Bayesian Inference for RL

Sergey Bartunov
DeepMind, HSE

Outline

Outline

- (very brief) Introduction to Reinforcement Learning

Outline

- (very brief) Introduction to Reinforcement Learning
- Solving MDP via variational inference

Outline

- (very brief) Introduction to Reinforcement Learning
- Solving MDP via variational inference
- Stable policy gradients

Outline

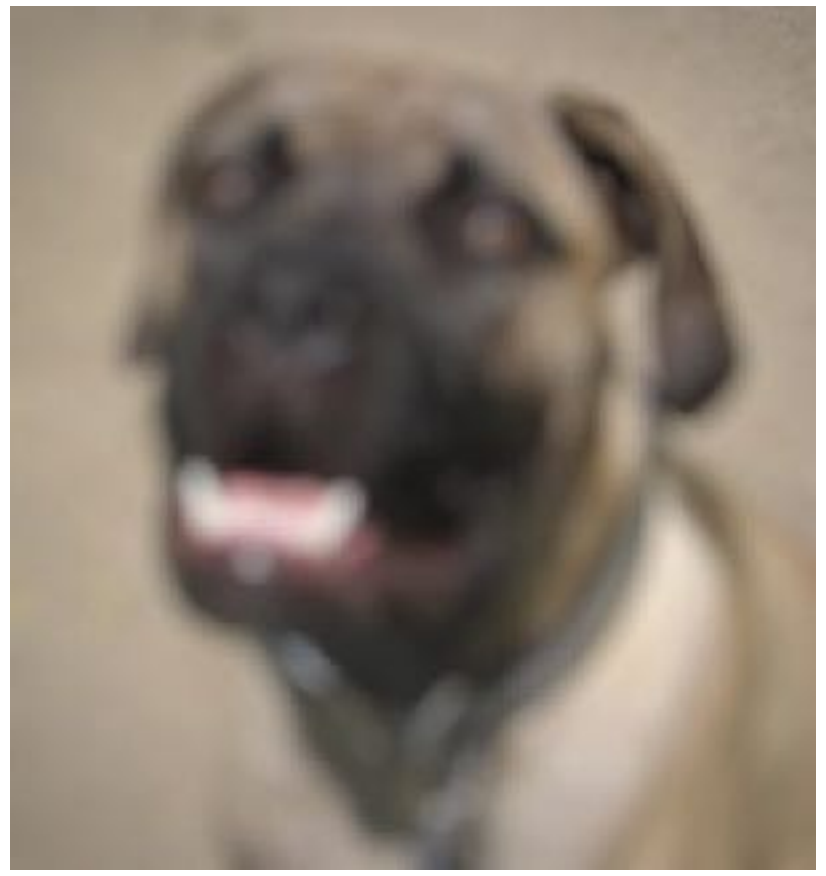
- (very brief) Introduction to Reinforcement Learning
- Solving MDP via variational inference
- Stable policy gradients
- Hierarchical RL with Options as auxiliary variables

Outline

- (very brief) Introduction to Reinforcement Learning
- Solving MDP via variational inference
- Stable policy gradients
- Hierarchical RL with Options as auxiliary variables
- Adversarial inference for model-based RL

Unsupervised Learning

Prediction



$p(x)$

Model

Ground-truth

Objective $\log p(x)$

Supervised Learning

Observation

$p(y|x)$

Model

Prediction

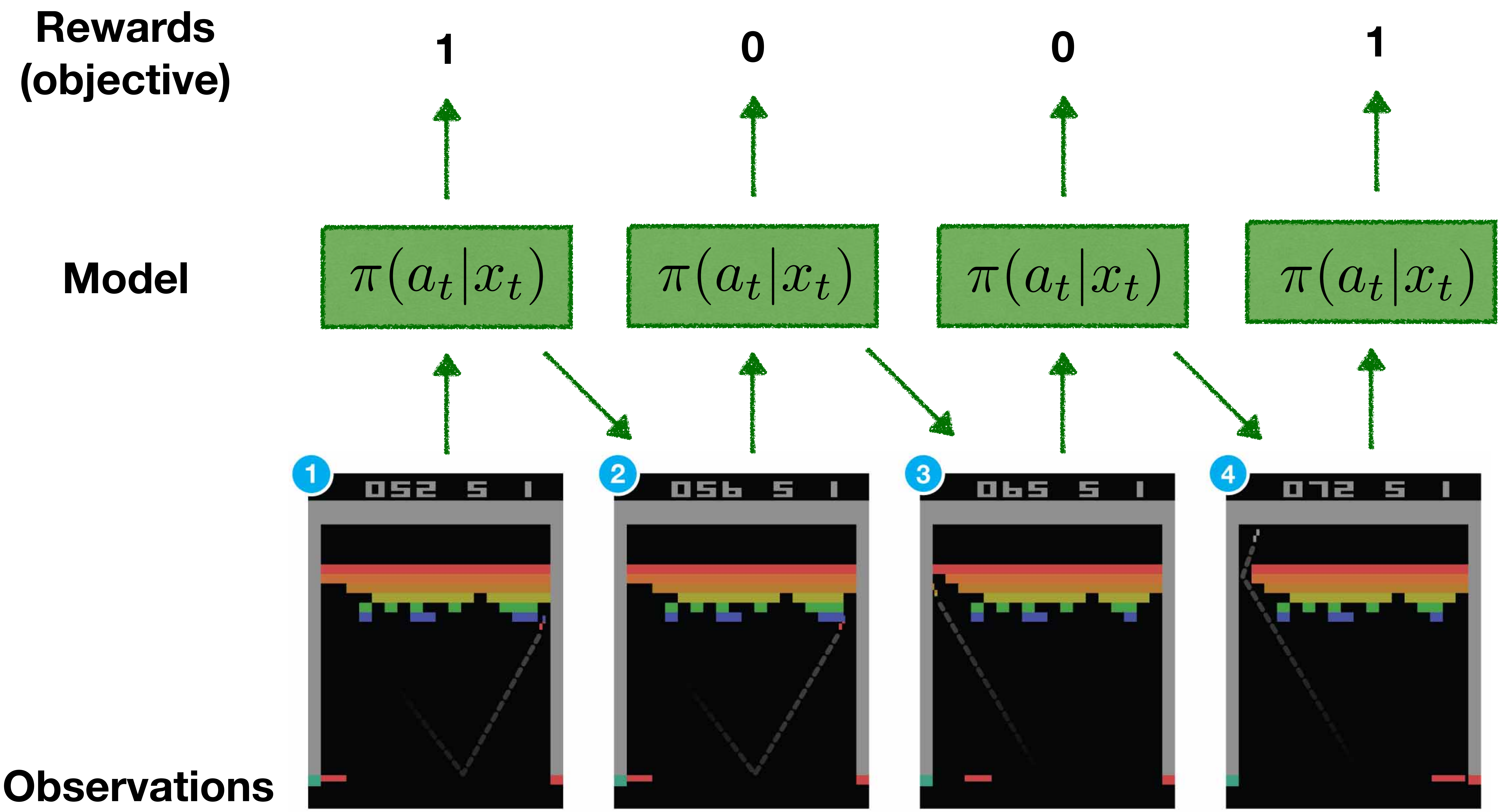
cat

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
- No right or wrong decisions
  - Only more or less optimal behaviors

Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- No right or wrong decisions
  - Only more or less optimal behaviors
- Potentially sparse, noisy and delayed rewards
  - Credit assignment problem

Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- No right or wrong decisions
  - Only more or less optimal behaviors
- Potentially sparse, noisy and delayed rewards
  - Credit assignment problem
- Exploration / exploitation trade-off

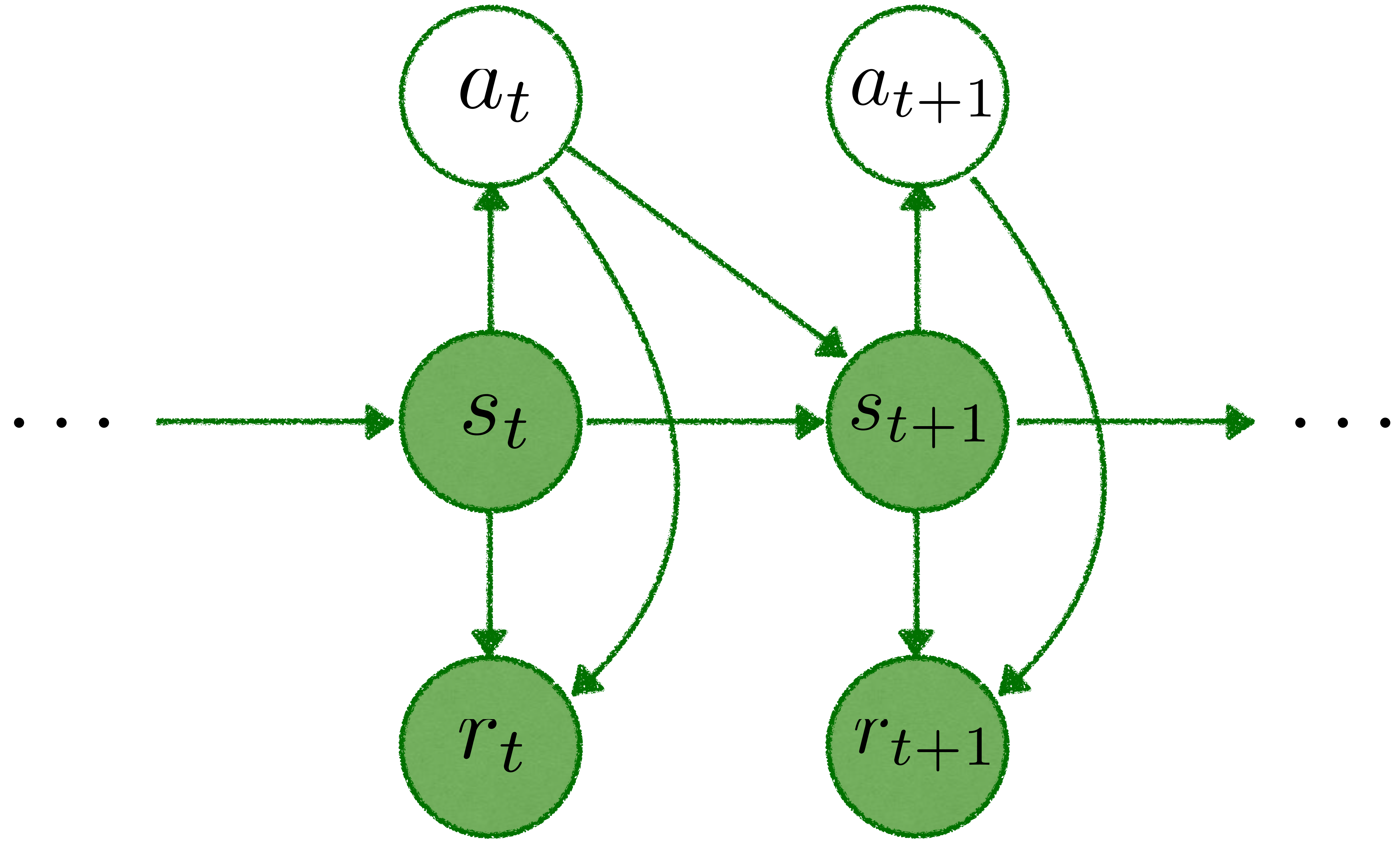
Sutton & Barto, 1998

Reinforcement Learning

- Sequential decision making
- No right or wrong decisions
  - Only more or less optimal behaviors
- Potentially sparse, noisy and delayed rewards
  - Credit assignment problem
- Exploration / exploitation trade-off
- Partial observability

Sutton & Barto, 1998

Markov Decision Process

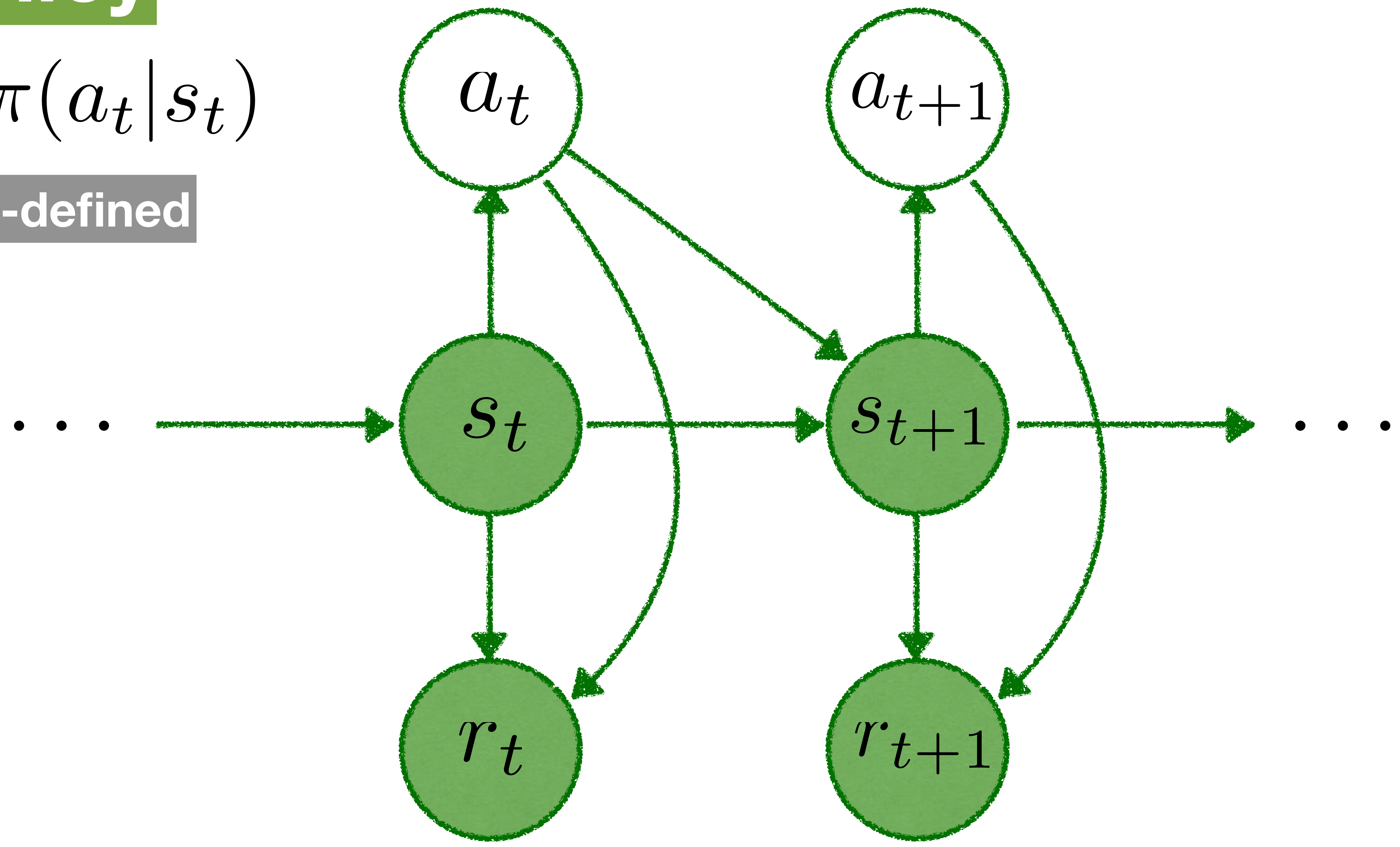


Markov Decision Process

Policy

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

Agent-defined

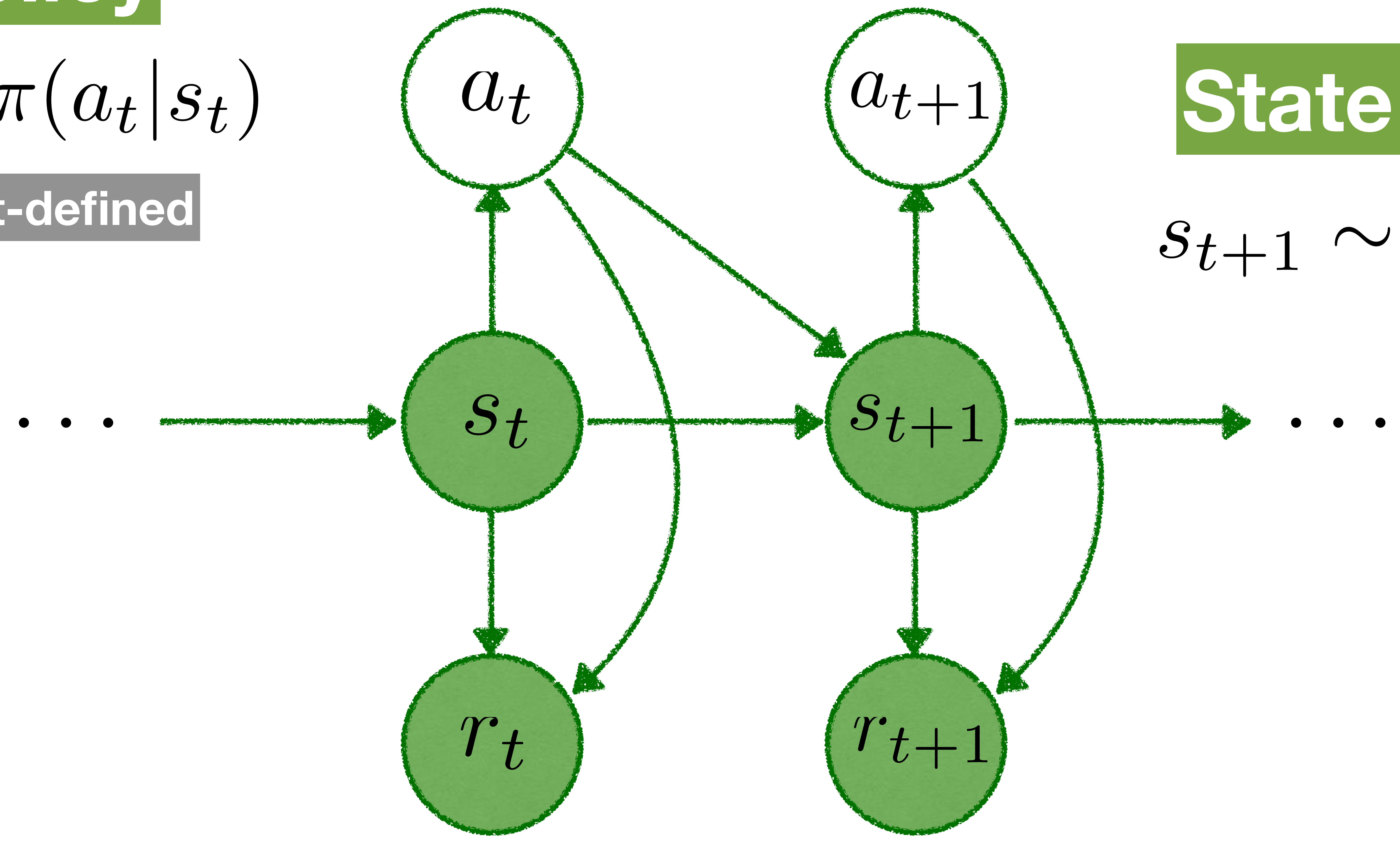


Markov Decision Process

Policy

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

Agent-defined



State transitions

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

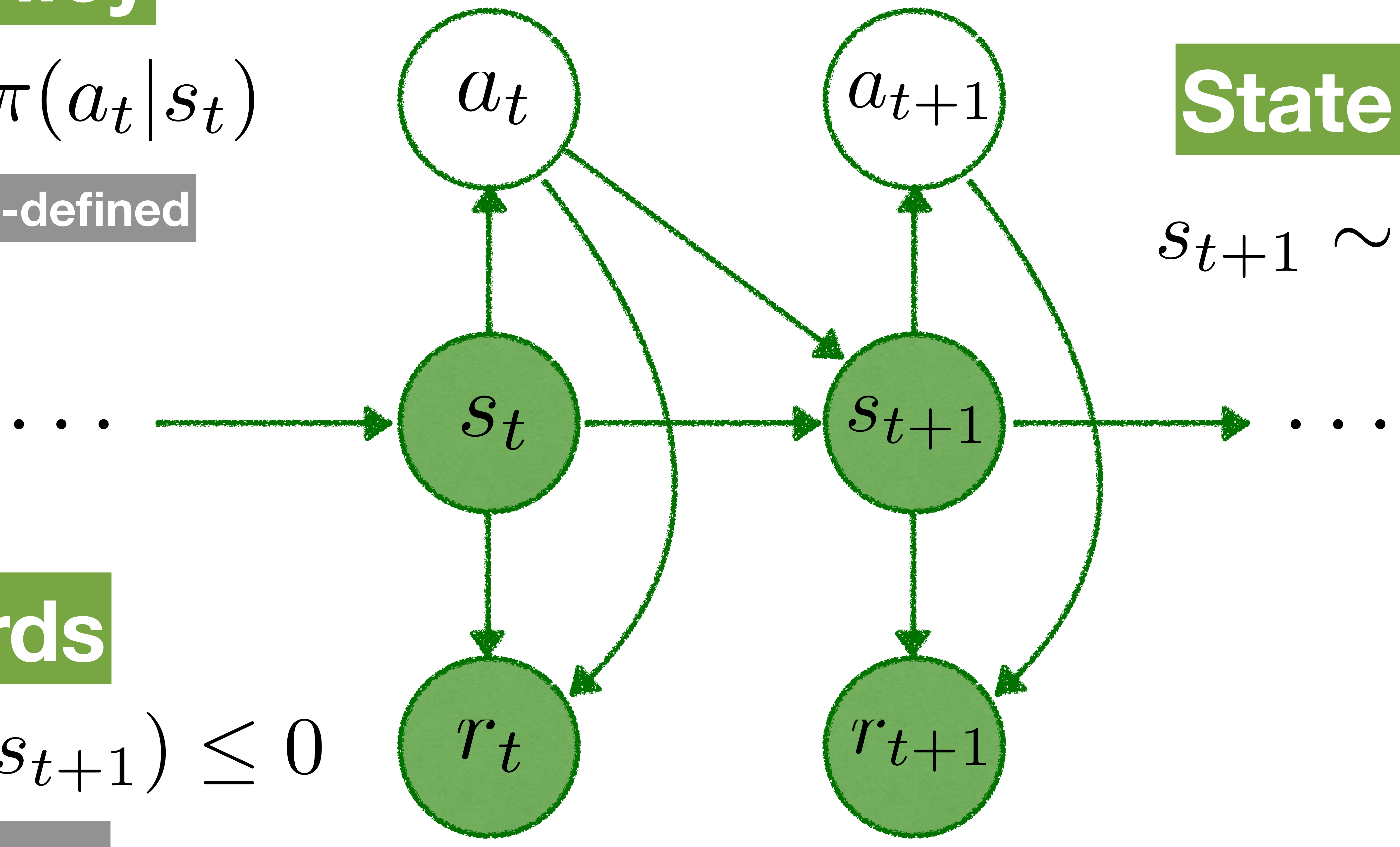
Likely unknown

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

$$r_t = r(s_t, a_t, s_{t+1}) \leq 0$$

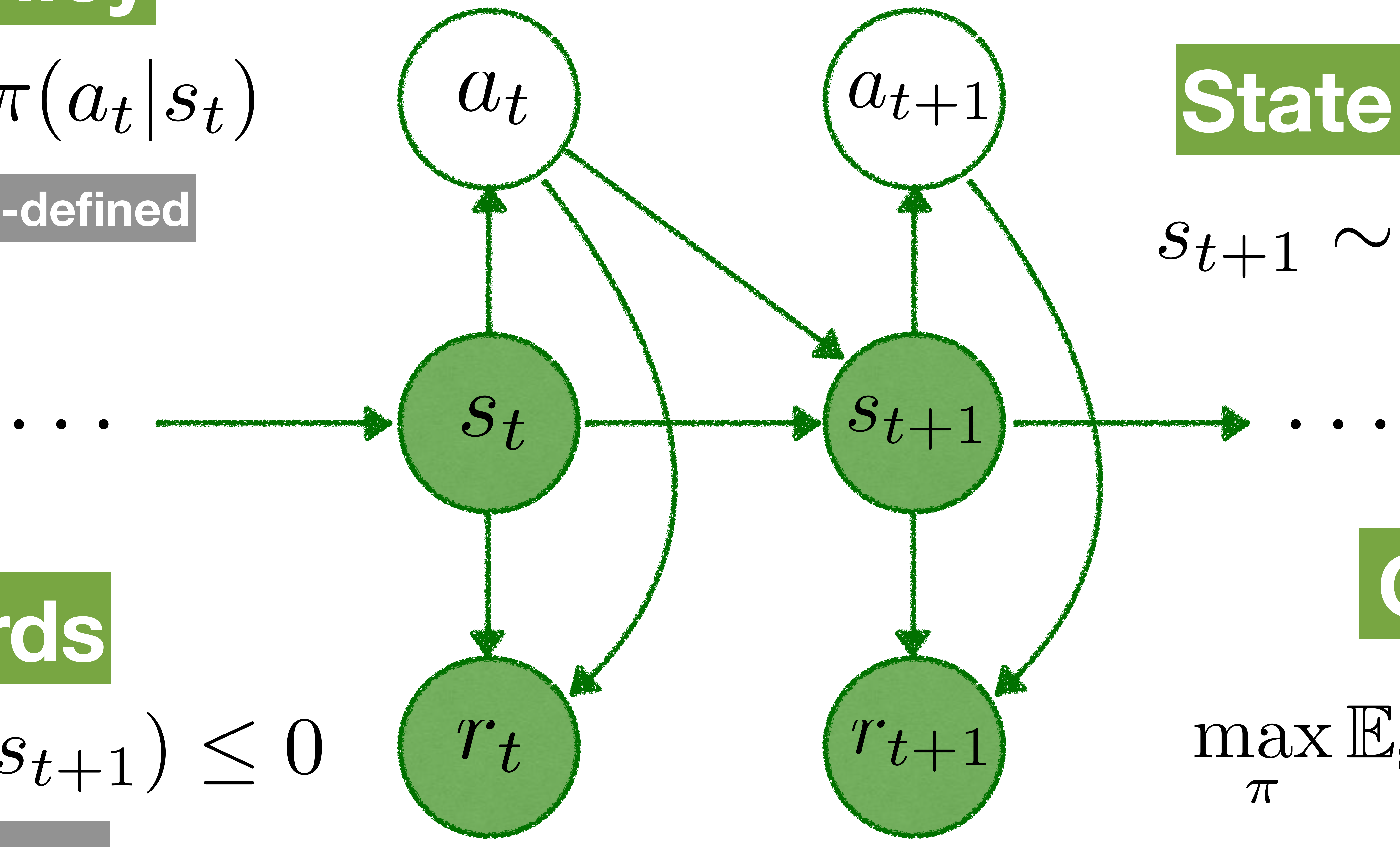
Likely unknown

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

$$r_t = r(s_t, a_t, s_{t+1}) \leq 0$$

Likely unknown

Goal

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

Variational inference

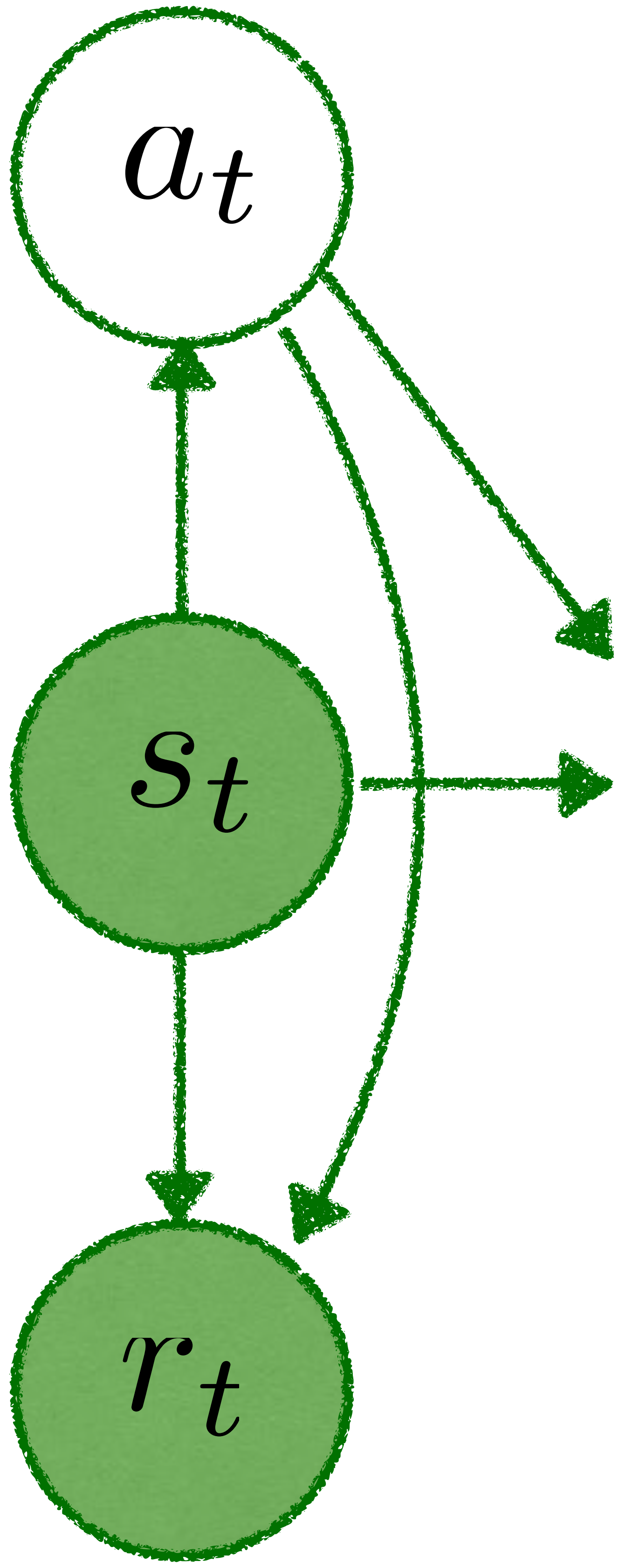
Latent-variable model

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

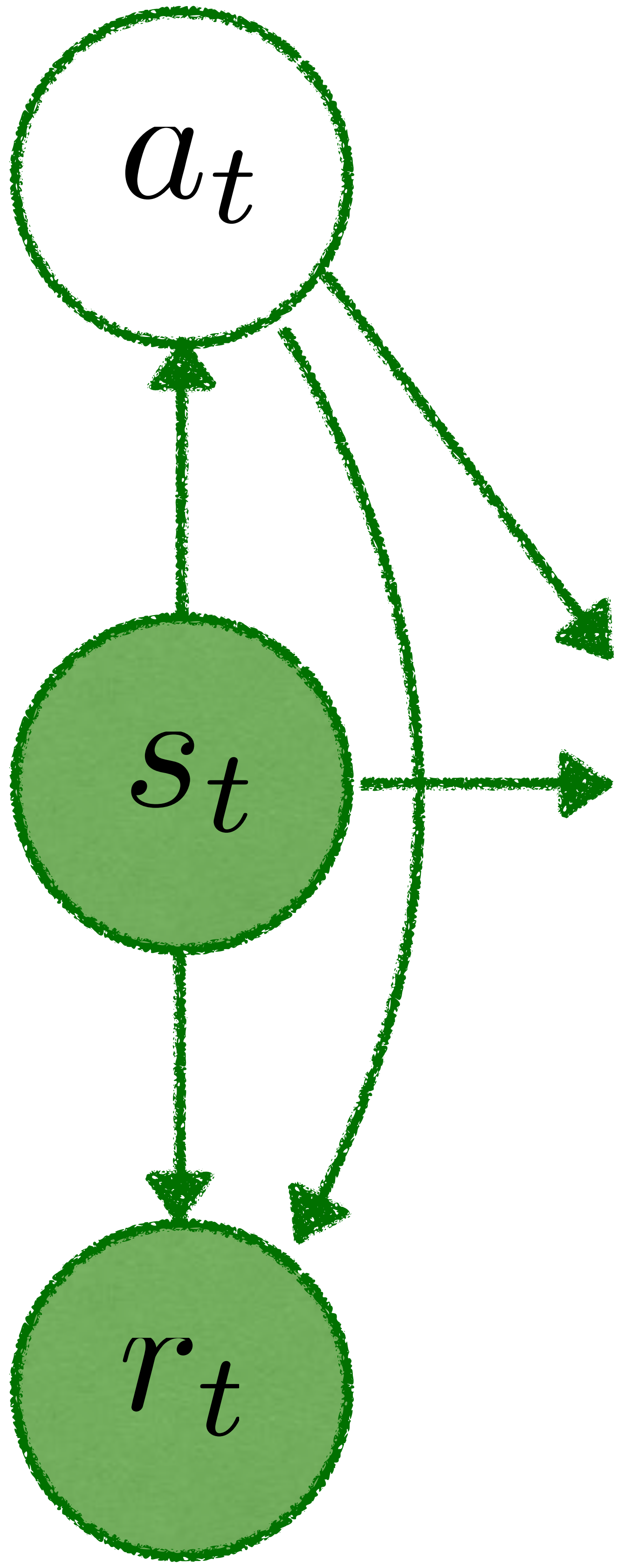
Variational lower bound

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

MDP as a probabilistic model



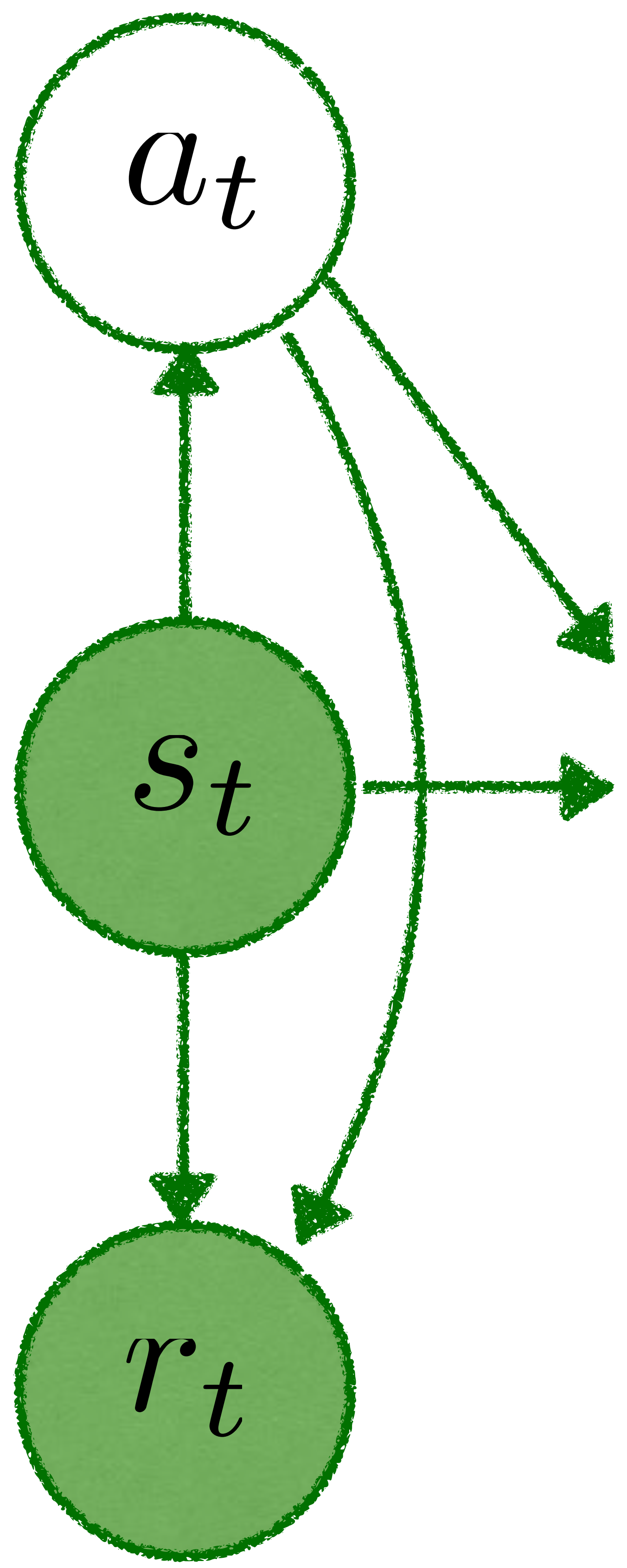
MDP as a probabilistic model



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



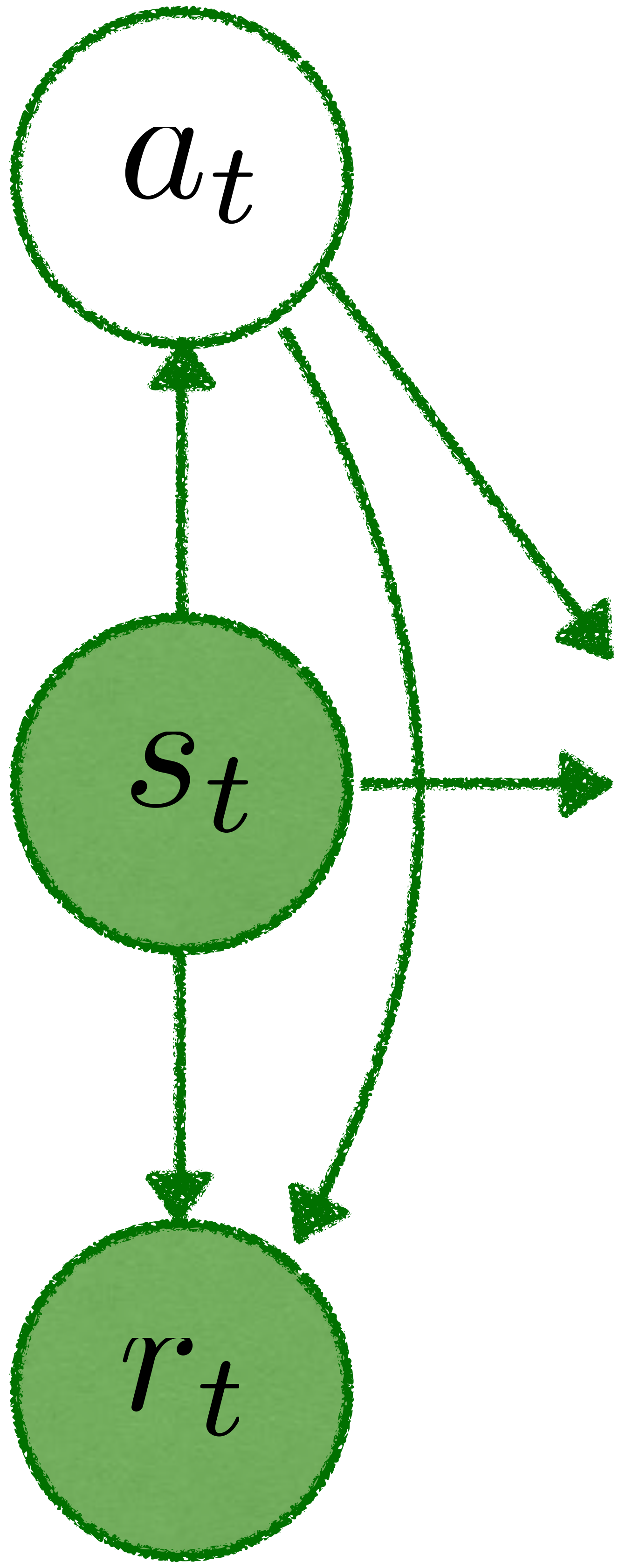
Prior (w.r.t. some policy)

