



Computer Engineering Department

# Reframing Control as an Inference Problem

**Mohammad Hossein Rohban, Ph.D.**

**Hosein Hasani**

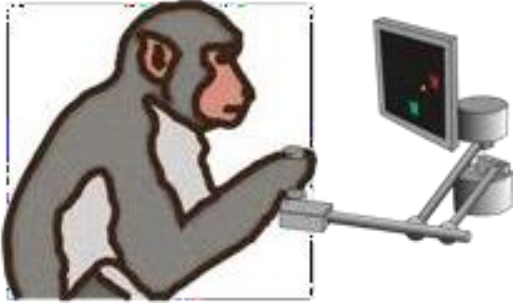
Spring 2023

Slides are adopted from CS 285, UC Berkeley.

# Lecture Outline

1. Does reinforcement learning provide a reasonable **model of human behavior**? Is there a better explanation?
  2. Can we derive optimal control, reinforcement learning, and planning as **probabilistic inference**?
  3. How does this **change** our RL algorithms?
  4. (next weeks) We'll see this is crucial for **inverse reinforcement** learning
- Goals:
    - Understand the connection between inference and control
    - Understand how specific RL algorithms can be instantiated in this framework
    - Understand why this might be a good idea

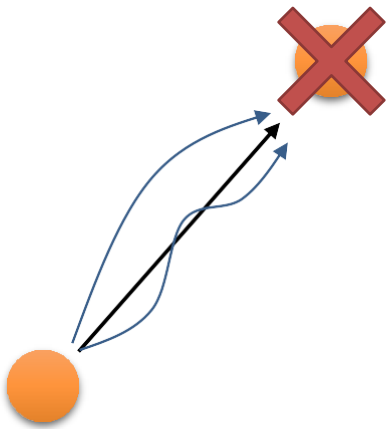
# What if the data is not optimal?



some mistakes matter more than others!

behavior is **stochastic**

but good behavior is still the most likely



# A probabilistic graphical model of decision making

~~$$\mathbf{a}_1, \dots, \mathbf{a}_T = \arg \max_{\mathbf{a}_1, \dots, \mathbf{a}_T} \sum_{t=1}^T r(\mathbf{s}_t, \mathbf{a}_t)$$~~

~~$$\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)$$~~

Can't explain near-optimal behaviours

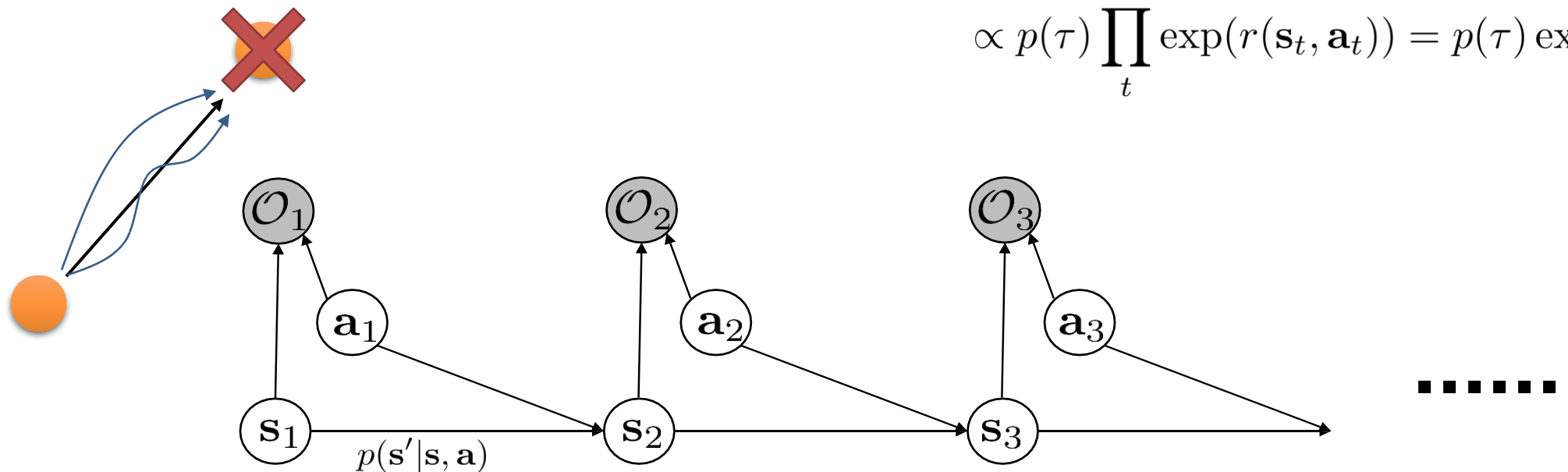
$$p(\underbrace{\mathbf{s}_{1:T}, \mathbf{a}_{1:T}}_{\tau}) = ?? \quad \text{no assumption of optimal behavior!}$$

