

535514: Reinforcement Learning

Lecture 5 – Regularized MDPs and Policy Optimization

Ping-Chun Hsieh

March 7, 2024

Announcements

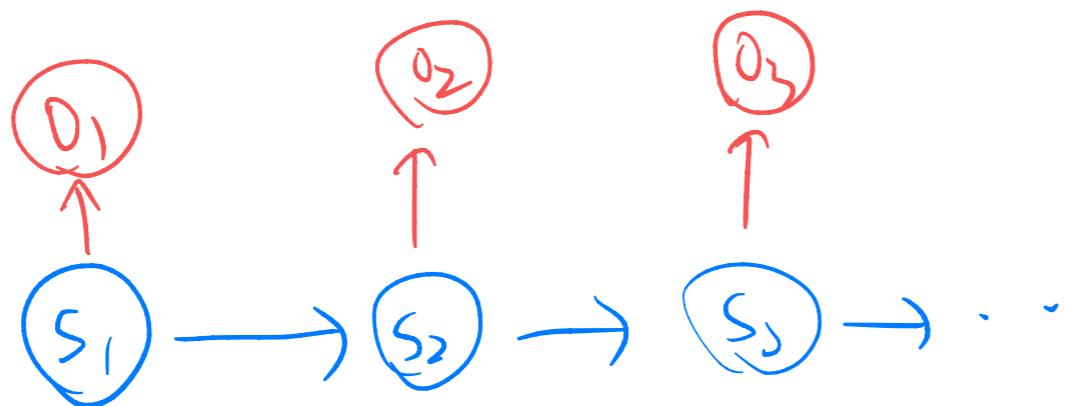
- HW1 has been posted on E3
 - Due date: 3/20, 9pm Taiwan time
- Team Matching Requests
 - <https://forms.gle/LX9TScJF2ycHiqG3A>

A Question from Richard Sutton at ICML 2019:



In RL, we typically need 3 things:
states, *actions*, and *rewards*...

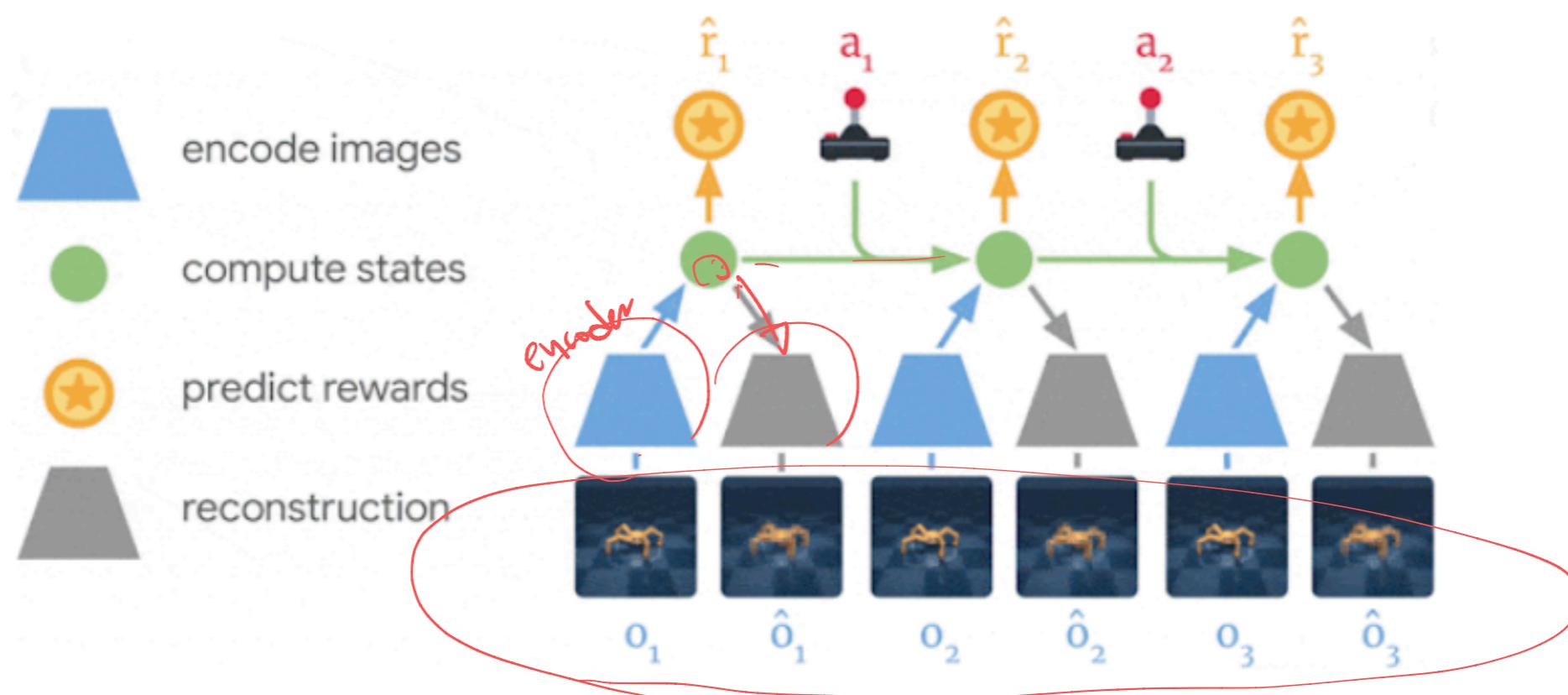
If we want to remove one thing and still learn a good policy, which one might not be necessary?