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

Likelihood

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

MDP as a probabilistic model



Prior (w.r.t. some policy)

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

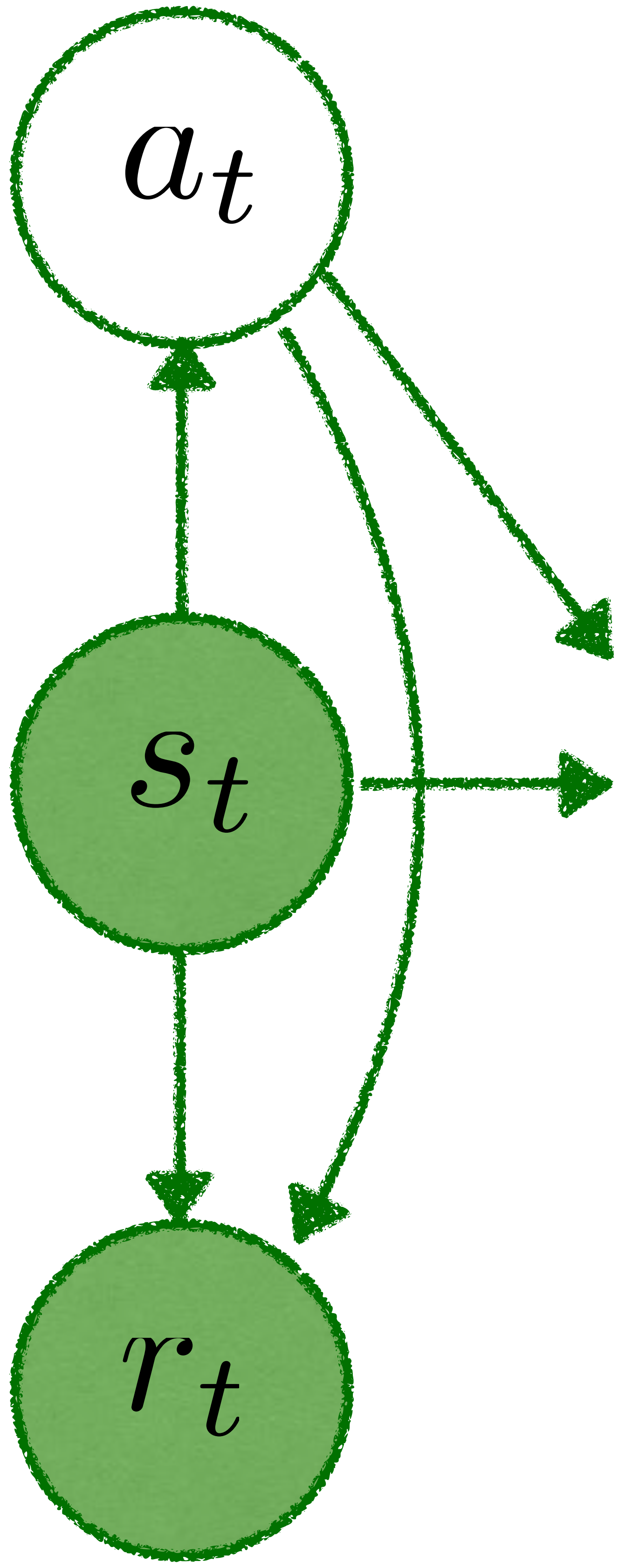
Likelihood

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

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)

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

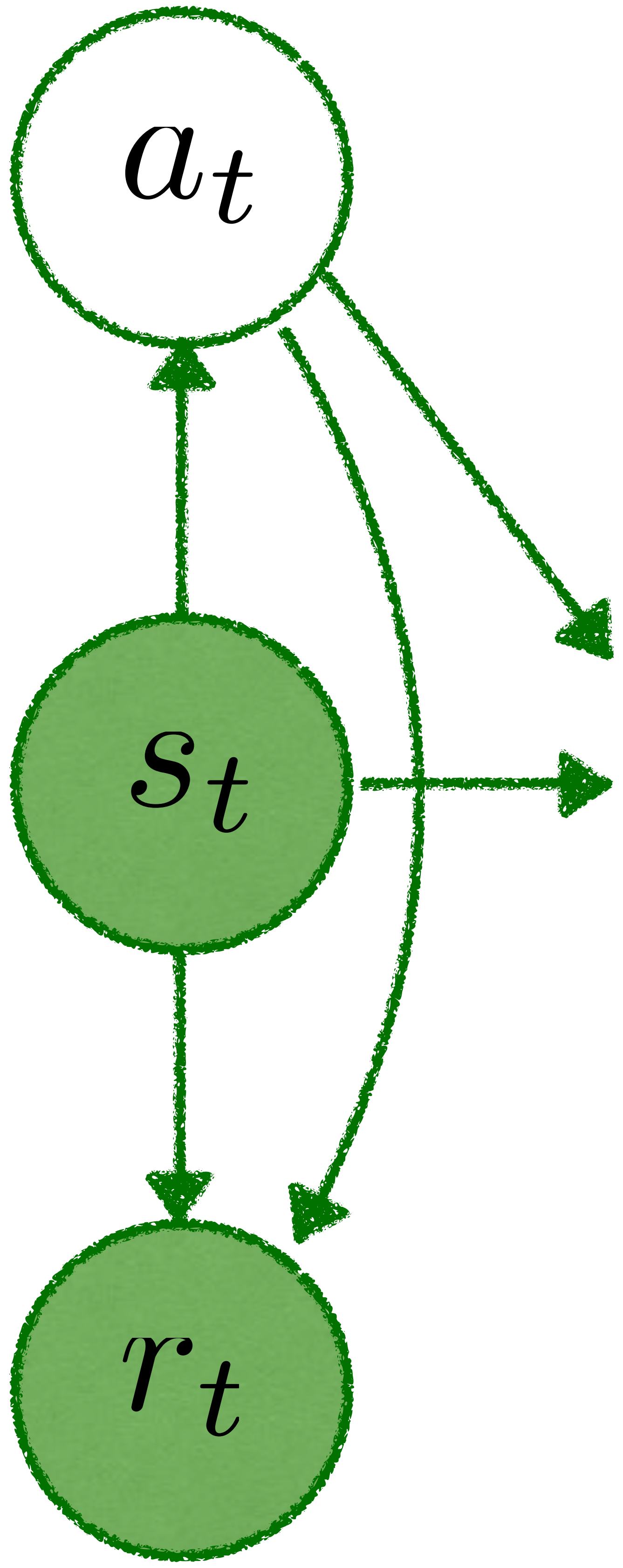
Likelihood

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

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

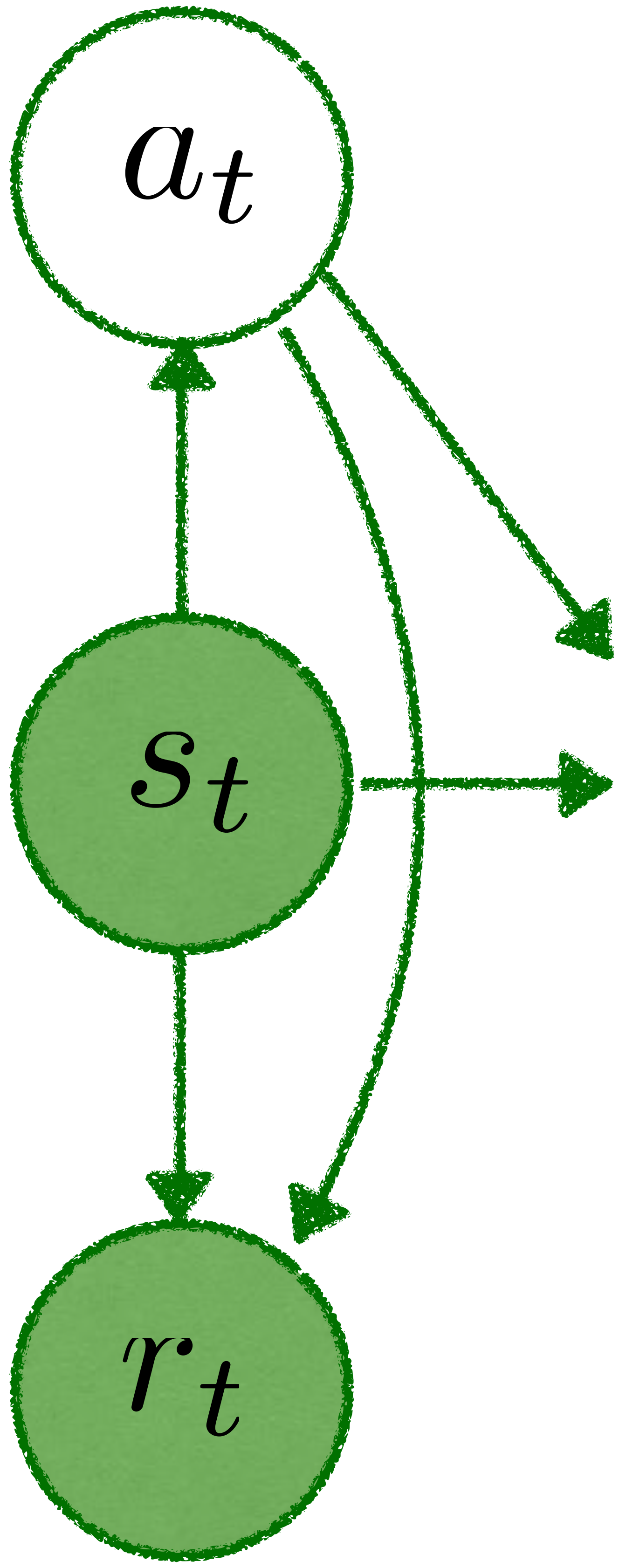
$$q_{\pi}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(s_1) \prod_{t=1}^{T-1} [\underbrace{\pi(a_t|s_t)}_{\textbf{Different}}\underbrace{p(s_{t+1}|s_t, a_t)}_{\textbf{Same}}]\, \pi(a_T|s_T)$$

Solving MDP via approximate inference



[1] Rawlik et al, 2012

Solving MDP via approximate inference

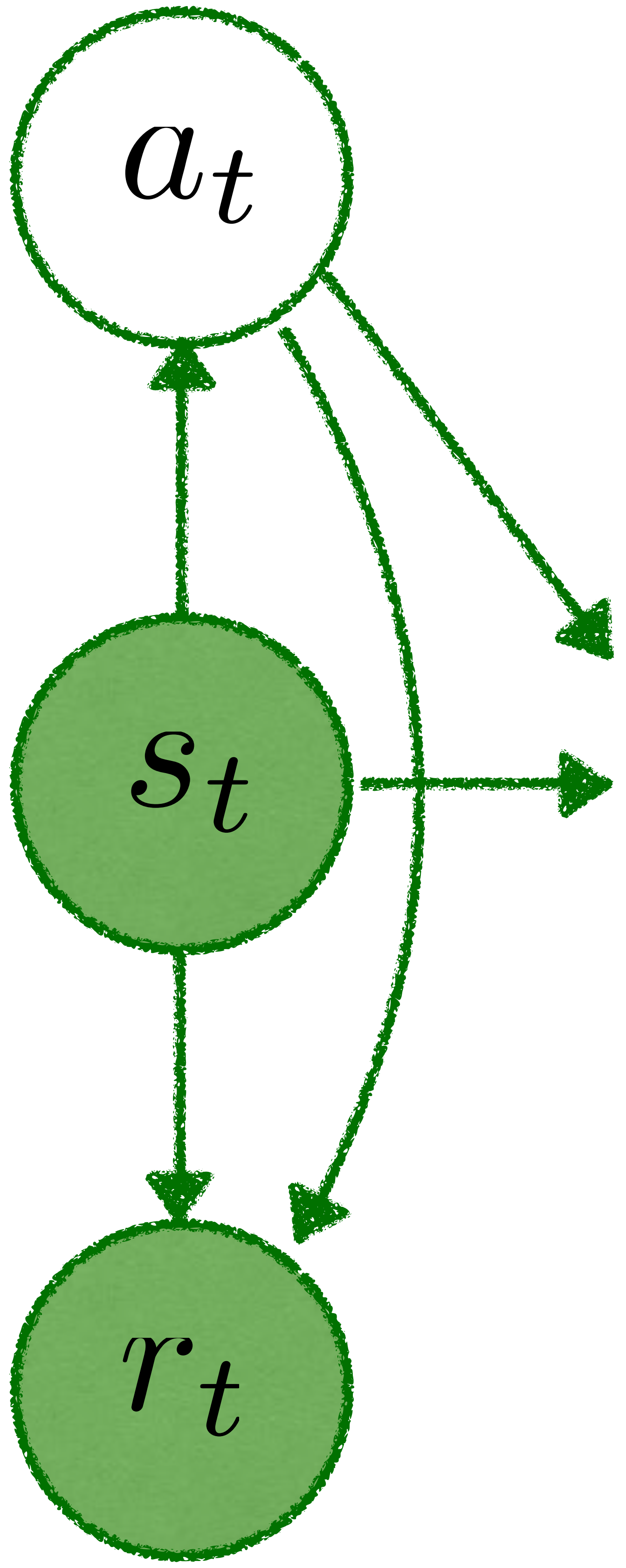


Marginal likelihood

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

[1] Rawlik et al, 2012

Solving MDP via approximate inference



Marginal likelihood

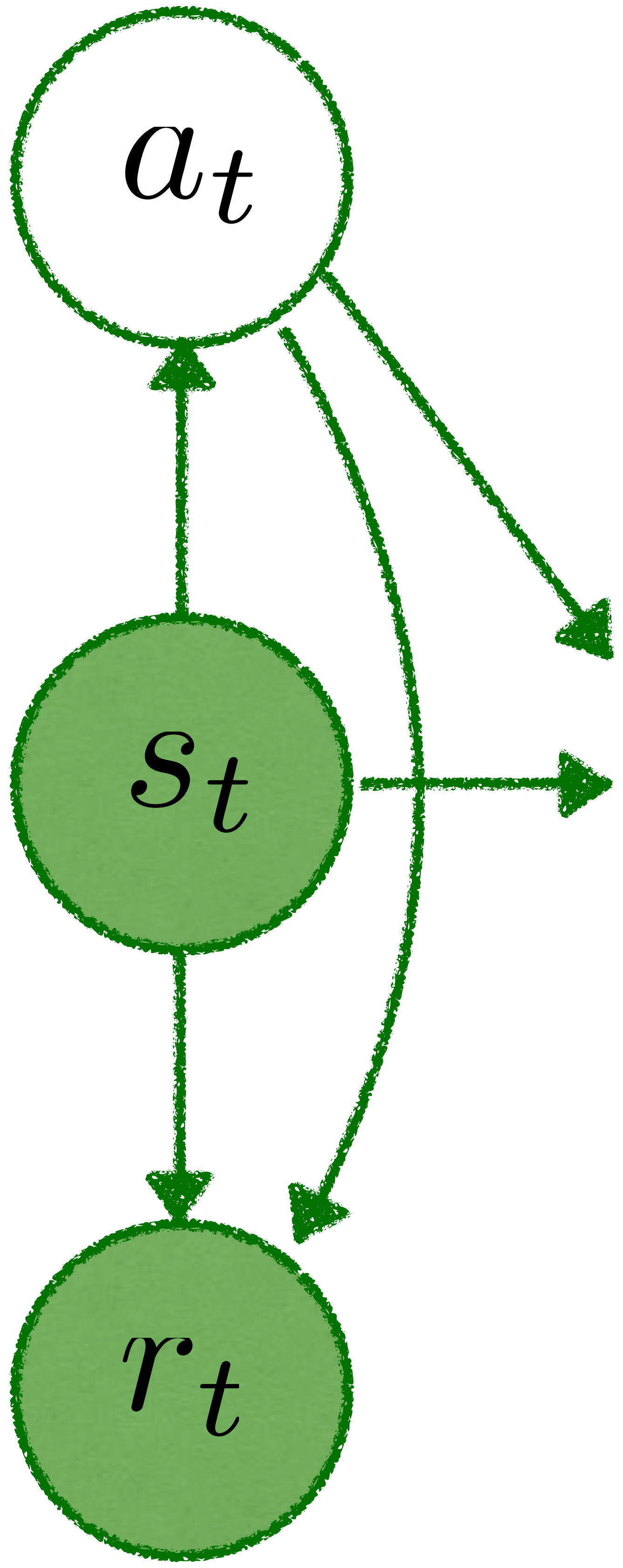
$$\log p(\hat{\mathbf{R}}_{1:T}) = \log \mathbb{E}_{p_{\pi_0}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T},\mathbf{a}_{1:T})$$

Variational lower bound

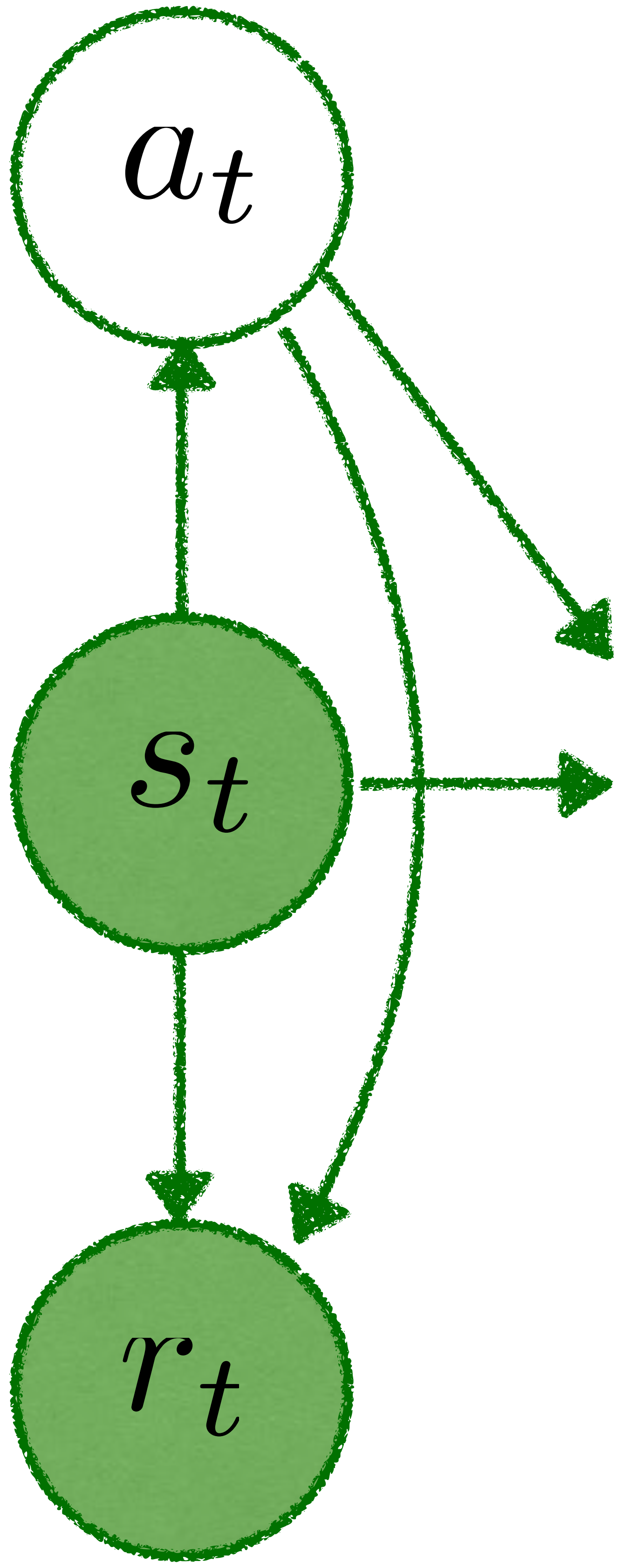
$$\begin{aligned}\log p(\hat{\mathbf{R}}_{1:T}) &= \log \mathbb{E}_{q_\pi(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} \frac{p_{\pi_0}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})}{q_\pi(\mathbf{s}_{1:T},\mathbf{a}_{1:T})} p(\hat{\mathbf{R}}_{1:T}|\mathbf{s}_{1:T},\mathbf{a}_{1:T}) \\ &\geq \mathbb{E}_{q_\pi(\mathbf{s}_{1:T},\mathbf{a}_{1:T})}\Big[\alpha \sum_{t=1}^{T} r_t\Big] - \mathrm{KL}(q_\pi(\mathbf{s}_{1:T},\mathbf{a}_{1:T})||p_{\pi_0}(\mathbf{s}_{1:T},\mathbf{a}_{1:T})) \\ &= \mathcal{L}(q_\pi, p_{\pi_0})\end{aligned}$$

[1] Rawlik et al, 2012

Policy gradients as inference



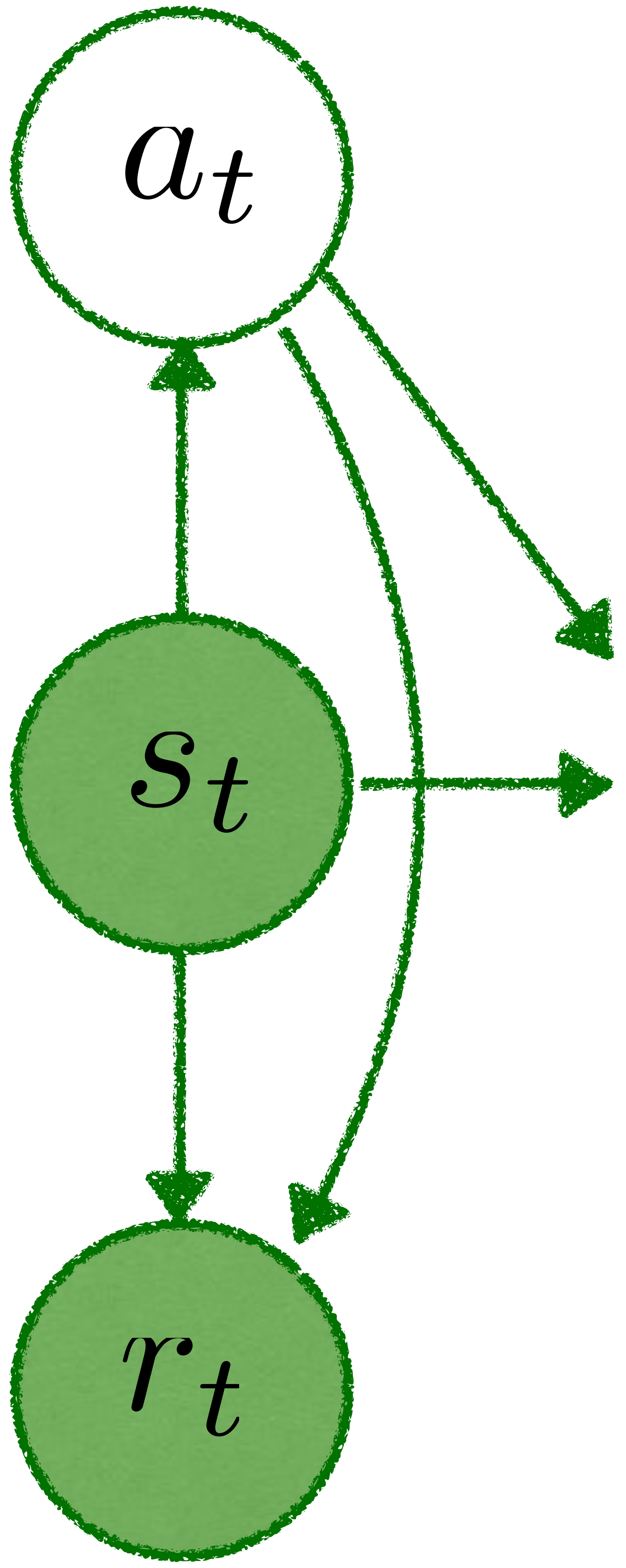
Policy gradients as inference



Variational lower bound

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

Policy gradients as inference

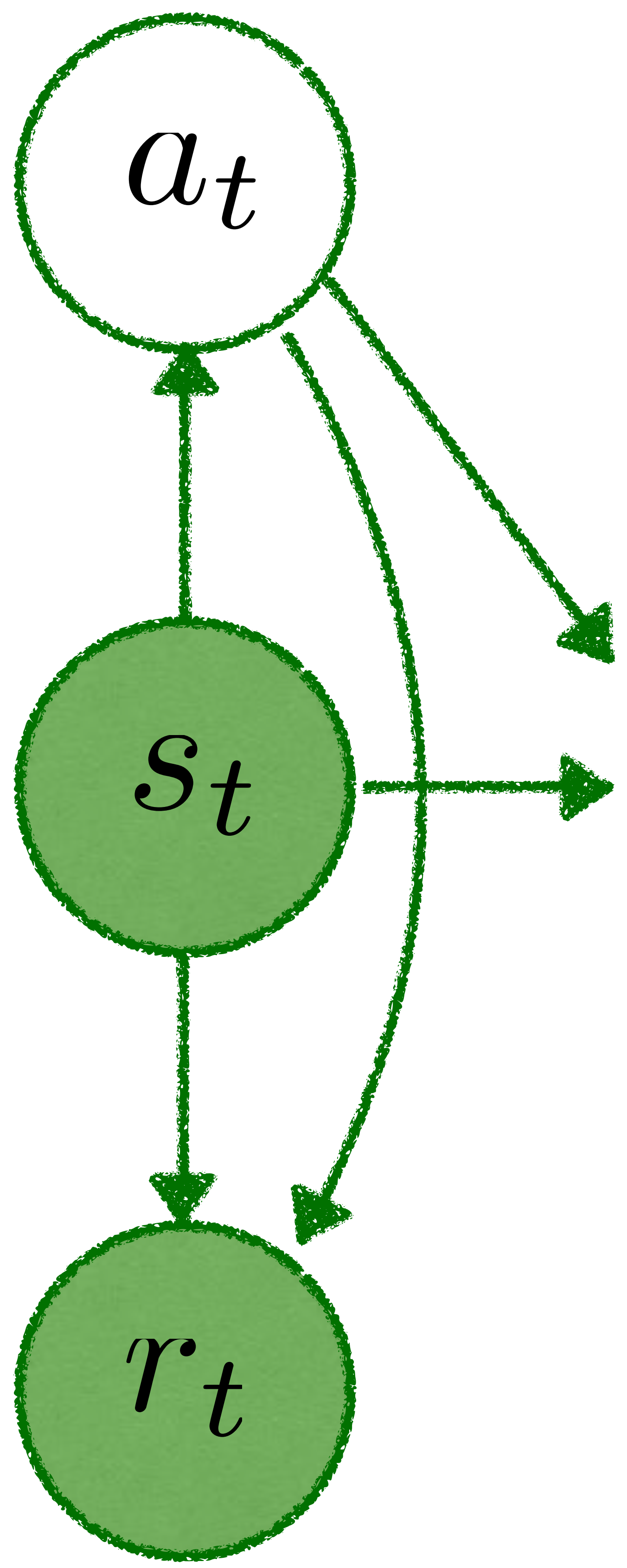


Variational lower bound

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

Recovering policy gradients algorithm

Policy gradients as inference



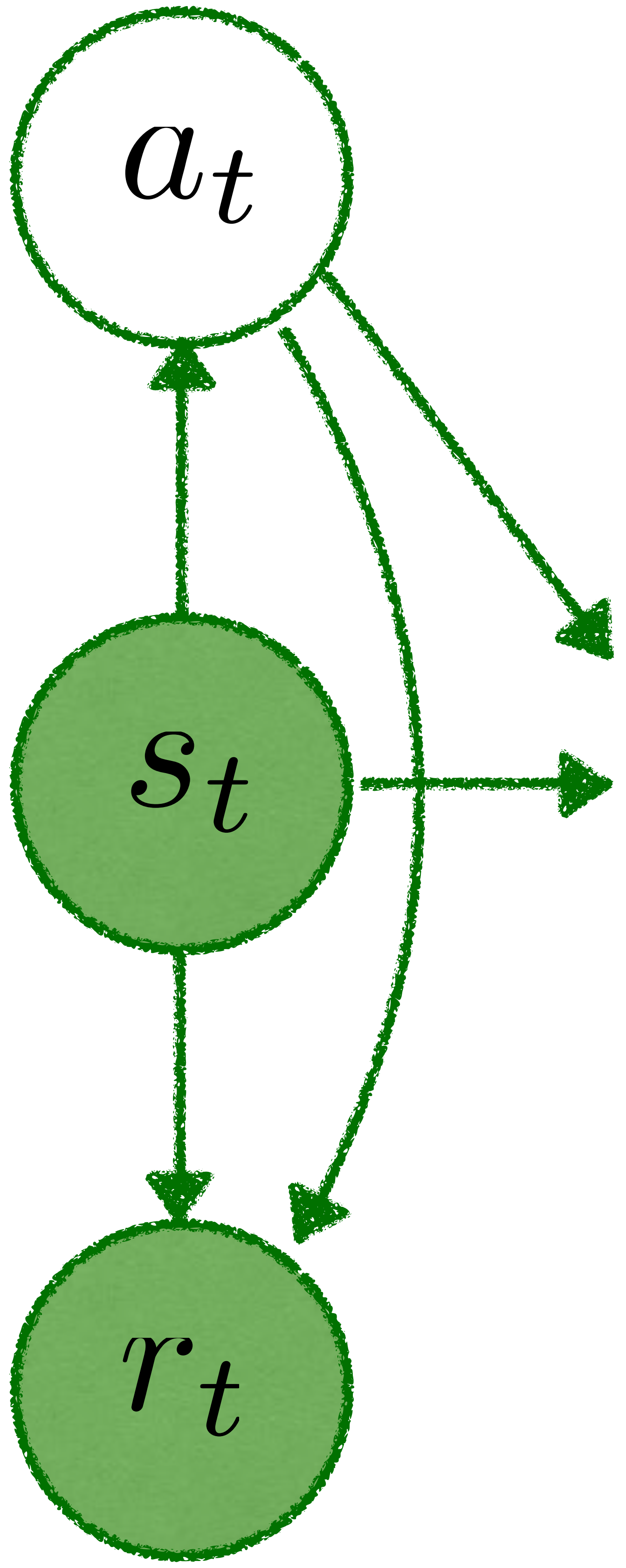
Variational lower bound

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

Recovering policy gradients algorithm

- Consider $\pi_0(a_t|s_t) = \mathrm{Uniform}(a_t)$

Policy gradients as inference



Variational lower bound

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

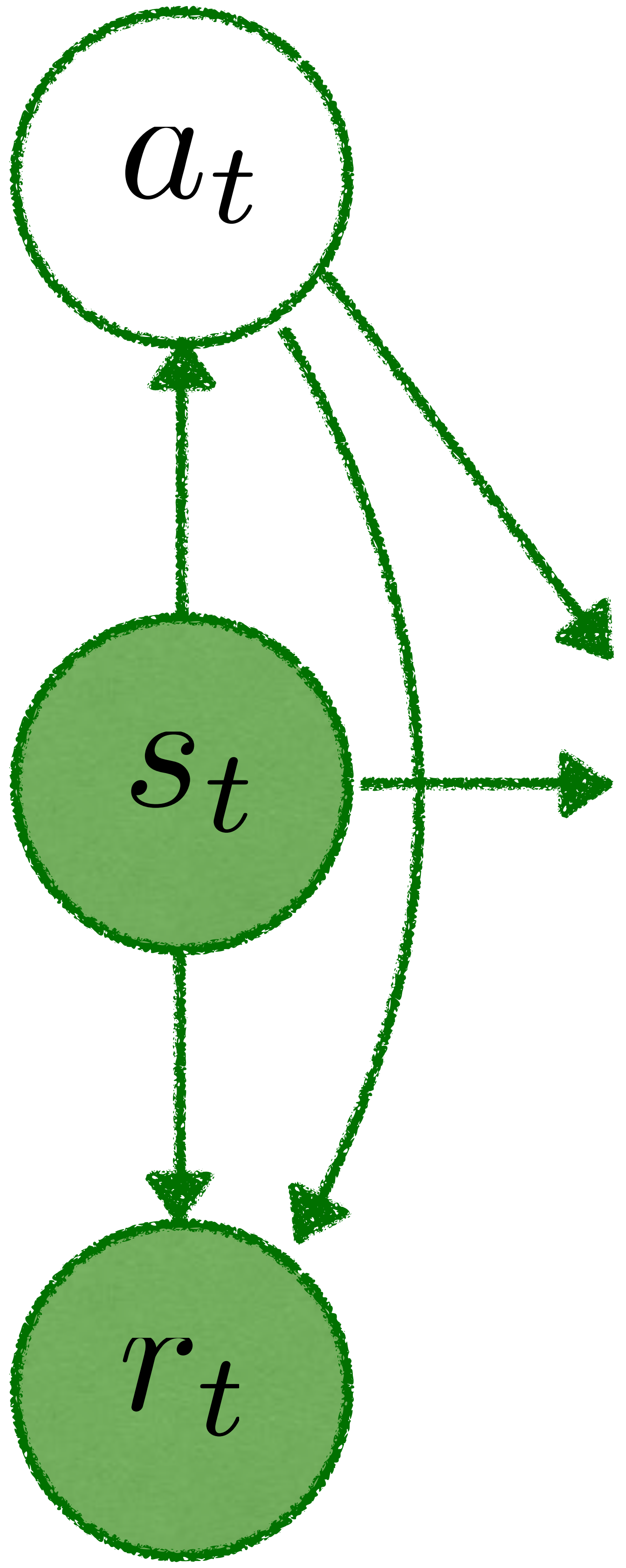
Recovering policy gradients algorithm

- Consider $\pi_0(a_t|s_t) = \mathrm{Uniform}(a_t)$
- $\mathrm{KL}(q_\pi(\mathbf{s}, \mathbf{a}), p_{\pi_0}(\mathbf{s}, \mathbf{a})) = \mathbb{E}_{q_\pi}\left[\log \frac{p(s_1)\prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t) \pi(a_T|s_T)}{p(s_1)\prod_{t=1}^{T} \pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t) \pi_0(a_T|s_T)}\right]$

$$= \mathbb{E}_{q_\pi}\left[\sum_{t=1}^{T} \log \pi(a_t|s_t) - \log \pi_0(a_t|s_t)\right]$$

$$= \mathbb{E}_{q_\pi}\left[-\mathcal{H}(\pi(\cdot|s_t))\right] + \text{const}$$

Policy gradients as inference



Variational lower bound

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

Recovering policy gradients algorithm

- Consider $\pi_0(a_t|s_t) = \mathrm{Uniform}(a_t)$
- $\mathrm{KL}(q_\pi(\mathbf{s}, \mathbf{a}), p_{\pi_0}(\mathbf{s}, \mathbf{a})) = \mathbb{E}_{q_\pi}\left[\log \frac{p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t) \pi(a_T|s_T)}{p(s_1) \prod_{t=1}^{T} \pi_0(a_t|s_t) p(s_{t+1}|s_t, a_t) \pi_0(a_T|s_T)}\right]$