$$p(\tau | \mathcal{O}_{1:T})$$

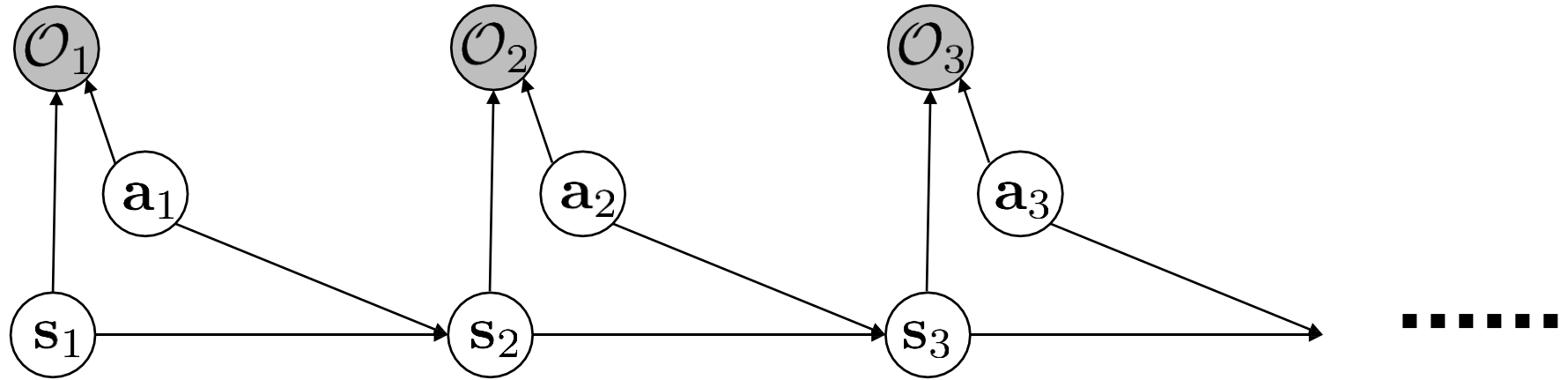
$$p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) = \exp(r(\mathbf{s}_t, \mathbf{a}_t))$$

$$p(\tau | \mathcal{O}_{1:T}) = \frac{p(\tau, \mathcal{O}_{1:T})}{p(\mathcal{O}_{1:T})}$$

$$\propto p(\tau) \prod_t \exp(r(\mathbf{s}_t, \mathbf{a}_t)) = p(\tau) \exp\left(\sum_t r(\mathbf{s}_t, \mathbf{a}_t)\right)$$