$$\underline{a_i} = f(\underline{s_i})$$

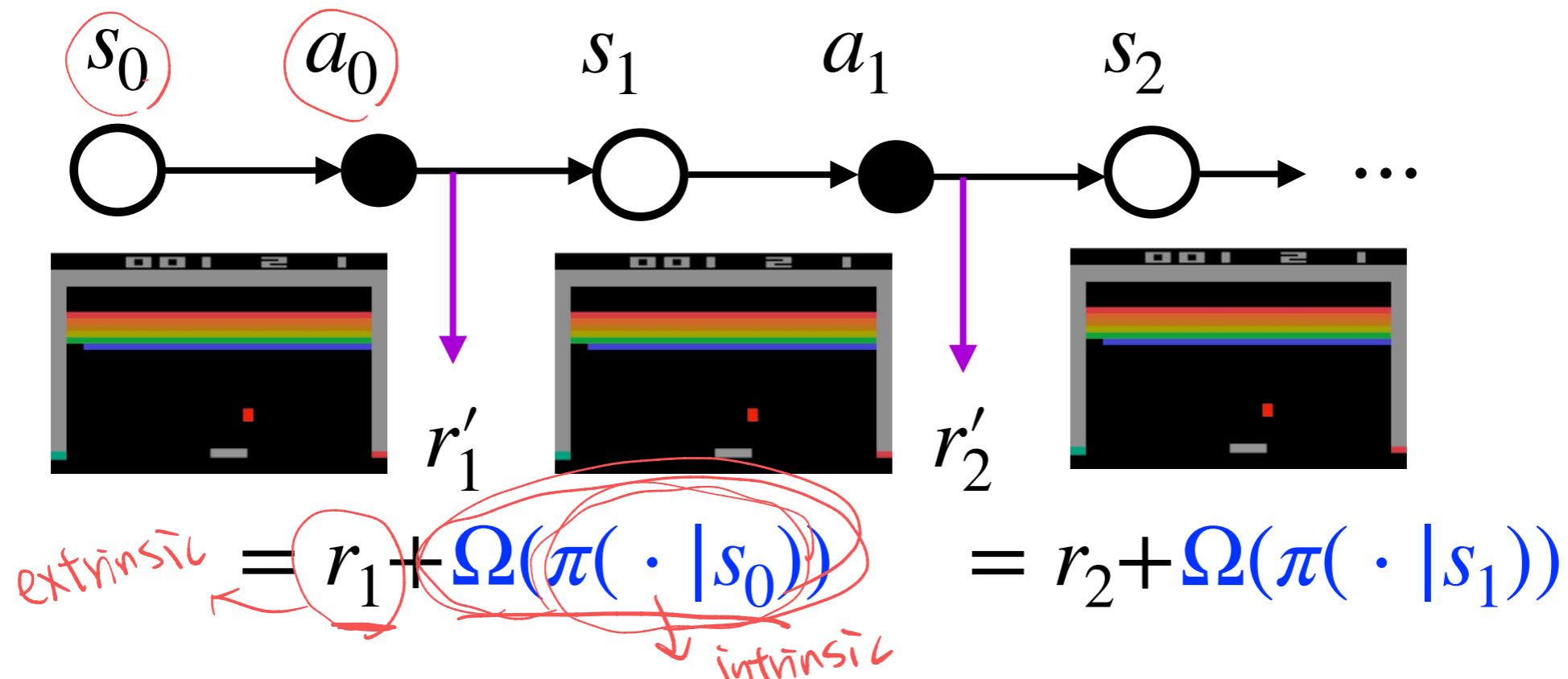
Model-Based RL via a Latent Dynamics Model

Example: Dreamer



Review: Regularized MDPs

Regularized MDP = Standard MDP + Regularized rewards!



- ▶ A regularized MDP can be specified by $(\mathcal{S}, \mathcal{A}, P, R, \Omega, \gamma)$
- ▶ $\Omega(\cdot)$: A function that maps an *action distribution* to a *real number*

Value Functions of Regularized MDPs

$\pi(a|s)$

$\pi(\cdot|s)$ f
 $f(\cdot)$

	Unregularized MDP	Regularized MDP
Return	$G_t := r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots$	$G_t := \underline{r_{t+1}} + \gamma (\underline{r_{t+2} + \Omega(\pi(\cdot s_{t+1}))}) + \gamma^2 (\underline{r_{t+3} + \Omega(\pi(\cdot s_{t+2}))}) + \dots$
Value function	$V^\pi(s) := \mathbb{E}[G_t s_t = s; \pi]$	$V_\Omega^\pi(s) := \mathbb{E}[G_t s_t = s; \pi] + \underline{\Omega(\pi(\cdot s))}$
Q function	$\underline{Q^\pi(s, a) := \mathbb{E}[G_t s_t = s, a_t = a; \pi]}$	$Q_\Omega^\pi(s, a) := \mathbb{E}[G_t s_t = s, a_t = a; \pi]$
Bellman expectation equations	$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a s) Q^\pi(s, a)$ $\underline{Q^\pi(s, a) = R_{s,a} + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a V^\pi(s')}$	$V_\Omega^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a s) \cdot \underline{Q_\Omega^\pi(s, a)} + \underline{\Omega(\pi(\cdot s))}$ $\underline{Q_\Omega^\pi(s, a) = R_{s,a} + \gamma \cdot \sum_{s' \in \mathcal{S}} P_{ss'}^a V_\Omega^\pi(s')}$

$$|A|=3$$
$$\pi(a^{(1)}|s) = 0.2$$

$$\pi(a^{(2)}|s) = 0.5$$

$$\pi(a^{(3)}|s) = 0.3$$

$$\pi(\cdot|s) \equiv [\pi(a^{(1)}|s), \pi(a^{(2)}|s), \pi(a^{(3)}|s)]$$

$$H(\pi(\cdot|s))$$

$$\Omega(\pi(\cdot|s))$$

Next Question: *How to find V_Ω^π ?*

Regularized Bellman Expectation Operator

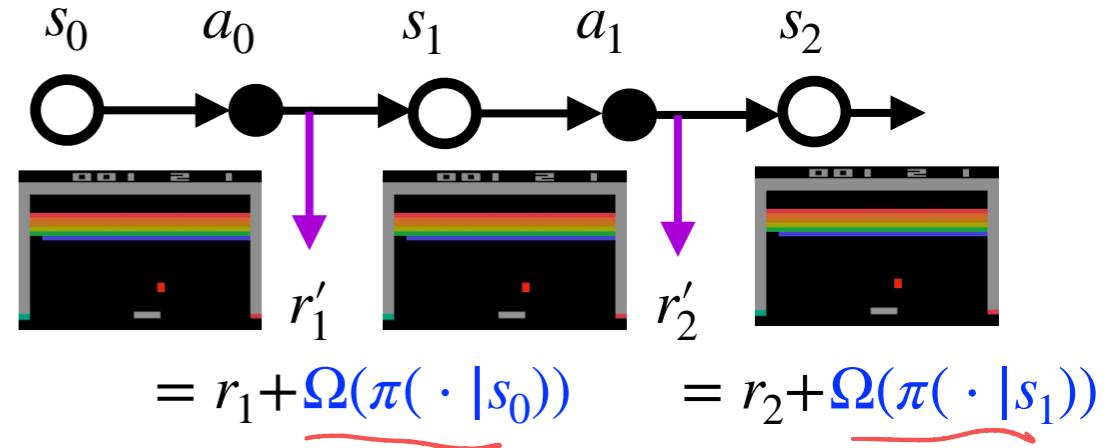
- ▶ Regularized Bellman Expectation **Operator**

$$\begin{aligned}[T_{\Omega}^{\pi} V](s) &:= [T^{\pi} V](s) + \underbrace{\Omega(\pi(\cdot | s))}_{\text{regularization term}} \\ &= \underbrace{R_s^{\pi} + \Omega(\pi(\cdot | s))}_{\text{regularized immediate reward}} + \gamma P_{ss'}^{\pi} V\end{aligned}$$

-
- ▶ **Question:** Is T_{Ω}^{π} a contraction? *Yes! (in L_{∞} -norm)*
 - ▶ Therefore, under T_{Ω}^{π} , there is a unique fixed point, which is the **regularized value function** V_{Ω}^{π}
 - ▶ To find V_{Ω}^{π} , we can use the (regularized) IPE method

