Machine Learning

A. G. Schwing & M. Telgarsky

University of Illinois at Urbana-Champaign, 2018

L25: Policy Gradient

Goals of this lecture

Goals of this lecture

- More about Reinforcement Learning Techniques

Goals of this lecture

- More about Reinforcement Learning Techniques
- Getting to know Policy Gradient

Goals of this lecture

- More about Reinforcement Learning Techniques
- Getting to know Policy Gradient
- Understanding its relation to other methods

Recap so far: Known MDP

Recap so far: Known MDP

- To compute V^*, Q^*, π^*: use **policy/value iteration or exhaustive**

Recap so far: Known MDP

- To compute V^*, Q^*, π^*: use **policy/value iteration or exhaustive**
- To evaluate fixed policy π: use policy evaluation

Recap so far: Known MDP

- To compute V^*, Q^*, π^*: use **policy/value iteration or exhaustive**
- To evaluate fixed policy π: use policy evaluation

Unknown MDP

Recap so far: Known MDP

- To compute V^*, Q^*, π^*: use **policy/value iteration or exhaustive**
- To evaluate fixed policy π: use policy evaluation

Unknown MDP

- Estimate transition probabilities using experience replay

Recap so far: Known MDP

- To compute V^*, Q^*, π^*: use **policy/value iteration or exhaustive**
- To evaluate fixed policy π: use policy evaluation

Unknown MDP

- Estimate transition probabilities using experience replay
- Q-learning

What else:

What else:

Directly optimize parametric policy $\pi_\theta(a|s)$

What else:

Directly optimize parametric policy $\pi_\theta(a|s)$

Why:

What else:

Directly optimize parametric policy $\pi_\theta(a|s)$

Why:

- π may be simpler than Q or V

What else:

Directly optimize parametric policy $\pi_\theta(a|s)$

Why:

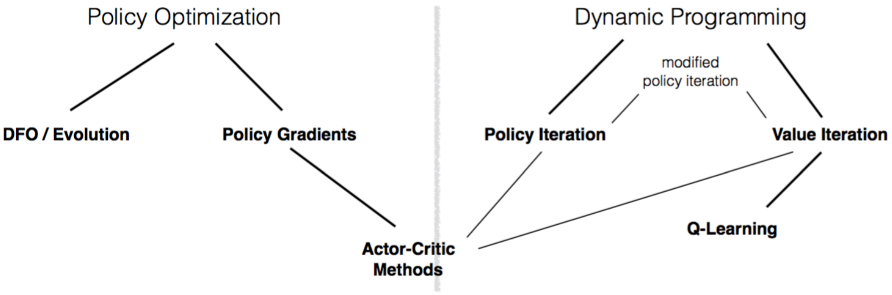
- π may be simpler than Q or V
- V doesn't prescribe actions: dynamics model + Bellman back-up needed

What else:

Directly optimize parametric policy $\pi_\theta(a|s)$

Why:

- π may be simpler than Q or V
- V doesn't prescribe actions: dynamics model + Bellman back-up needed
- Q requires efficient maximization: issue in continuous/high-dimensional action spaces



John Schulman & Pieter Abbeel – OpenAI + UC Berkeley

Variant: Likelihood ratio policy gradient

Variant: Likelihood ratio policy gradient

- Rollout, state-action sequence: $\tau = (s_0, a_0, s_1, a_1, \ldots)$

Variant: Likelihood ratio policy gradient

- Rollout, state-action sequence: $\tau = (s_0, a_0, s_1, a_1, \ldots)$
- Expected reward: $R(\tau) = \sum_t R(s_t, a_t)$

$$U(\theta) = \mathbb{E}\left[\sum_t R(s_t, a_t); \pi_\theta\right] = \sum_\tau P(\tau; \theta) R(\tau)$$

Variant: Likelihood ratio policy gradient

- Rollout, state-action sequence: $\tau = (s_0, a_0, s_1, a_1, \ldots)$
- Expected reward: $R(\tau) = \sum_t R(s_t, a_t)$

$$U(\theta) = \mathbb{E}\left[\sum_t R(s_t, a_t); \pi_\theta\right] = \sum_\tau P(\tau; \theta) R(\tau)$$

Goal:

Variant: Likelihood ratio policy gradient

- Rollout, state-action sequence: $\tau = (s_0, a_0, s_1, a_1, \ldots)$
- Expected reward: $R(\tau) = \sum_t R(s_t, a_t)$

$$U(\theta) = \mathbb{E}\left[\sum_t R(s_t, a_t); \pi_\theta\right] = \sum_\tau P(\tau;\theta)R(\tau)$$