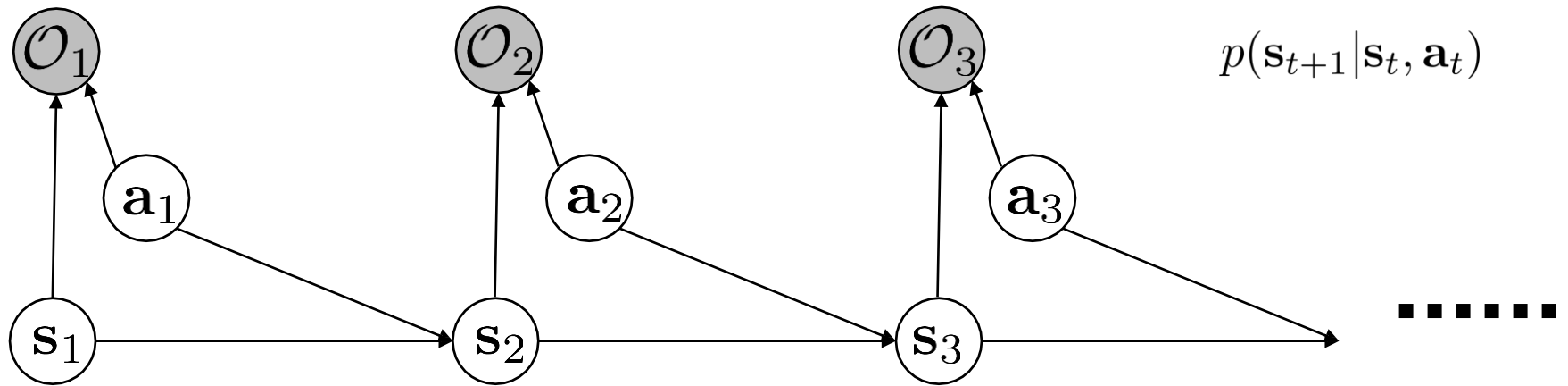
# Why is this interesting?



- Can **model suboptimal behavior** (important for inverse RL)
- Can **apply inference algorithms** to solve control and planning problems
- Provides an explanation for **why stochastic behavior might be preferred** (useful for exploration and transfer learning)

# Control as Inference

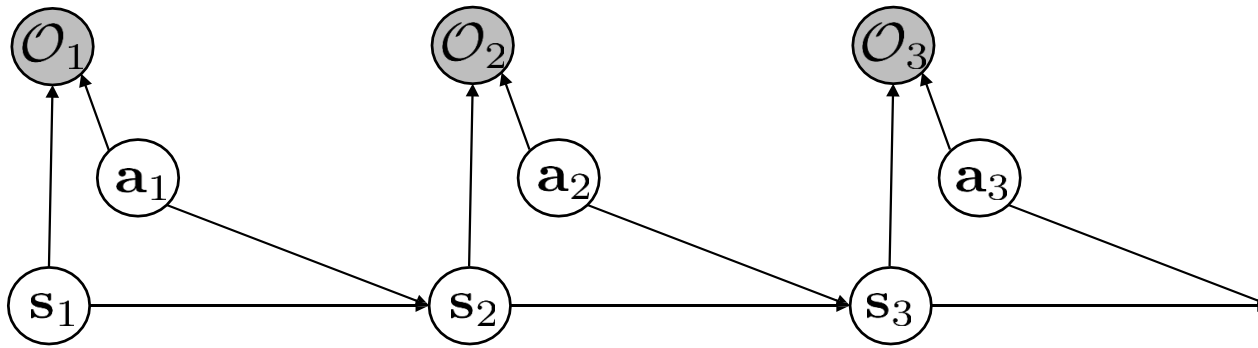
# Inference = planning



## how to do inference?

1. compute backward messages  $\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$
2. compute policy  $p(\mathbf{a}_t | \mathbf{s}_t, \mathcal{O}_{1:T})$
3. compute forward messages  $\alpha_t(\mathbf{s}_t) = p(\mathbf{s}_t | \mathcal{O}_{1:t-1})$