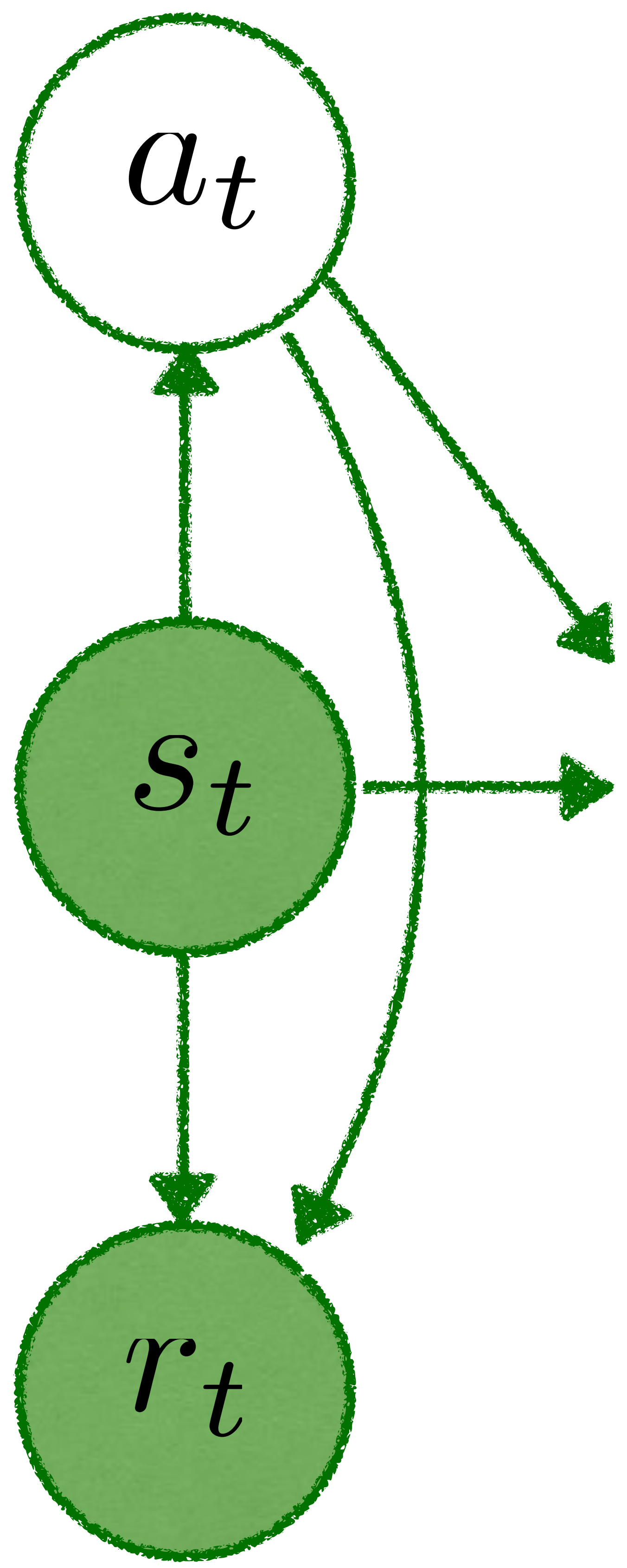
$$= \mathbb{E}_{q_\pi}\left[\sum_{t=1}^{T} \log \pi(a_t|s_t) - \log \pi_0(a_t|s_t)\right]$$

$$= \mathbb{E}_{q_\pi}\left[-\mathcal{H}(\pi(\cdot|s_t))\right] + \mathrm{const}$$

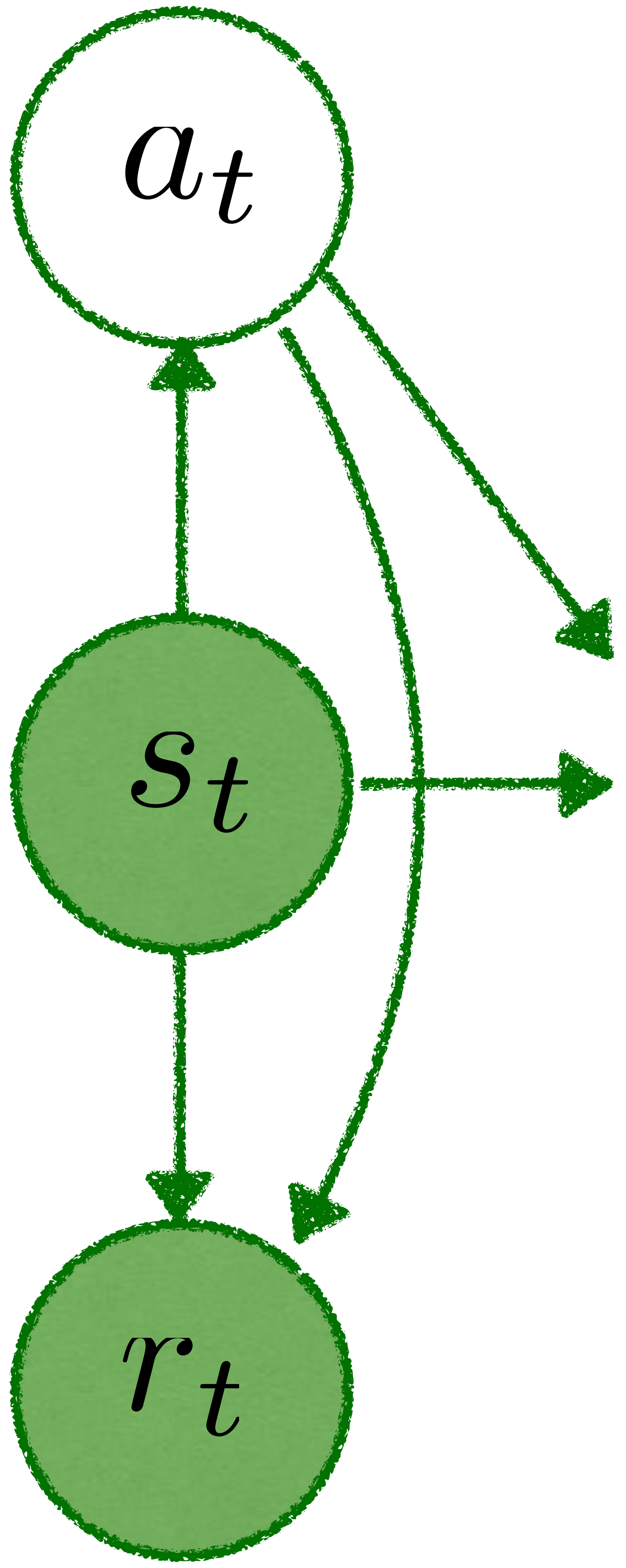
$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T} \alpha r_t + \mathcal{H}(\pi(\cdot|s_t))\right] + \mathrm{const}$$

Policy gradients as stochastic EM

$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T}\Big(\alpha \sum_{k=0}^{T-t} r_{t+k}\Big)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t))\right]$$



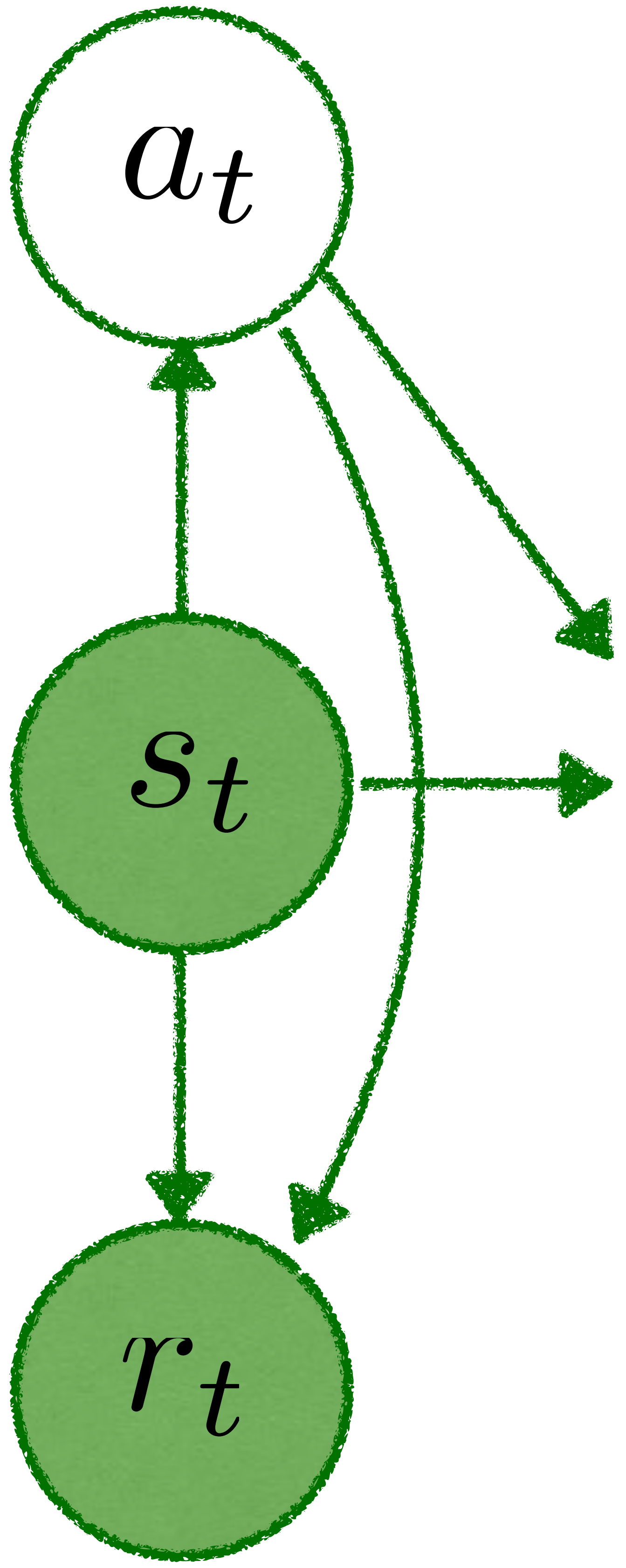
Policy gradients as stochastic EM



$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T} \left(\alpha \sum_{k=0}^{T-t} r_{t+k}\right) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t))\right]$$

Stochastic policy gradient

Policy gradients as stochastic EM



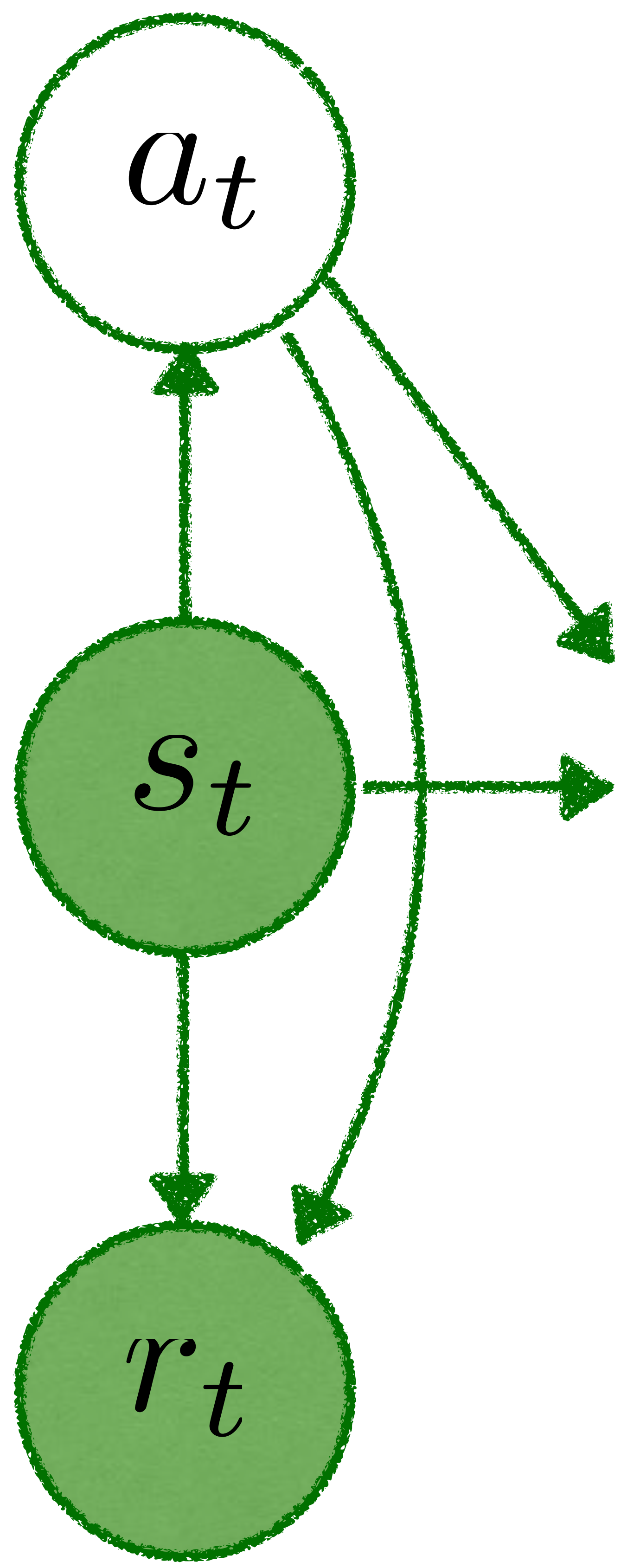
$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\sum_{t=1}^{T} \Big(\alpha \sum_{k=0}^{T-t} r_{t+k}\Big) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t)) \right]$$

Stochastic policy gradient

- 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 stochastic EM

$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T}\left(\alpha \sum_{k=0}^{T-t} r_{t+k}\right)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t))\right]$$

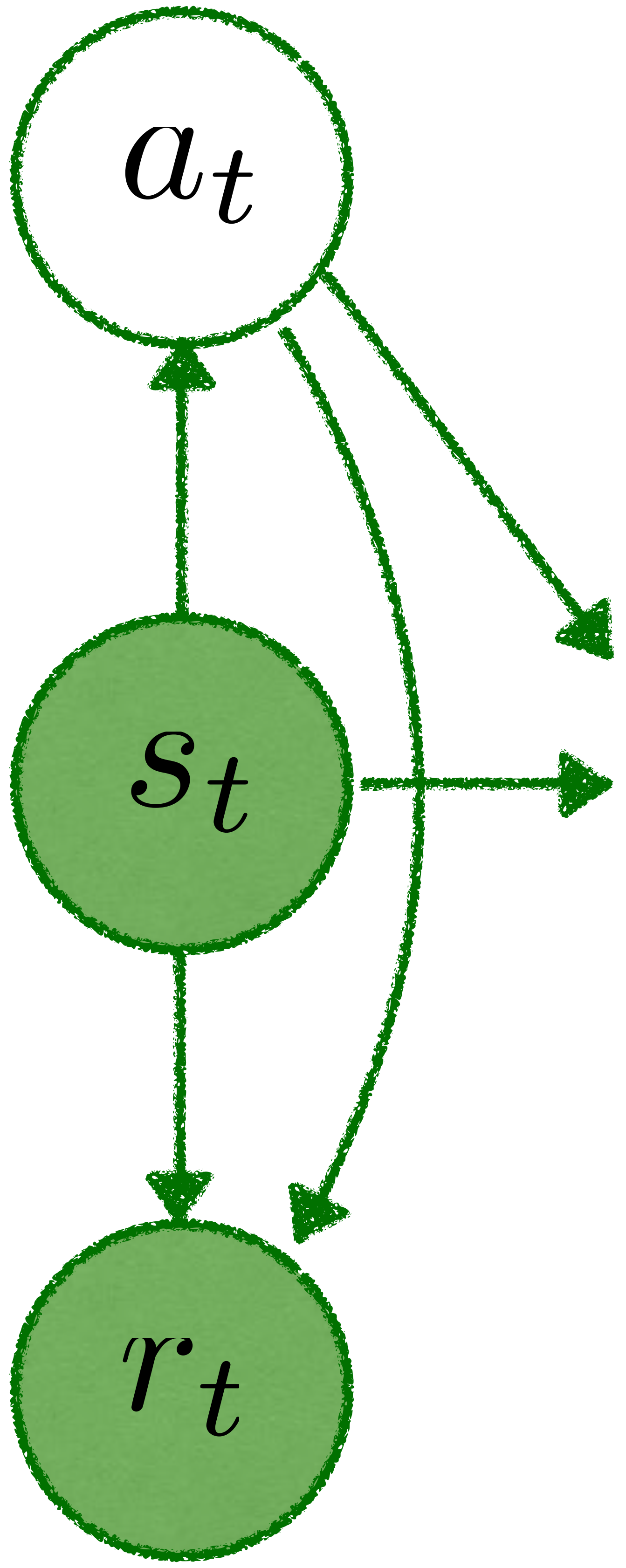


Stochastic policy gradient

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

$$\tilde{\nabla}_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \sum_{t=1}^{T}\left(\alpha \sum_{k=0}^{T-t} \hat{r}_{t+k}\right)\nabla_\pi \log \pi(\hat{a}_t|\hat{s}_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|\hat{s}_t))$$

Policy gradients as stochastic EM



$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\sum_{t=1}^{T} (\alpha \sum_{k=0}^{T-t} r_{t+k}) \nabla_\pi \log \pi(a_t | s_t) + \nabla_\pi \mathcal{H}(q(\cdot | s_t)) \right]$$

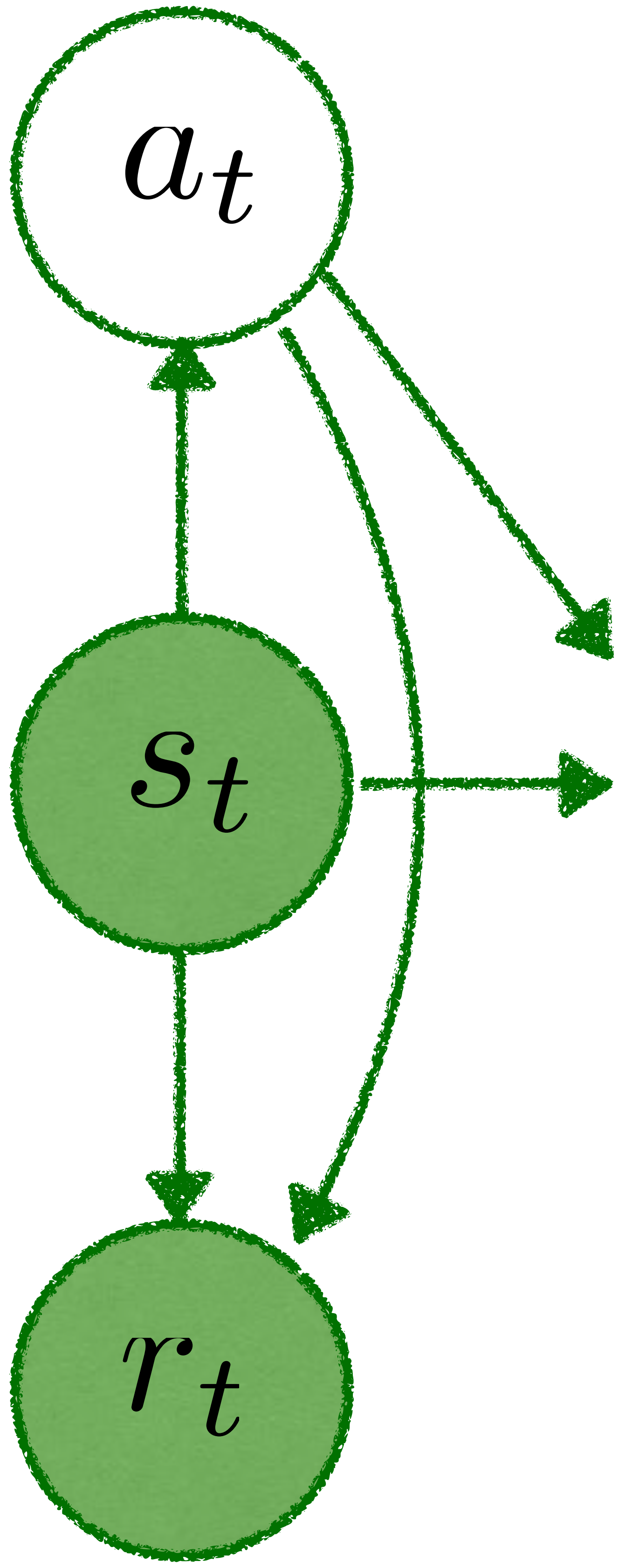
Stochastic policy gradient

- 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:

$$\tilde{\nabla}_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \sum_{t=1}^{T} (\alpha \sum_{k=0}^{T-t} \hat{r}_{t+k}) \nabla_\pi \log \pi(\hat{a}_t | \hat{s}_t) + \nabla_\pi \mathcal{H}(\pi(\cdot | \hat{s}_t))$$

- Model-free learning algorithm

Policy gradients as stochastic EM



$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\sum_{t=1}^{T}\left(\alpha \sum_{k=0}^{T-t} r_{t+k}\right)\nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t))\right]$$

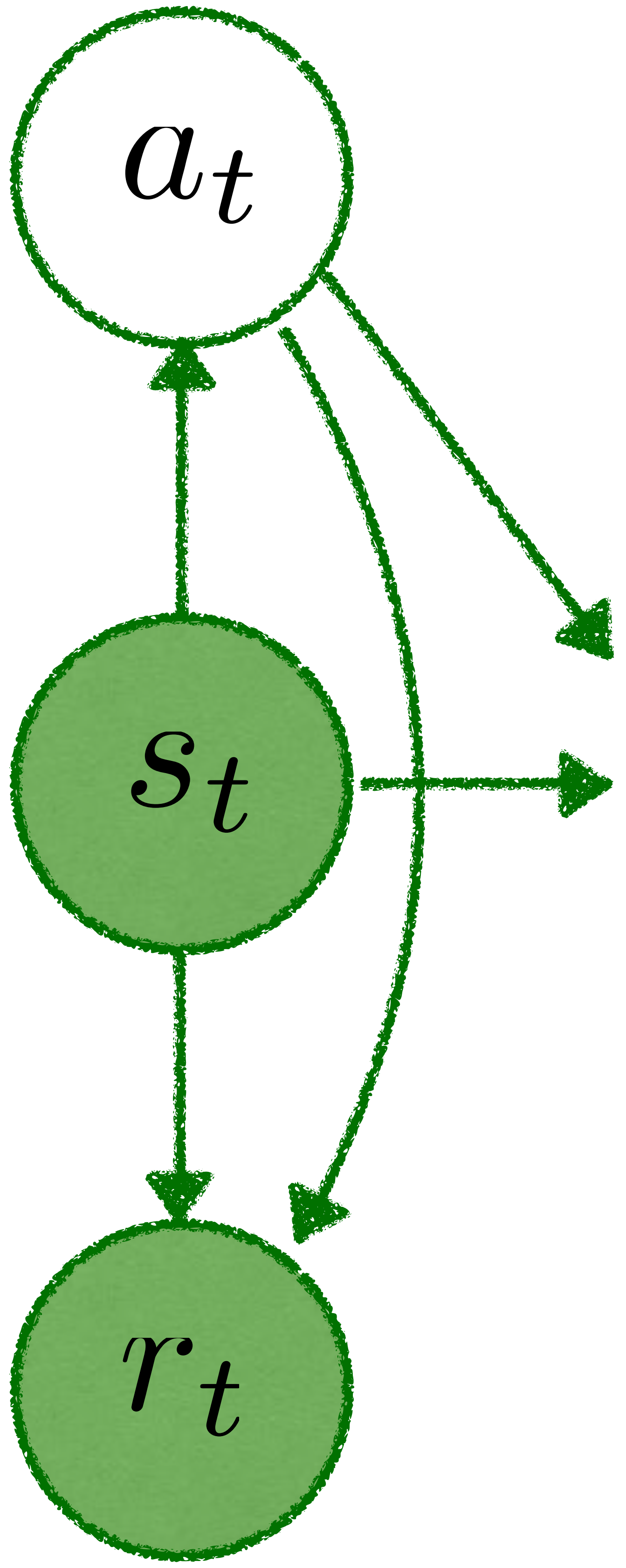
Stochastic policy gradient

- 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:

$$\tilde{\nabla}_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \sum_{t=1}^{T}\left(\alpha \sum_{k=0}^{T-t} \hat{r}_{t+k}\right)\nabla_\pi \log \pi(\hat{a}_t|\hat{s}_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|\hat{s}_t))$$

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

Policy gradients as stochastic EM



$$\nabla_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})} \left[\sum_{t=1}^{T} \left(\alpha \sum_{k=0}^{T-t} r_{t+k}\right) \nabla_\pi \log \pi(a_t|s_t) + \nabla_\pi \mathcal{H}(q(\cdot|s_t)) \right]$$

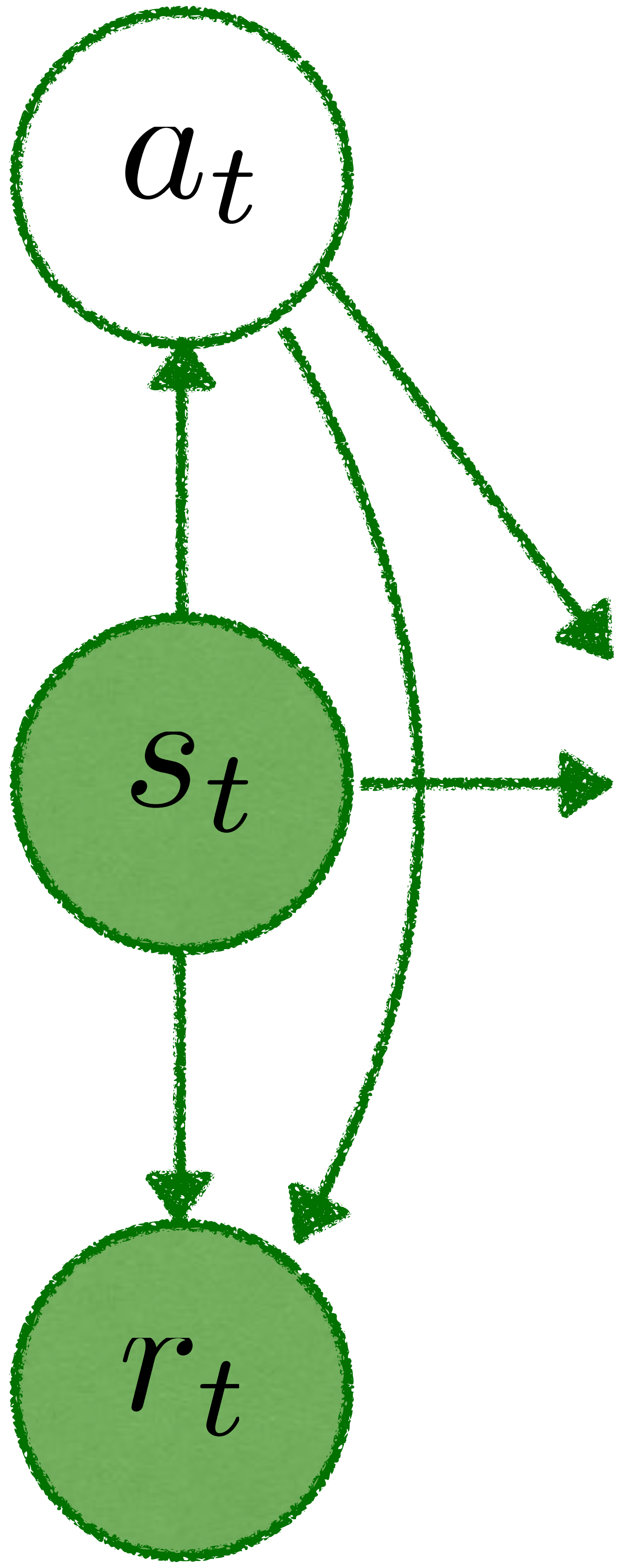
Stochastic policy gradient

- 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:

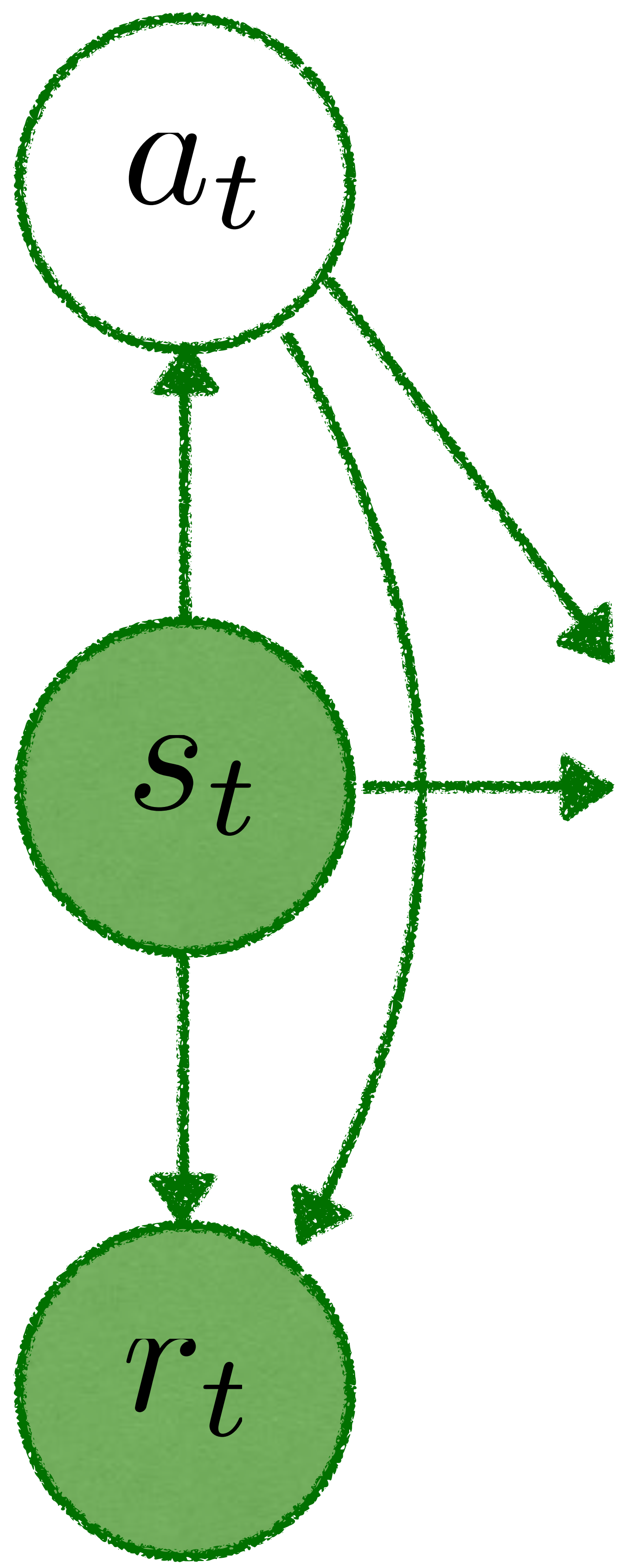
$$\tilde{\nabla}_\pi \mathcal{L}(q_\pi, p_{\pi_0}) = \sum_{t=1}^{T} \left(\alpha \sum_{k=0}^{T-t} \hat{r}_{t+k}\right) \nabla_\pi \log \pi(\hat{a}_t|\hat{s}_t) + \nabla_\pi \mathcal{H}(\pi(\cdot|\hat{s}_t))$$

- Model-free learning algorithm
- Known as REINFORCE [2] Williams, 1992
- Basis for the modern RL algorithms [3] Mnih et al, 2016

