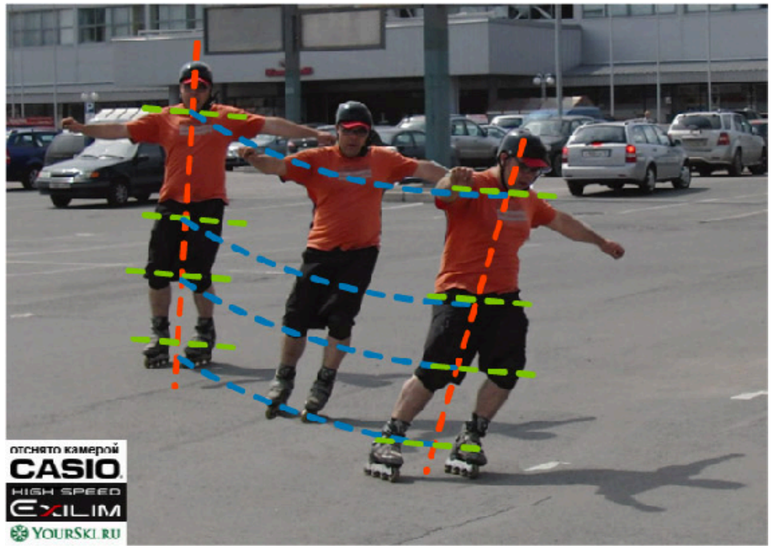
Hierarchical Reinforcement Learning

https://en.wikipedia.org/wiki/Roller_skating

Task: roller skating

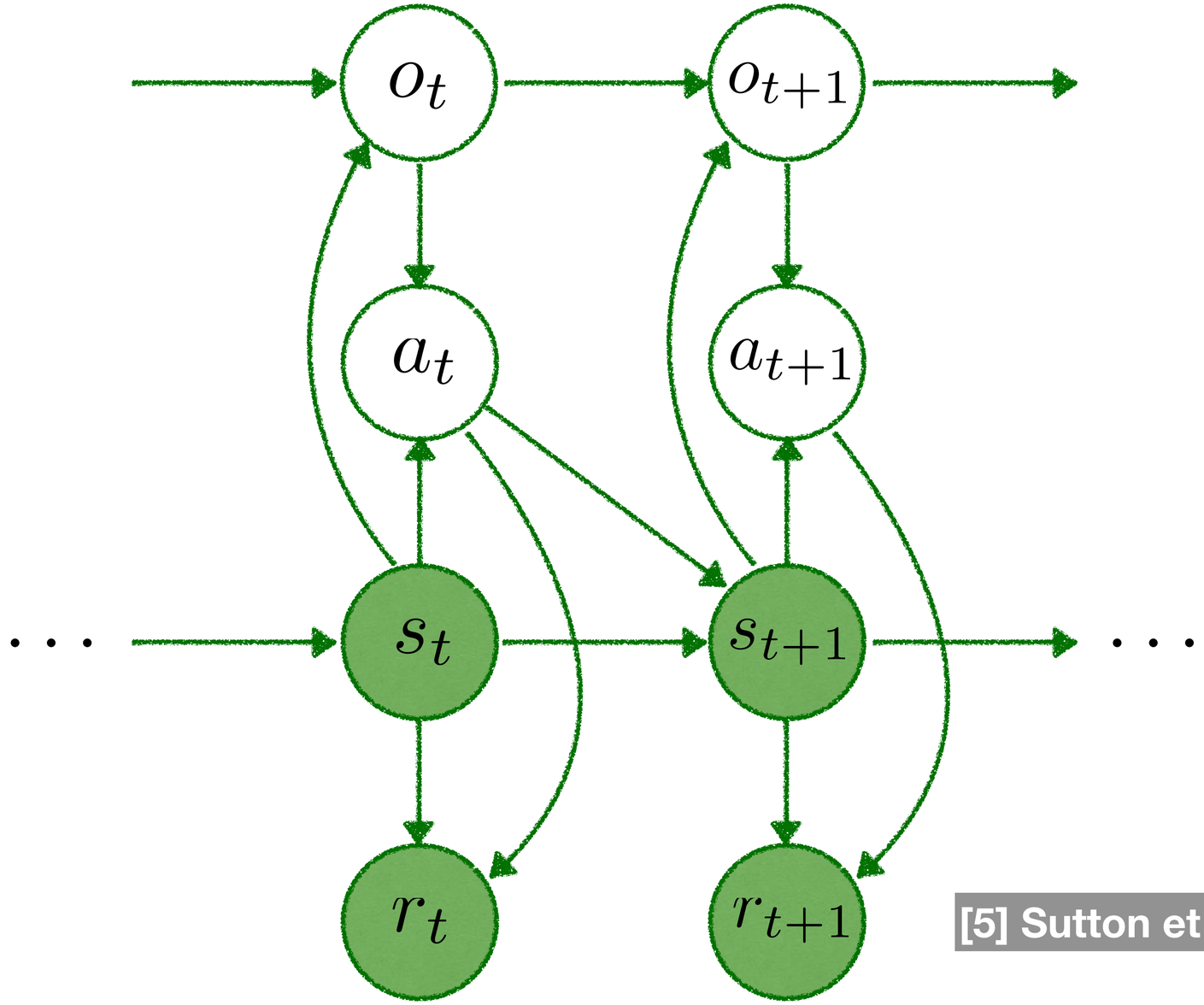


Low-level actions
muscles control



Macro-actions
turn, gain speed, slow down

RL with Options

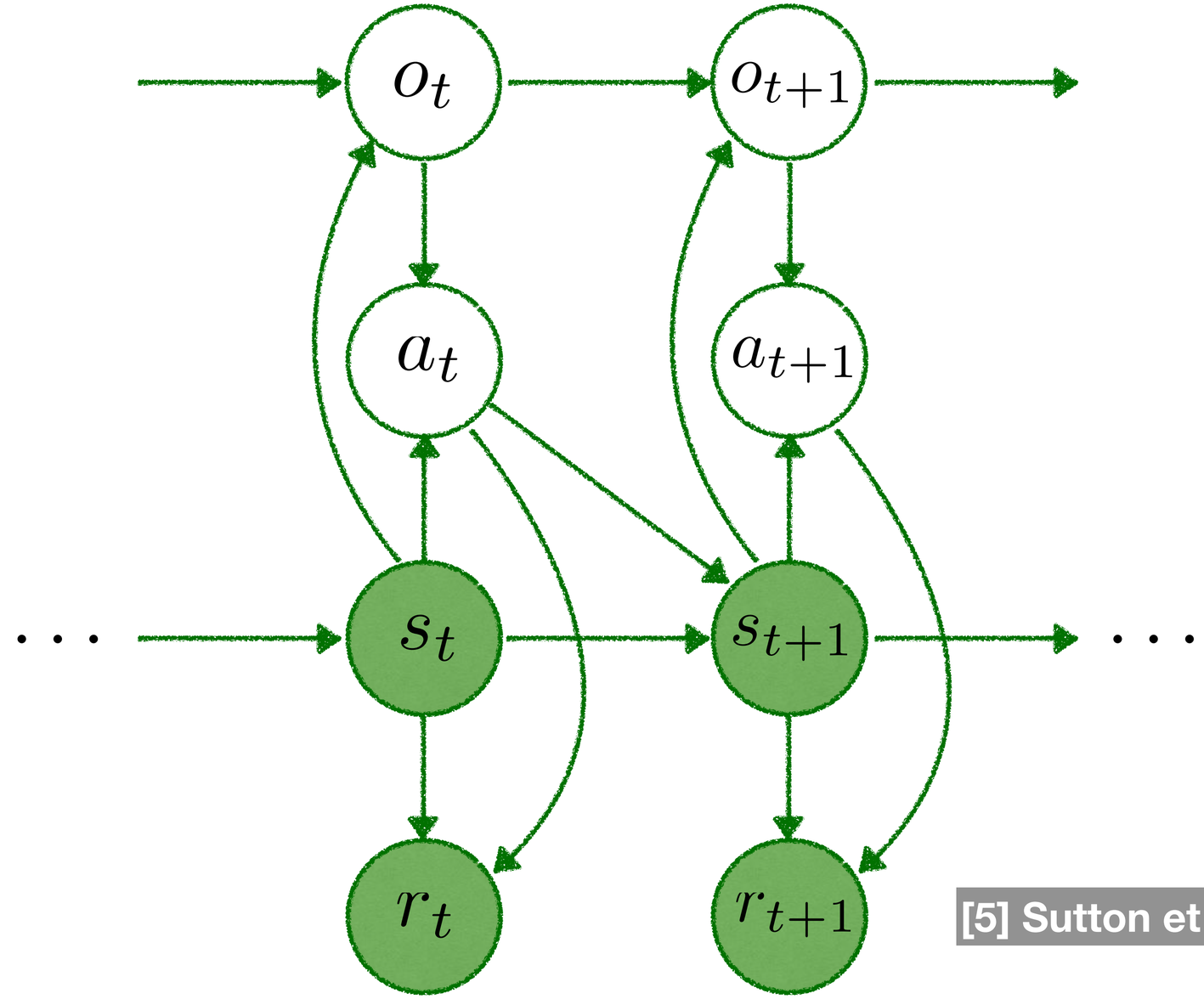


[5] Sutton et al, 1999

RL with Options

Policy

$$a_t \sim \pi(a_t|s_t, o_t)$$



[5] Sutton et al, 1999

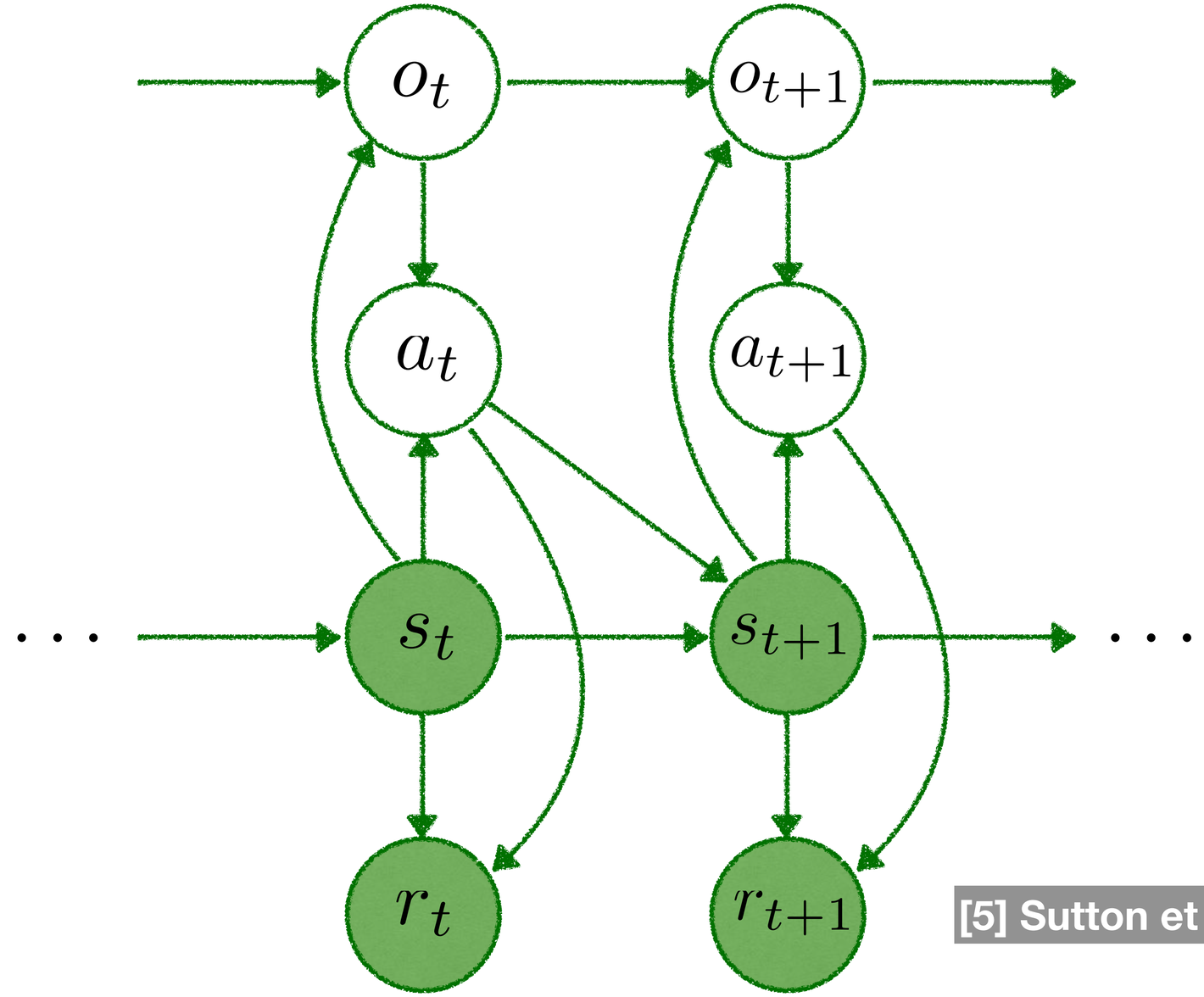
RL with Options

Option-policy

$$o_t \sim \pi(o_t | s_t, o_{t-1})$$

Policy