Goal:

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

$$\max_{\theta} U(\theta) = \max_{\theta} \sum_{\tau} P(\tau; \theta) R(\tau)$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Related work:

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta) R(\tau)$$

Related work:

- Aleksandrov, Sysoyev & Shemaneva; 1968
- Rubinstein; 1969
- Glynn; 1986
- Williams; 1992 –> Reinforce
- Baxter & Bartlett; 2001

$$\max_{\theta} U(\theta) = \max_{\theta} \sum_{\tau} P(\tau;\theta) R(\tau)$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \\
&= \\
&= \\
&= \\
&=
\end{aligned}$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \\
&= \\
&= \\
&=
\end{aligned}$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \\
&= \\
&=
\end{aligned}$$

Gradient descent: $$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \\
&=
\end{aligned}$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \\
&=
\end{aligned}$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau P(\tau;\theta)\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \\
&= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)
\end{aligned}$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \\
&= \sum_\tau P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau)
\end{aligned}$$

Approx. with empirical estimate for sample paths under policy π_θ:

Gradient descent:

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau) \\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau) \\
&= \sum_\tau P(\tau;\theta)\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \\
&= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)
\end{aligned}$$

Approx. with empirical estimate for sample paths under policy π_θ:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)R(\tau^{(i)})$$

$$\max_\theta U(\theta) = \max_\theta \sum_\tau P(\tau;\theta)R(\tau)$$

Gradient descent:

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_\tau P(\tau;\theta)R(\tau)\\
&= \sum_\tau \nabla_\theta P(\tau;\theta)R(\tau)\\
&= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta)R(\tau)\\
&= \sum_\tau P(\tau;\theta)\frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau)\\
&= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta)R(\tau)
\end{aligned}$$

Approx. with empirical estimate for sample paths under policy π_θ:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^m \nabla_\theta \log P(\tau^{(i)};\theta)R(\tau^{(i)})$$ approx. impossible w/o trick

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)})$$

Important property:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)})$$

Important property:

Valid even if R is discontinuous

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)})$$

Important property:

Valid even if R is discontinuous

Intuition:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

Important property:

Valid even if R is discontinuous

Intuition:

- Increase probability of paths τ with positive R

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) R(\tau^{(i)})$$

Important property:

Valid even if R is discontinuous

Intuition:

- Increase probability of paths τ with positive R
- Decrease probability of paths τ with negative R

No need for dynamics model:

No need for dynamics model:

$$\nabla_\theta \log P(\tau;\theta) =$$

$$=$$

$$=$$

$$=$$

No need for dynamics model:

$$\begin{aligned}
\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log \left[\prod_t \underbrace{P(s_{t+1}|s_t, a_t)}_{\text{dynamics model}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}}\right] \\
&= \\
&= \\
&=
\end{aligned}$$

No need for dynamics model:

$$
\begin{aligned}
\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log \left[\prod_t \underbrace{P(s_{t+1}|s_t,a_t)}_{\text{dynamics model}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}}\right] \\
&= \nabla_\theta \left[\sum_t \log P(s_{t+1}|s_t,a_t) + \sum_t \log \pi_\theta(a_t|s_t)\right] \\
&= \\
&=
\end{aligned}
$$

No need for dynamics model:

$$
\begin{aligned}
\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log \left[\prod_t \underbrace{P(s_{t+1}|s_t,a_t)}_{\text{dynamics model}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}}\right] \\
&= \nabla_\theta \left[\sum_t \log P(s_{t+1}|s_t,a_t) + \sum_t \log \pi_\theta(a_t|s_t)\right] \\
&= \nabla_\theta \sum_t \log \pi_\theta(a_t|s_t) \\
&=
\end{aligned}
$$

No need for dynamics model:

$$\begin{aligned}
\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log \left[\prod_t \underbrace{P(s_{t+1}|s_t,a_t)}_{\text{dynamics model}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}}\right] \\
&= \nabla_\theta \left[\sum_t \log P(s_{t+1}|s_t,a_t) + \sum_t \log \pi_\theta(a_t|s_t)\right] \\
&= \nabla_\theta \sum_t \log \pi_\theta(a_t|s_t) \\
&= \sum_t \underbrace{\nabla_\theta \log \pi_\theta(a_t|s_t)}_{\text{no dynamics model required}}
\end{aligned}$$