Improving policy optimization

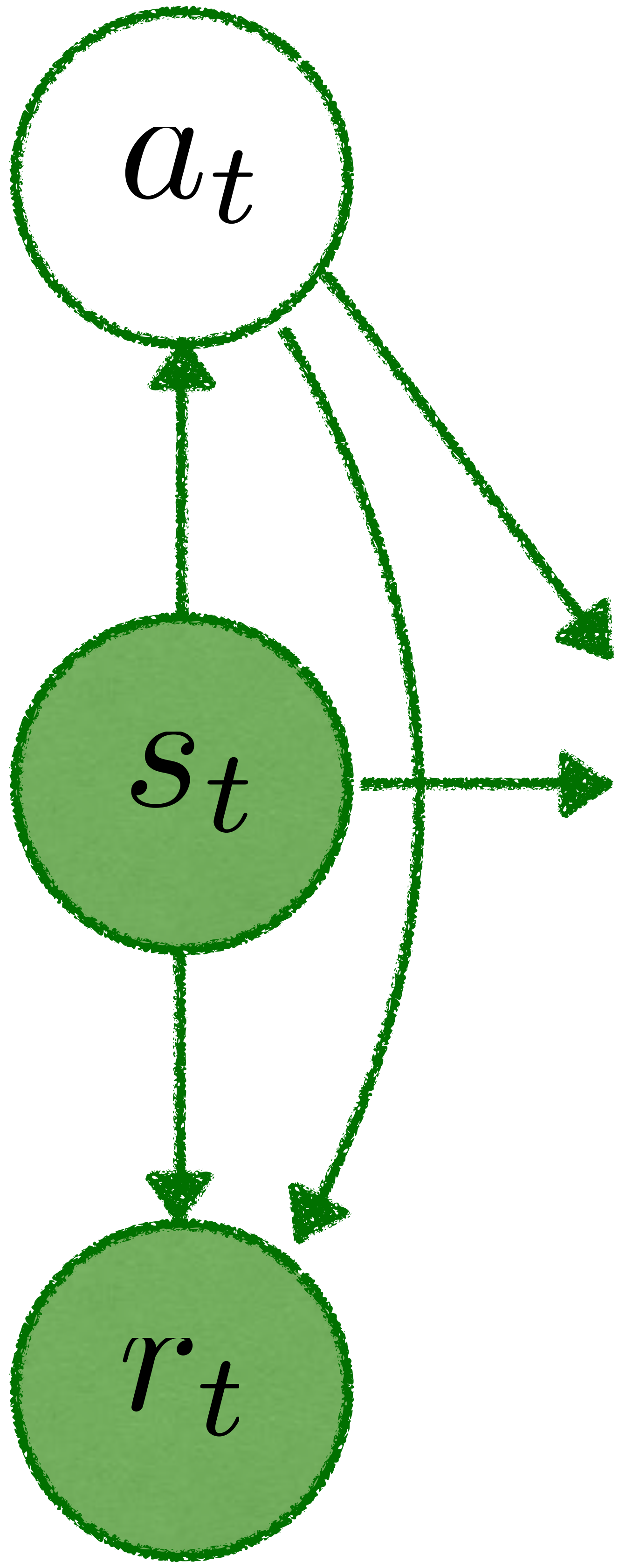


Improving policy optimization

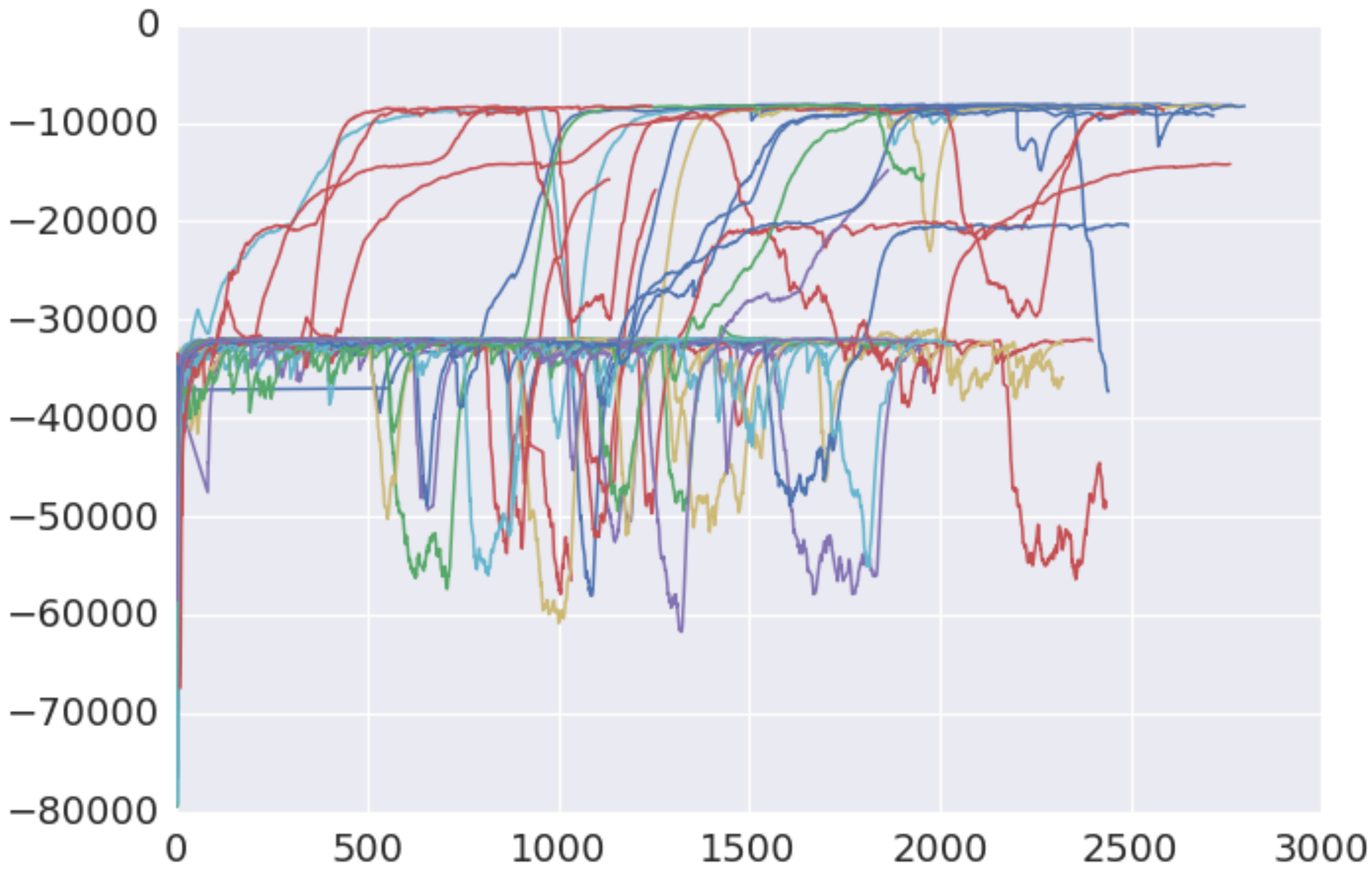


- Reward optimization can be a **very** hard optimization problem

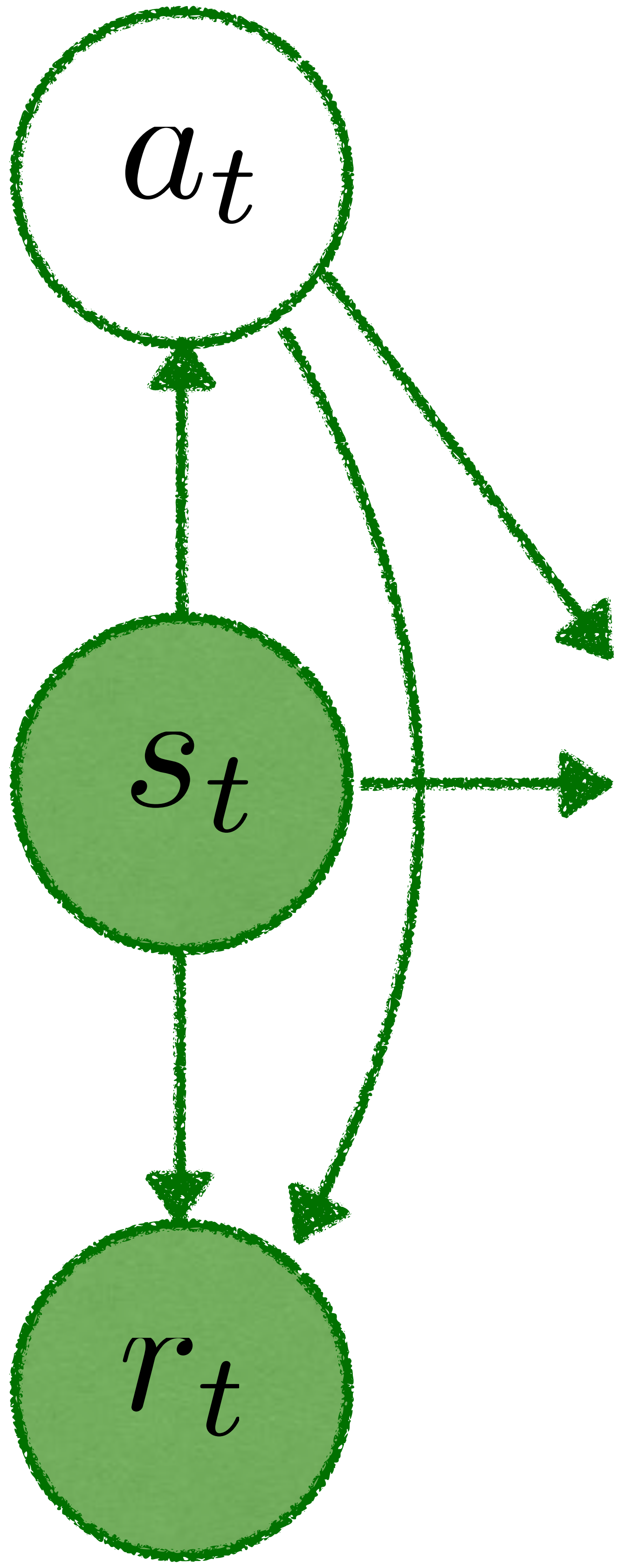
Improving policy optimization



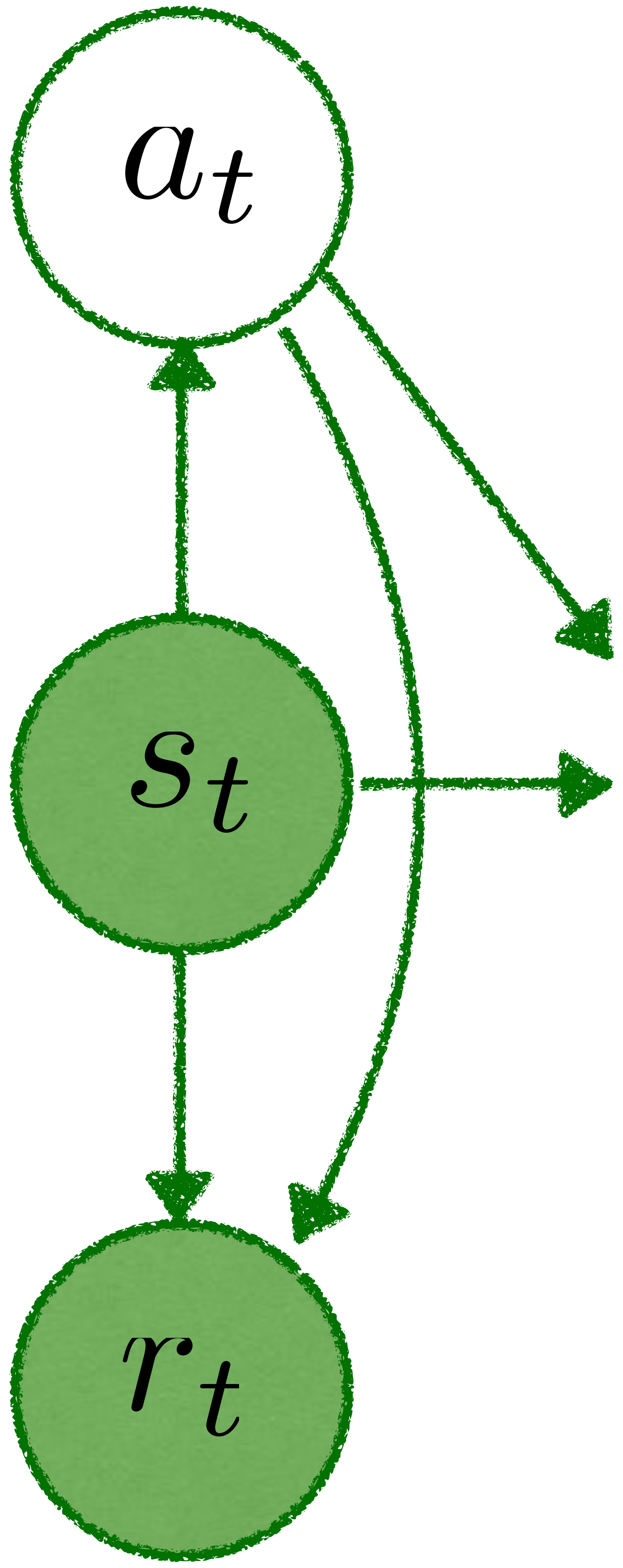
- Reward optimization can be a **very** hard optimization problem
- Typical rewards vs time plot:



Improving policy optimization

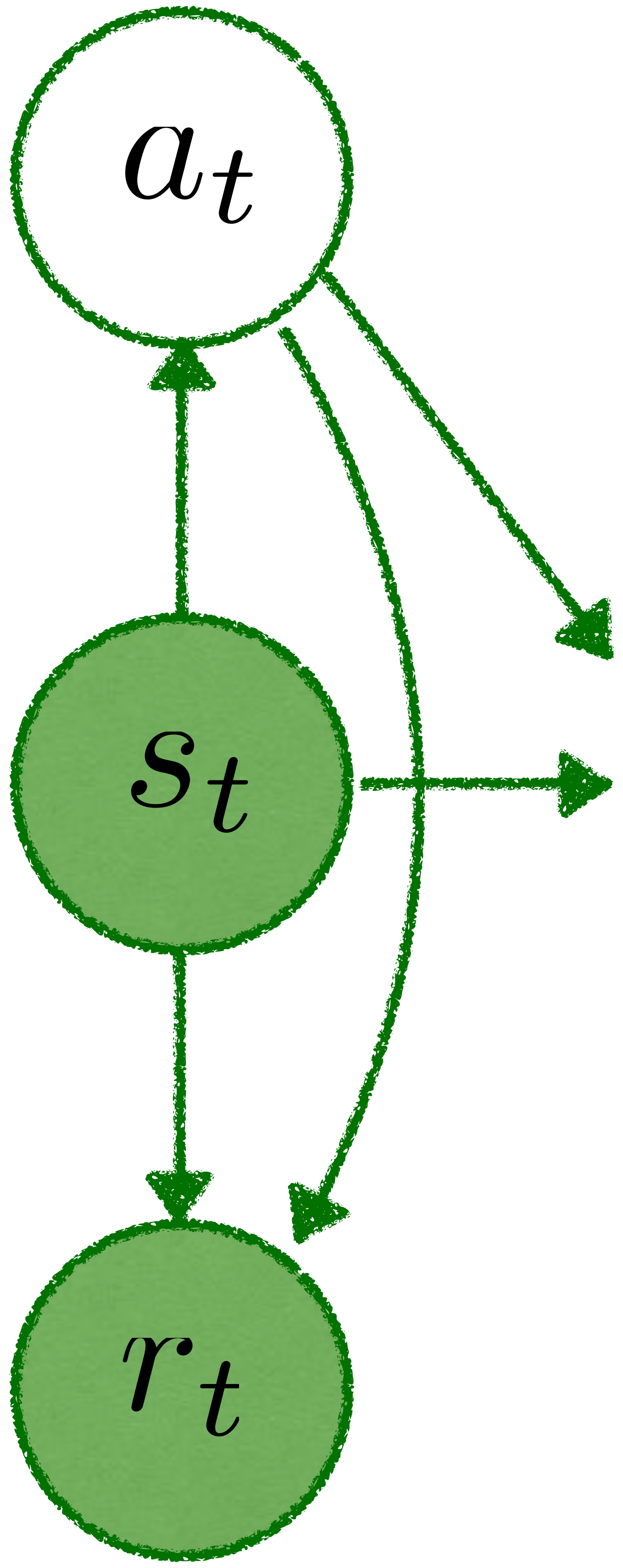


Improving policy optimization



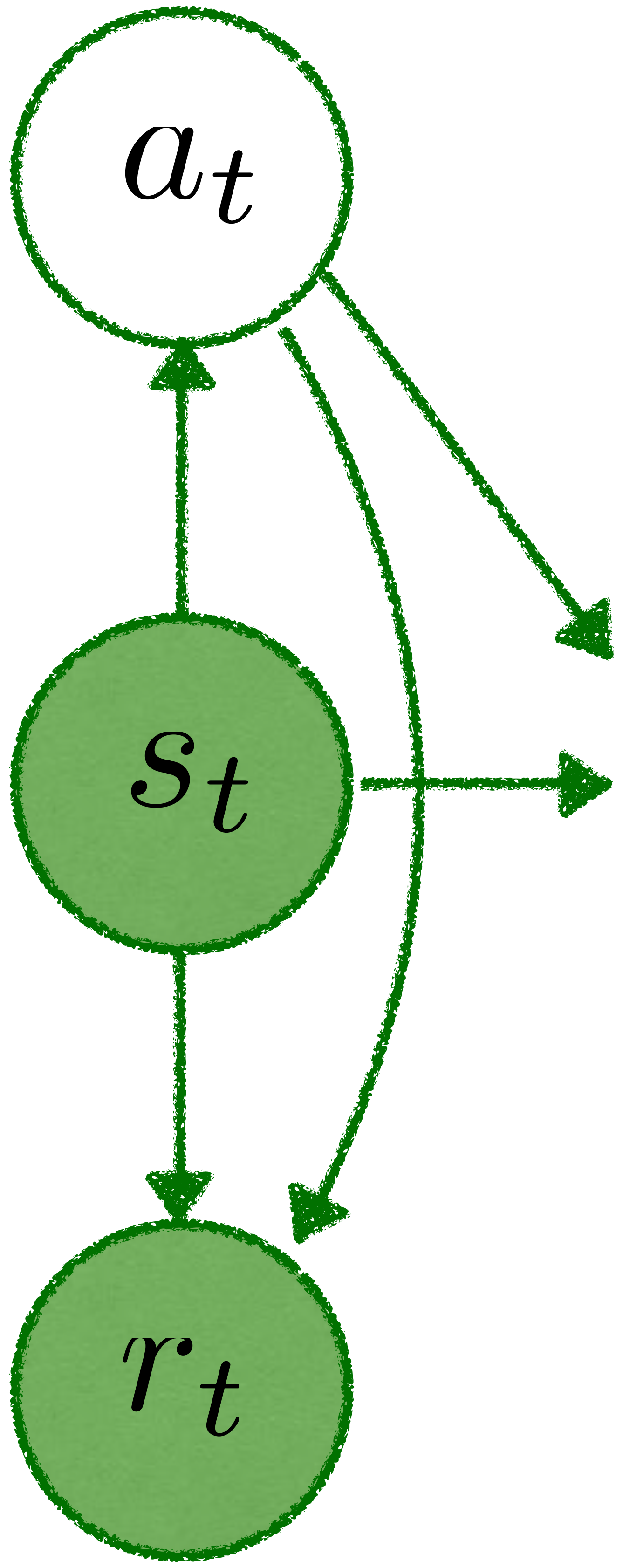
- Policy can change too fast

Improving policy optimization



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

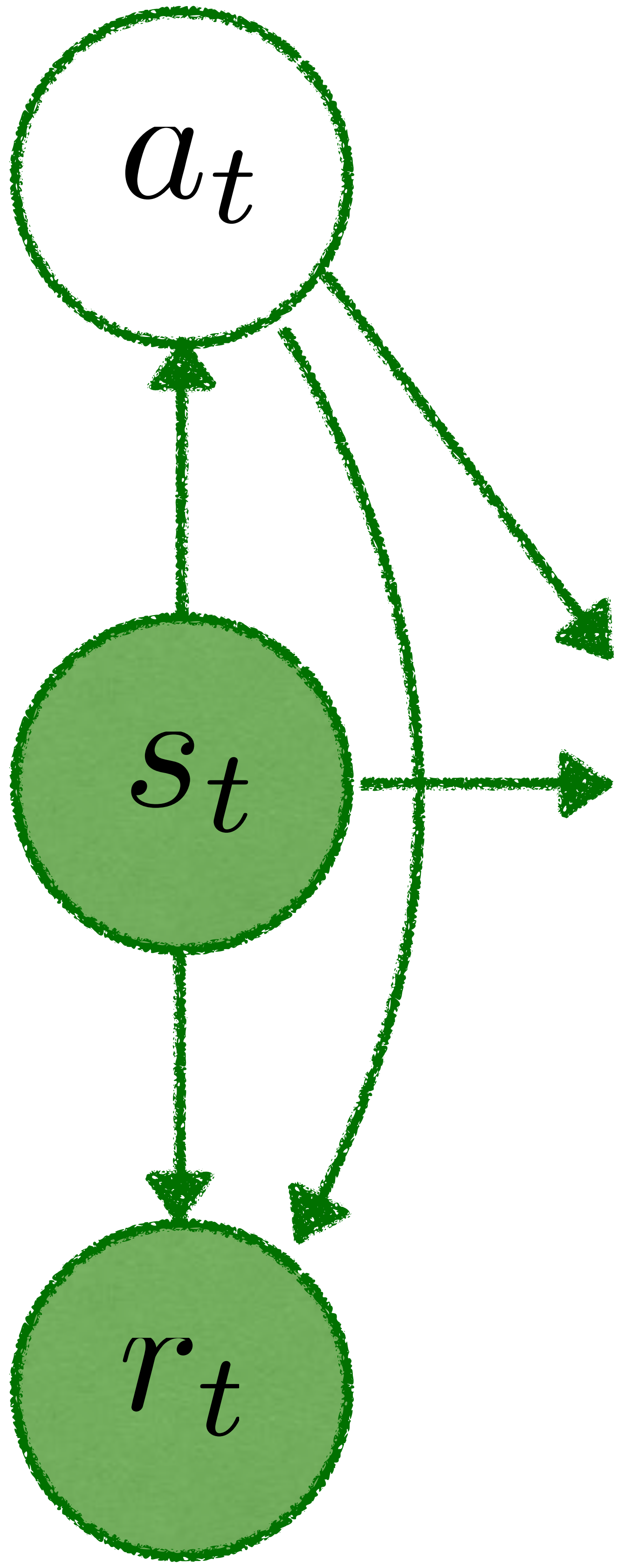
Improving policy optimization



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

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

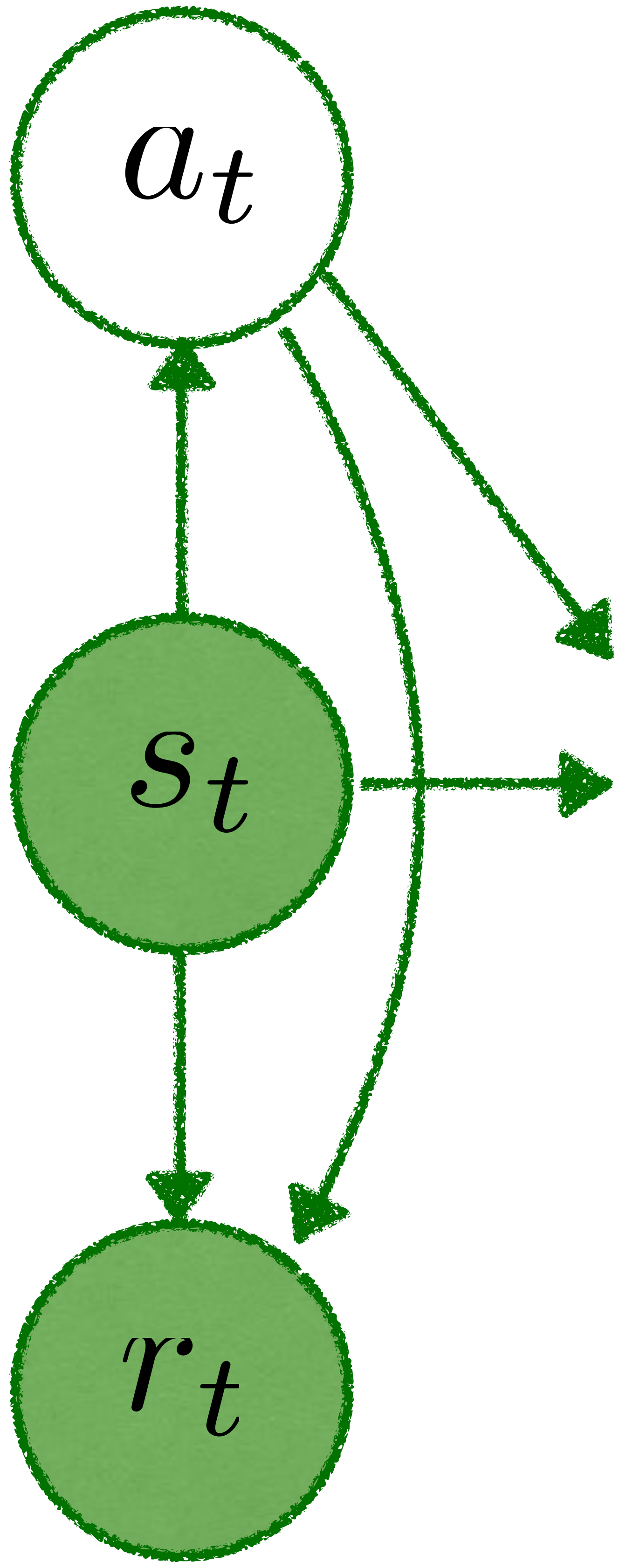
Improving policy optimization



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

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

Improving policy optimization

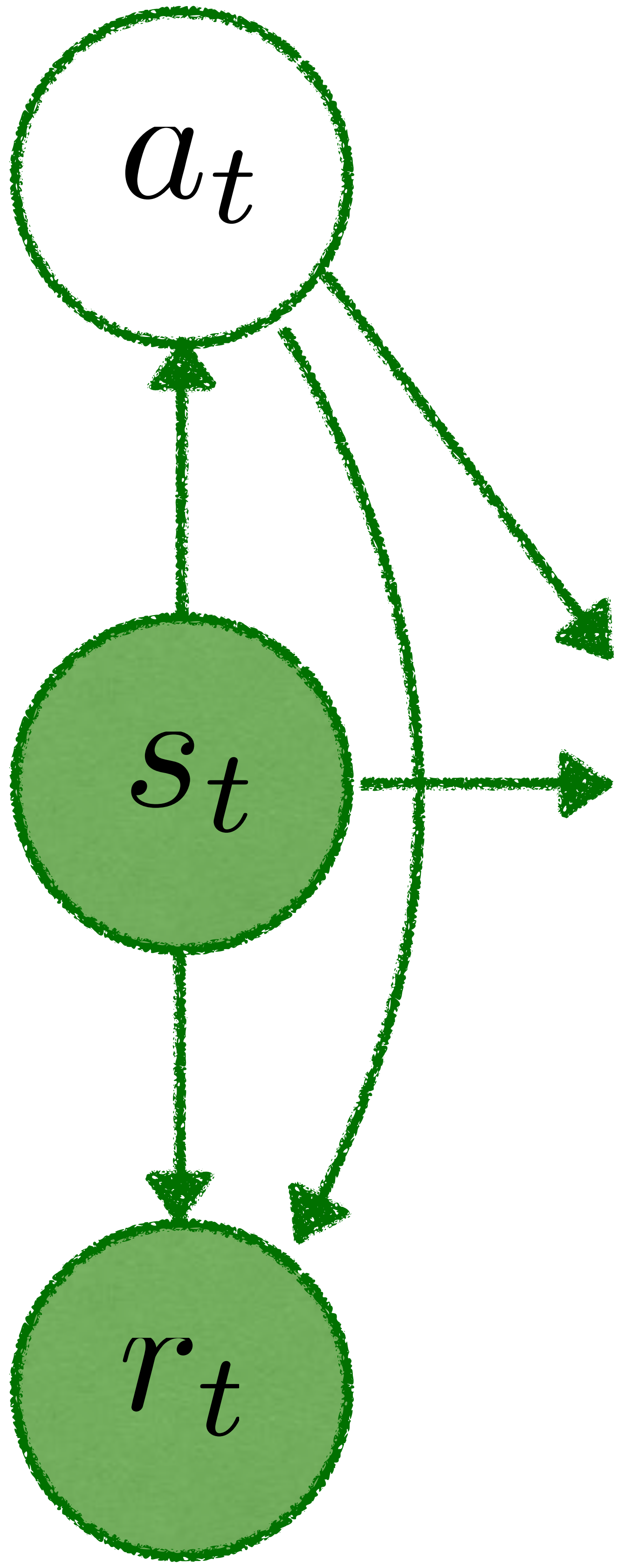


- Policy can change too fast
- We want to make small, reliable improvements
- Our lower bound has already all necessary ingredients

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

- Looking a bit closer: $\mathrm{KL}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T})} \sum_{t=1}^{T} \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))$

Improving policy optimization

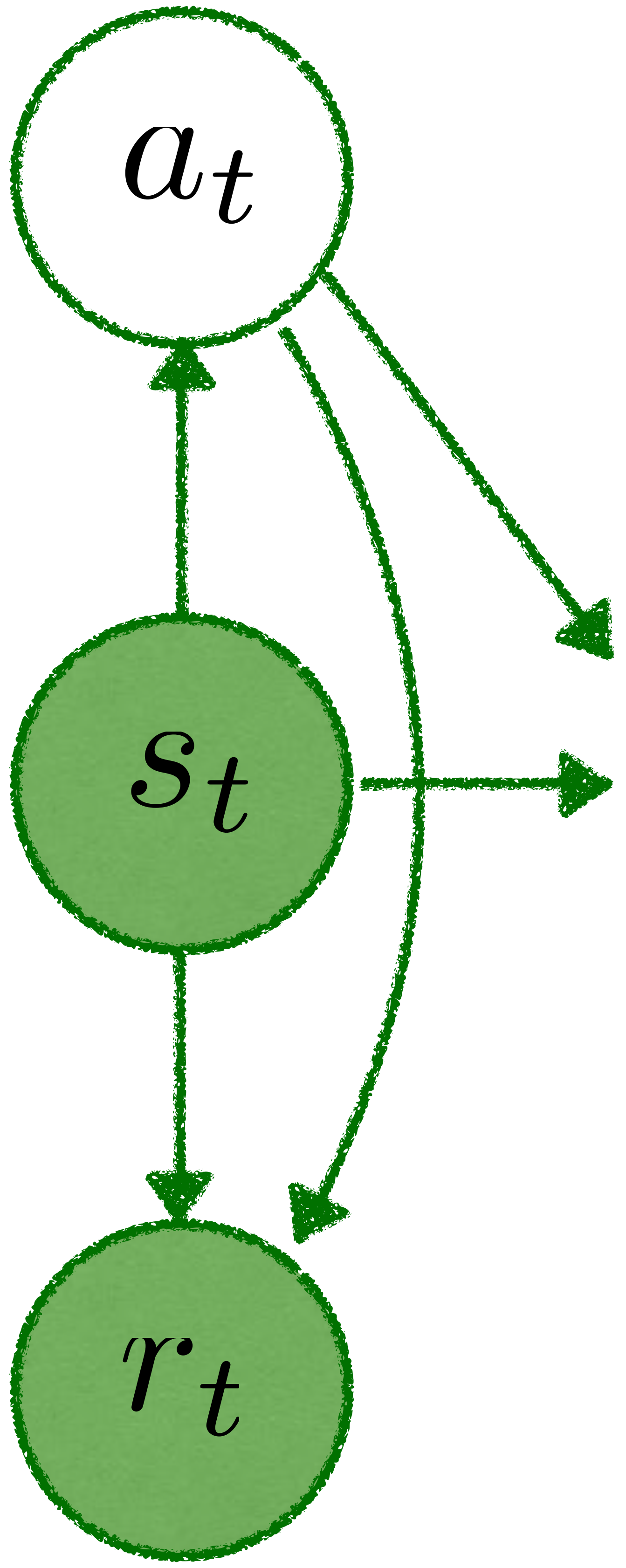


- Policy can change too fast
- We want to make small, reliable improvements
- Our lower bound has already all necessary ingredients

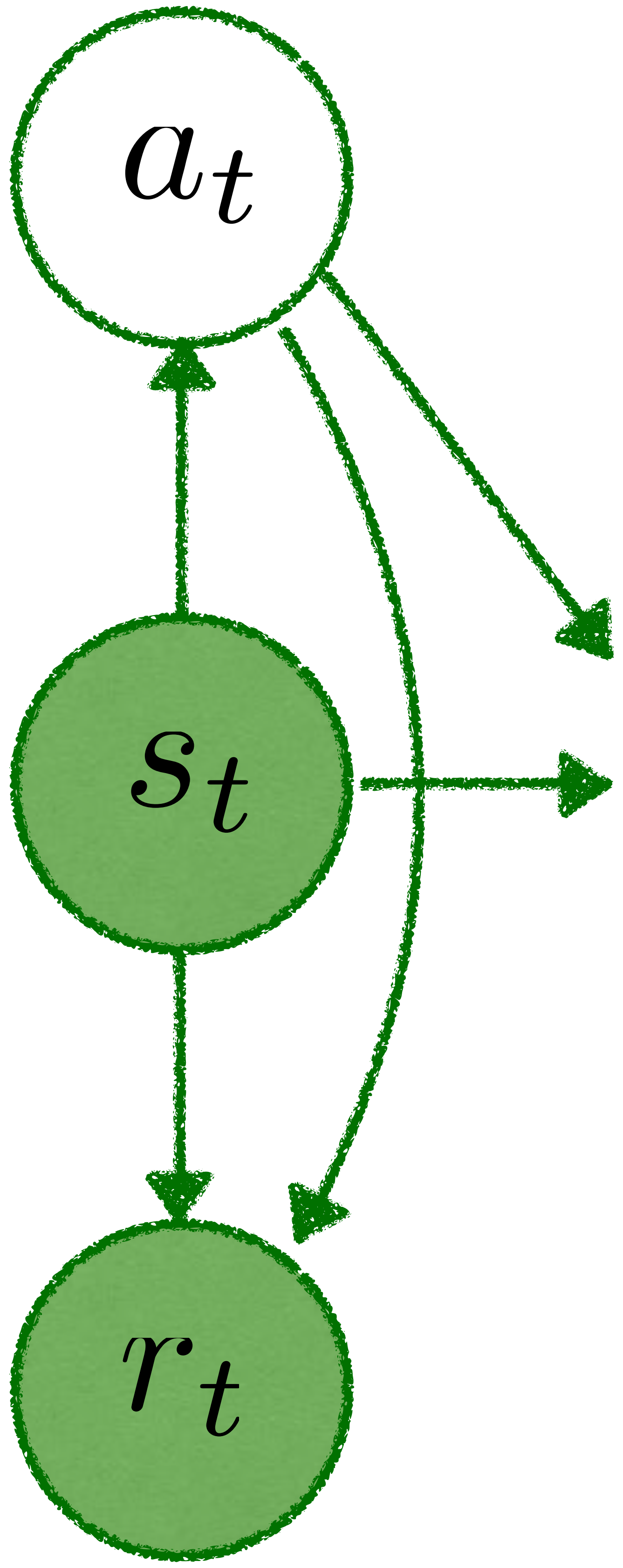
$$\mathcal{L}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})}\left[\alpha \sum_{t=1}^{T} r_t\right] - \boxed{\mathrm{KL}(q_\pi(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) || \underline{p_{\pi_0}}(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}))}$$

- Looking a bit closer: $\mathrm{KL}(q_\pi, p_{\pi_0}) = \mathbb{E}_{q_\pi(\mathbf{s}_{1:T})} \sum_{t=1}^{T} \mathrm{KL}(\pi(\cdot|s_t) || \pi_0(\cdot|s_t))$
- Use this constraint to prevent too rapid changes in the policy

Stable iterative algorithm

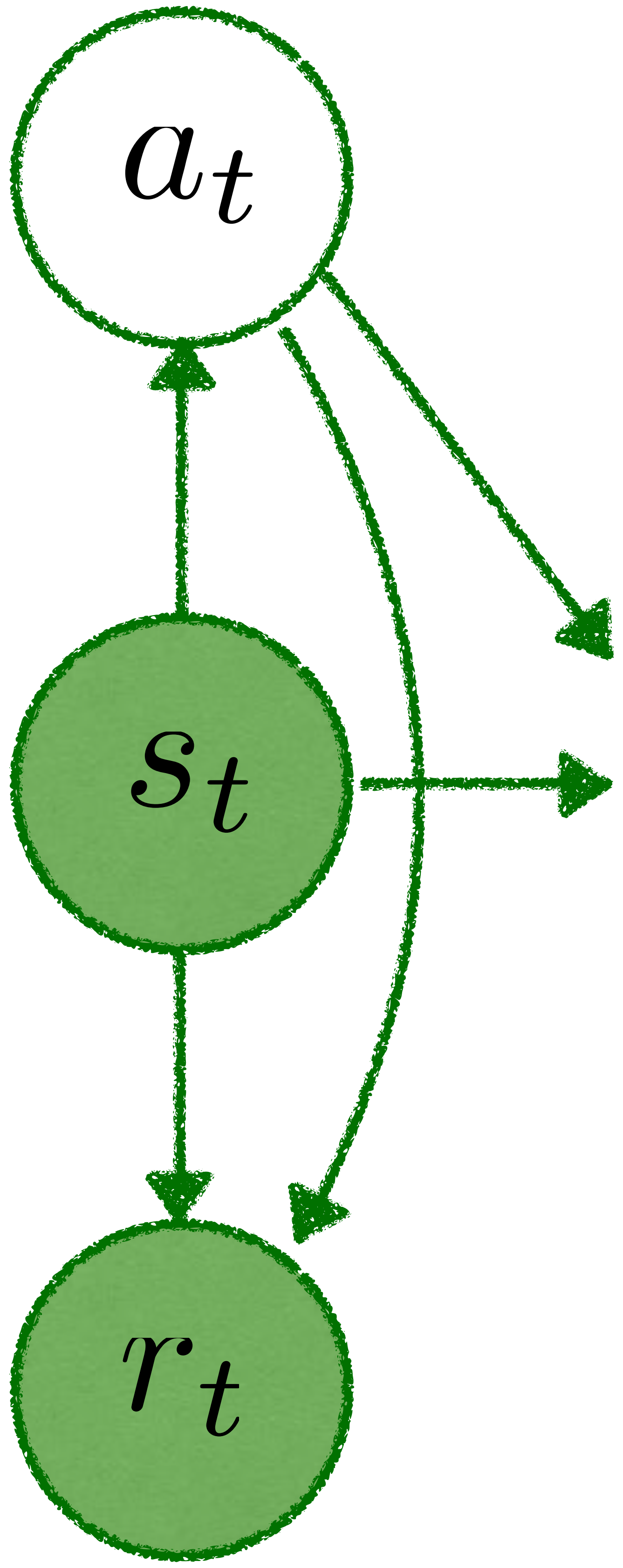


Stable iterative algorithm



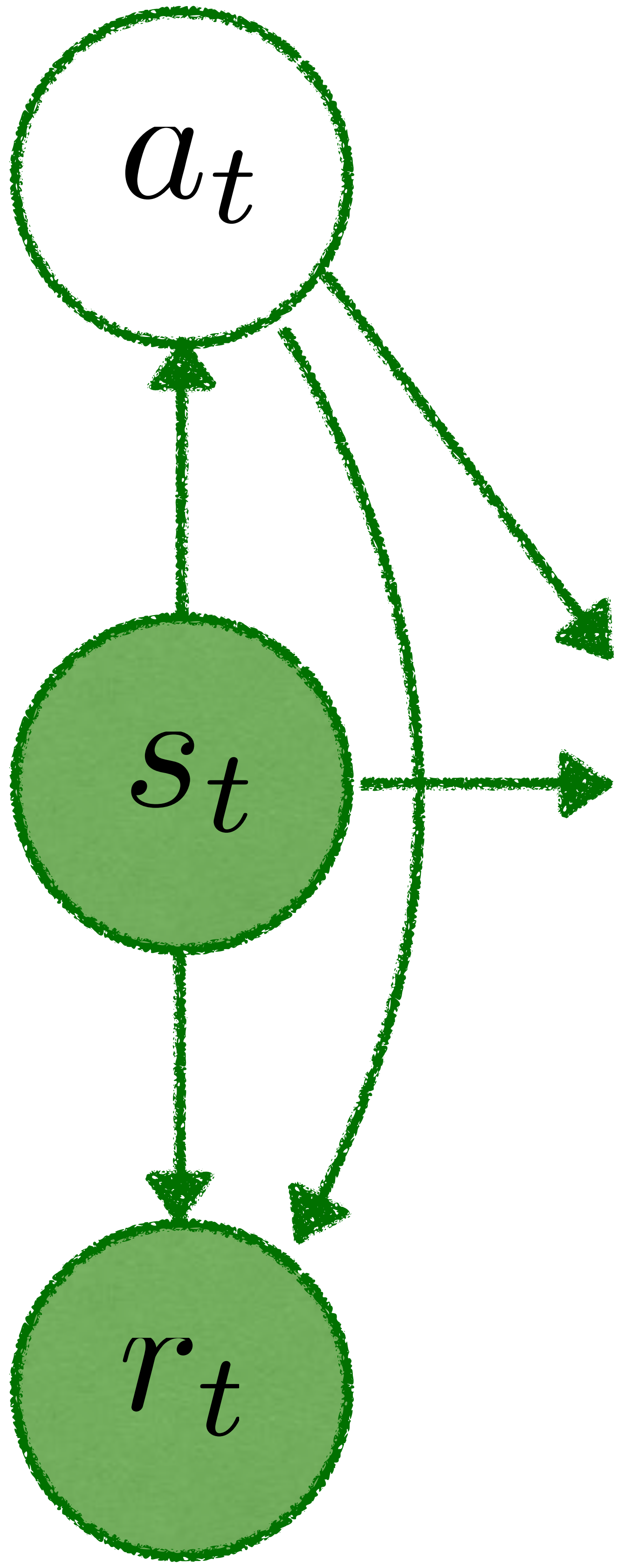
- Initialize π_0

Stable iterative algorithm



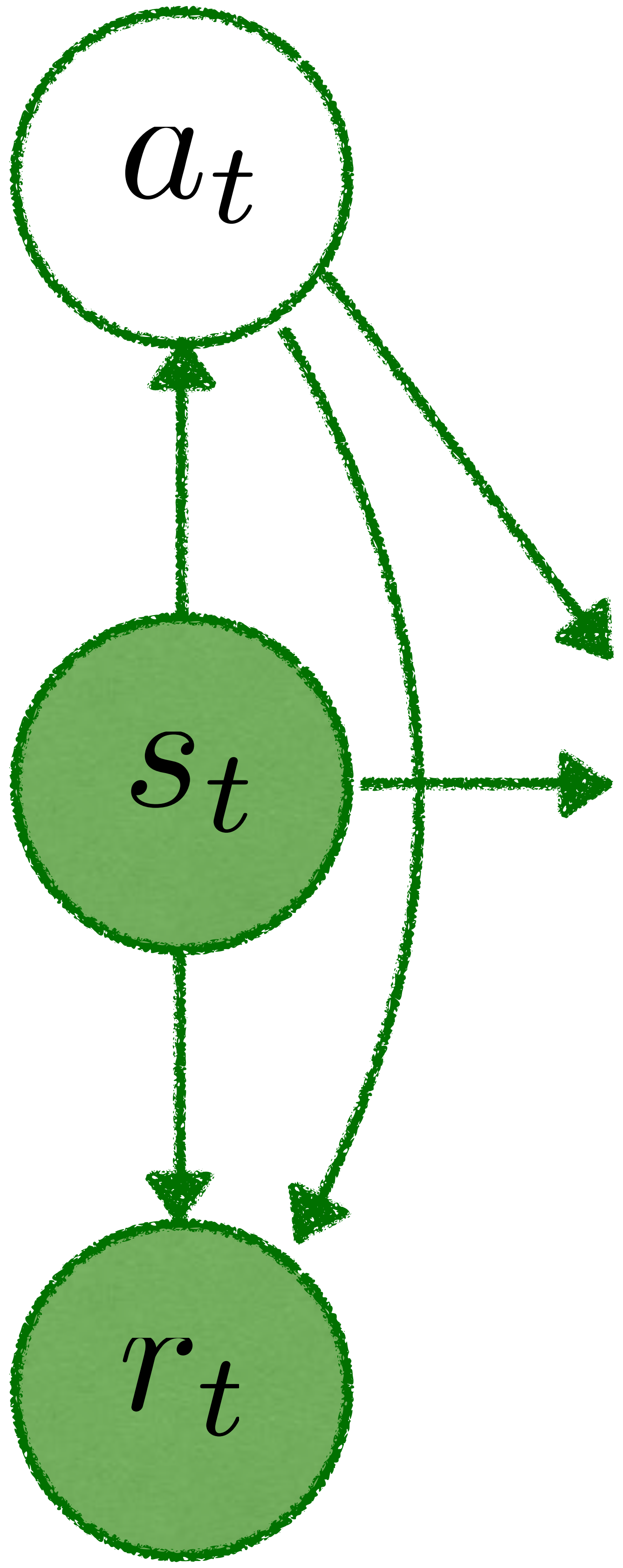
- Initialize π_0
- For each j = 1, 2, …:

Stable iterative algorithm



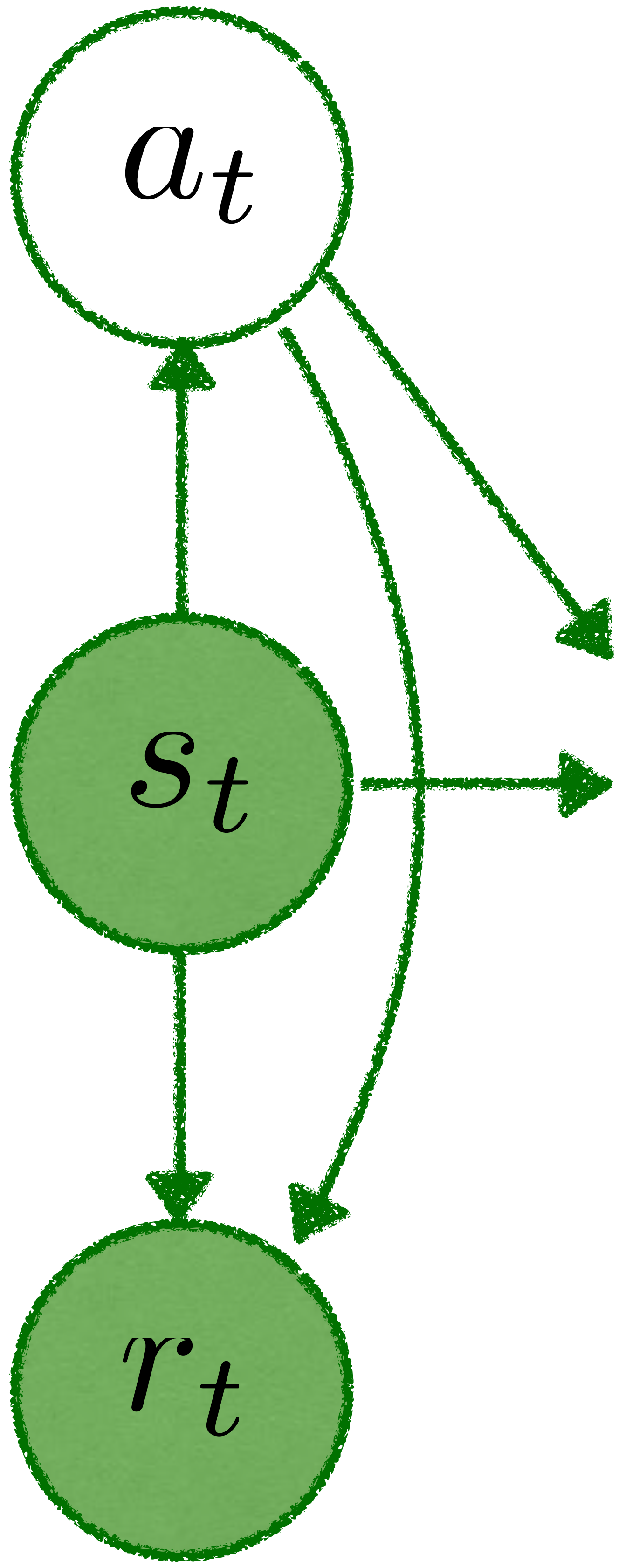
- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$

Stable iterative algorithm



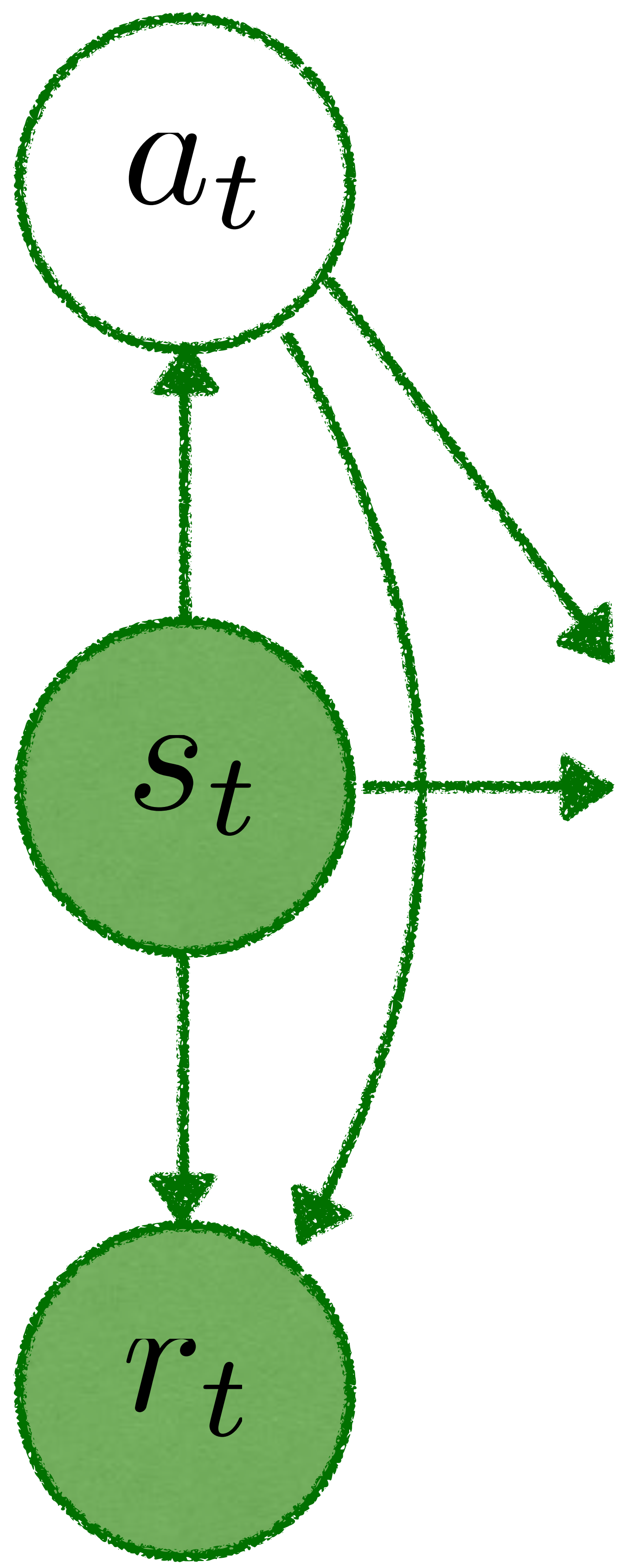
- Initialize π_0
- For each j = 1, 2, …:
  - Obtain $\pi_j \approx \arg\max_\pi \mathcal{L}(q_\pi, p_{\pi_{j-1}})$
  - Repeat

Stable iterative algorithm



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

Stable iterative algorithm



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

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

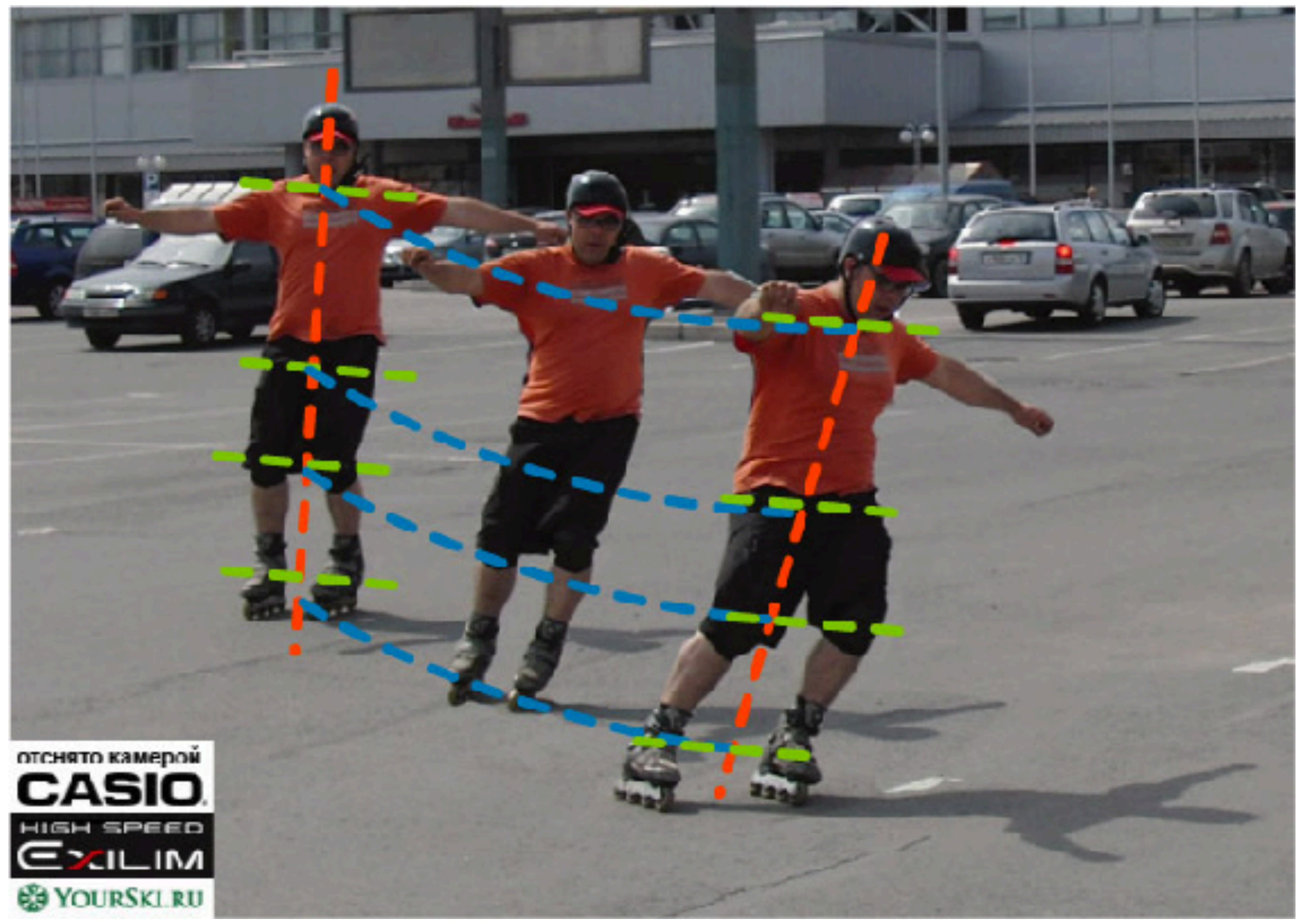
Low-level actions
muscles control

Hierarchical Reinforcement Learning

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

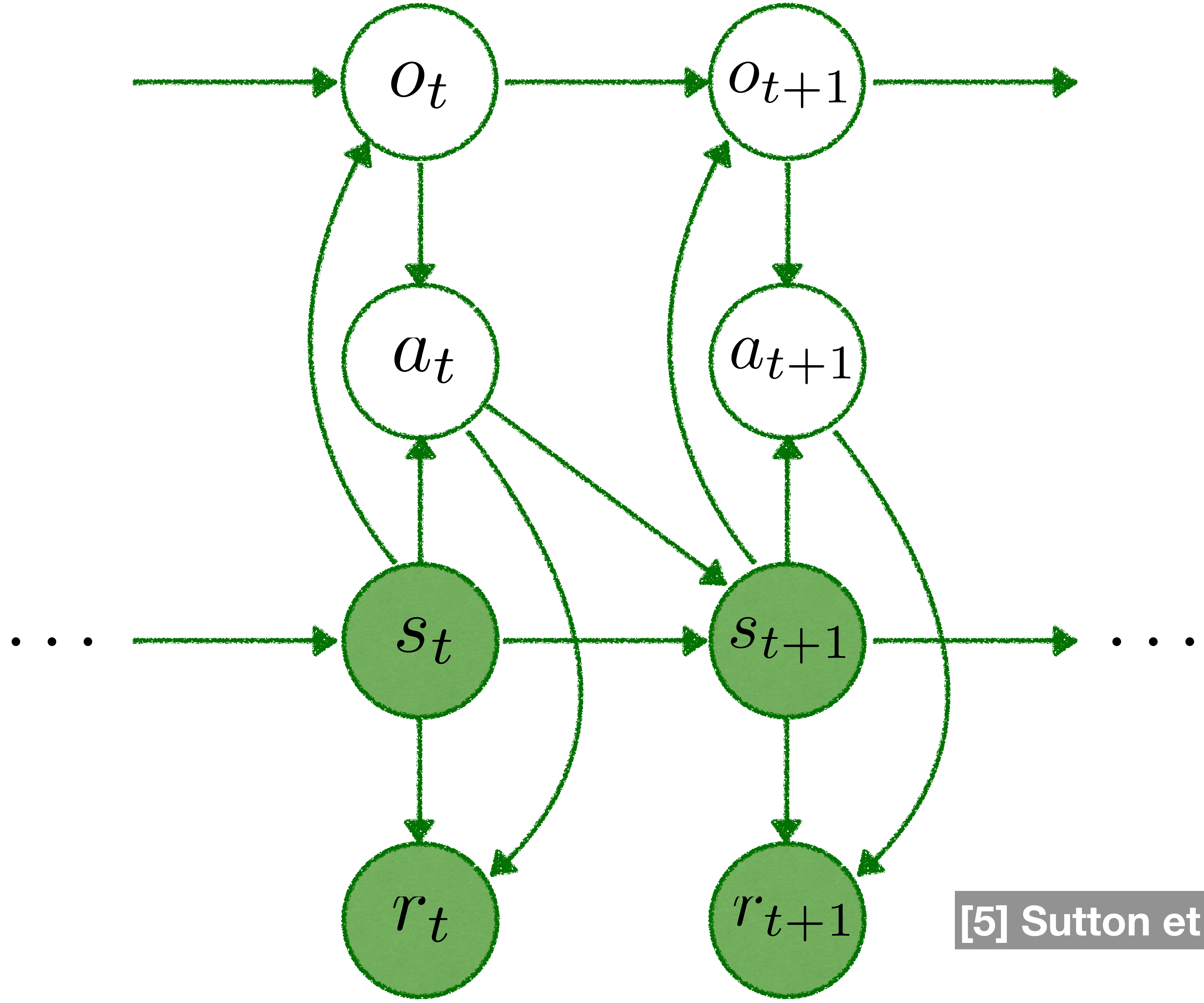
Task: roller skating

Low-level actions
muscles control



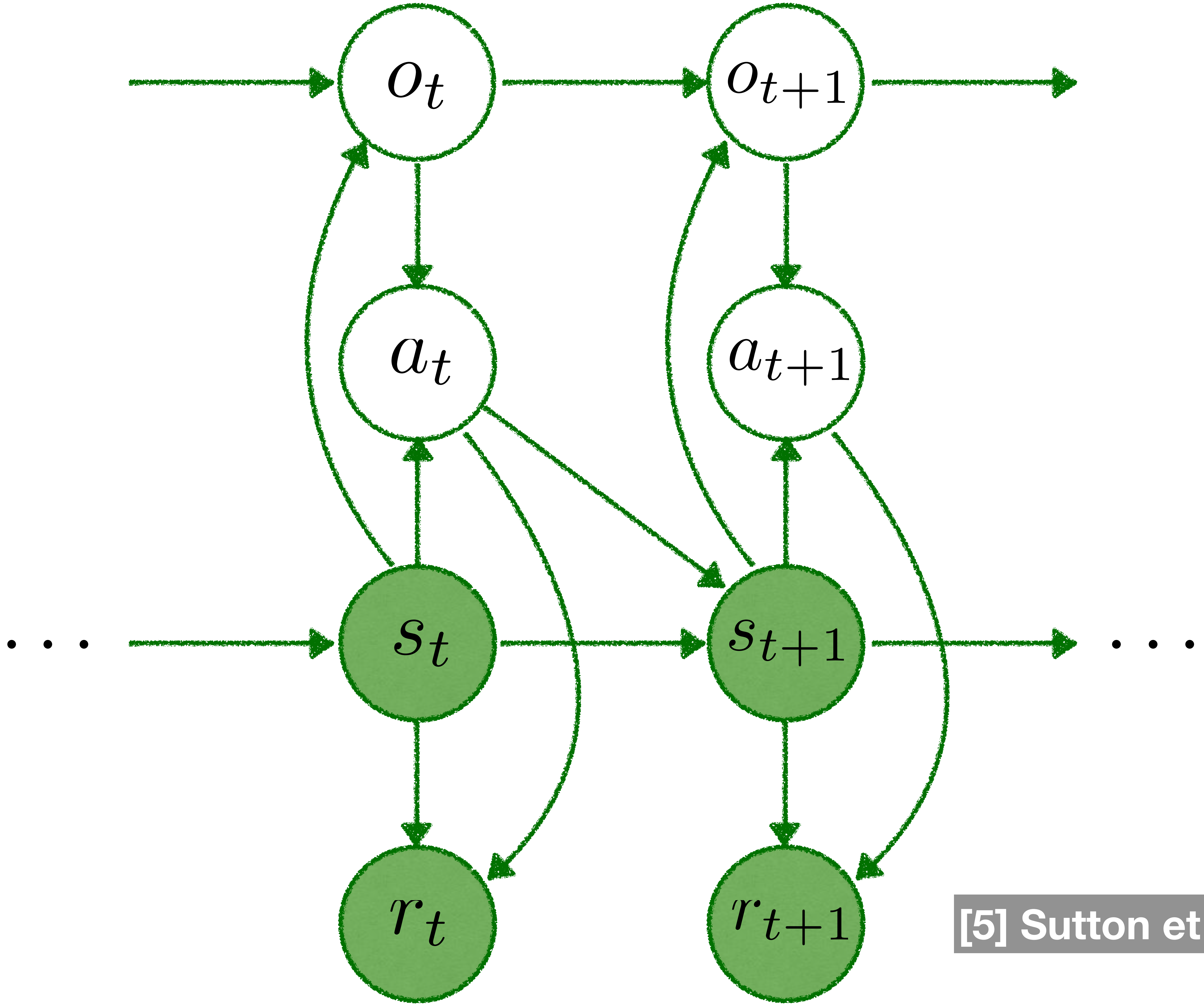
Macro-actions
turn, gain speed, slow down

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

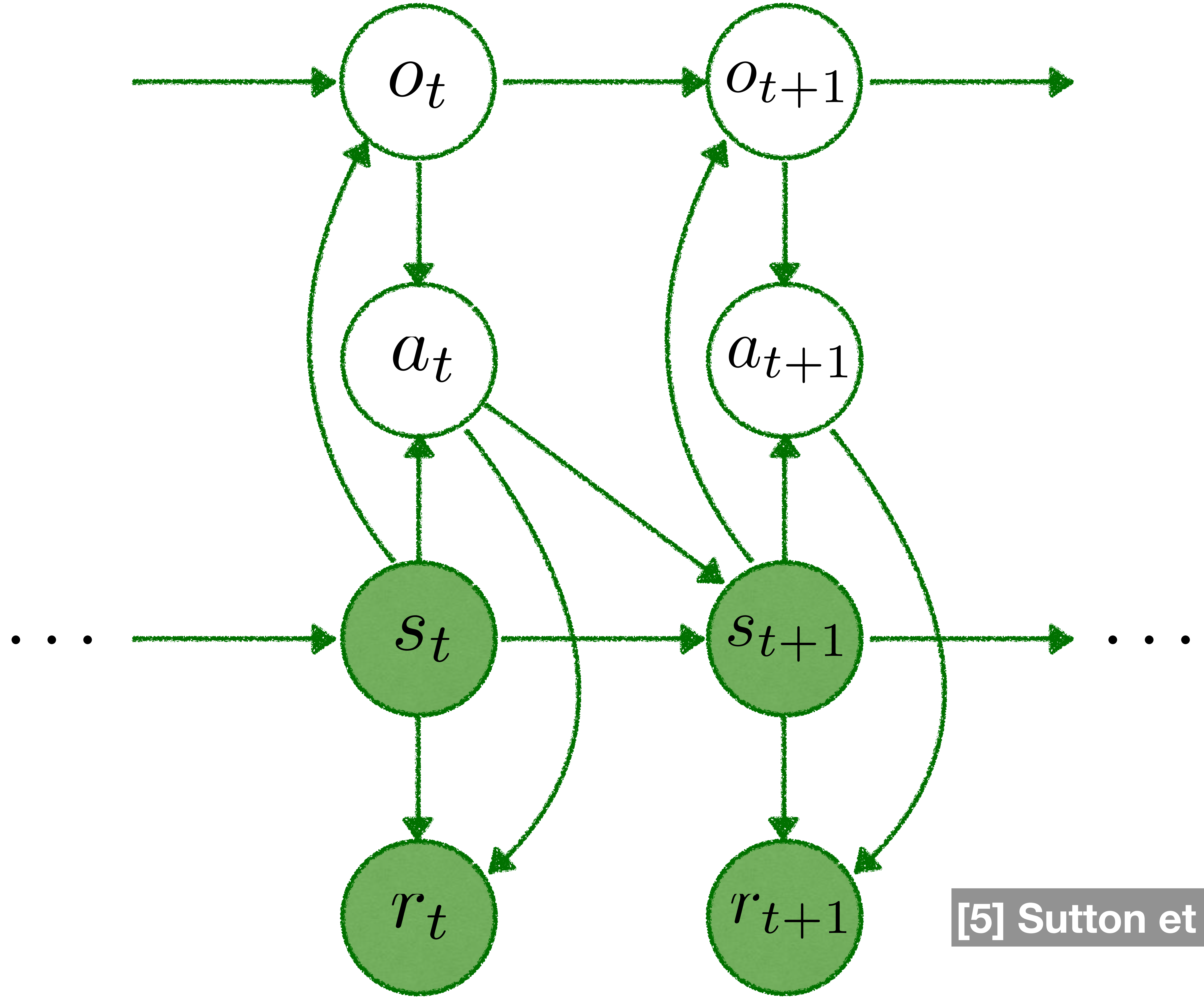
RL with Options

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

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

Auxiliary variables in variational inference

Augmented model and the lower bound

- Latent variable z, observations x

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

Auxiliary variables in variational inference

Augmented model and the lower bound

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

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

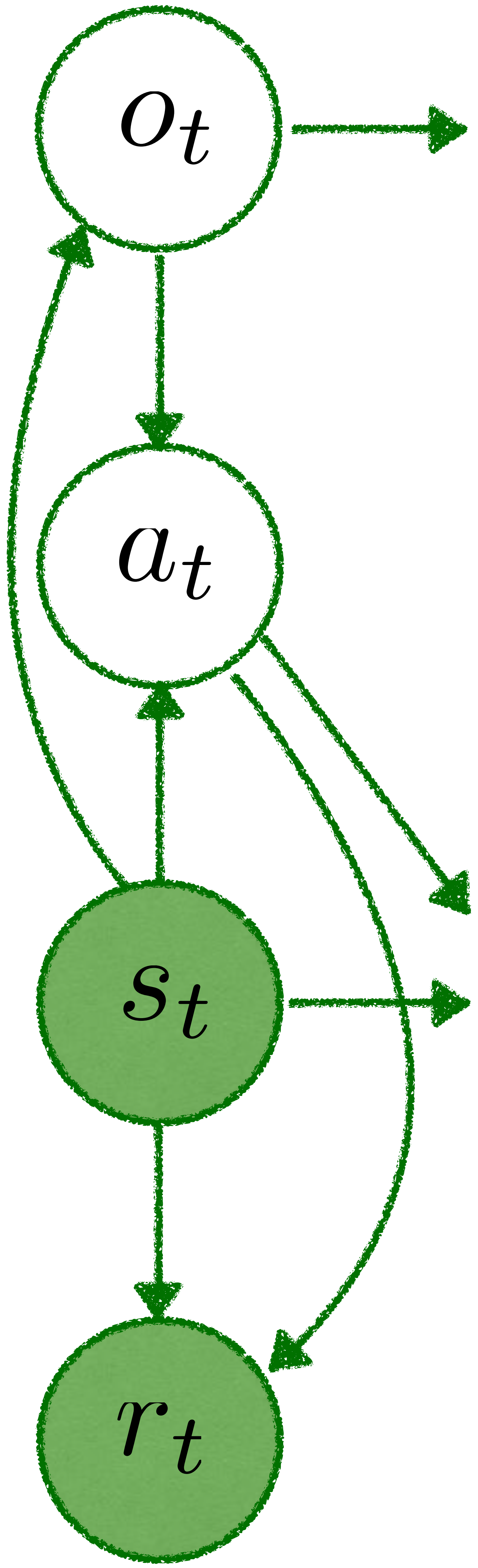
Auxiliary variables in variational inference

Augmented model and the lower bound

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

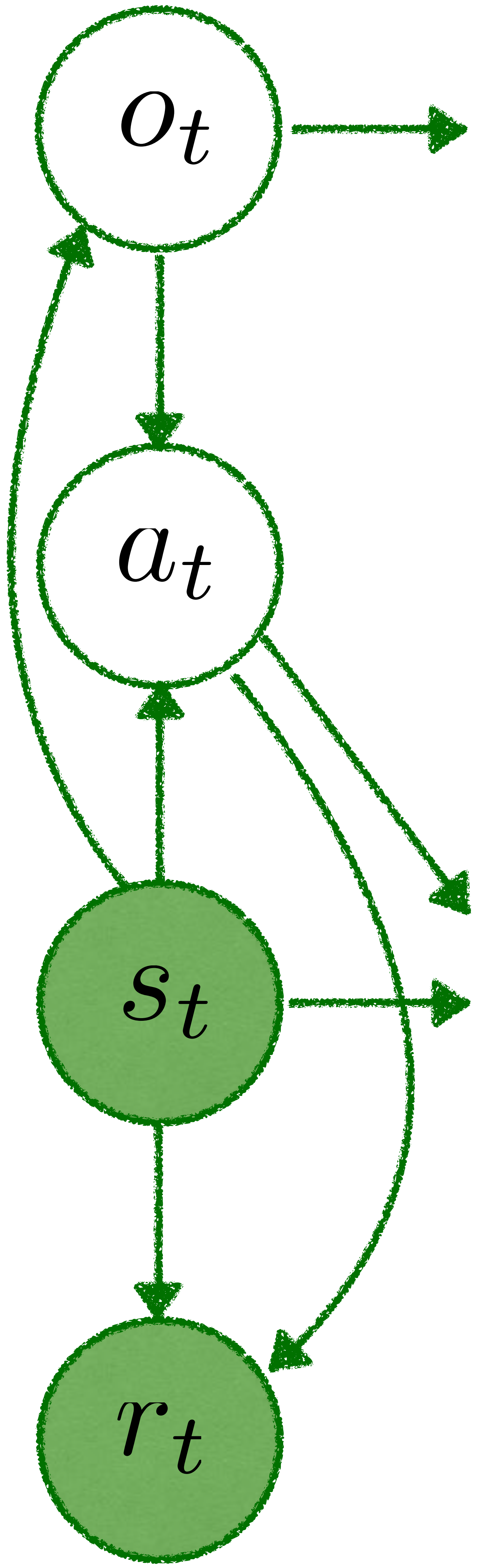
$$
\begin{aligned}
\log p(x) &= \log \int p(z)p(x|z)\tilde{q}(t|x,z)dzdt \\
&= \log \int\int q(t)q(z|t)\frac{p(z)\tilde{q}(t|x,z)}{q(t)q(z|t)}p(x|z)dtdz \\
&\geq \int\int q(t)q(z|t)\log\frac{p(z)\tilde{q}(t|x,z)}{q(t)q(z|t)}p(x|z)dtdz \\
&= \mathbb{E}_{q(t,z)}\left[\log p(x|z) - \log\frac{q(t)}{\tilde{q}(t|x,z)}\right]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

Options as auxiliary variables



Temporally extended execution

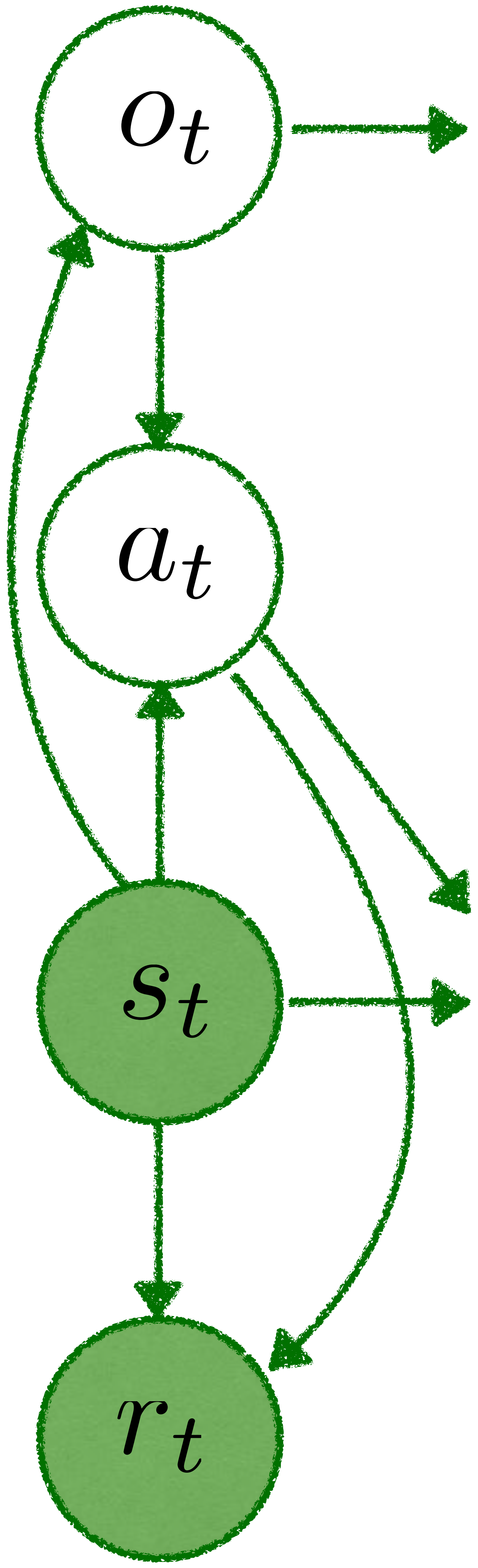
Options as auxiliary variables



Temporally extended execution

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

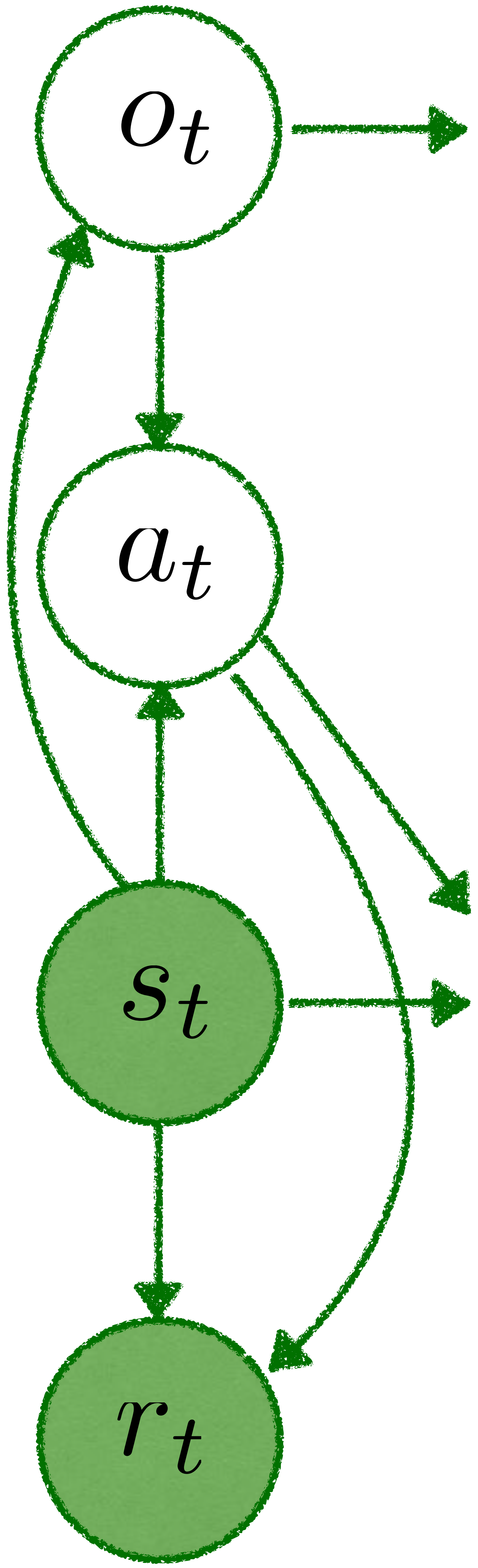
Options as auxiliary variables



Temporally extended execution

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

Options as auxiliary variables

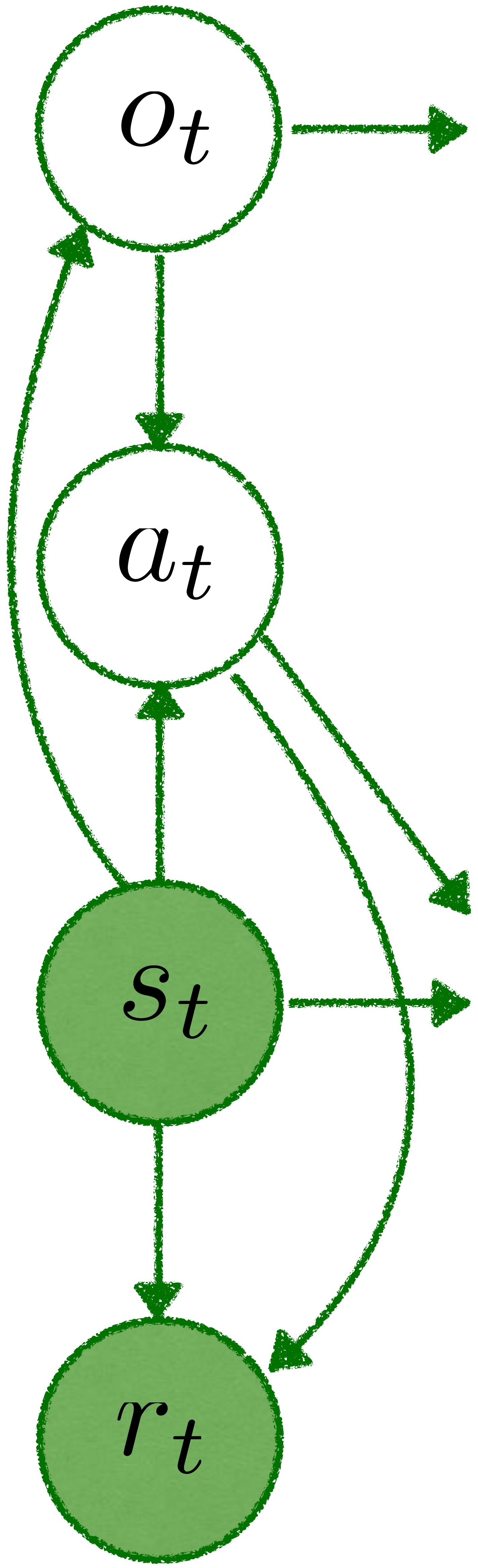


Temporally extended execution

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

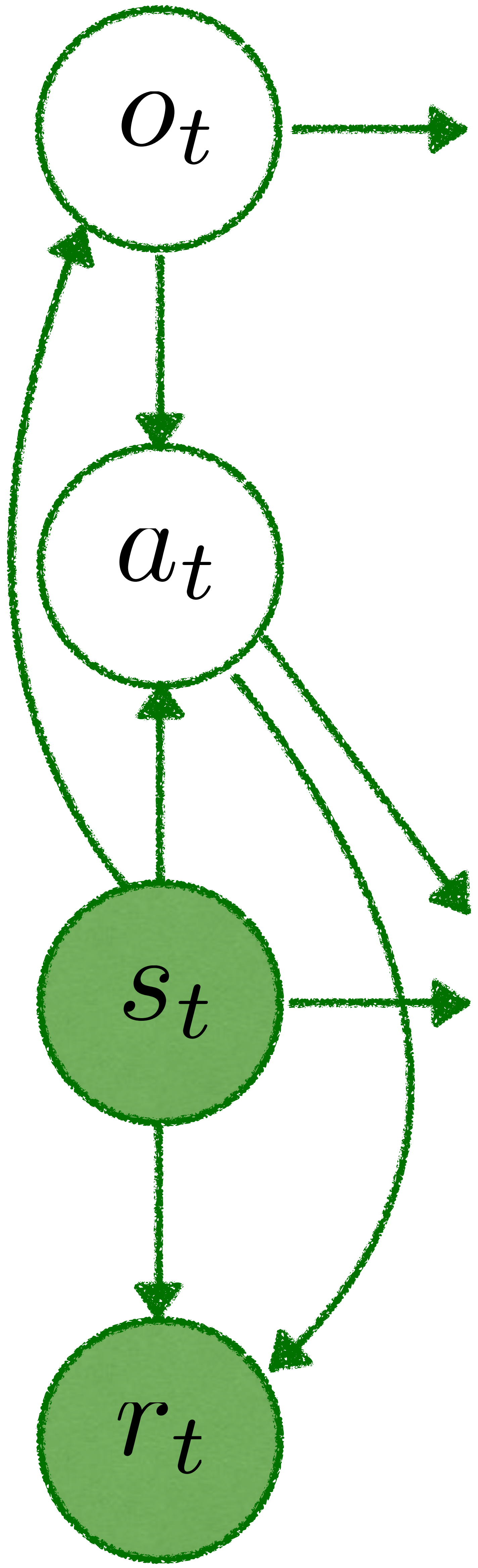
- Options enumerate policies: $\pi(a_t|s_t, o_t) = \pi_{z_t}(a_t|s_t)$
- Probability of continuing an option: $q_{\text{cont}} = q_{\text{cont}}(s_t, o_{t-1})$
- Choosing a policy:

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

Augmented approximate posterior

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

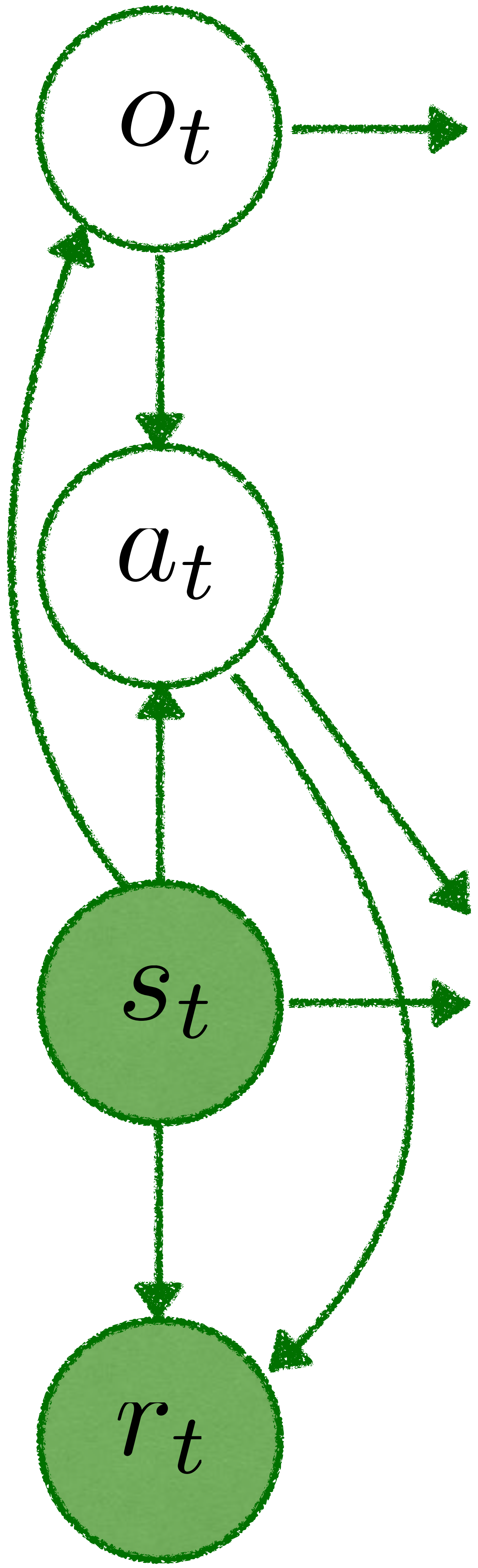
Options as auxiliary variables



Augmented prior

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

Options as auxiliary variables

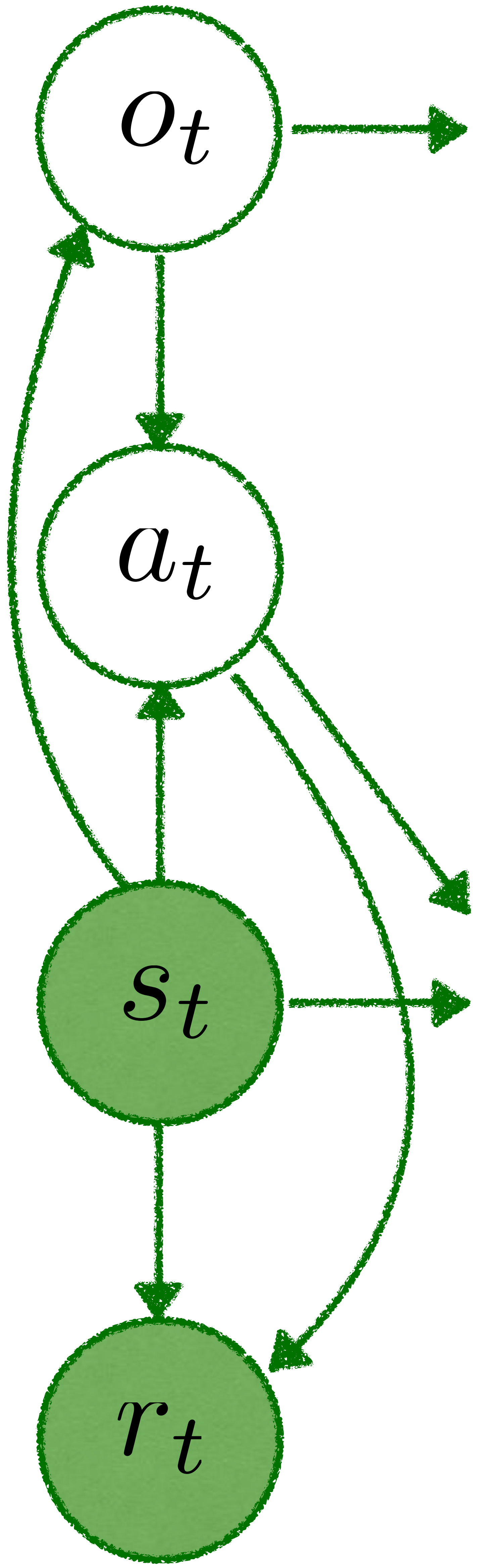


Augmented prior

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

- Reverse model is introduced

Options as auxiliary variables

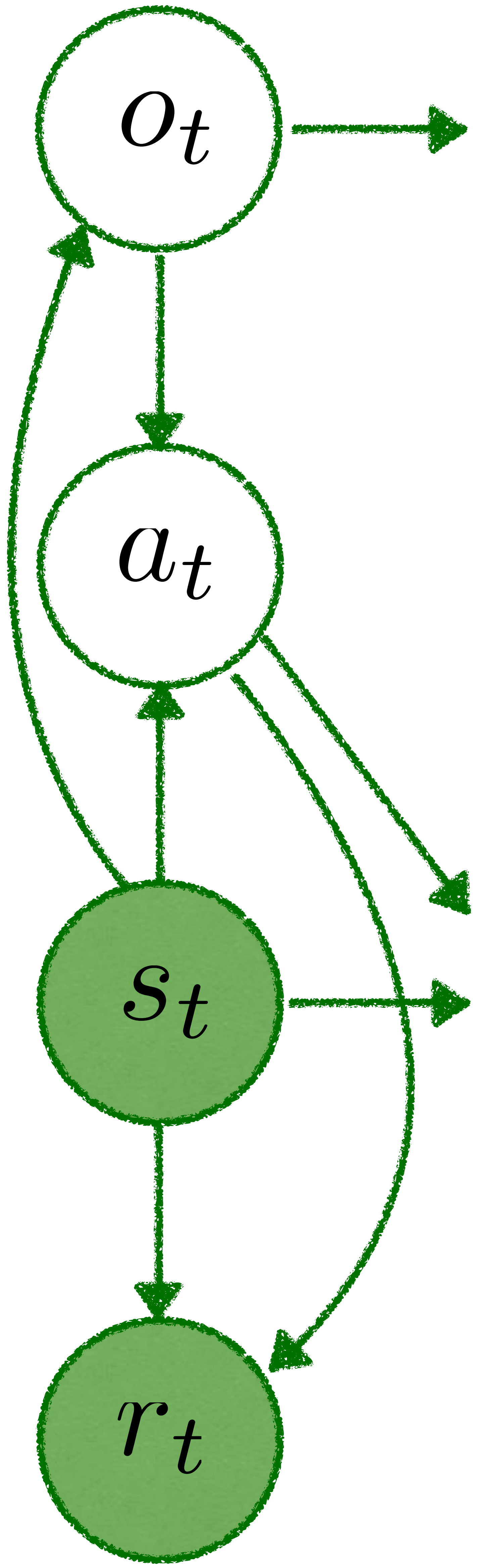


Augmented prior

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

- 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

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

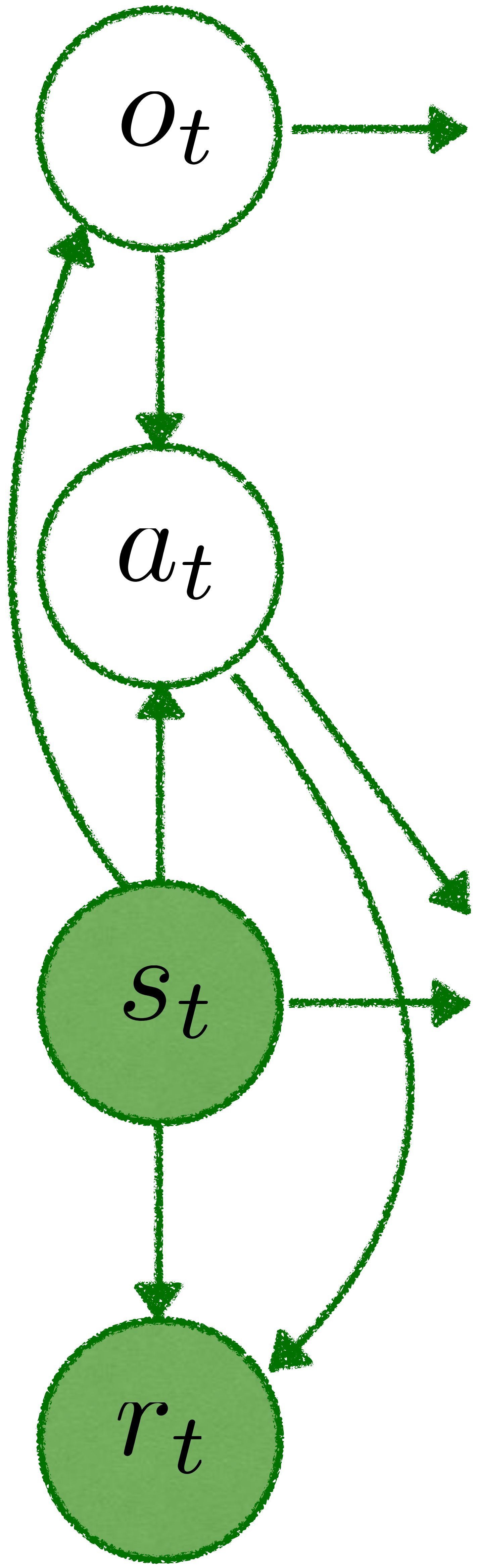
- Reverse model is introduced
- Assume $\tilde{q}(\mathbf{o}|\mathbf{s},\mathbf{a}) = \prod_{t=1}^{T}\tilde{q}(o_t|\mathbf{o}_{<t},\mathbf{s},\mathbf{a})$

Variational lower bound

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

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

Options as auxiliary variables

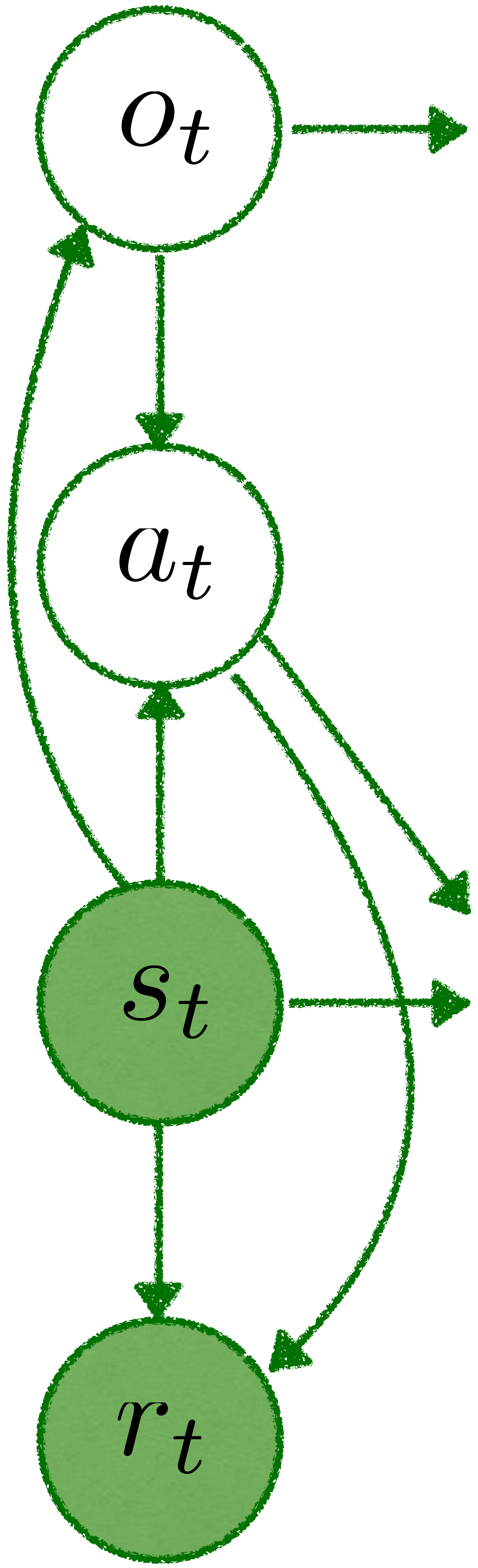


Reverse model

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

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

Options as auxiliary variables

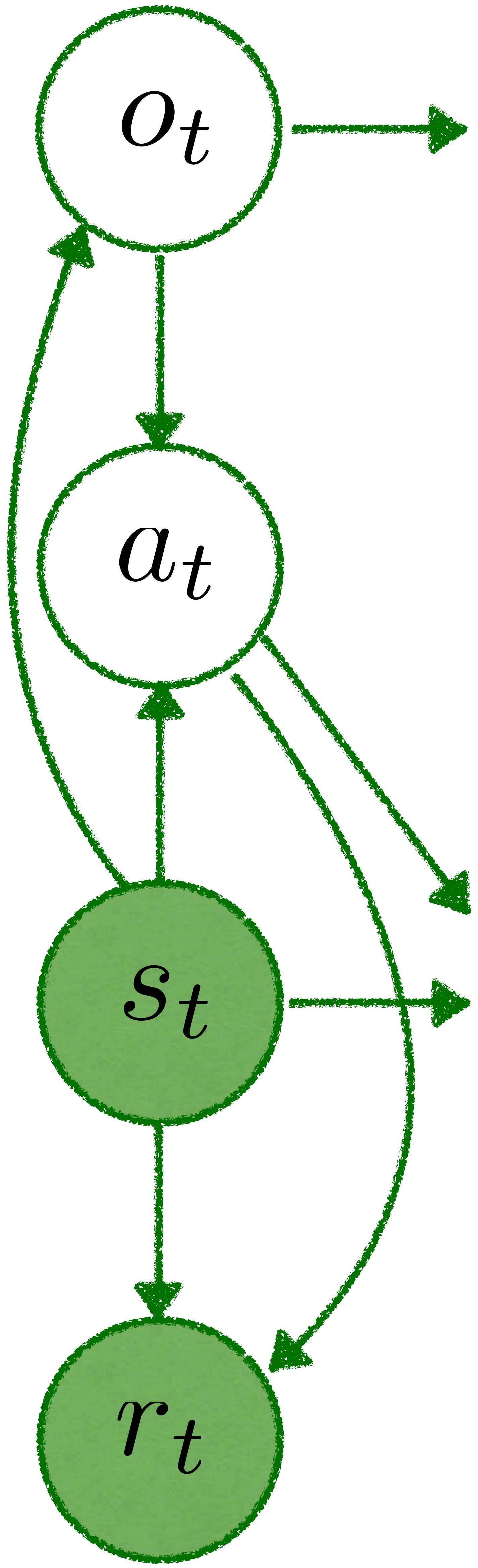


Reverse model

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

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

Options as auxiliary variables



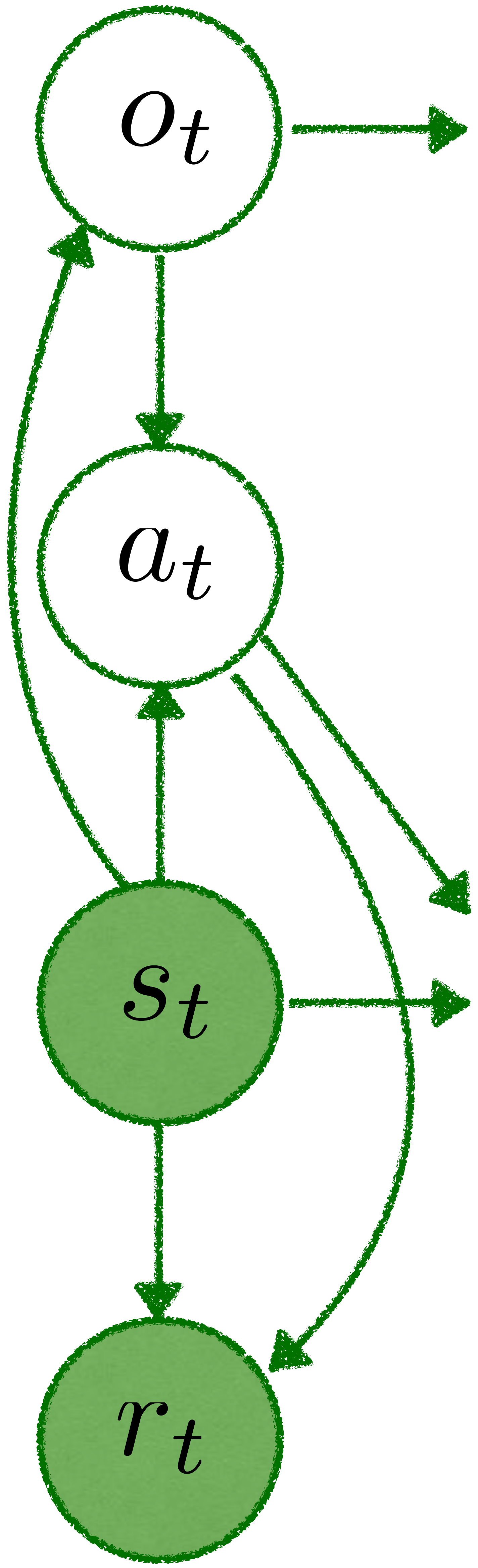
Reverse model

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

- Not used in the original Options framework

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

Options as auxiliary variables



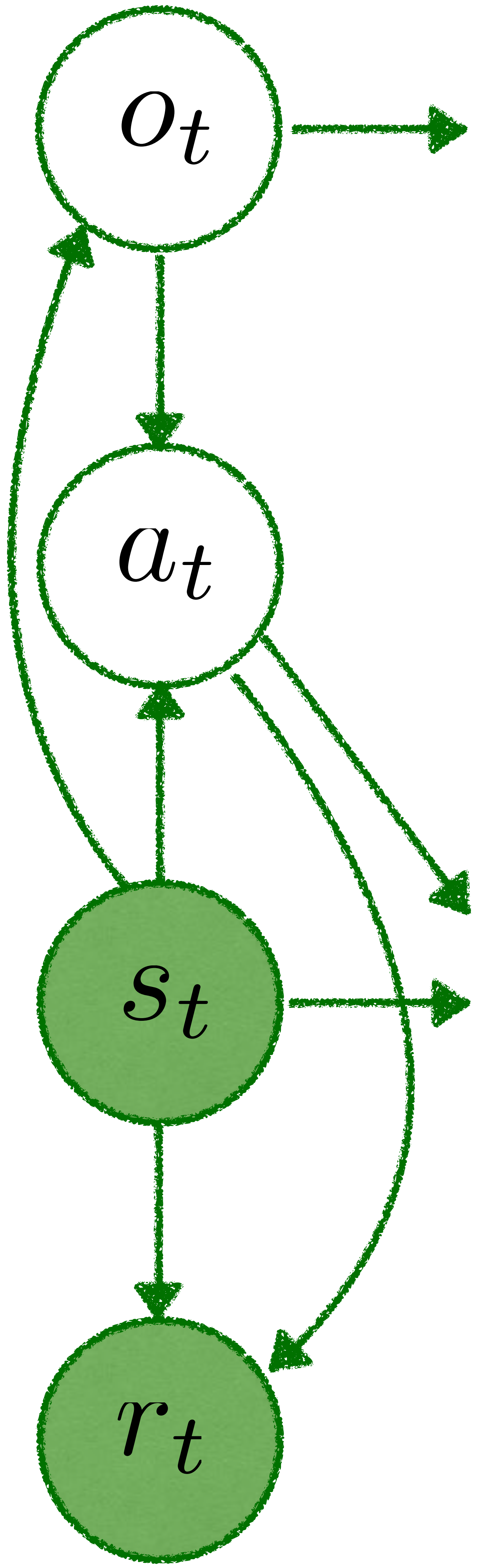
Reverse model

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

- Not used in the original Options framework
- Without reverse model options are easier to ignore

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

Options as auxiliary variables



Reverse model

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

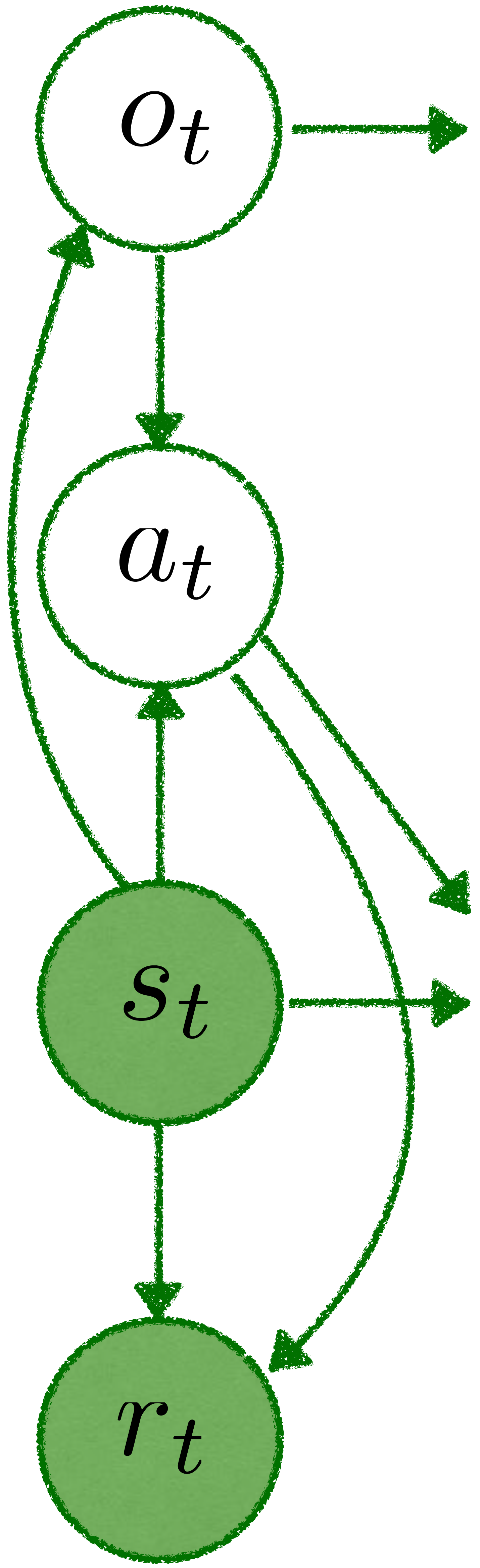
- Not used in the original Options framework
- Without reverse model options are easier to ignore

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

- Different forms are possible, e.g. forward:

$$\tilde{q}(\mathbf{o}|\mathbf{s},\mathbf{a}) = \prod_{t=1}^{T} \tilde{q}(o_t|\mathbf{o}_{>t},\mathbf{s},\mathbf{a})$$

Options as auxiliary variables



Reverse model

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

- Not used in the original Options framework
- Without reverse model options are easier to ignore

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

- Different forms are possible, e.g. forward:

$$\tilde{q}(\mathbf{o}|\mathbf{s},\mathbf{a}) = \prod_{t=1}^{T} \tilde{q}(o_t|\mathbf{o}_{>t},\mathbf{s},\mathbf{a})$$

- Practically, one can use K-step lookups

$$\tilde{q}(\mathbf{o}|\mathbf{s},\mathbf{a}) = \prod_{t=1}^{T} \tilde{q}(o_t|\mathbf{o}_{t:t+K},\mathbf{s}_{t-K:t+K},\mathbf{a}_{t-K:t+K})$$

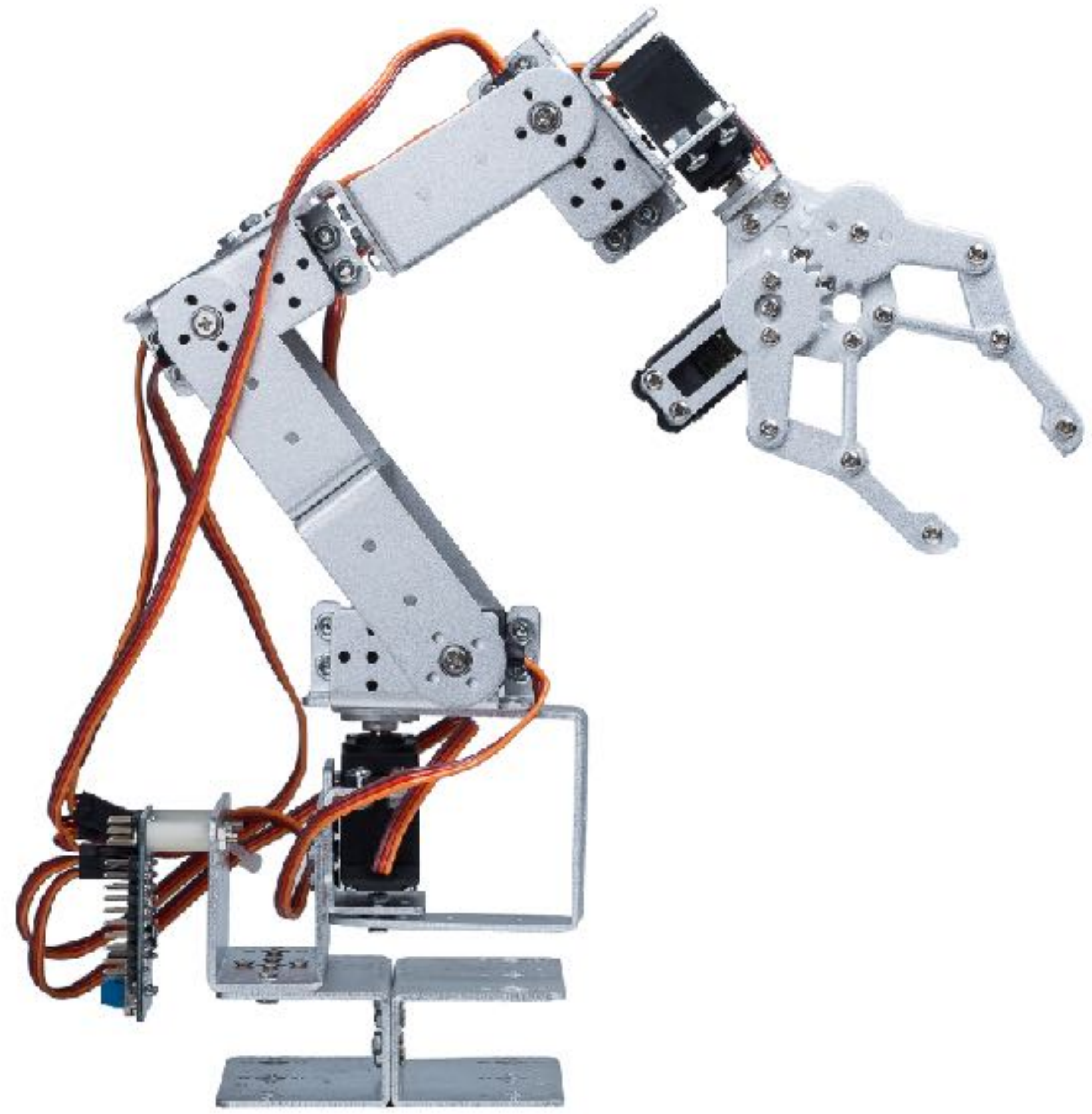
Model-based RL

Model-based RL

Simulator
(cheap)

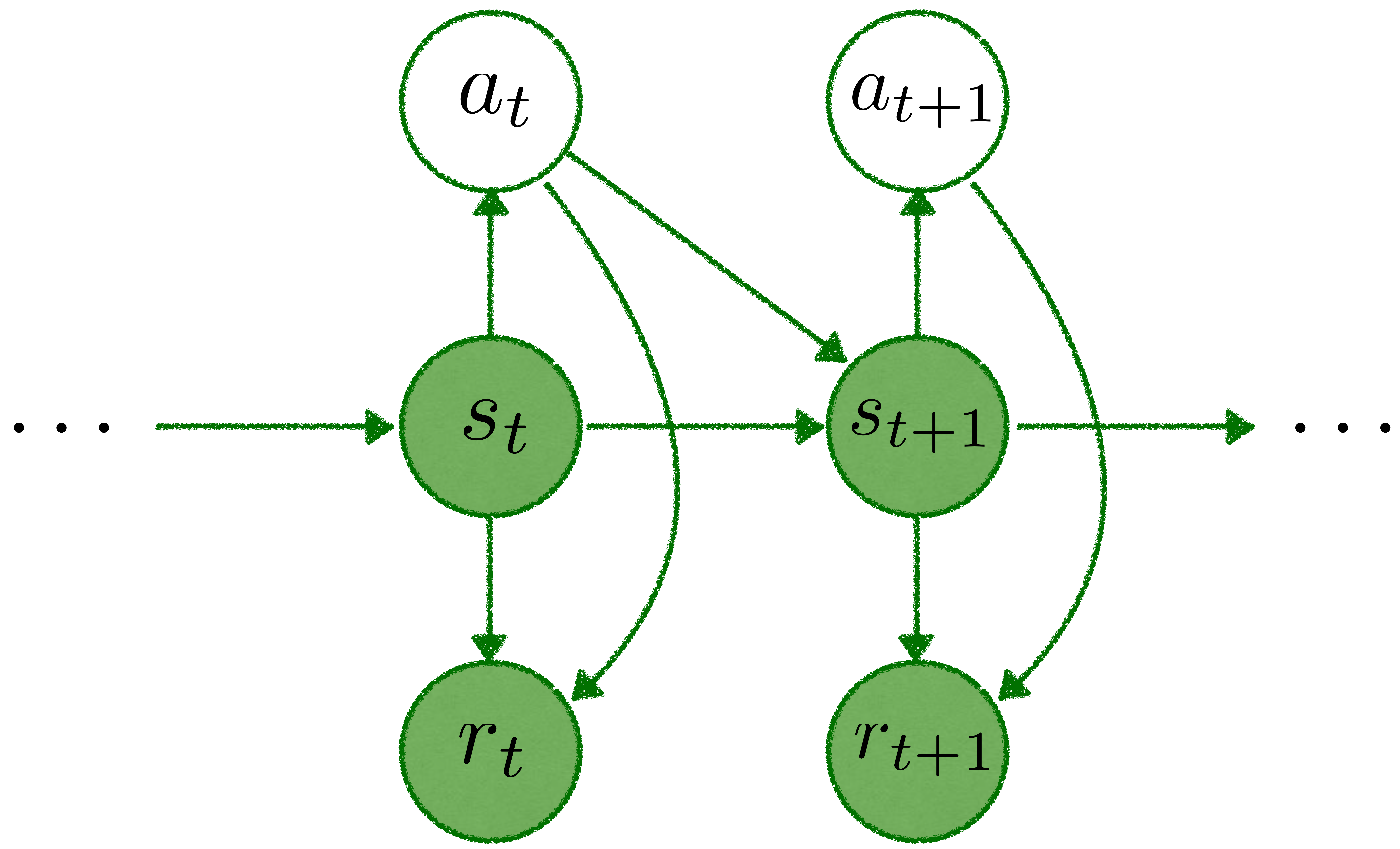
Model-based RL

Simulator
(cheap)



Real world
(expensive)

Learning model of environment

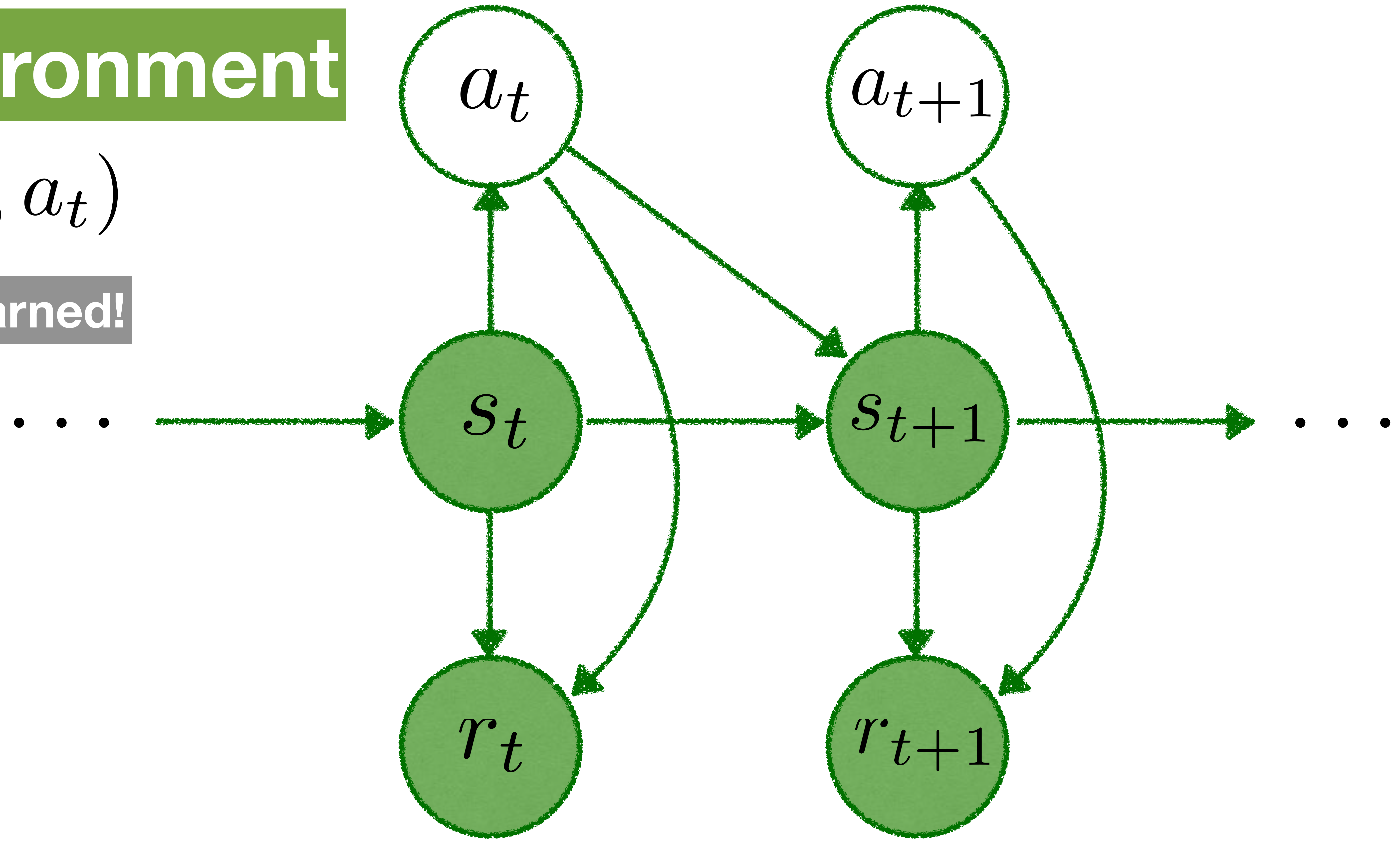


Learning model of environment

Model of environment

$$q(s_{t+1}|s_t, a_t)$$

Can be learned!