$$a_t \sim \pi(a_t | s_t, o_t)$$



[5] Sutton et al, 1999

Auxiliary variables in variational inference

Augmented model and the lower bound

Auxiliary variables in variational inference

Augmented model and the lower bound

- Latent variable z, observations x

Auxiliary variables in variational inference

Augmented model and the lower bound

- Latent variable z, observations x
- Introduce auxiliary variables t

Auxiliary variables in variational inference

Augmented model and the lower bound

- Latent variable z, observations x
- Introduce auxiliary variables t
- Define a hierarchical approximation $q(z) = \int q(t)q(z|t)dt$

Auxiliary variables in variational inference

Augmented model and the lower bound

- Latent variable z, observations x
- Introduce auxiliary variables t
- Define a hierarchical approximation $q(z) = \int q(t)q(z|t)dt$

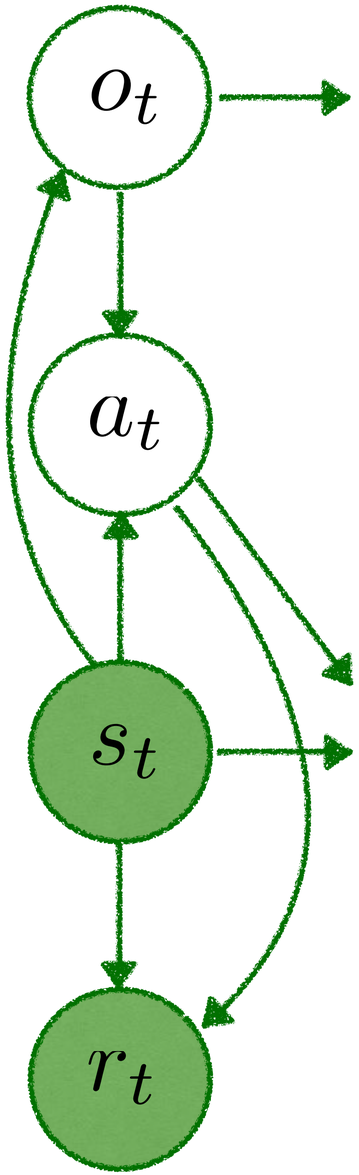
$$\log p(x) = \log \int p(z)p(x|z)\tilde{q}(t|x,z)dzdt$$

$$= \log \int \int q(t)q(z|t)\frac{p(z)\tilde{q}(t|x,z)}{q(t)q(z|t)}p(x|z)dtdz$$

$$\geq \int \int q(t)q(z|t)\log \frac{p(z)\tilde{q}(t|x,z)}{q(t)q(z|t)}p(x|z)dtdz$$