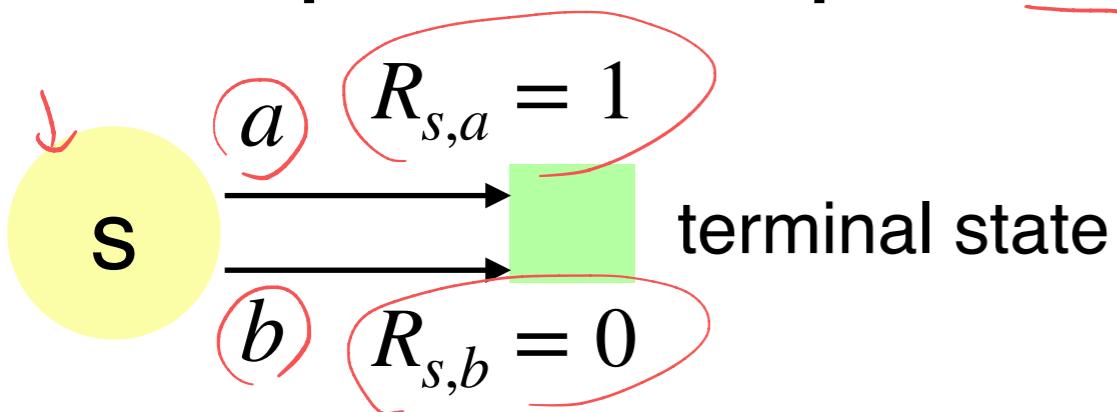
Next Question: *How to define “optimality” for regularized MDPs?*

Optimal Value Functions of *Regularized MDPs*



	Unregularized MDP	Regularized MDP
Optimal value functions	$V^*(s) := \max_{\pi \in \Pi} V^\pi(s)$	$V_\Omega^*(s) := \max_{\pi \in \Pi} V_\Omega^\pi(s)$
Optimal Q functions	$Q^*(s, a) := \max_{\pi \in \Pi} Q^\pi(s, a)$	$\underline{Q}_\Omega^*(s, a) := \max_{\pi \in \Pi} \underline{Q}_\Omega^\pi(s, a)$

Example: A Simple One-State Regularized MDP



Suppose we choose entropy regularizer
 $\Omega(\pi(\cdot | s)) \equiv H(\pi(\cdot | s))$

► **Question 1:** $\underline{V}_{\Omega}^*(s) = ?$ $\underline{Q}_{\Omega}^*(s, a) = ?$ $\underline{Q}_{\Omega}^*(s, b) = ?$

► **Question 2:** Which policy π can achieve $V_{\Omega}^*(s)$, i.e., $V_{\Omega}^{\pi}(s) = V_{\Omega}^*(s)$?

$$\underline{Q}_{\Omega}^*(s, a) = 1, \quad \underline{Q}_{\Omega}^*(s, b) = 0.$$

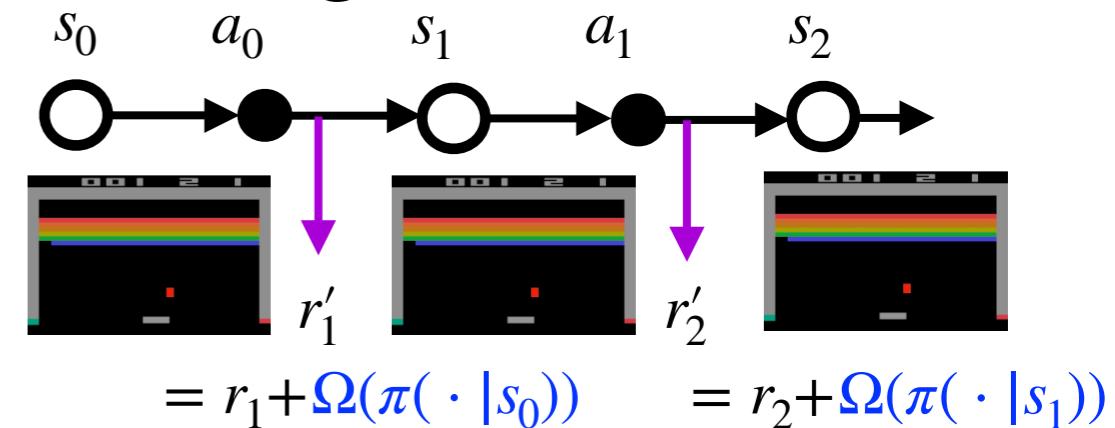
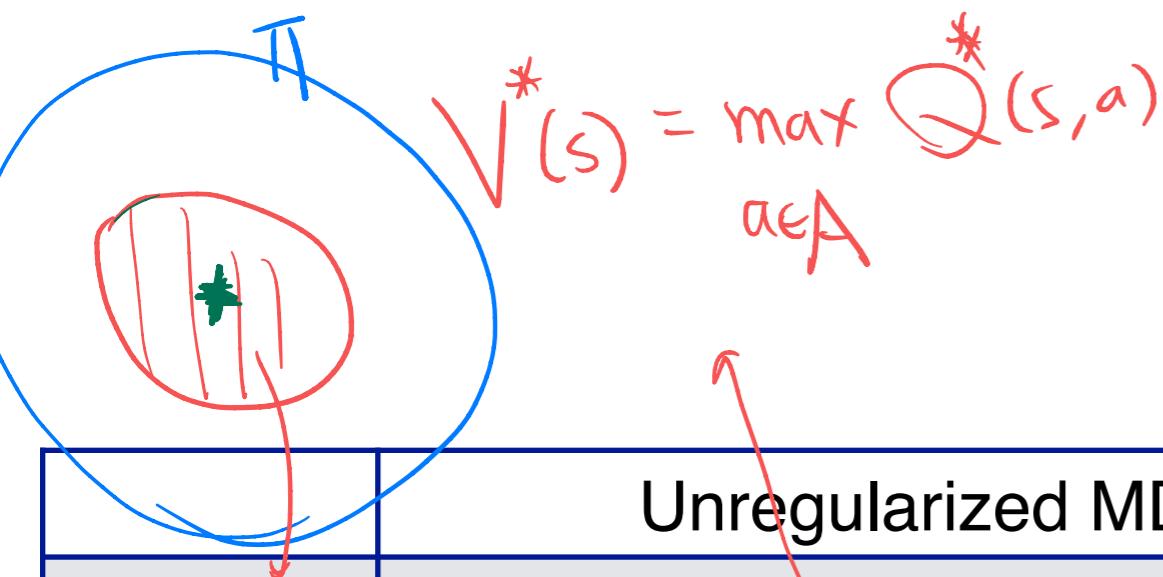
$$V_{\Omega}^{\pi}(s) = \left(\frac{R_{s,a} \cdot \pi_a + R_{s,b} \cdot \pi_b}{1} \right) + \left(-\pi_a \ln \pi_a - (1-\pi_a) \ln (1-\pi_a) \right)$$

$$\frac{\partial V_{\Omega}^{\pi}(s)}{\partial \pi_a} = 1 + \left(-1 - \ln \pi_a + 1 + \ln (1-\pi_a) \right)$$

$$= 1 + \ln \left(\frac{1-\pi_a}{\pi_a} \right) = 0 \Rightarrow \pi_a = \frac{e}{1+e}$$

$$\boxed{\frac{\exp(R_{s,a})}{\exp(R_{s,a}) + \exp(R_{s,b})}}$$

Bellman Optimality Equations of *Regularized MDPs*



	Unregularized MDP	Regularized MDP
Bellman optimality equations	$V^*(s) = \max_{a \in \mathcal{A}} R_s^a + \gamma P_s^a V^*$ $= \max_{\pi \in \Pi} R_s^\pi + \gamma P_s^\pi V^*$	$V_\Omega^*(s) = \max_{\pi \in \Pi} R_s^\pi + \gamma P_s^\pi V_\Omega^*$
	$Q^*(s, a) = R_s^a + \gamma E_{s' \sim P(\cdot s, a)}[V^*(s')]$	$Q_\Omega^*(s, a) = R_s^a + \gamma E_{s' \sim P(\cdot s, a)}[V_\Omega^*(s')]$

- ▶ **Next Question:** How to obtain $V_\Omega^*(s)$ and $Q_\Omega^*(s, a)$?