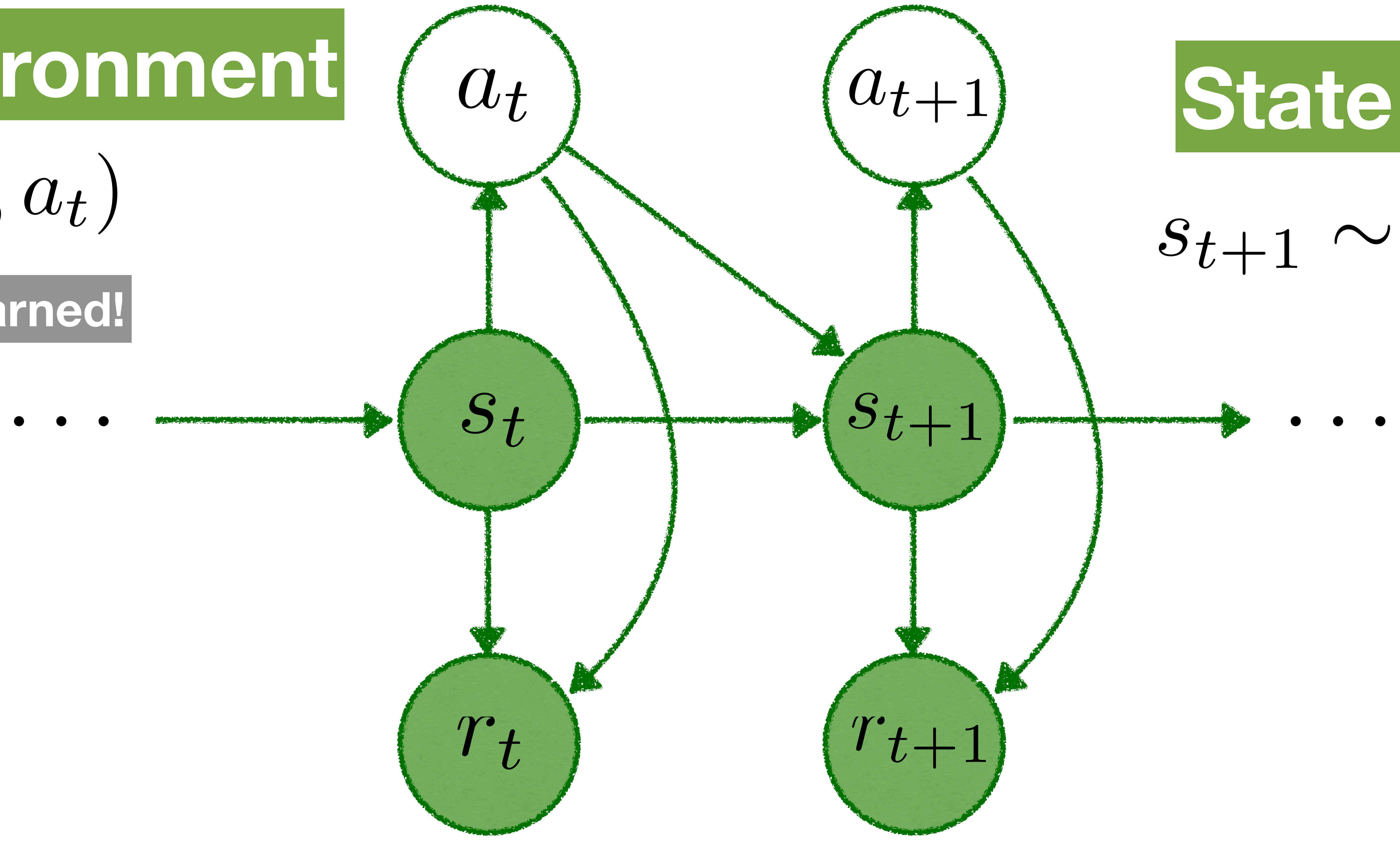


Learning model of environment

Model of environment

$$q(s_{t+1}|s_t, a_t)$$

Can be learned!



State transitions

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

Likely unknown

Learning model of environment

Model of environment

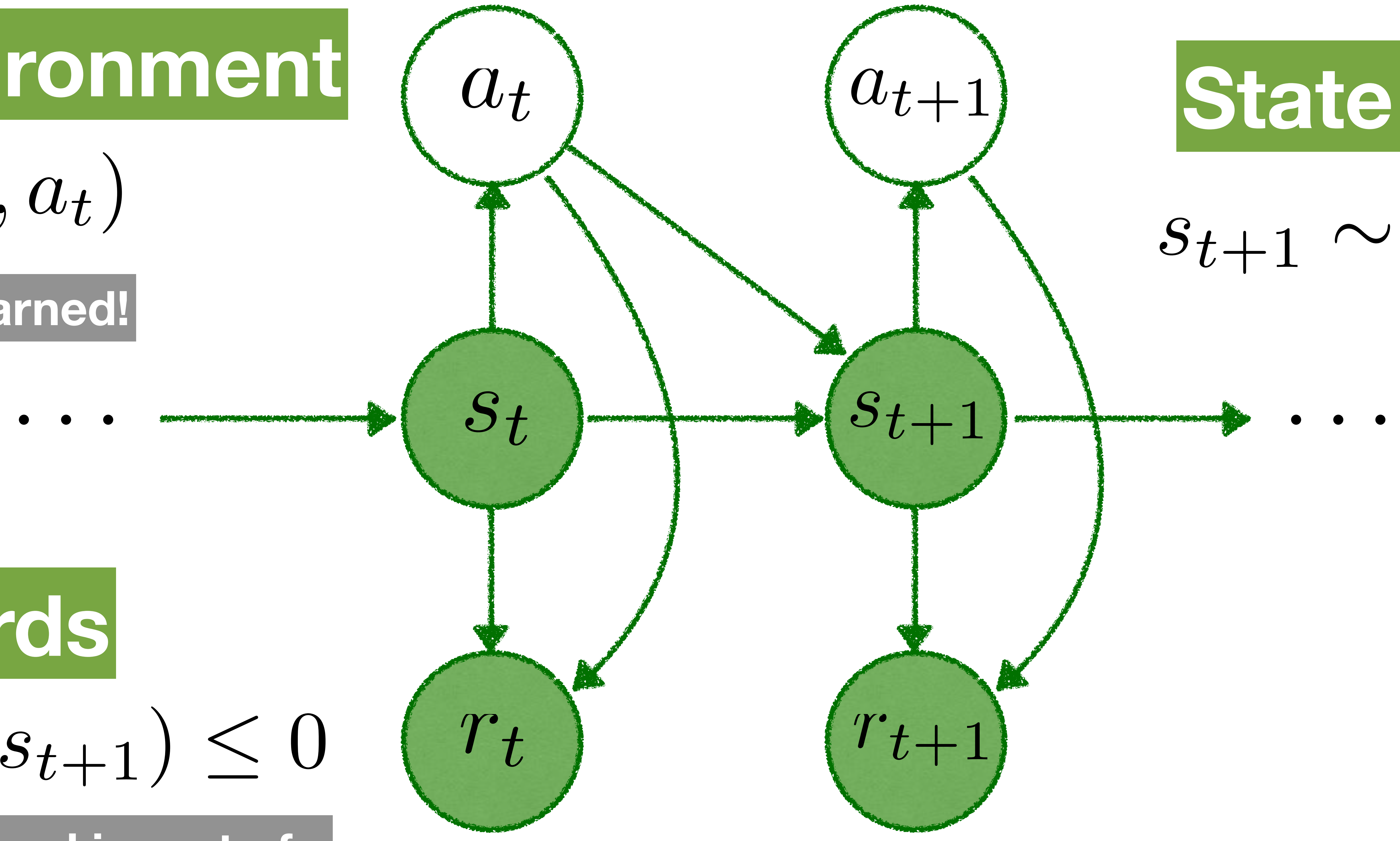
$$q(s_{t+1}|s_t, a_t)$$

Can be learned!

Rewards

$$r_t = r(s_t, a_t, s_{t+1}) \leq 0$$

Assume that reward is part of the observation



State transitions

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

Likely unknown

Learning model of environment

Learning model of environment

Prior

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

Learning model of environment

Prior

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

Assumed environment model

Learning model of environment

Prior

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

Assumed environment model

Likelihood

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

Assume that reward is part of the observation

Learning model of environment

Prior

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

Assumed environment model

Likelihood

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

Assume that reward is part of the observation

Approximate posterior

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

Learning model of environment

Prior

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

Assumed environment model

Likelihood

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

Assume that reward is part of the observation

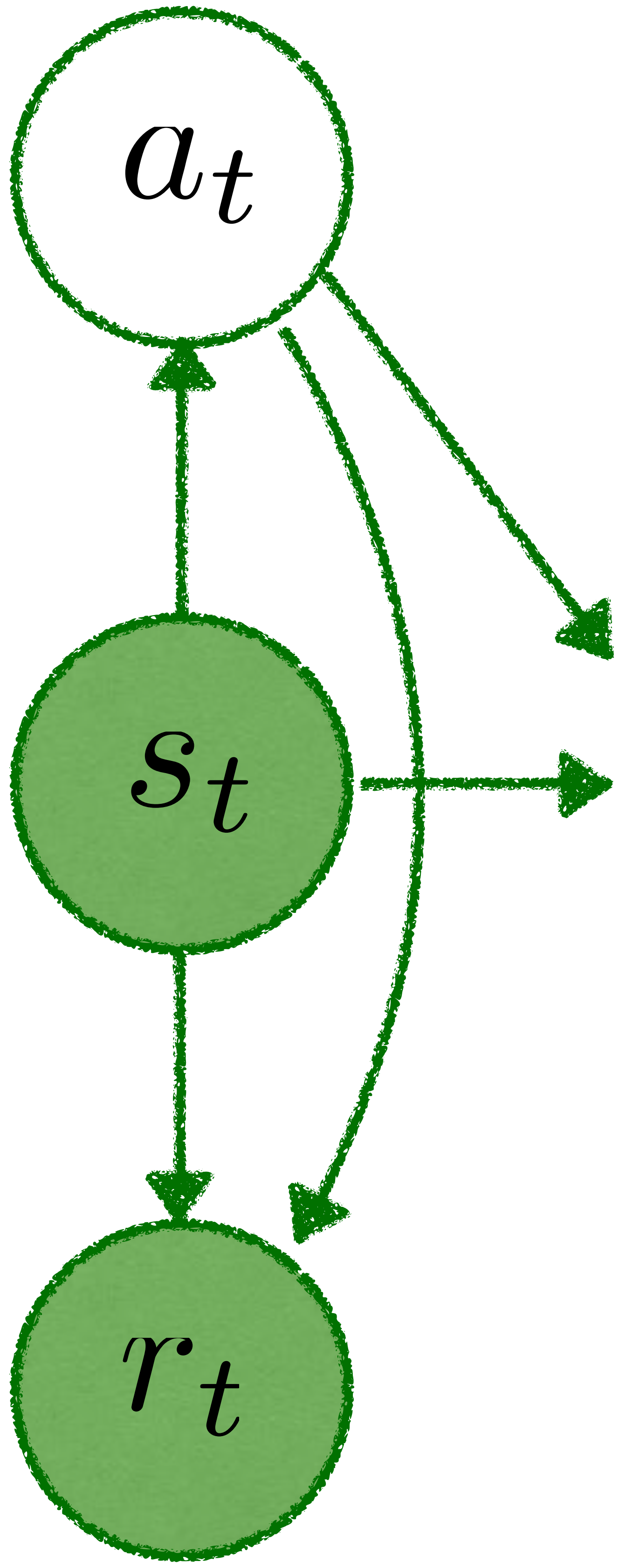
Approximate posterior

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

Real model

Learning model of environment

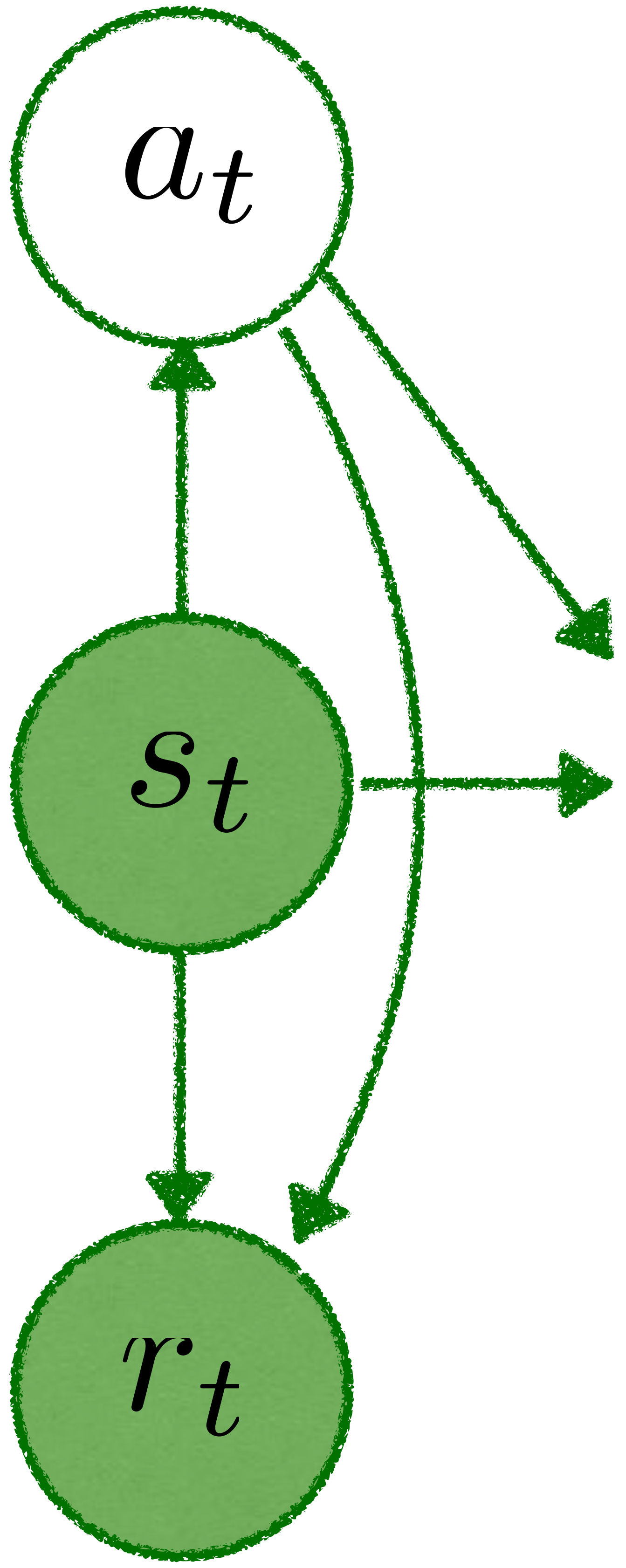
Variational lower bound



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

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

Learning model of environment



Variational lower bound

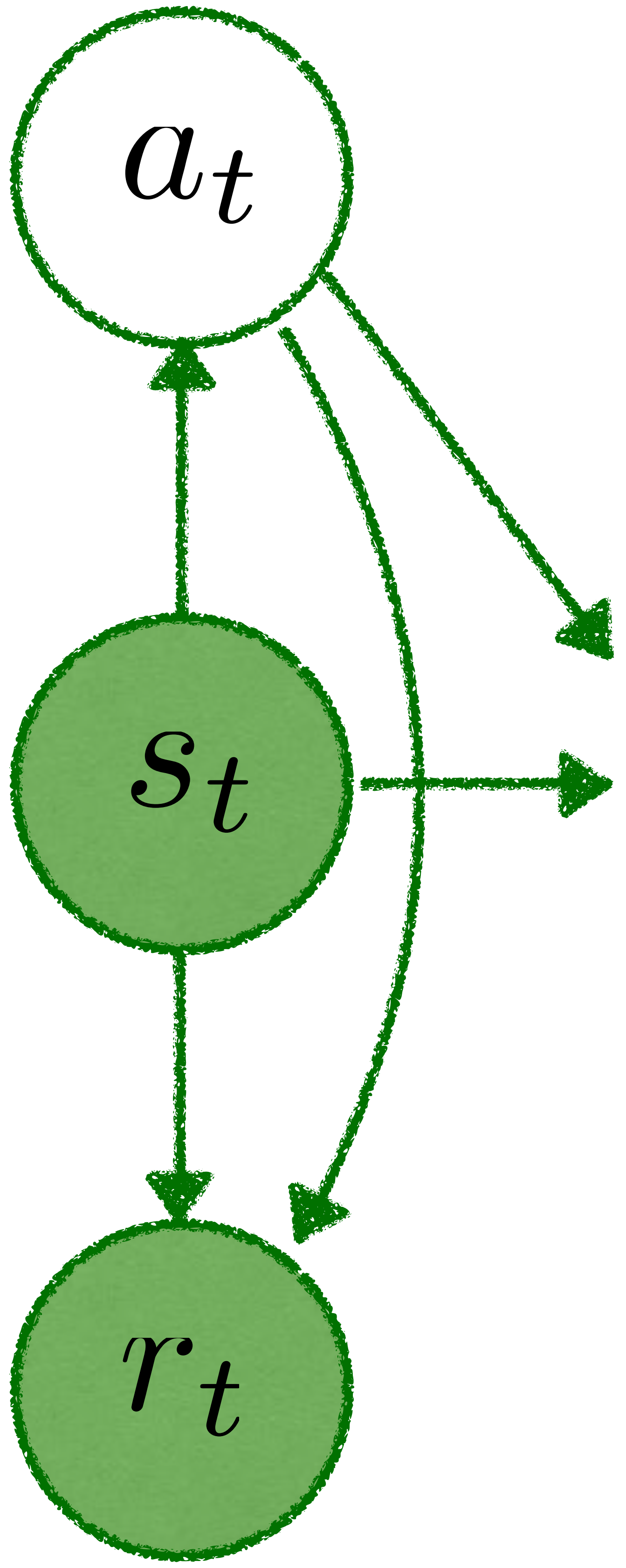
$$\mathcal{L}(p_\pi, q_{\pi_0}) = \mathbb{E}_{p_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(p(\cdot|s_t, a_t)||q(\cdot|s_t, a_t))\right]$$

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

- Optimize the lower bound for both policy and model

Learning model of environment

Variational lower bound

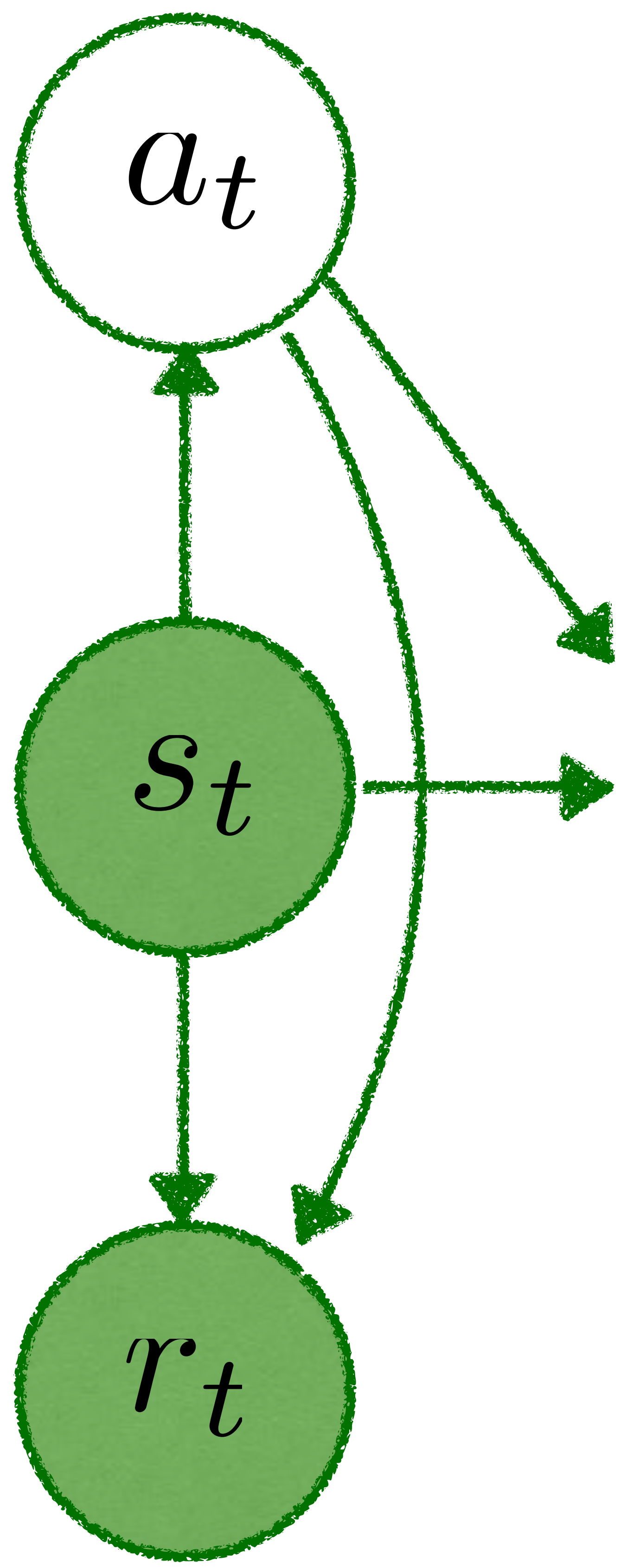


$$\mathcal{L}(p_\pi, q_{\pi_0}) = \mathbb{E}_{p_\pi(\mathbf{s},\mathbf{a})}\left[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(p(\cdot|s_t, a_t)||q(\cdot|s_t, a_t))\right]$$

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

- Optimize the lower bound for both policy and model
- Share parameters between policy and model

Learning model of environment



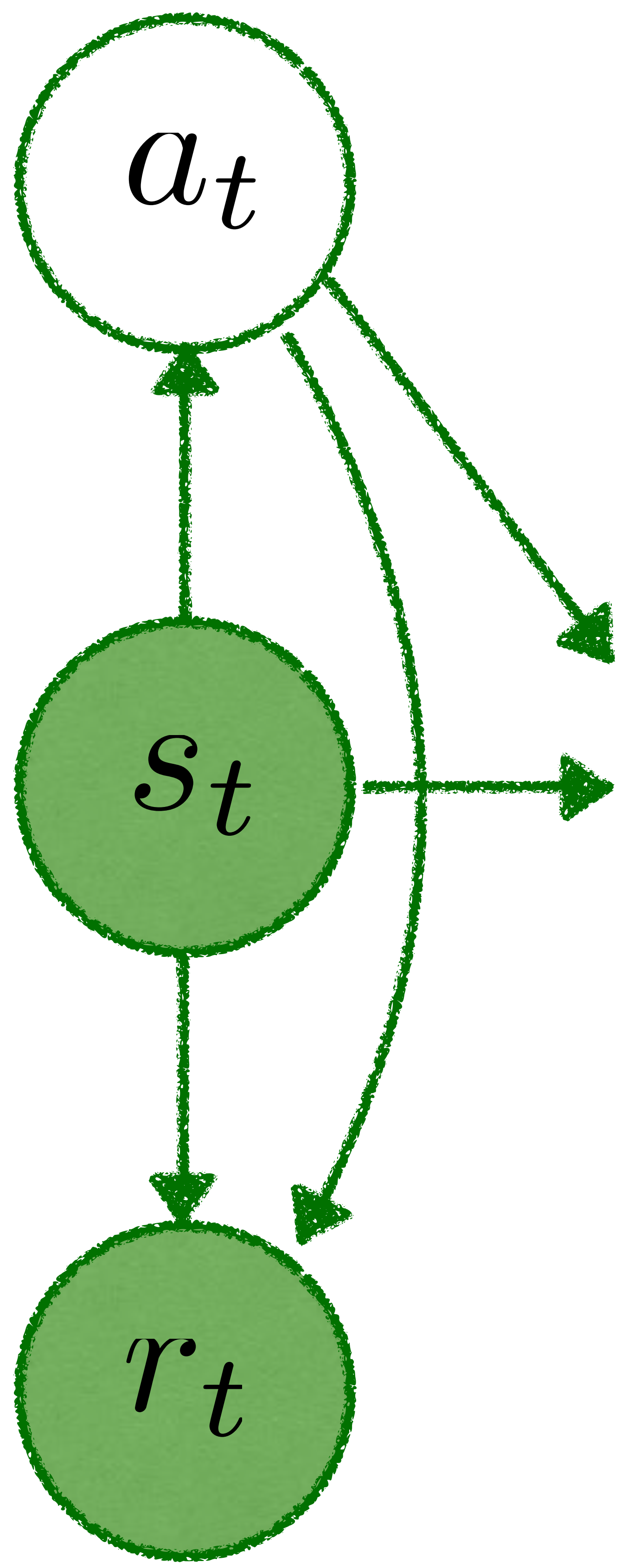
Variational lower bound

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

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

- Optimize the lower bound for both policy and model
- Share parameters between policy and model
- Use simulated rollouts in the policy

Learning model of environment



Variational lower bound

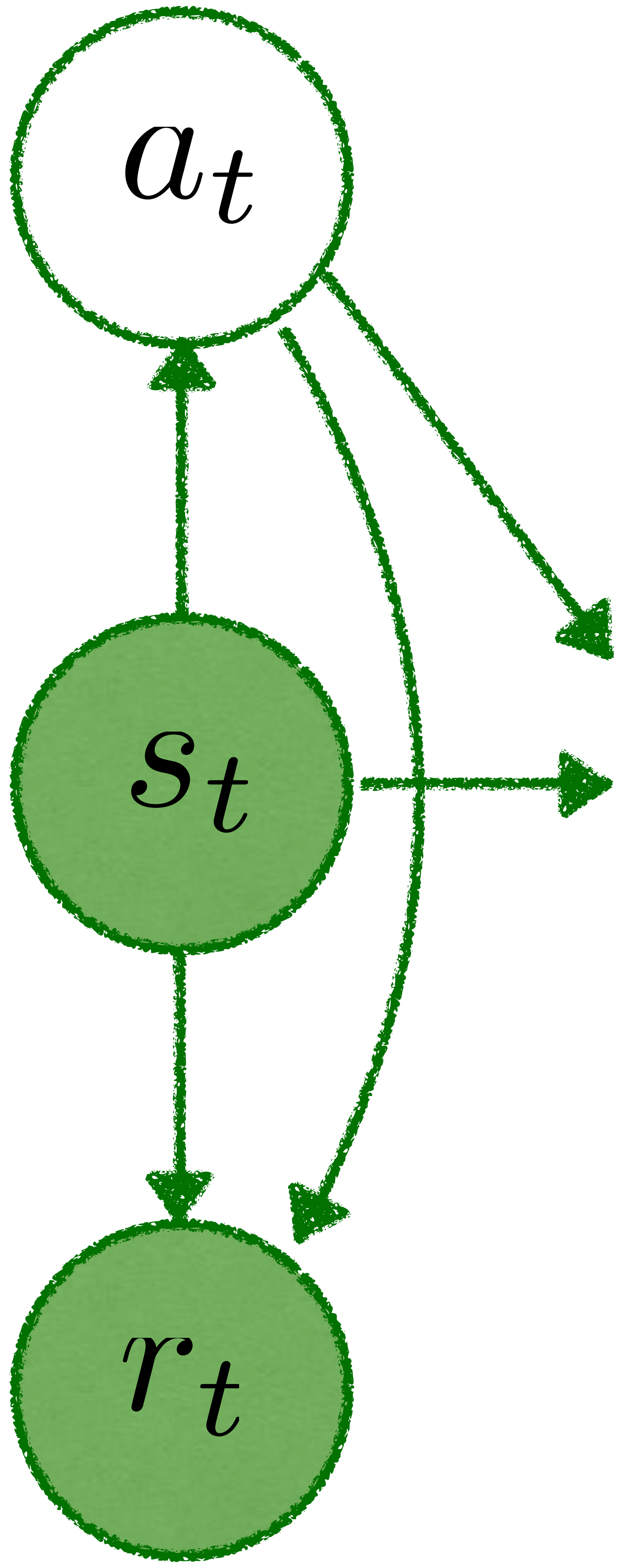
$$\mathcal{L}(p_\pi, q_{\pi_0}) = \mathbb{E}_{p_\pi(\mathbf{s},\mathbf{a})}\Big[\alpha \sum_{t=1}^{T} r_t - \mathrm{KL}(p(\cdot|s_t, a_t)||q(\cdot|s_t, a_t))\Big]$$

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

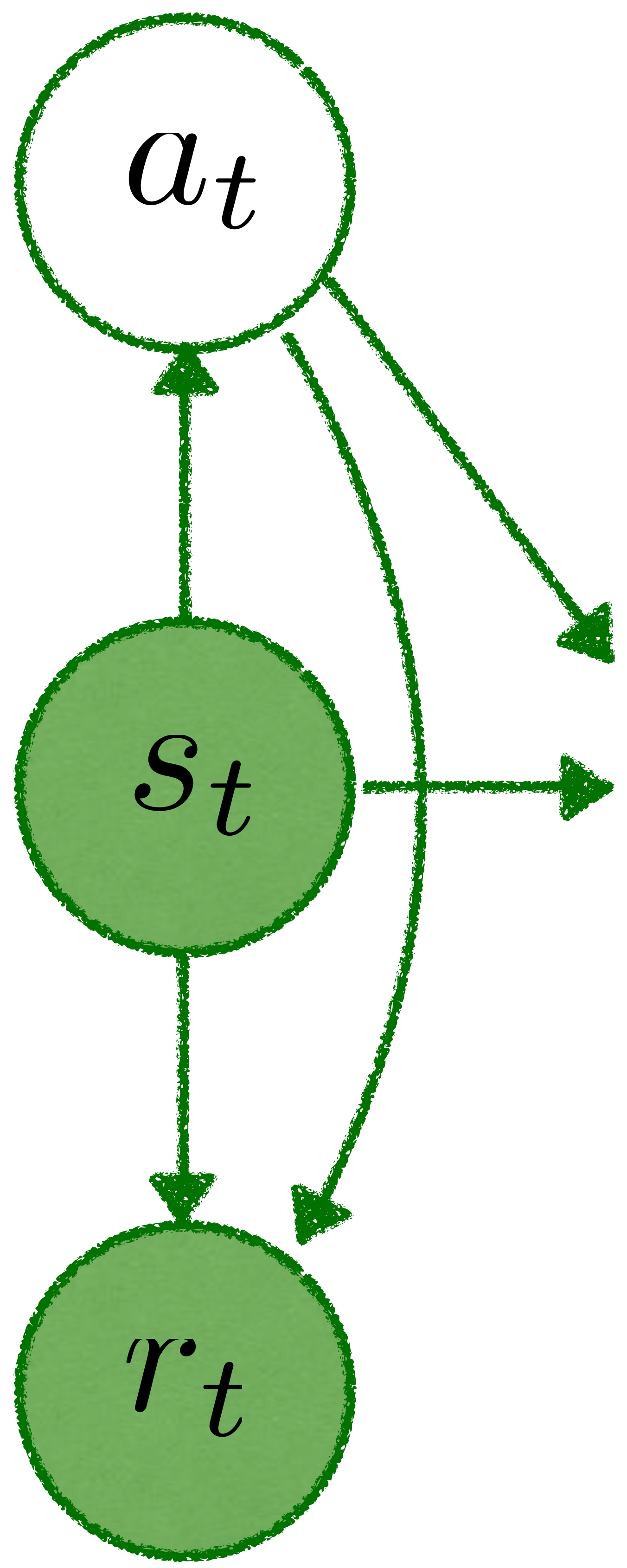
- Optimize the lower bound for both policy and model
- Share parameters between policy and model
- Use simulated rollouts in the policy

Advanced usage in [7] Weber et al, 2017

Learning from model of environment

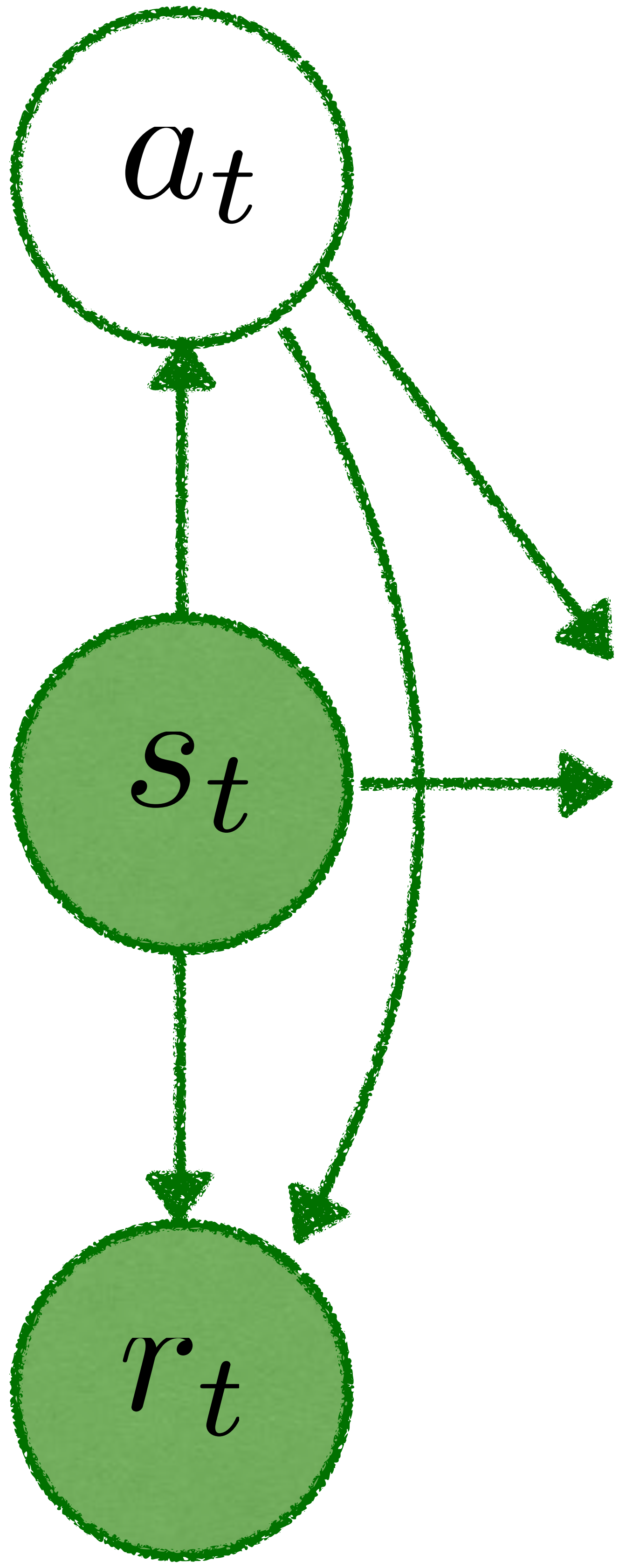


Learning from model of environment



- We would like to learn from simulations, i.e. $q_\pi(\mathbf{s}, \mathbf{a})$

Learning from model of environment

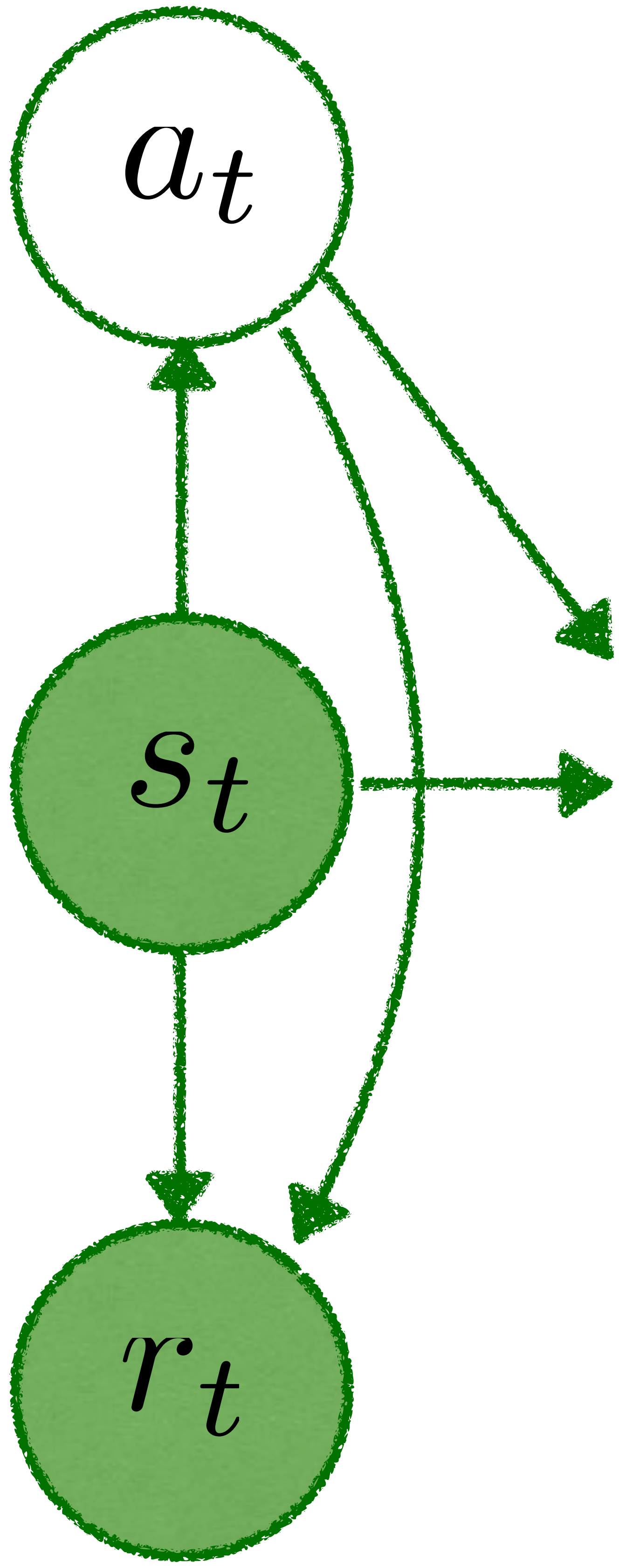


- We would like to learn from simulations, i.e. $q_\pi(\mathbf{s}, \mathbf{a})$

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

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

Learning from model of environment



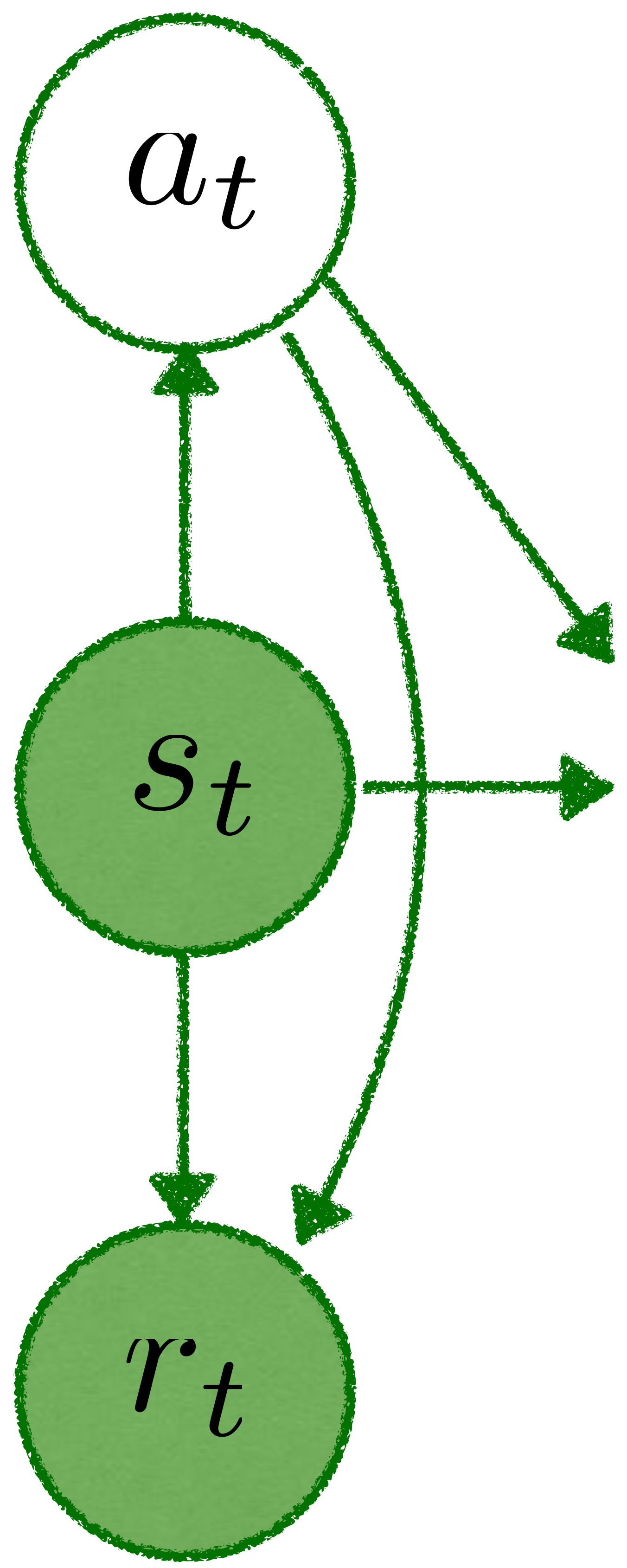
- We would like to learn from simulations, i.e. $q_\pi(\mathbf{s}, \mathbf{a})$