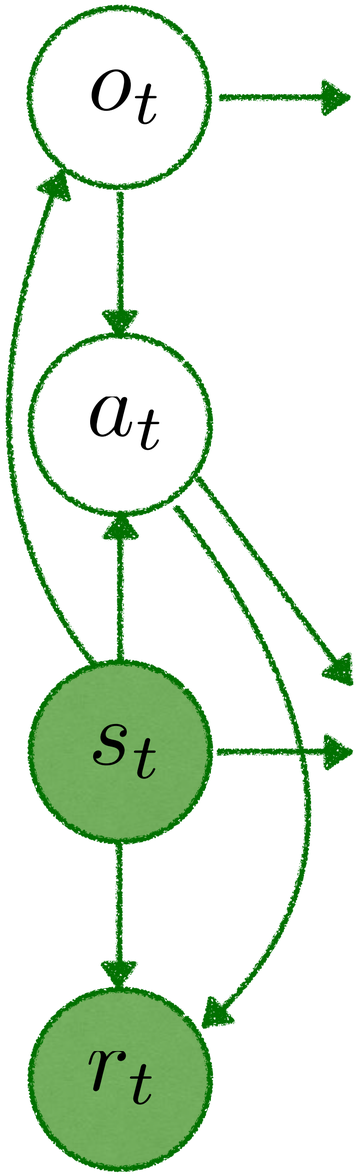
$$= \mathbb{E}_{q(t,z)}\left[\log p(x|z) - \log \frac{q(t)}{\tilde{q}(t|x,z)}\right] - \mathbb{E}_{q(t)}\mathrm{KL}(q(z|t)||p(z))$$

Options as auxiliary variables



Temporally extended execution

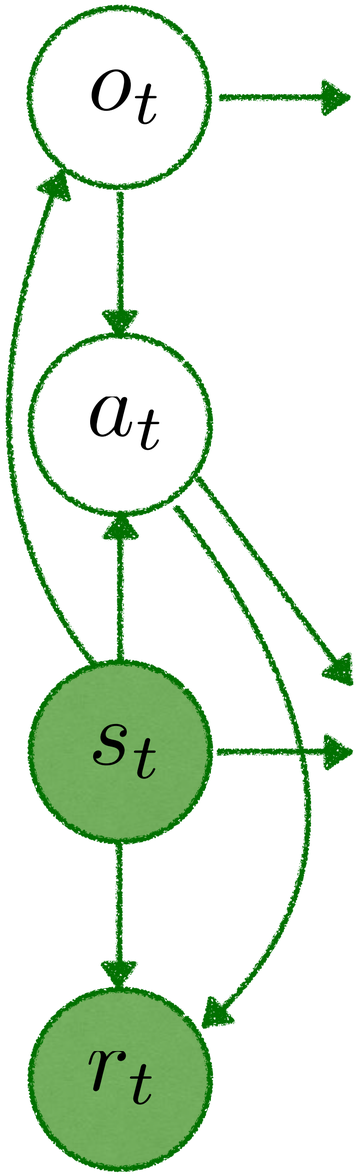
Options as auxiliary variables



Temporally extended execution

- Options enumerate policies: $\pi(a_t|s_t, o_t) = \pi_{o_t}(a_t|s_t)$

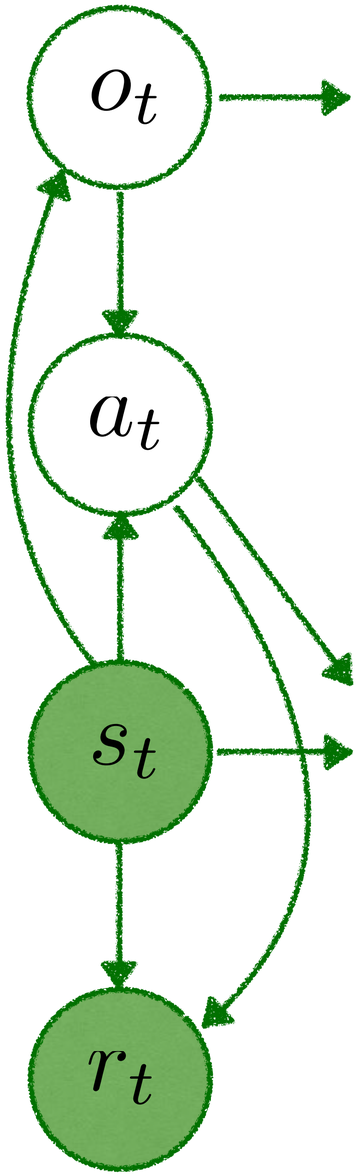
Options as auxiliary variables



Temporally extended execution

- Options enumerate policies: $\pi(a_t|s_t, o_t) = \pi_{o_t}(a_t|s_t)$
- Probability of continuing an option: $q_{\text{cont}} = q_{\text{cont}}(s_t, o_{t-1})$

Options as auxiliary variables

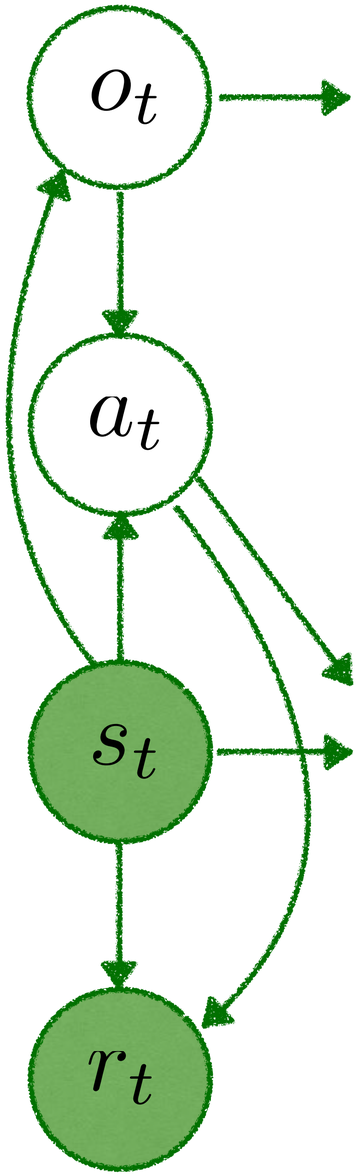


Temporally extended execution

- Options enumerate policies: $\pi(a_t|s_t, o_t) = \pi_{o_t}(a_t|s_t)$
- Probability of continuing an option: $q_{\text{cont}} = q_{\text{cont}}(s_t, o_{t-1})$
- Choosing a policy:

$$\pi(o_t|s_t, o_{t-1}) = q_{\text{cont}}\delta(o_t - o_{t-1}) + (1 - q_{\text{cont}})q(o_t|s_t, o_{t-1})$$

Options as auxiliary variables



Temporally extended execution

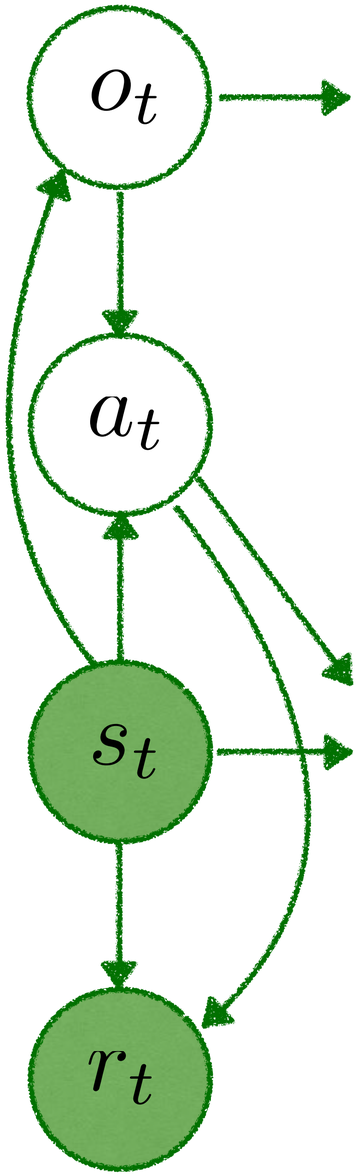
- Options enumerate policies: $\pi(a_t|s_t, o_t) = \pi_{o_t}(a_t|s_t)$
- Probability of continuing an option: $q_{\text{cont}} = q_{\text{cont}}(s_t, o_{t-1})$
- Choosing a policy:

$$\pi(o_t|s_t, o_{t-1}) = q_{\text{cont}}\delta(o_t - o_{t-1}) + (1 - q_{\text{cont}})q(o_t|s_t, o_{t-1})$$

Augmented approximate posterior

$$q(\mathbf{s}, \mathbf{o}, \mathbf{a}) = p(s_1)\pi(o_1|s_1) \prod_{t=1}^{T-1} \left[\pi(a_t|o_t)p(s_{t+1}|s_t, a_t)\pi(o_{t+1}|o_t, s_{t+1})\right]\pi(a_T|s_T, o_T)$$

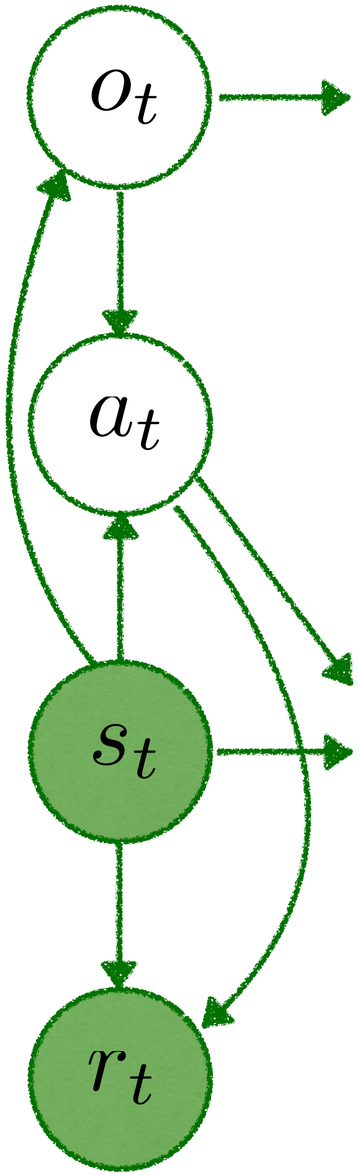
Options as auxiliary variables



Augmented prior

$$p_{\pi_0}(\mathbf{s}, \mathbf{o}, \mathbf{a}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T) \tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a})$$