(See HW1 problem for more details)

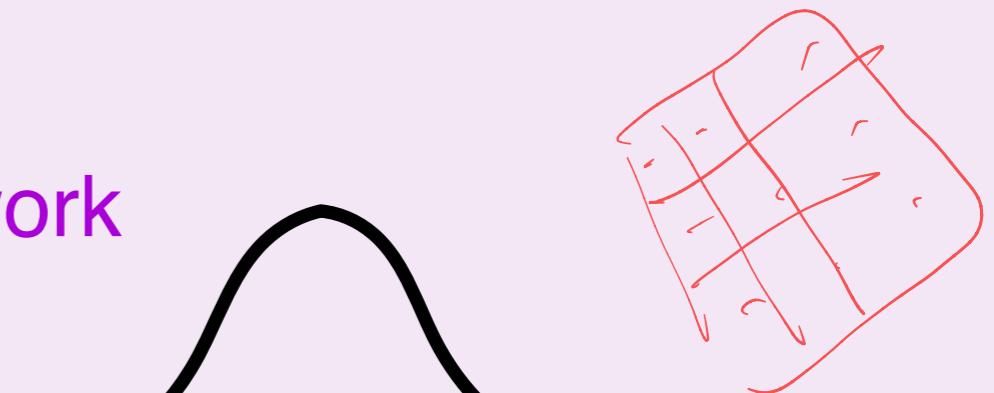
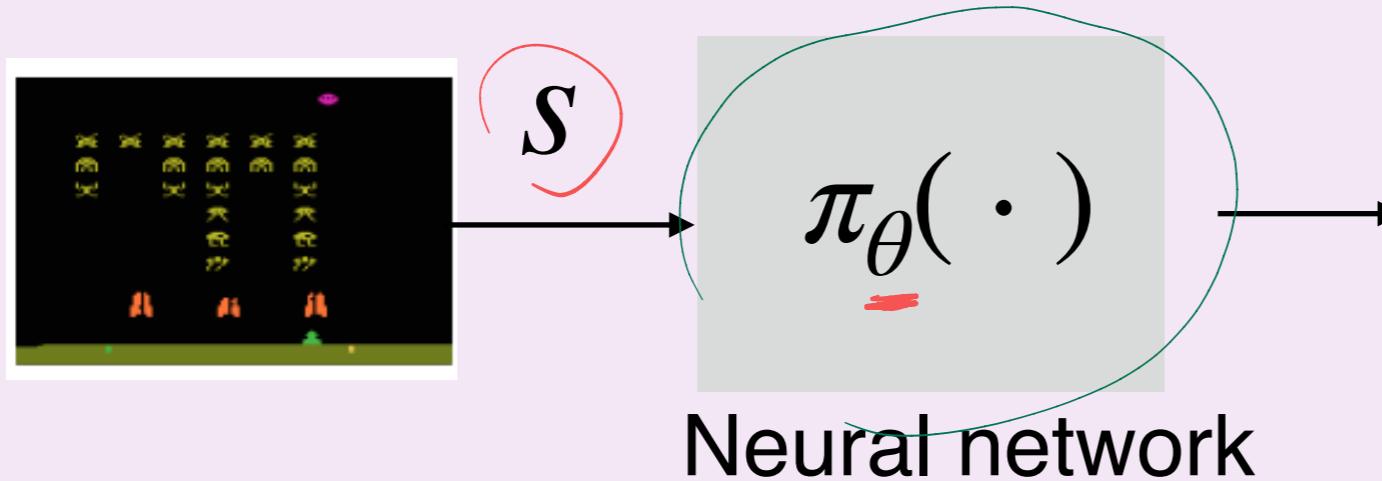
Next Topic: Policy Optimization

Policy Optimization Framework:

- Consider a parametric policy class: $\{\pi_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\}$
- View RL as an optimization problem: $\max_{\theta} V^{\pi_\theta}(\mu) := \mathbb{E}_{s_0 \sim \mu}[V^{\pi_\theta}(s_0)]$

Example:

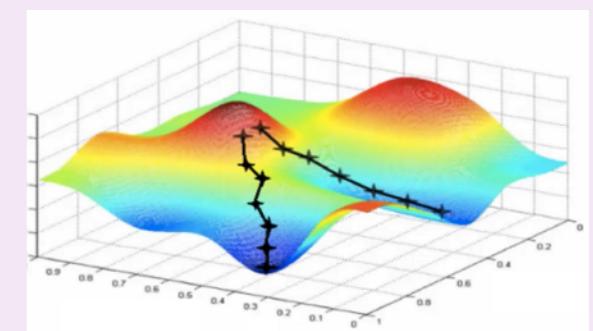
- Parametrize the policy by a neural network



A deterministic action or an action distribution

- Update π_θ iteratively, e.g., by gradient ascent (GA)

$$\theta \leftarrow \theta + \alpha \nabla_\theta V^{\pi_\theta}(\mu)$$



- Question: Why using $V^\pi(\mu)$ as the objective (instead of each individual $V^\pi(s)$)?

$\overline{\pi}^*$: $V^{\overline{\pi}^*}(s) \geq V^\pi(s)$, for all s , for all π

$$V^{\pi^*}(s) \geq V^\pi(s) \quad \text{for all } \pi$$

$$\text{for all } s$$

$$V^{\pi_0}(s) = \frac{1}{2} V^{\pi_0}(s^{(1)}) + \frac{1}{2} V^{\pi_0}(s^{(2)})$$

\geq
 π^*

$$V^{\pi_0}(s) = \text{Objective}$$

Theorem: If a policy $\tilde{\pi}$ achieves that

$V^{\tilde{\pi}}(s^{(1)}) = 10 > V^\pi(s^{(1)})$, for all π

$V^{\tilde{\pi}}(s^{(2)}) = 20 > V^\pi(s^{(2)})$, for all π

then $\hat{V}^{\tilde{\pi}}(s^{(1)}) > V^\pi(s^{(1)})$

$\hat{V}^{\tilde{\pi}}(s^{(2)}) > V^\pi(s^{(2)})$

3 Challenges in Policy Optimization

(C1) How to parametrize the policy?

(C2) How to solve the optimization problem in a model-free manner?

1. Policy gradient (PG)
2. Model-free prediction (aka model-free policy evaluation)

(C3) Any theoretical guarantee of PG?