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

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

- We might know $q(s_{t+1}|s_t, a_t)$

Learning from model of environment



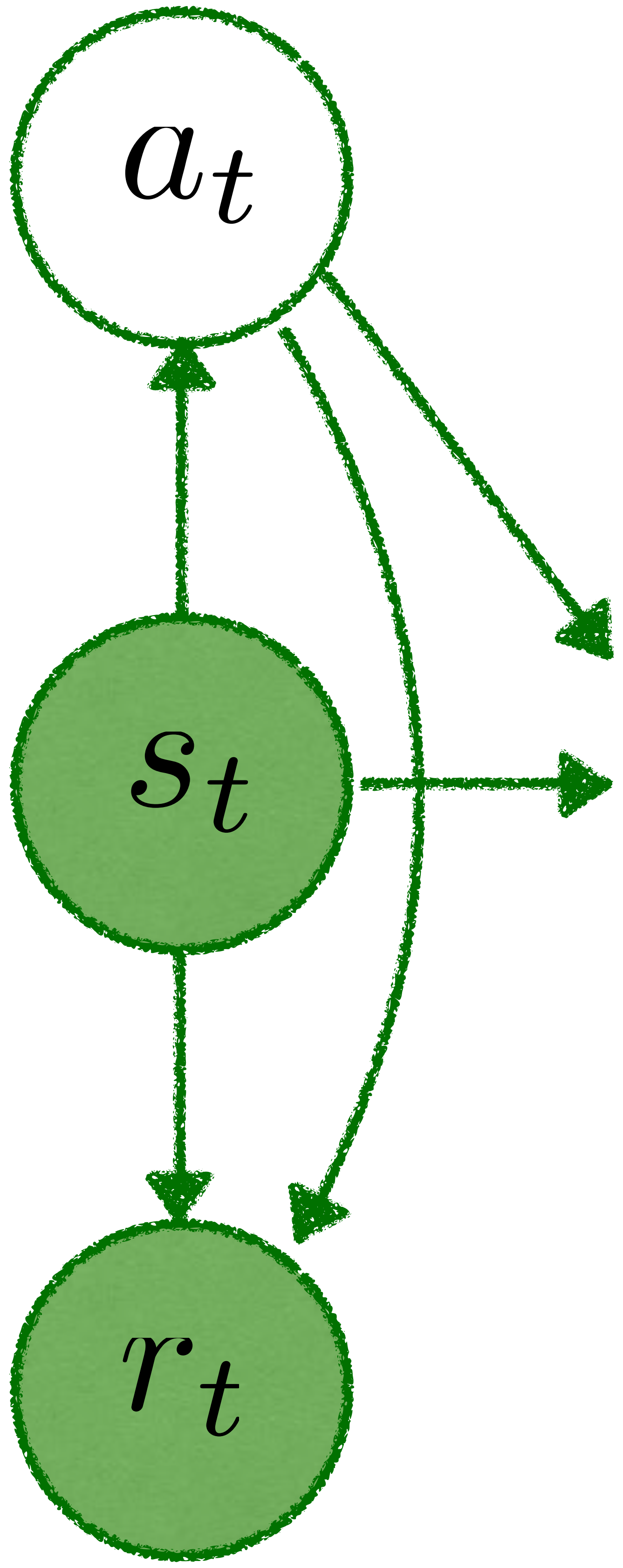
- We would like to learn from simulations, i.e. $q_\pi(\mathbf{s}, \mathbf{a})$

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

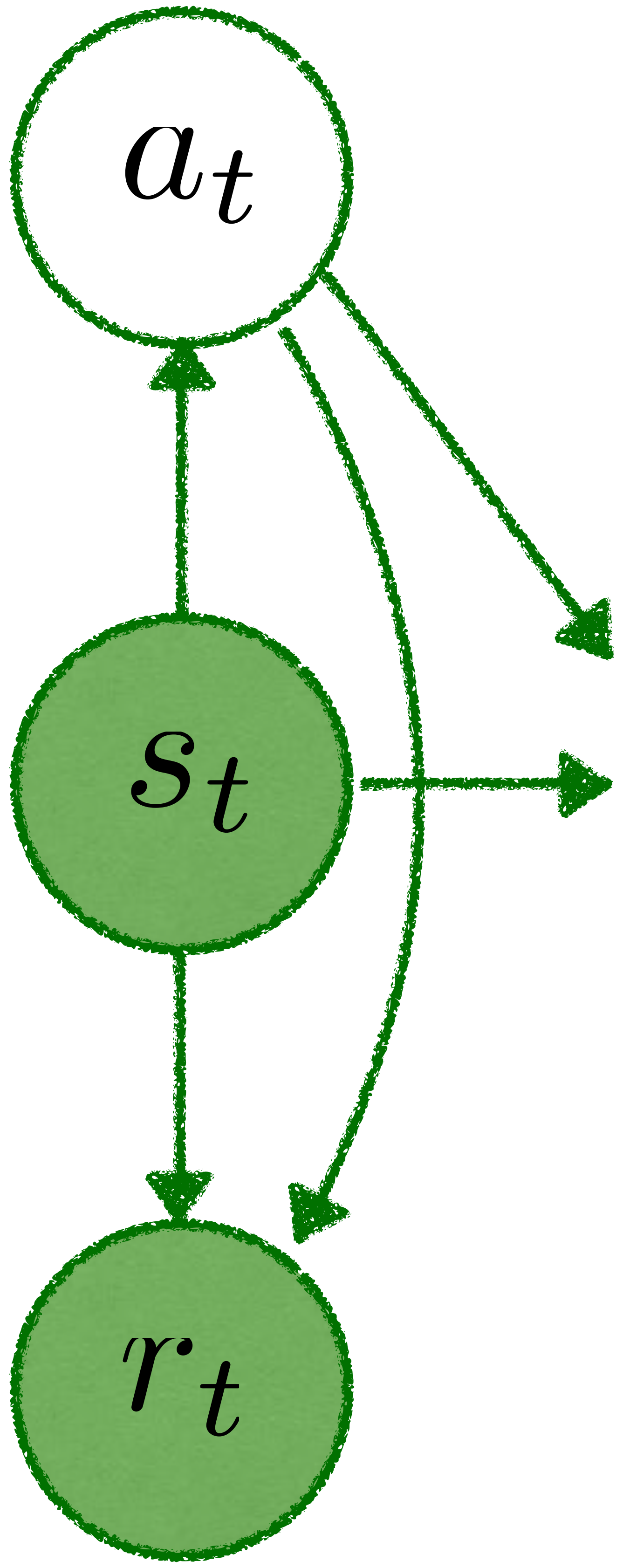
$$- \mathbb{E}_{q_\pi(\mathbf{s})}\Big[\sum_{t=1}^{T} \mathrm{KL}(\pi(\cdot|s_t)||\pi_0(\cdot|s_t))\Big]$$

- We might know $q(s_{t+1}|s_t, a_t)$
- But we don't know $p(s_{t+1}|s_t, a_t)$

Density ratio trick

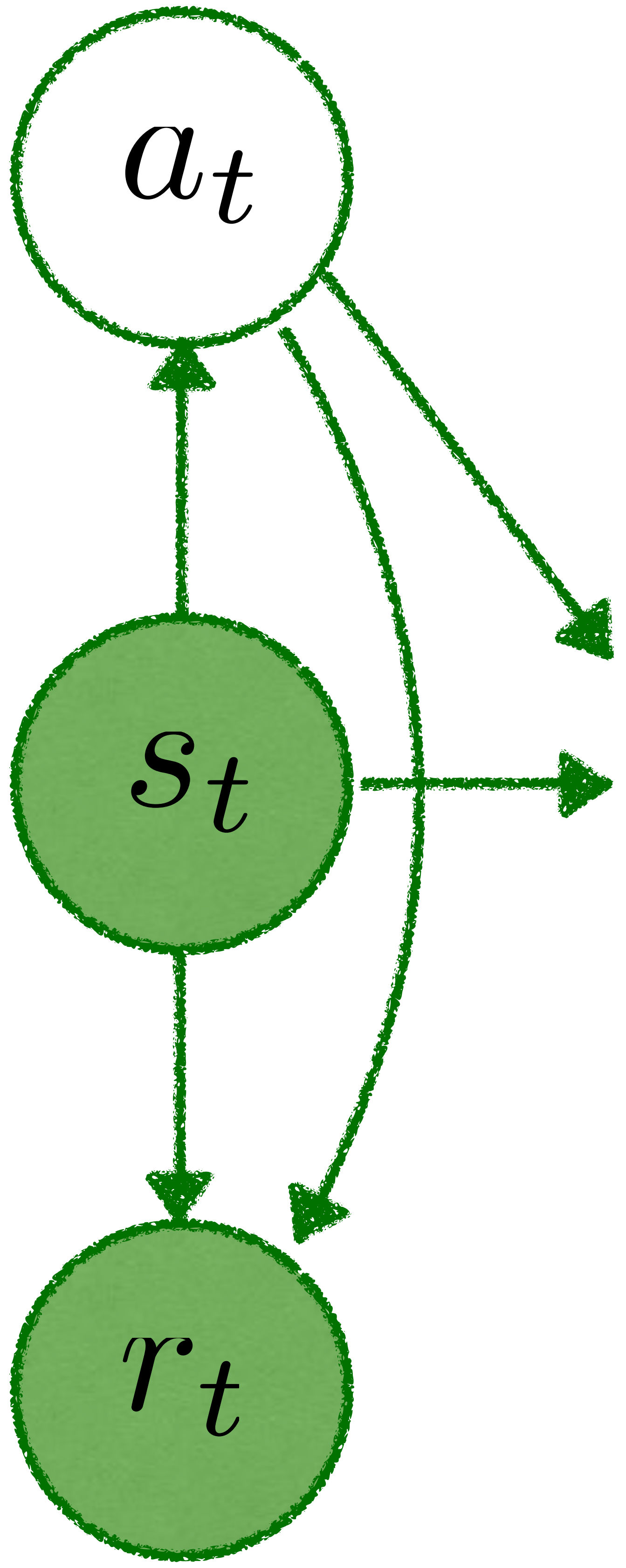


Density ratio trick



- Denote $p(s_{t+1}|y_{t+1} = 1) = p(s_{t+1}|s_t, a_t)$
  $p(s_{t+1}|y_{t+1} = 0) = q(s_{t+1}|s_t, a_t)$

Density ratio trick



- Denote $p(s_{t+1}|y_{t+1}=1) = p(s_{t+1}|s_t, a_t)$

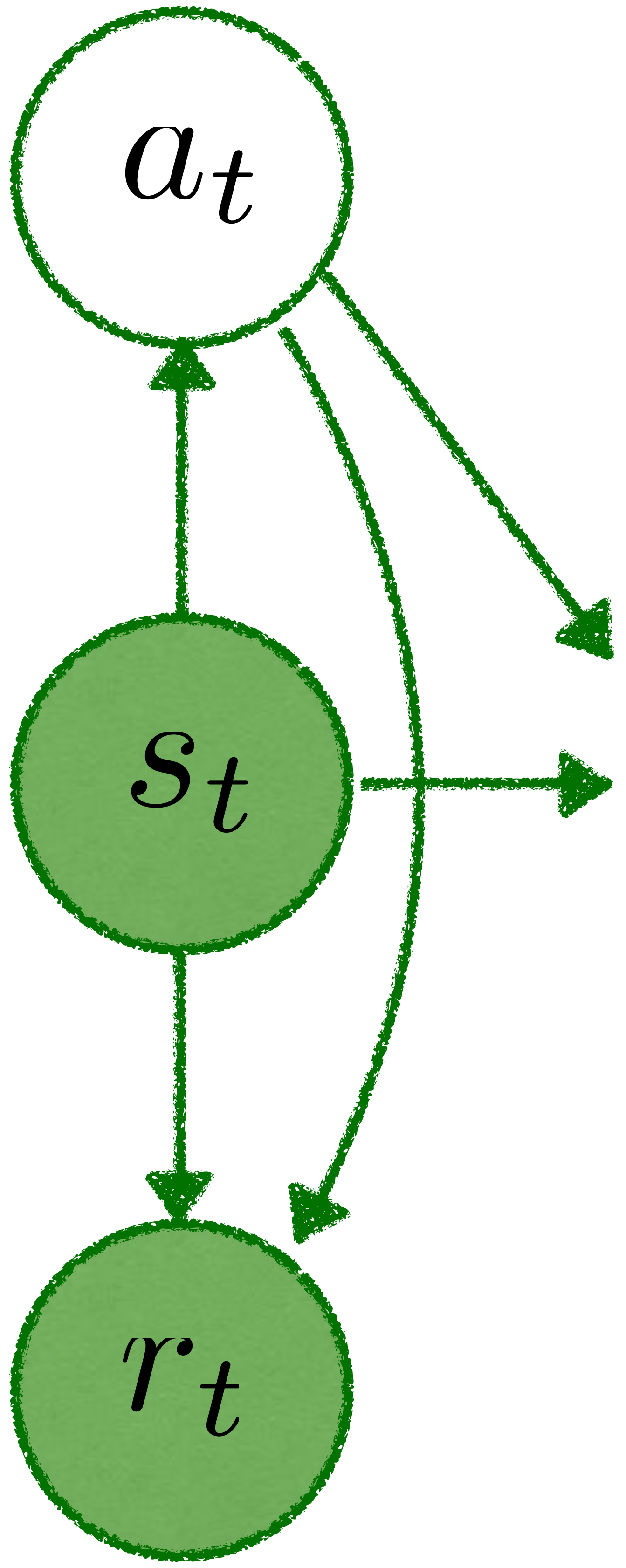
  $p(s_{t+1}|y_{t+1}=0) = q(s_{t+1}|s_t, a_t)$

- The density ratio can be written as

$$\frac{p(s_{t+1}|s_t, a_t)}{q(s_{t+1}|s_t, a_t)} = \frac{p(s_{t+1}|y_{t+1}=1)}{p(s_{t+1}|y_{t+1}=0)} = \frac{p(y_{t+1}=1|s_{t+1})p(s_{t+1})p(y_{t+1}=0)}{p(y_{t+1}=0|s_{t+1})p(s_{t+1})p(y_{t+1}=1)}$$

$$= \frac{p(y_{t+1}=1|s_{t+1})}{p(y_{t+1}=0|s_{t+1})} = \frac{p(y_{t+1}=1|s_{t+1})}{1 - p(y_{t+1}=1|s_{t+1})}$$

Density ratio trick



- Denote $p(s_{t+1}|y_{t+1}=1) = p(s_{t+1}|s_t, a_t)$

$$p(s_{t+1}|y_{t+1}=0) = q(s_{t+1}|s_t, a_t)$$

- The density ratio can be written as

$$\frac{p(s_{t+1}|s_t, a_t)}{q(s_{t+1}|s_t, a_t)} = \frac{p(s_{t+1}|y_{t+1}=1)}{p(s_{t+1}|y_{t+1}=0)} = \frac{p(y_{t+1}=1|s_{t+1})p(s_{t+1})p(y_{t+1}=0)}{p(y_{t+1}=0|s_{t+1})p(s_{t+1})p(y_{t+1}=1)}$$

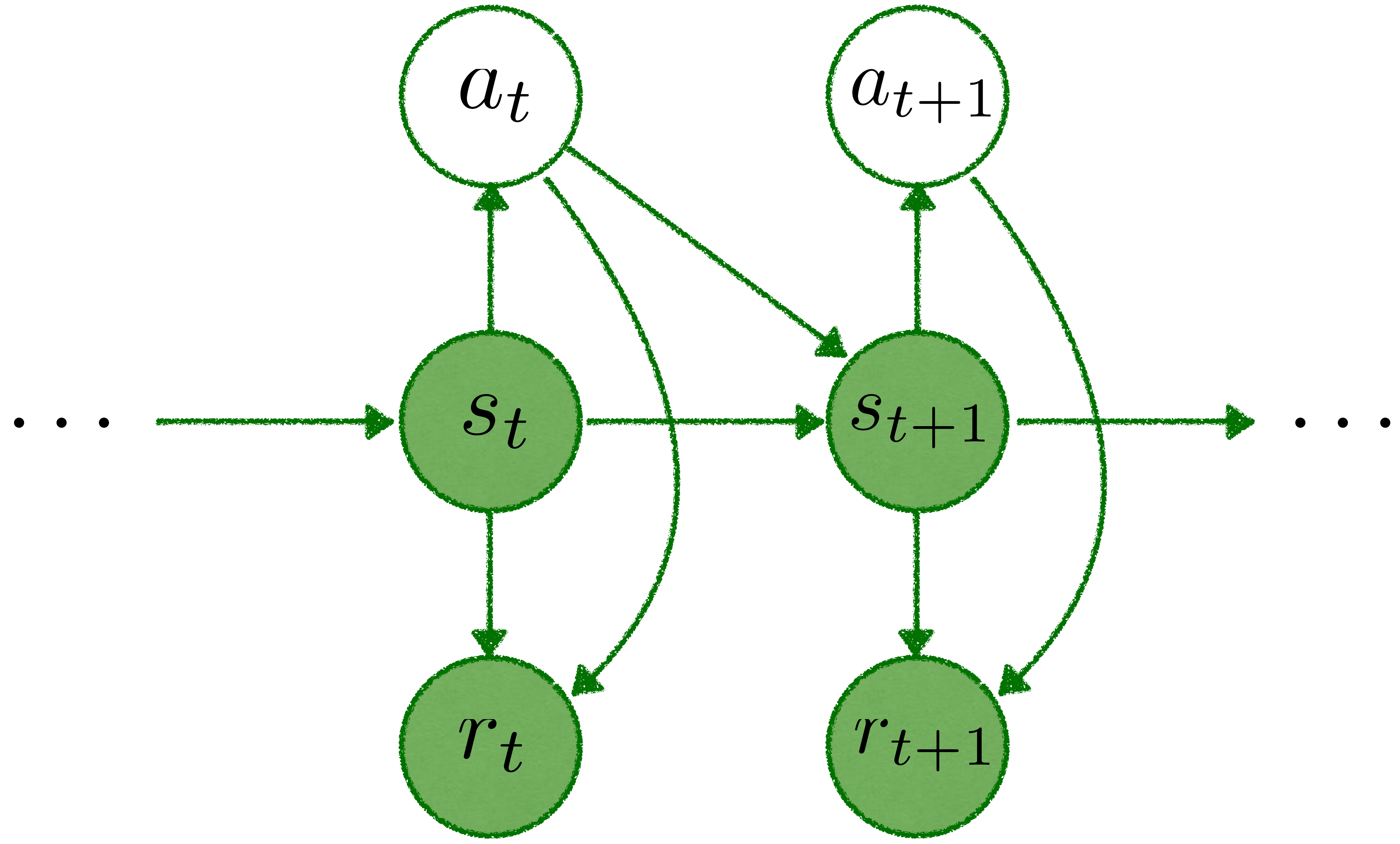
$$= \frac{p(y_{t+1}=1|s_{t+1})}{p(y_{t+1}=0|s_{t+1})} = \frac{p(y_{t+1}=1|s_{t+1})}{1-p(y_{t+1}=1|s_{t+1})}$$

- We can train a classifier to **discriminate** between imagined and real transitions

[9] Huszár, 2017

Model-based learning

[10] Ho & Efron, 2017



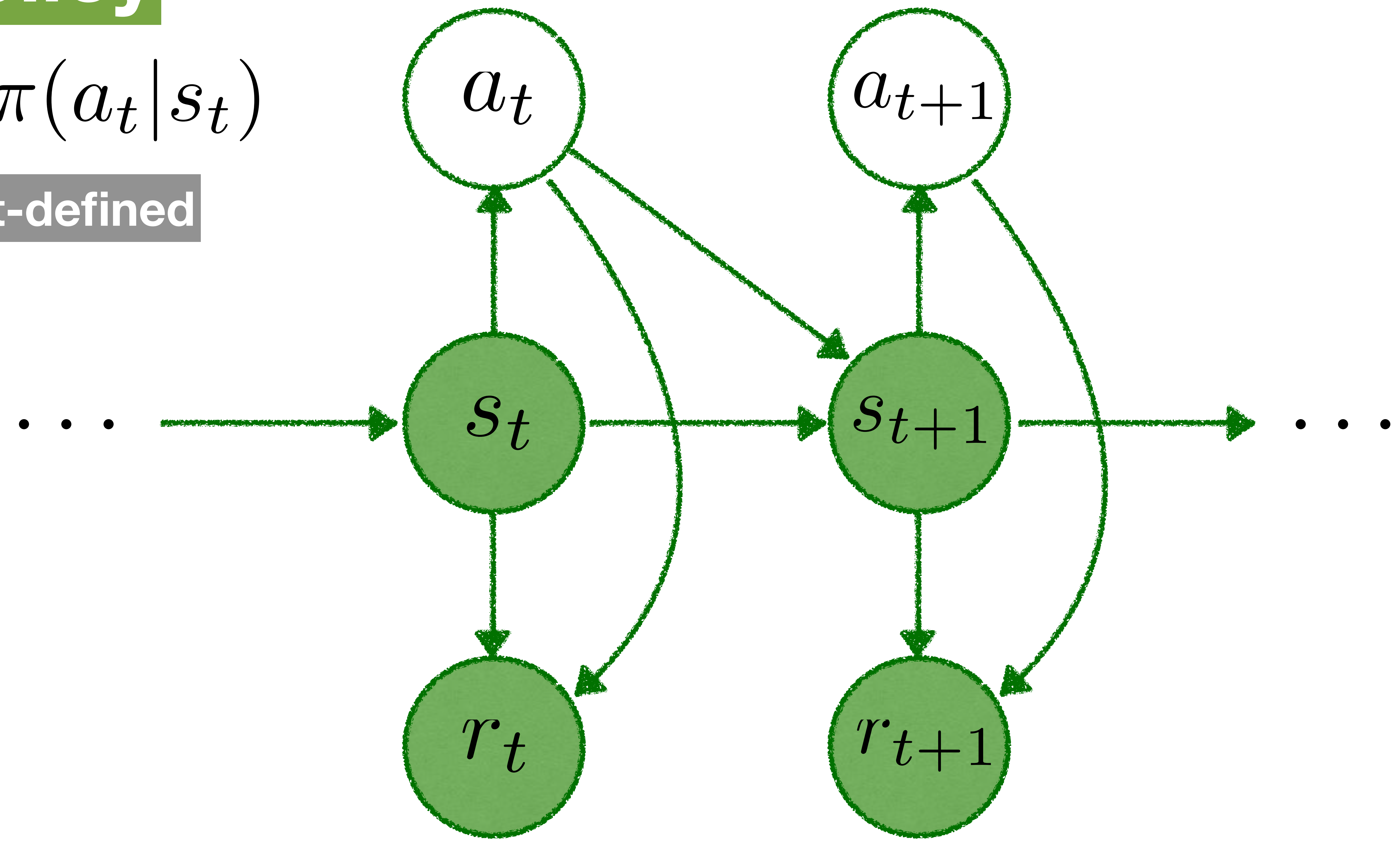
Model-based learning

[10] Ho & Efron, 2017

Policy

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

Agent-defined



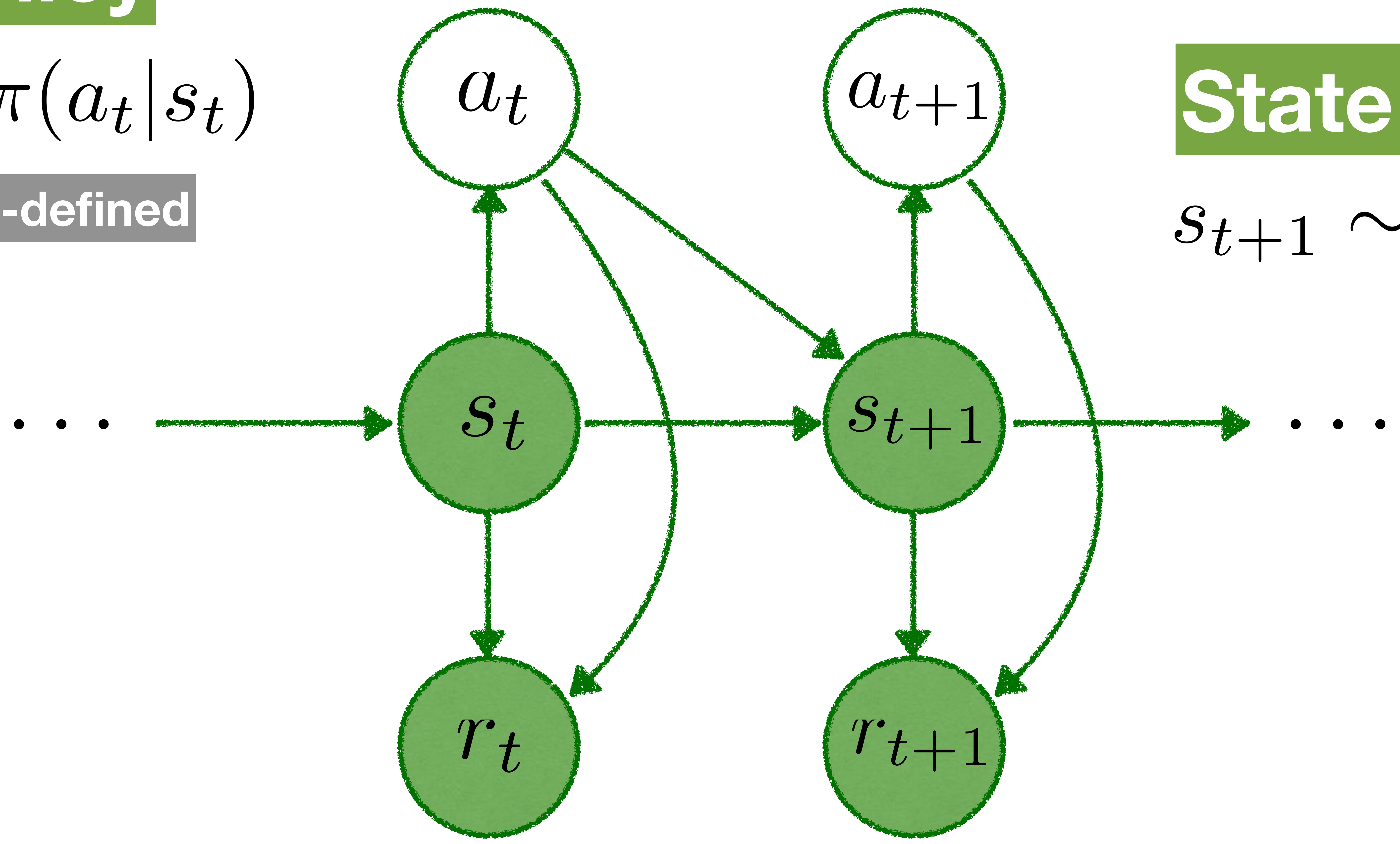
Model-based learning

[10] Ho & Efron, 2017

Policy

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

Agent-defined



State transitions

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

Known

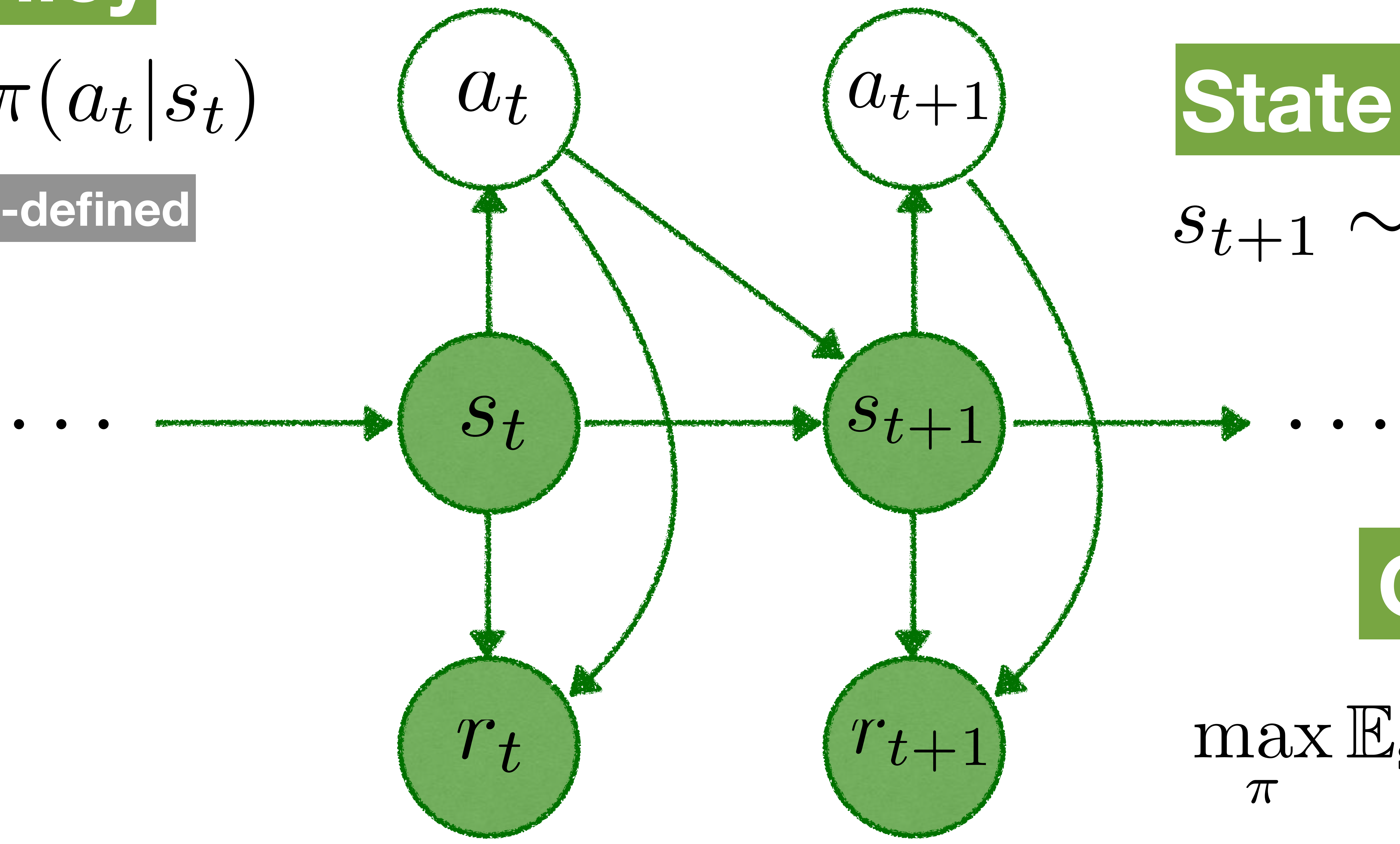
Model-based learning

[10] Ho & Efron, 2017

Policy

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

Agent-defined



State transitions

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

Known

Goal

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

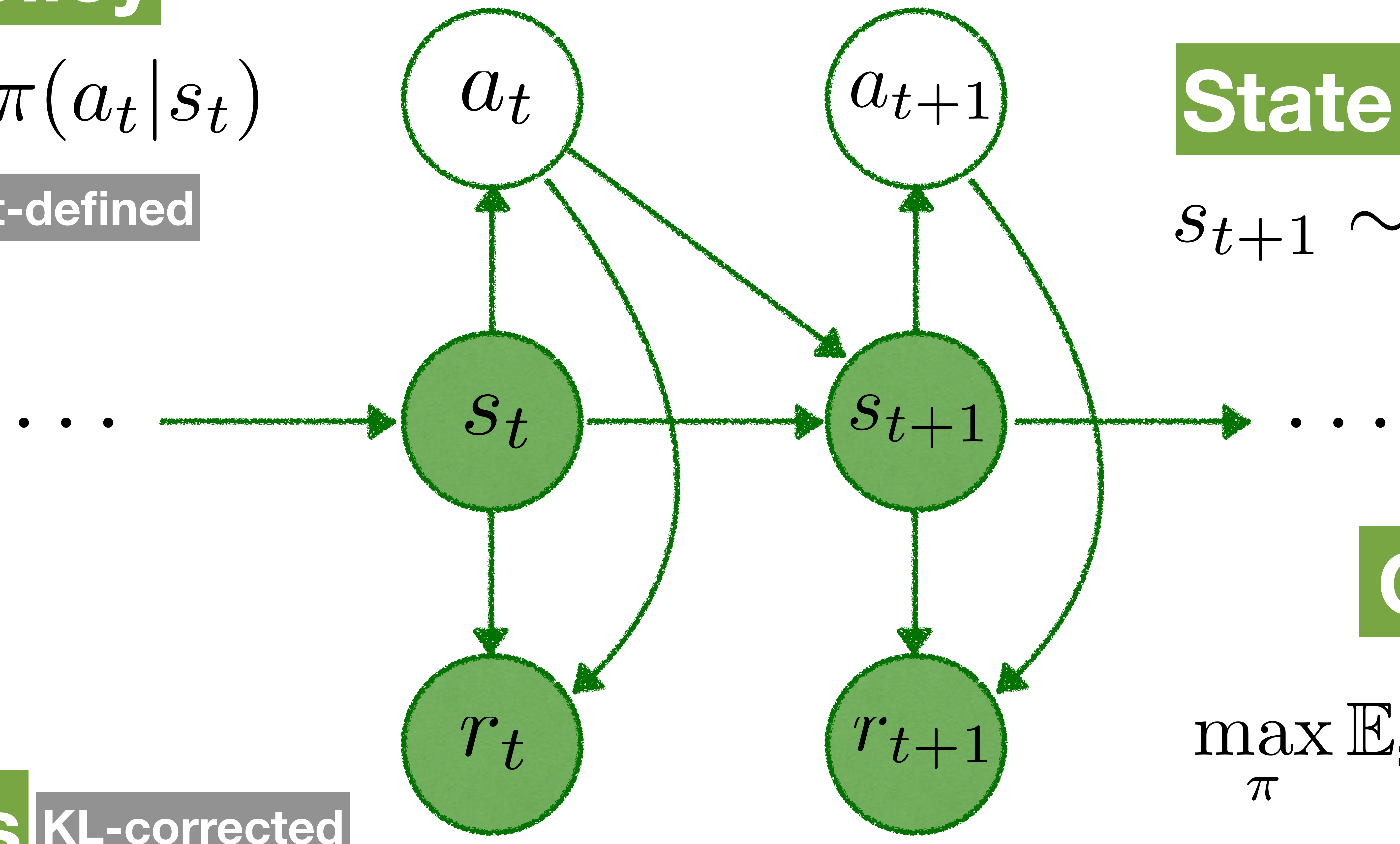
Model-based learning

[10] Ho & Efron, 2017

Policy

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

Agent-defined



State transitions

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

Known

Goal

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

Rewards KL-corrected

$$r_t = r(s_t, a_t, s_{t+1}) - \mathrm{KL}(q(\cdot | s_t, a_t) || p(\cdot | s_t, a_t))$$

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).