Options as auxiliary variables

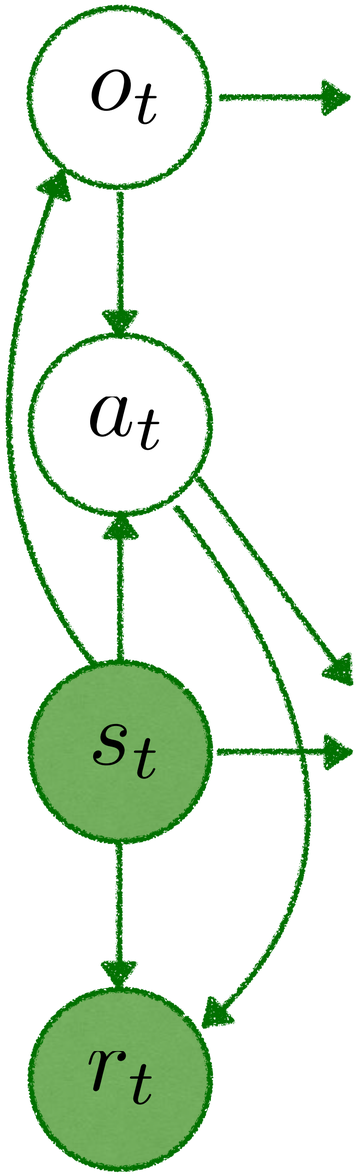


Augmented prior

$$p_{\pi_0}(\mathbf{s}, \mathbf{o}, \mathbf{a}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T) \tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a})$$

- Reverse model is introduced

Options as auxiliary variables

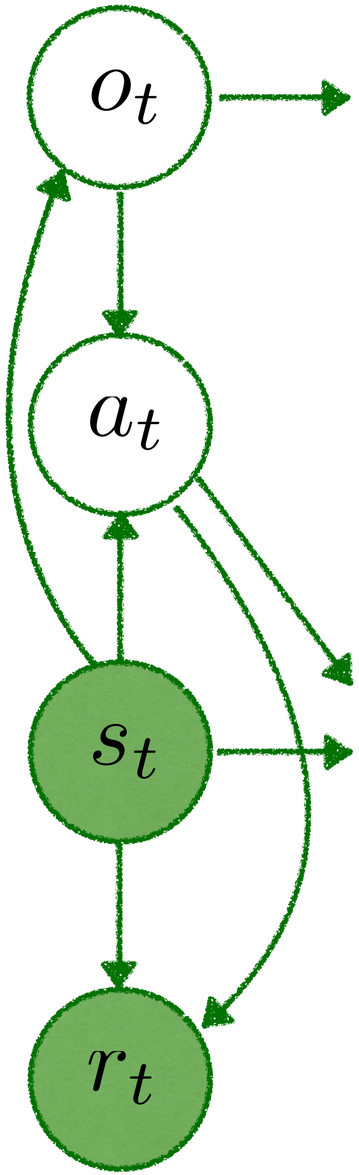


Augmented prior

$$p_{\pi_0}(\mathbf{s}, \mathbf{o}, \mathbf{a}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T) \tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a})$$

- Reverse model is introduced
- Assume $\tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a}) = \prod_{t=1}^{T} \tilde{q}(o_t|\mathbf{o}_{<t}, \mathbf{s}, \mathbf{a})$

Options as auxiliary variables



Augmented prior

$$p_{\pi_0}(\mathbf{s}, \mathbf{o}, \mathbf{a}) = p(s_1) \prod_{t=1}^{T-1} \left[\pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t)\right] \pi_0(a_T|s_T) \tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a})$$

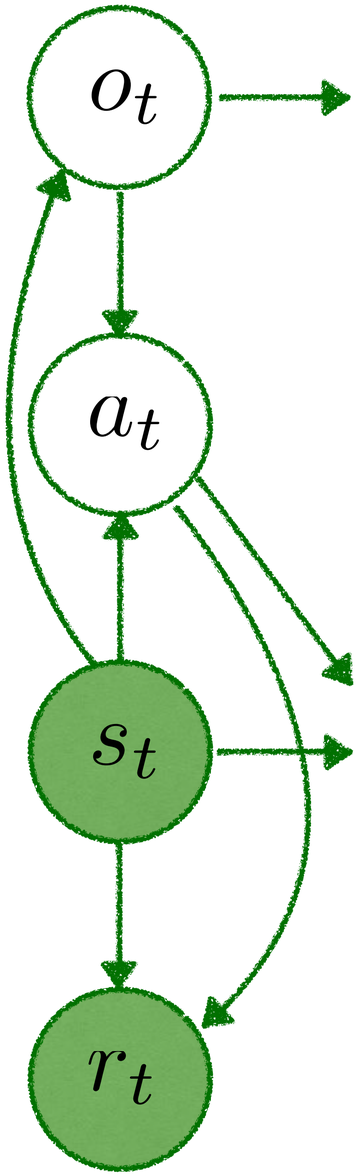
- Reverse model is introduced
- Assume $\tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a}) = \prod_{t=1}^{T} \tilde{q}(o_t|\mathbf{o}_{<t}, \mathbf{s}, \mathbf{a})$

Variational lower bound

$$\log p(\hat{\mathbf{R}} = 1|\mathbf{s}, \mathbf{o}, \mathbf{a}) = \log \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{o},\mathbf{a})} \left[p(\hat{\mathbf{R}} = 1|\mathbf{s}, \mathbf{a}) \frac{p_{\pi_0}(\mathbf{s}, \mathbf{a}) \tilde{q}(\mathbf{o}|\mathbf{s}, \mathbf{a})}{q_\pi(\mathbf{s}, \mathbf{o}, \mathbf{a})} \right]$$