No need for dynamics model:

$$\begin{aligned}
\nabla_\theta \log P(\tau;\theta) &= \nabla_\theta \log \left[\prod_t \underbrace{P(s_{t+1}|s_t, a_t)}_{\text{dynamics model}} \underbrace{\pi_\theta(a_t|s_t)}_{\text{policy}} \right] \\
&= \nabla_\theta \left[\sum_t \log P(s_{t+1}|s_t, a_t) + \sum_t \log \pi_\theta(a_t|s_t) \right] \\
&= \nabla_\theta \sum_t \log \pi_\theta(a_t|s_t) \\
&= \sum_t \underbrace{\nabla_\theta \log \pi_\theta(a_t|s_t)}_{\text{no dynamics model required}}
\end{aligned}$$

Consequently:

$$\nabla_\theta \log P(\tau;\theta) = \sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)$$

From

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

From

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

to

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)}) \right) R(\tau^{(i)})$$

From

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

to

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right) R(\tau^{(i)})$$

Practically important:

From

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

to

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right) R(\tau^{(i)})$$

Practically important:

- Baseline

From

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

to

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right) R(\tau^{(i)})$$

Practically important:

- Baseline
- Temporal structure

Baseline:

Baseline: issue when $R(\tau^{(i)}) > 0$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) \left(R(\tau^{(i)}) - b \right)$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \\ &= \\ &= \end{aligned}$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta) b \\ &= \\ &= \end{aligned}$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)}; \theta) \left(R(\tau^{(i)}) - b \right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned} \mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) b \\ &= \sum_\tau \nabla_\theta P(\tau;\theta) b \\ &= \end{aligned}$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta) b \\
&= \sum_\tau \nabla_\theta P(\tau;\theta) b \\
&= \nabla_\theta \left(\sum_\tau P(\tau;\theta)\right) b
\end{aligned}$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta) b \\
&= \sum_\tau \nabla_\theta P(\tau;\theta) b \\
&= \nabla_\theta \left(\sum_\tau P(\tau;\theta)\right) b = 0
\end{aligned}$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta) b \\ &= \sum_\tau \nabla_\theta P(\tau;\theta) b \\ &= \nabla_\theta \left(\sum_\tau P(\tau;\theta)\right) b = 0\end{aligned}$$

Choices of b: e.g.,

$$b = \mathbb{E}\left[R(\tau)\right] = \frac{1}{m}\sum_{i=1}^{m} R(\tau^{(i)})$$

Baseline: issue when $R(\tau^{(i)}) > 0$ can be fixed with

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta)\left(R(\tau^{(i)}) - b\right)$$

Why is subtraction of baseline b okay?

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta P(\tau;\theta) b\right] &= \sum_\tau P(\tau;\theta)\nabla_\theta \log P(\tau;\theta) b \\
&= \sum_\tau \nabla_\theta P(\tau;\theta) b \\
&= \nabla_\theta \left(\sum_\tau P(\tau;\theta)\right) b = 0
\end{aligned}$$

Choices of b: e.g., (others are available, e.g., Greensmith et al. (2004))

$$b = \mathbb{E}\left[R(\tau)\right] = \frac{1}{m}\sum_{i=1}^{m} R(\tau^{(i)})$$

Temporal structure:

Temporal structure:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum_t R(s_t^{(i)}, a_t^{(i)}) - b\right)$$

Temporal structure:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum_t R(s_t^{(i)}, a_t^{(i)}) - b\right)$$

Future actions don't depend on past rewards: lower variance via

Temporal structure:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum_t R(s_t^{(i)}, a_t^{(i)}) - b\right)$$

Future actions don't depend on past rewards: lower variance via

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\left(\sum_{\hat{t}\geq t} R(s_{\hat{t}}^{(i)}, a_{\hat{t}}^{(i)}) - b(s_{\hat{t}}^{(i)})\right)$$

Temporal structure:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum_t R(s_t^{(i)}, a_t^{(i)}) - b\right)$$

Future actions don't depend on past rewards: lower variance via

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\left(\sum_{\hat{t}\geq t} R(s_{\hat{t}}^{(i)}, a_{\hat{t}}^{(i)}) - b(s_{\hat{t}}^{(i)})\right)$$

Good choices for *b*:

Temporal structure:

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\right)\left(\sum_t R(s_t^{(i)}, a_t^{(i)}) - b\right)$$

Future actions don't depend on past rewards: lower variance via

$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m}\sum_{i=1}^{m}\sum_t \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})\left(\sum_{\color{red}{\hat{t}\geq t}} R(s_{\hat{t}}^{(i)}, a_{\hat{t}}^{(i)}) - b(s_{\hat{t}}^{(i)})\right)$$