Typical Examples of Parametric Policies

Discrete action space

- ▶ Softmax policies:
- ▶ Linear softmax policies:
- ▶ Neural softmax policies:

$$\pi_{\theta}(a | s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}$$
$$\pi_{\theta}(a | s) = \frac{\exp(\theta^T \phi_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta^T \phi_{s,a'})}$$
$$\pi_{\theta}(a | s) = \frac{\exp(f_{\theta}(s, a))}{\sum_{a' \in \mathcal{A}} \exp(f_{\theta}(s, a'))}$$

Continuous action space

- ▶ Gaussian policies:

$$a \sim \mathcal{N}(f_{\theta}(s), \sigma^2)$$

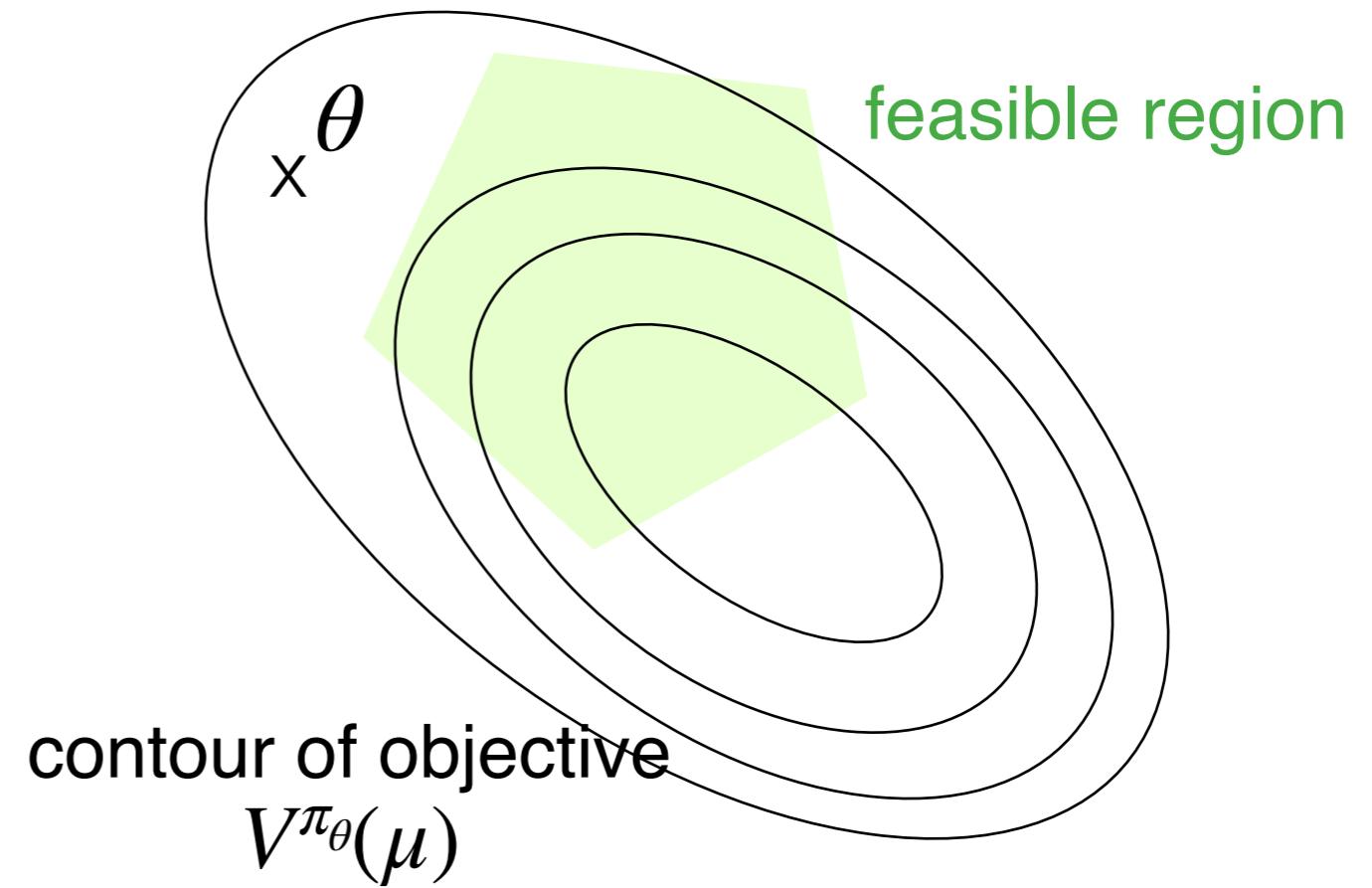
Questions: Are these parametric policies general enough?

How About Direct Tabular Parametrization?

- ▶ **Question:** Can we do direct parametrization, i.e., $\pi_\theta(a | s) = \theta_{s,a}$?
- ▶ Possible, but now we face **constrained** policy optimization

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\mu)$$

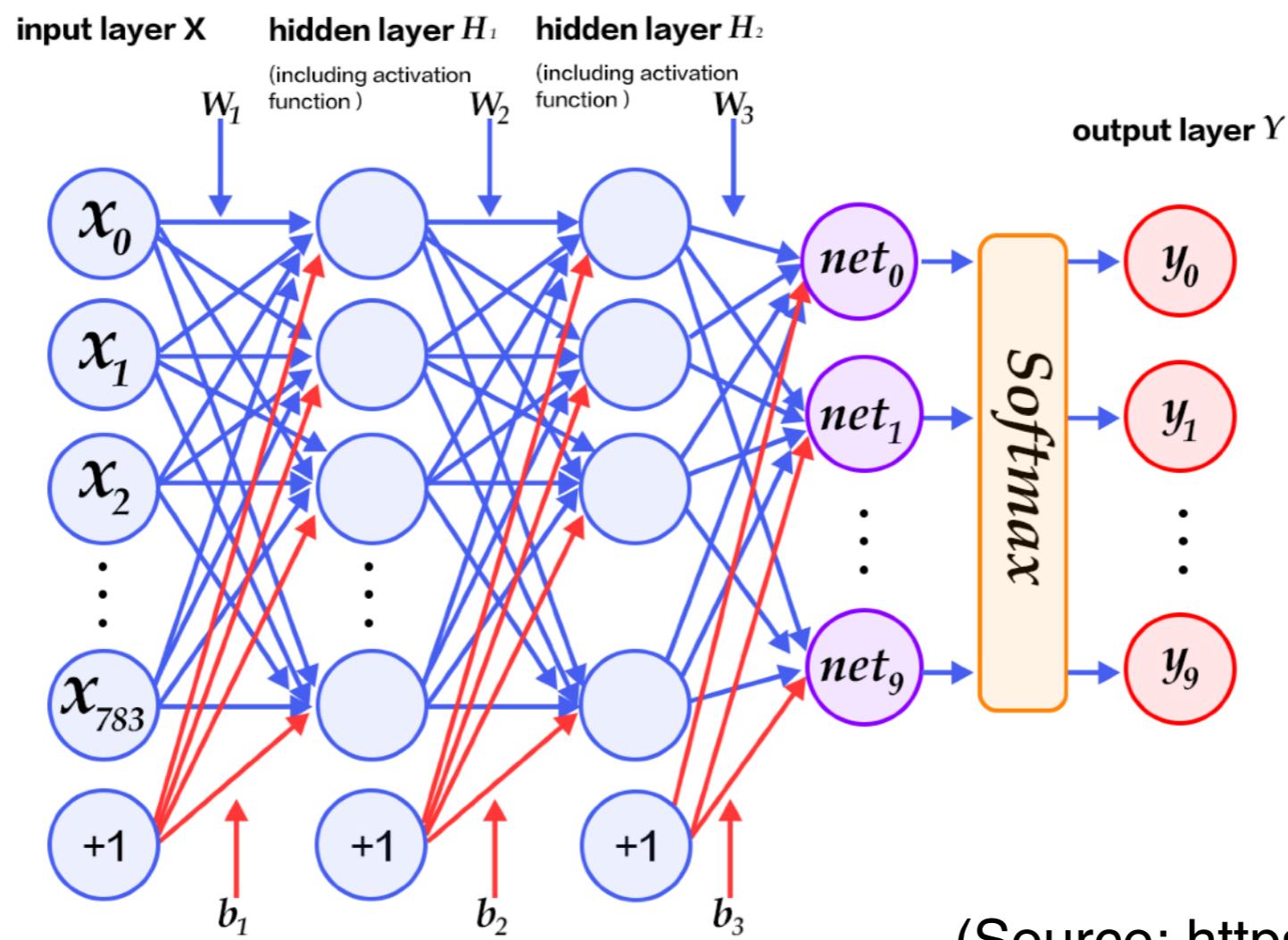
subject to $\sum_a \theta_{s,a} = 1, \forall s$