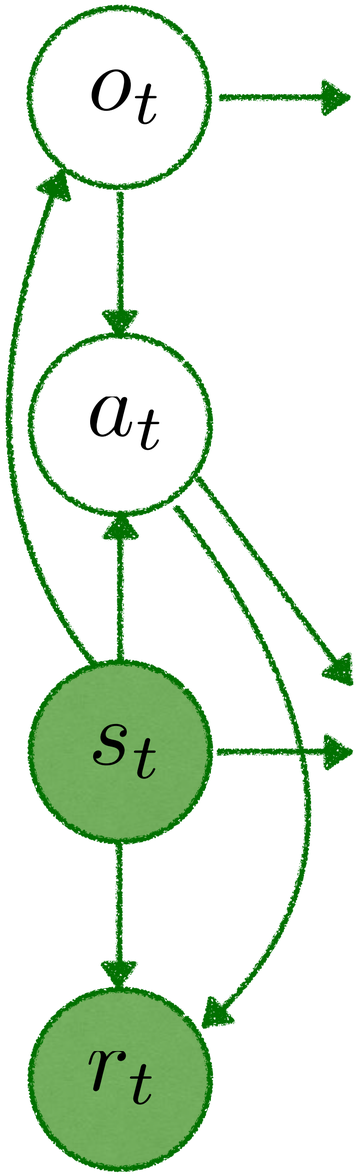
$$\geq \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{o},\mathbf{a})} \left[\sum_{t=1}^{T} \alpha r_t - \mathrm{KL}(\pi_{o_t}(\cdot|s_t)||\pi_0(\cdot|s_t)) - \log \frac{\pi(o_t|o_{t-1}, s_t)}{\tilde{q}(o_t|\mathbf{o}_{<t}, \mathbf{s}, \mathbf{a})} \right]$$

Options as auxiliary variables



Reverse model

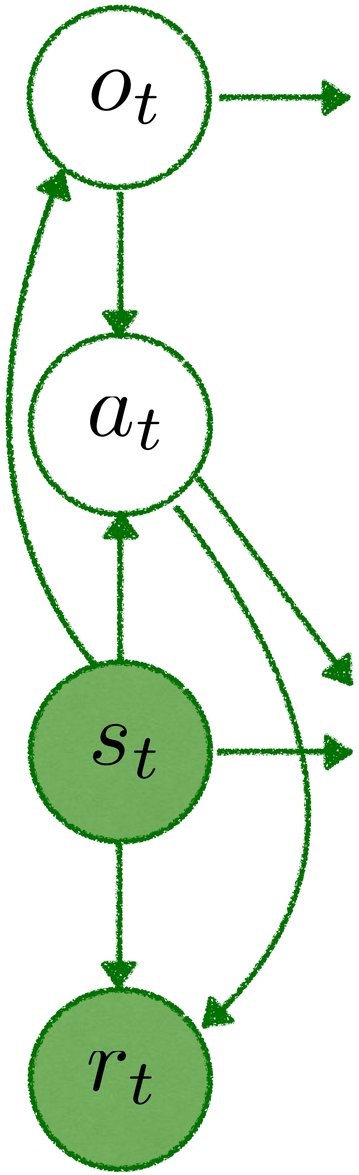
Options as auxiliary variables



Reverse model

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) - \mathrm{KL}(\pi(\cdot|\mathbf{s})||\tilde{q}(\cdot|\mathbf{s},\mathbf{a}))\right]$$

Options as auxiliary variables

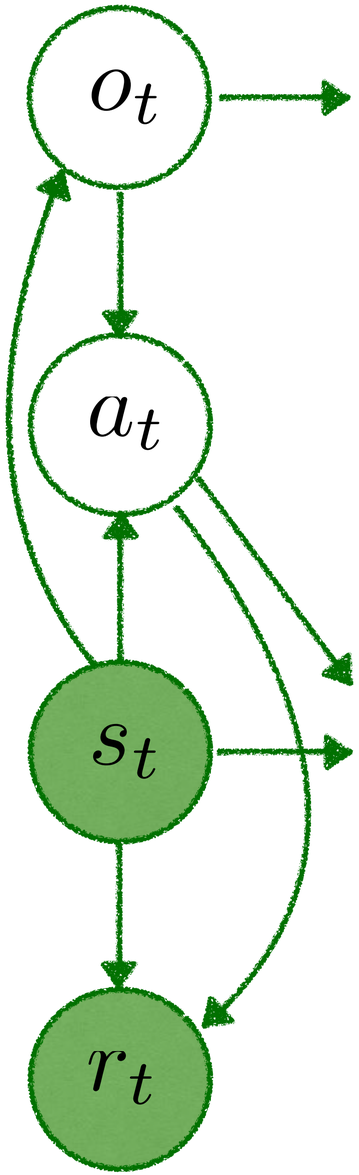


Reverse model

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) - \mathrm{KL}(\pi(\cdot|\mathbf{s})||\tilde{q}(\cdot|\mathbf{s},\mathbf{a}))\right]$$

- Not used in the original options framework

Options as auxiliary variables



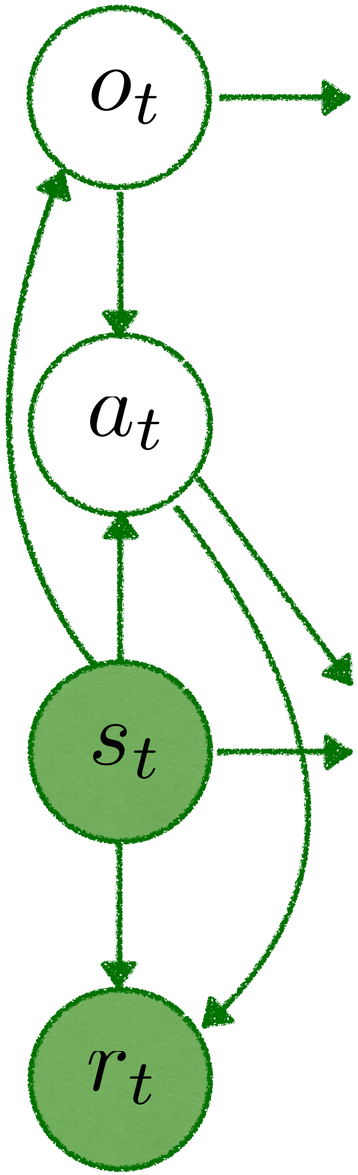
Reverse model

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) - \mathrm{KL}(\pi(\cdot|\mathbf{s})||\tilde{q}(\cdot|\mathbf{s},\mathbf{a}))\right]$$