Good choices for b:

$$b(s_t) = \mathbb{E}\left[r_t + r_{t+1} + \ldots\right]$$

Algorithm: Reinforce aka vanilla Policy Gradient

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b
- For iteration = 1, 2, $\ldots$

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b
- For iteration = 1, 2, . . .
  - Collect a set of trajectories $\tau^{(i)}$ by executing policy π_θ

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b
- For iteration = 1, 2, ...
  - Collect a set of trajectories $\tau^{(i)}$ by executing policy π_θ
  - Compute reward and bias

$$R_t^{(i)} = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}}$$

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b
- For iteration = 1, 2, ...
  - Collect a set of trajectories $\tau^{(i)}$ by executing policy π_θ
  - Compute reward and bias

$$R_t^{(i)} = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}}$$

  - Re-fit the baseline b

Algorithm: Reinforce aka vanilla Policy Gradient

- Initial θ, b
- For iteration = 1, 2, ...
  - Collect a set of trajectories $\tau^{(i)}$ by executing policy π_θ
  - Compute reward and bias

$$R_t^{(i)} = \sum_{\hat{t} \geq t} \gamma^{\hat{t}-t} r_{\hat{t}}$$

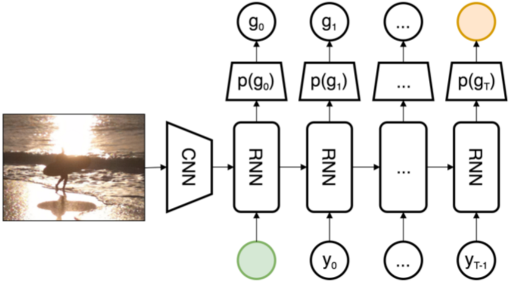
  - Re-fit the baseline b
  - Update the policy using the policy gradient estimate $\hat{g}$

Applications:

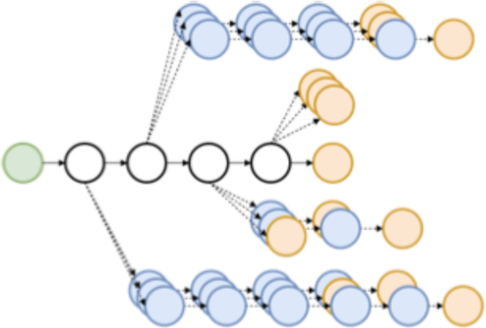
- S. Liu, Z. Zhu, N. Ye, S. Guadarrama, K. Murphy
- Improved Image Captioning via Policy Gradient optimization of SPIDEr
- 2016

Image Captioning

Image Captioning



Sampling a caption:



$$\nabla_\theta V_\theta(s_0) \approx \sum_{t=1}^{T} \sum_{g_t} [\pi_\theta(g_t|s_t) \nabla_\theta \log \pi_\theta(g_t|s_t) \times (Q_\theta(s_t, g_t) - B_\phi(s_t))] \quad (7)$$

$$L_\phi = \sum_t E_{s_t} E_{g_t} (Q_\theta(s_t, g_t) - B_\phi(s_t))^2 \quad (8)$$

Algorithm 1: PG training algorithm

Input: $\mathcal{D} = \{(\mathbf{x}^n, \mathbf{y}^n) : n = 1 : N\}$;
Train $\pi_\theta(g_{1:T}|x)$ using MLE on $\mathcal{D}$;
Train B_ϕ using MC estimates of Q_θ on a small subset of $\mathcal{D}$;
for each epoch **do**
    for example (x^n, y^n) **do**
        Generate sequence $g_{1:T} \sim \pi_\theta(\cdot|x^n)$;
        for $t = 1 : T$ **do**
            Compute $Q(g_{1:t-1}, g_t)$ for g_t with K Monte Carlo rollouts, using (6);
            Compute estimated baseline $B_\phi(g_{1:t-1})$;
        Compute $\mathcal{G}_\theta = \nabla_\theta V_\theta(s_0)$ using (7);
        Compute $\mathcal{G}_\phi = \nabla_\phi L_\phi$;
        SGD update of θ using $\mathcal{G}_\theta$;
        SGD update of ϕ using $\mathcal{G}_\phi$;

Quiz:

Quiz:

- Why Policy Gradient?

Quiz:

- Why Policy Gradient?
- Techniques to improve vanilla Policy Gradient?

Important topics of this lecture

Important topics of this lecture

- Getting a feeling for reinforcement learning

Important topics of this lecture

- Getting a feeling for reinforcement learning
- Understanding how to use reinforcement learning

Important topics of this lecture

- Getting a feeling for reinforcement learning
- Understanding how to use reinforcement learning

Thank you!