- ▶ In this case, **projection** is needed to ensure a valid probability distribution (often computationally heavy)

Neural Softmax Policies

- ▶ Neural softmax policies: $\pi_{\theta}(a | s) = \frac{\exp(f_{\theta}(s, a))}{\sum_{a' \in \mathcal{A}} \exp(f_{\theta}(s, a'))}$
- ▶ θ denotes all the weights and biases of a neural net



(Source: https://files_ifi_uzh_ch_ddis/)

Policy Gradient (PG)

Historical Accounts on PG

Policy Gradient Methods for Reinforcement Learning with Function Approximation

Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour
AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932

Abstract

Function approximation is essential to reinforcement learning, but the standard approach of approximating a value function and determining a policy from it has so far proven theoretically intractable. In this paper we explore an alternative approach in which the policy is explicitly represented by its own function approximator, independent of the value function, and is updated according to the gradient of expected reward with respect to the policy parameters. Williams's REINFORCE method and actor-critic methods are examples of this approach. Our main new result is to show that the gradient can be written in a form suitable for estimation from experience aided by an approximate action-value or advantage function. Using this result, we prove for the first time that a version of policy iteration with arbitrary differentiable function approximation is convergent to a locally optimal policy.



Richard Sutton
(U of Alberta & DeepMind) **Satinder Singh**
(UMich & DeepMind)



A seminal paper on PG in NIPS 1999

Review: Some Useful Notations

- ▶ **Question:** To apply GA, what do we need?
- ▶ Policy gradient! $\nabla_{\theta} V^{\pi_{\theta}}(\mu)$ (μ : distribution of initial state)

Some useful notations

- ▶ 1. Sample return $G(\tau)$ along a trajectory $\tau = (s_0, a_0, r_1, \dots)$

$$G(\tau) := \sum_{t=0}^{\infty} \gamma^t \cdot r_{t+1}(\tau)$$

- ▶ 2. Value function written in $R(\tau)$

$$V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} [G(\tau)] = \sum_{\tau} G(\tau) P_{\mu}^{\pi_{\theta}}(\tau)$$

Review: Some Useful Notations (Cont.)

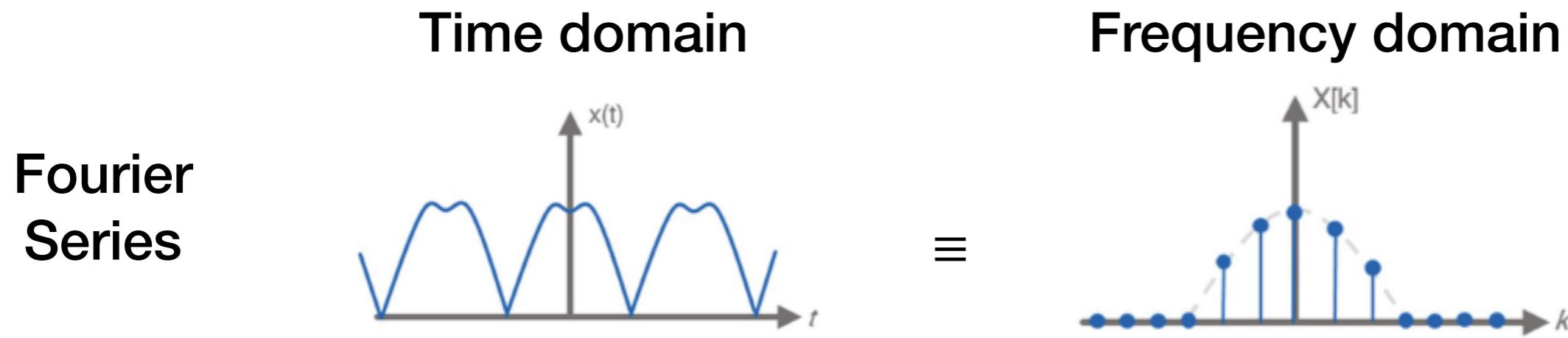
- ▶ 3. Discounted state visitation distribution (aka “occupancy measure”)

$$d_{s_0}^\pi(s) := (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid s_0, \pi)$$

$$d_\mu^\pi(s) := \mathbb{E}_{s_0 \sim \mu} [d_{s_0}^\pi(s)]$$

- ▶ **Question:** What's the purpose of $(1 - \gamma)$ in the above?

State Visitation Distribution: An Analogy



RL
Objective
Function

Time domain

$$\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 \sim \mu; \pi_{\theta} \right] = \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}, a \sim \pi(\cdot | s)} [R(s, a)]$$

Distribution domain

(We will use this form several times in the sequel)

The Policy Gradient: Inherent Difficulty

- ▶ **Issue:** Computing $\nabla_{\theta} V^{\pi_{\theta}}(\mu)$ might seem highly non-trivial...
 - ▶ **Recall:** $V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} [G(\tau)] = \sum_{\tau} G(\tau) P_{\mu}^{\pi_{\theta}}(\tau)$
-
- ▶ $V^{\pi_{\theta}}(\mu)$ is affected by θ in two ways:
 1. θ determines $\pi_{\theta}(\cdot | s)$
 2. The actions taken by π_{θ} affect the state visitation accordingly

Expressions of The Policy Gradient

- ▶ **Expressions of Policy Gradient (aka Policy Gradient Theorem):**

(P1) Total reward: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[G(\tau) \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$