- Not used in the original options framework
- Helps learning *interpretable* options

$$\log p(\hat{\mathbf{R}} = 1|\mathbf{s},\mathbf{o},\mathbf{a}) \geq \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{o},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha r_t - \mathrm{KL}(\pi_{o_t}(\cdot|s_t)||\pi_0(\cdot|s_t)) - \log \frac{\pi(o_t|o_{t-1},s_t)}{\tilde{q}(o_t|\mathbf{o}_{<t},\mathbf{s},\mathbf{a})}\right]$$

Options as auxiliary variables



Reverse model

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) - \mathrm{KL}(\pi(\cdot|\mathbf{s})||\tilde{q}(\cdot|\mathbf{s},\mathbf{a}))\right]$$

- Not used in the original options framework
- Helps learning *interpretable* options

$$\log p(\hat{\mathbf{R}} = 1|\mathbf{s},\mathbf{o},\mathbf{a}) \geq \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{o},\mathbf{a})}\left[\sum_{t=1}^{T} \alpha r_t - \mathrm{KL}(\pi_{o_t}(\cdot|s_t)||\pi_0(\cdot|s_t)) - \log \frac{\pi(o_t|o_{t-1},s_t)}{\tilde{q}(o_t|\mathbf{o}_{<t},\mathbf{s},\mathbf{a})}\right]$$

- Without reverse model options are easier to ignore

$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s},\mathbf{o},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t)) + \mathcal{H}(\pi(o_t|s_t,o_{t-1}))\right]$$

Further reading

Further reading

- Uncertainty over parameters

Further reading

- Uncertainty over parameters
- Exploration

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL
- Intrinsic motivation

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL
- Intrinsic motivation
- Other picks on Hierarchical RL

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL
- Intrinsic motivation
- Other picks on Hierarchical RL
- Other picks on Bayesian RL

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL
- Intrinsic motivation
- Other picks on Hierarchical RL
- Other picks on Bayesian RL
  - there are quite a few

Further reading

- Uncertainty over parameters
- Exploration
- Distributional RL
- Intrinsic motivation
- Other picks on Hierarchical RL
- Other picks on Bayesian RL
  - there are quite a few

See references!

Thank you!

Questions?

References

[0] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

[1] Rawlik, K., Toussaint, M., & Vijayakumar, S. (2012, July). On stochastic optimal control and reinforcement learning by approximate inference. In *Robotics: science and systems*.

[2] Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, *8*(3-4), 229-256.

[3] Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., ... & Kavukcuoglu, K. (2016, June). Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning* (pp. 1928-1937).

[4] Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)* (pp. 1889-1897).

References

[5] Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, *112*(1-2), 181-211.

[6] Jaderberg, M., Mnih, V., Czarnecki, W. M., Schaul, T., Leibo, J. Z., Silver, D., & Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. *arXiv preprint arXiv: 1611.05397*.

[7] Weber, T., Racanière, S., Reichert, D. P., Buesing, L., Guez, A., Rezende, D. J., ... & Pascanu, R. (2017). Imagination-Augmented Agents for Deep Reinforcement Learning. *arXiv preprint arXiv:1707.06203*.

[8] Schulman, John, et al. "Proximal Policy Optimization Algorithms." *arXiv preprint arXiv: 1707.06347* (2017).

[9] Huszár, Ferenc. "Variational Inference using Implicit Distributions." *arXiv preprint arXiv: 1702.08235* (2017).

References

[10] Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in Neural Information Processing Systems*. 2016.

[11] Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2015). Bayesian reinforcement learning: A survey. Foundations and Trends® in Machine Learning, 8(5-6), 359-483.

[12] Houthooft, R., Chen, X., Duan, Y., Schulman, J., De Turck, F., & Abbeel, P. (2016). Vime: Variational information maximizing exploration. In Advances in Neural Information Processing Systems (pp. 1109-1117).

[13] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. *arXiv preprint arXiv:1707.06887*.

[14] Gregor, K., Rezende, D. J., & Wierstra, D. (2016). Variational Intrinsic Control. *arXiv preprint arXiv:1611.07507*.

[15] Mohamed, S., & Rezende, D. J. (2015). Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems* (pp. 2125-2133).

References

[16] Neu, Gergely, Anders Jonsson, and Vicenç Gómez. "A unified view of entropy-regularized markov decision processes." *arXiv preprint arXiv:1705.07798* (2017).

[17] Haarnoja, Tuomas, et al. "Reinforcement learning with deep energy-based policies." *arXiv preprint arXiv:1702.08165* (2017).

[18] Schulman, John, Xi Chen, and Pieter Abbeel. "Equivalence between policy gradients and soft q-learning." arXiv preprint arXiv:1704.06440 (2017).

[19] Levine, Sergey. "Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review." *arXiv preprint arXiv:1805.00909* (2018).

[20] Burda, Yuri, Roger Grosse, and Ruslan Salakhutdinov. "Importance weighted autoencoders." *arXiv preprint arXiv:1509.00519* (2015).