(P2) REINFORCE: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$

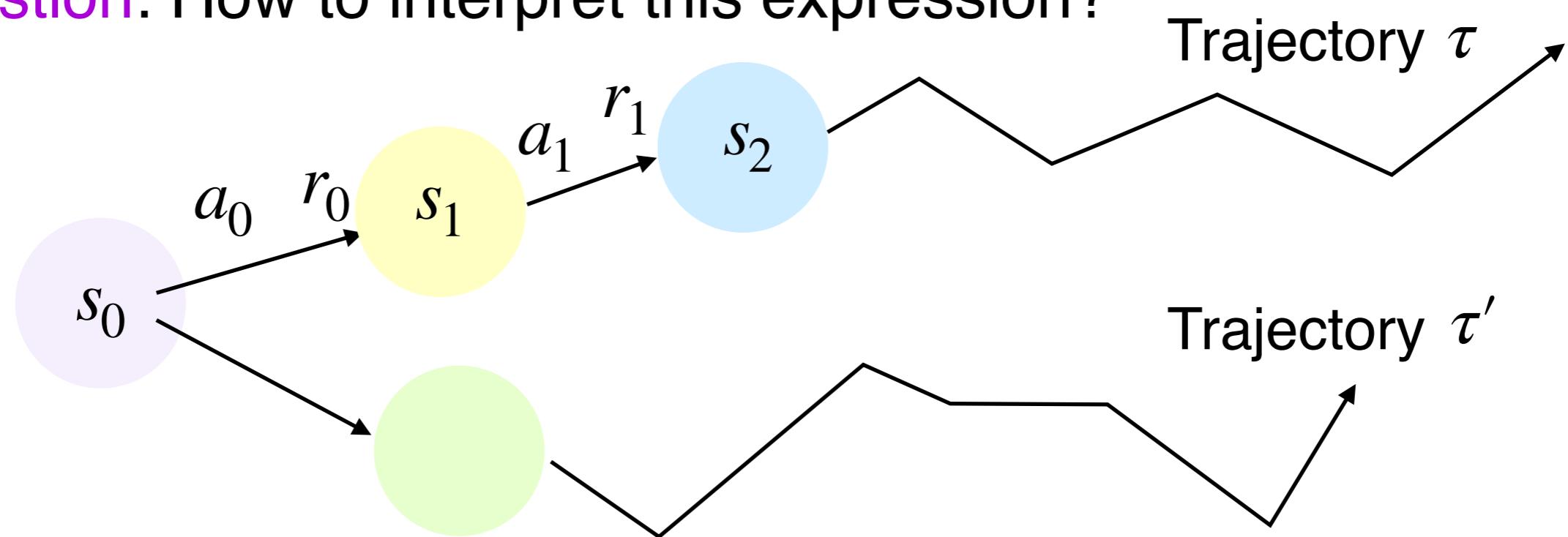
(P3) Q-value and discounted state visitation:

$$\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot | s)} \left[Q^{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a | s) \right]$$

- ▶ **Question:** Why do we care about various expressions?
- ▶ **Each expression gives us an RL algorithm!**

(P1) Total Reward Policy Gradient

- Want to show: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[G(\tau) \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$
- Question:** How to interpret this expression?



1. $G(\tau) > 0$:

2. $G(\tau) < 0$:

- Question:** Can you find out any potential issue?

(P1) Total Reward Policy Gradient (Cont.)

- Want to show: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[G(\tau) \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$
- Recall: $V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} [G(\tau)] = \sum_{\tau} G(\tau) P_{\mu}^{\pi_{\theta}}(\tau)$

$$\begin{aligned}\nabla_{\theta} V^{\pi_{\theta}}(\mu) &= \sum_{\tau} G(\tau) \nabla_{\theta} P_{\mu}^{\pi_{\theta}}(\tau) \\ &= \sum_{\tau} G(\tau) \left(P_{\mu}^{\pi_{\theta}}(\tau) \cdot \nabla_{\theta} \log P_{\mu}^{\pi_{\theta}}(\tau) \right) \\ &= \sum_{\tau} G(\tau) \left(P_{\mu}^{\pi_{\theta}}(\tau) \cdot \nabla_{\theta} \log(\mu(s_0) \pi_{\theta}(a_0 | s_0) P(s_1 | s_0, a_0) \pi_{\theta}(a_1 | s_1) \dots) \right) \\ &= \sum_{\tau} G(\tau) \left(P_{\mu}^{\pi_{\theta}}(\tau) \cdot \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right) \\ &= \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[G(\tau) \sum_{t=0}^{\infty} \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]\end{aligned}$$

(Not a rigorous proof)

Technical Subtlety in the Proof of (P1):

- ▶ For τ of infinite length, $\nabla_\theta P_\mu^{\pi_\theta}(\tau)$ needs to be handled more carefully
- ▶ **Idea:** Consider “truncated trajectories” and take the limit

1. Without loss of generality, assume $R(s, a) \in [0, 1]$

Define $\tau^{(k)}$ to be the truncated version of τ up to step k

Define $V_k^{\pi_\theta}(\mu) = \sum_{\tau^{(k)}} G(\tau^{(k)}) P_\mu^{\pi_\theta}(\tau^{(k)})$

2. Then, $\lim_{k \rightarrow \infty} V_k^{\pi_\theta}(\mu) = V^{\pi_\theta}(\mu)$ (by Monotone convergence theorem)

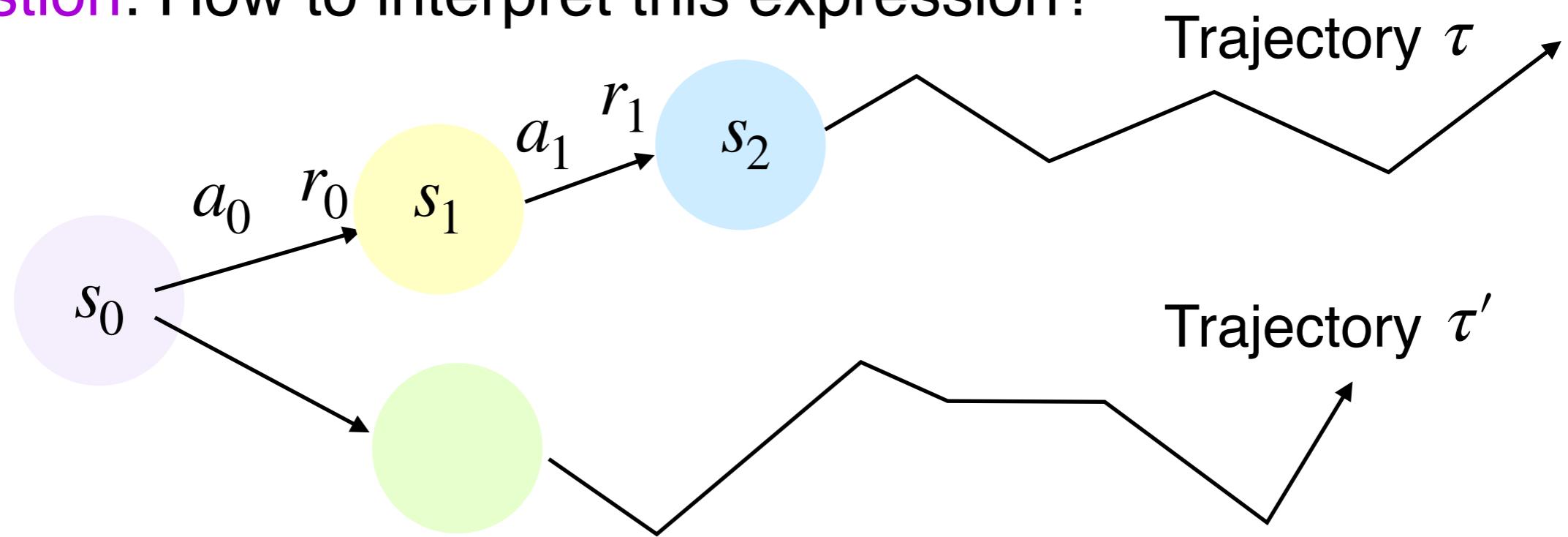
3. Next, we need to verify that $\lim_{k \rightarrow \infty} \nabla_\theta V_k^{\pi_\theta}(\mu) = \nabla_\theta V^{\pi_\theta}(\mu)$ by

showing “uniform convergence” of $\nabla_\theta V_k^{\pi_\theta}(\mu)$ [Rudin 1976, Thm 7.17]

(Note that it is not always valid to interchange ∇ and \lim)

(P2) REINFORCE Policy Gradient

- Want to show: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$
- Question:** How to interpret this expression?



1. $Q(s, a) > 0$:

2. $Q(s, a) < 0$:

- Question:** Can you find out any potential issue?

(P2) REINFORCE Policy Gradient

- Want to show: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$
- Idea: prove by induction

Recall: $V^{\pi_{\theta}}(s_0) = \sum_{a_0 \in \mathcal{A}} \pi_{\theta}(a_0 | s_0) Q^{\pi_{\theta}}(s_0, a_0)$

$$\begin{aligned} \nabla_{\theta} V^{\pi_{\theta}}(s_0) &= \sum_{a_0} \nabla_{\theta} (\pi_{\theta}(a_0 | s_0) Q^{\pi_{\theta}}(s_0, a_0)) \\ &= \underbrace{\sum_{a_0} \nabla_{\theta} (\pi_{\theta}(a_0 | s_0)) Q^{\pi_{\theta}}(s_0, a_0)}_{(1)} + \underbrace{\sum_{a_0} \pi_{\theta}(a_0 | s_0) \nabla (Q^{\pi_{\theta}}(s_0, a_0))}_{(2)} \end{aligned}$$

$$(1) = \sum_{a_0} \pi_{\theta}(a_0 | s_0) \cdot \nabla_{\theta} (\log \pi_{\theta}(a_0 | s_0)) \cdot Q^{\pi_{\theta}}(s_0, a_0)$$

$$= \mathbb{E}_{\tau \sim P_{s_0}^{\pi_{\theta}}} \left[\nabla_{\theta} (\log \pi_{\theta}(a_0 | s_0)) \cdot Q^{\pi_{\theta}}(s_0, a_0) \right]$$

(P2) REINFORCE Policy Gradient (Cont.)

$$\begin{aligned}(2) &= \sum_{a_0} \pi_\theta(a_0 | s_0) \nabla (Q^{\pi_\theta}(s_0, a_0)) \\&= \sum_{a_0} \pi_\theta(a_0 | s_0) \nabla_\theta \left(r(s_0, a_0) + \gamma \sum_{s_1} P(s_1 | s_0, a_0) \cdot V^{\pi_\theta}(s_1) \right) \\&= \sum_{a_0} \pi_\theta(a_0 | s_0) \left(\gamma \sum_{s_1} P(s_1 | s_0, a_0) \cdot \nabla_\theta V^{\pi_\theta}(s_1) \right) \\&= \mathbb{E}_{\tau \sim P_{s_0}^{\pi_\theta}} [\gamma \nabla_\theta V^{\pi_\theta}(s_1)]\end{aligned}$$

By combining (1) and (2):

$$\nabla_\theta V^{\pi_\theta}(s_0) = \mathbb{E}_{\tau \sim P_{s_0}^{\pi_\theta}} [\nabla_\theta (\log \pi_\theta(a_0 | s_0)) \cdot Q^{\pi_\theta}(s_0, a_0)] + \mathbb{E}_{\tau \sim P_{s_0}^{\pi_\theta}} [\gamma \nabla_\theta V^{\pi_\theta}(s_1)]$$

This is a recursion!

(P2) REINFORCE Policy Gradient (Cont.)

Given the recursion:

$$\nabla_{\theta} V^{\pi_{\theta}}(s_0) = \mathbb{E}_{\tau \sim P_{s_0}^{\pi_{\theta}}} \left[\nabla_{\theta} (\log \pi_{\theta}(a_0 | s_0)) \cdot Q^{\pi_{\theta}}(s_0, a_0) \right] + \mathbb{E}_{\tau \sim P_{s_0}^{\pi_{\theta}}} \left[\gamma \nabla_{\theta} V^{\pi_{\theta}}(s_1) \right]$$

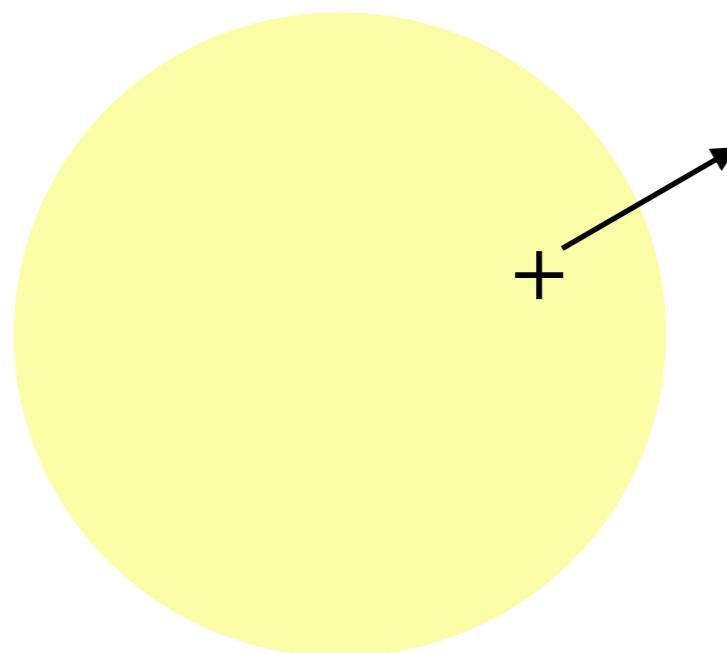
- We can verify that:

$$\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_{\theta}}(s_t, a_t) \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right]$$

(P3) Q-Value and Discounted State Visitation

- ▶ Want: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot | s)} \left[Q^{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a | s) \right]$
- ▶ **Question:** How to interpret this expression?

Sample space of (s, a)



1. $Q(s, a) > 0$:
2. $Q(s, a) < 0$:

- ▶ **Question:** Why is this expression useful?

(P3) Q-Value and Discounted State Visitation (Cont.)

- ▶ Want: $\nabla_{\theta} V^{\pi_{\theta}}(\mu) = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}} \mathbb{E}_{a \sim \pi_{\theta}(\cdot | s)} [Q^{\pi_{\theta}}(s, a) \nabla_{\theta} \log \pi_{\theta}(a | s)]$
- ▶ Proof idea: Use the following lemma and (P2)
- ▶ **Lemma (From Trajectories to Visitation):**

For any function $f(s, a) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}_{\tau \sim P_{\mu}^{\pi_{\theta}}} \left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t) \right] = \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d_{\mu}^{\pi_{\theta}}, a \sim \pi_{\theta}(\cdot | s)} [f(s, a)]$$

- ▶ Proof: Expand RHS and recollect terms to get LHS (HW1)