# Backward messages



$$\underline{p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t)} \propto \exp(r(\mathbf{s}_t, \mathbf{a}_t))$$

$$\underline{p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)}$$

■■■■■■■

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$$

$$= \int p(\mathcal{O}_{t:T}, \mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) d\mathbf{s}_{t+1}$$

for  $t = T - 1$  to 1:

$$= \int p(\mathcal{O}_{t+1:T} | \mathbf{s}_{t+1}) \underline{p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} \underline{p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t)} d\mathbf{s}_{t+1} \longrightarrow \beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} [\beta_{t+1}(\mathbf{s}_{t+1})]$$

$$\underline{p(\mathcal{O}_{t+1:T} | \mathbf{s}_{t+1})} = \int \underline{p(\mathcal{O}_{t+1:T} | \mathbf{s}_{t+1}, \mathbf{a}_{t+1})} \underline{p(\mathbf{a}_{t+1} | \mathbf{s}_{t+1})} d\mathbf{a}_{t+1} \longrightarrow \beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t | \mathbf{s}_t)} [\beta_t(\mathbf{s}_t, \mathbf{a}_t)]$$

$$\beta_t(\mathbf{s}_t)$$

$$\beta_t(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})$$

which actions are likely *a priori*  
(assume uniform for now)



# A closer look at the backward pass

for  $t = T - 1$  to 1:

$$\underline{\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} [\beta_{t+1}(\mathbf{s}_{t+1})]}$$

$$\underline{\beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t | \mathbf{s}_t)} [\beta_t(\mathbf{s}_t, \mathbf{a}_t)]}$$


$$\text{let } V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$$

$$\text{let } Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t$$

$$V_t(\mathbf{s}_t) \rightarrow \max_{\mathbf{a}_t} Q_t(\mathbf{s}_t, \mathbf{a}_t) \text{ as } Q_t(\mathbf{s}_t, \mathbf{a}_t) \text{ gets bigger!}$$

value iteration algorithm:

- 
1. set  $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')] ]$
  2. set  $V(\mathbf{s}) \leftarrow \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

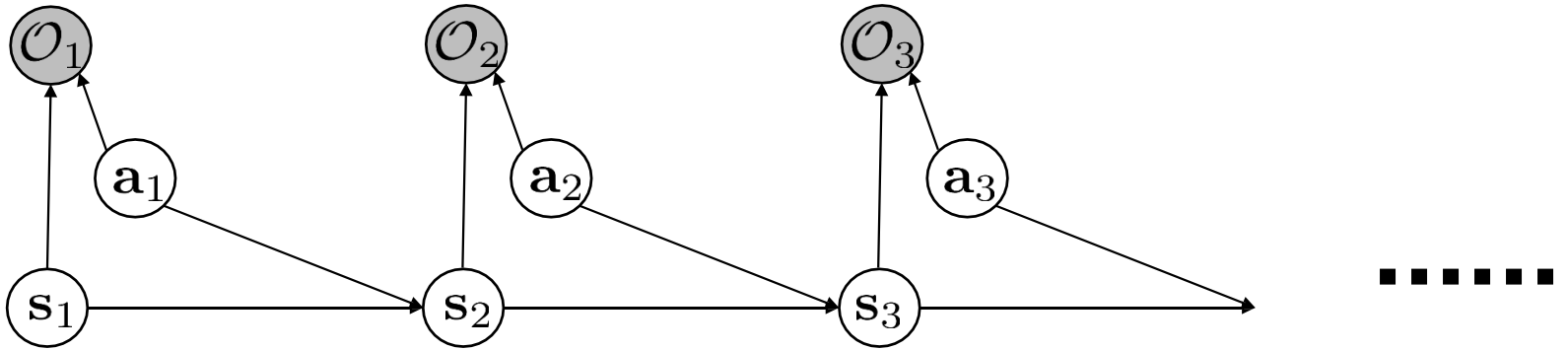
“optimistic” transition  
(not a good idea!)

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \overbrace{\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]}$$

$$\text{deterministic transition: } Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + V_{t+1}(\mathbf{s}_{t+1})$$

we'll come back to the stochastic case later!

# Backward pass summary



$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$$

probability that we can be optimal at steps  $t$  through  $T$   
given that we take action  $\mathbf{a}_t$  in state  $\mathbf{s}_t$

for  $t = T - 1$  to 1:

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} [\beta_{t+1}(\mathbf{s}_{t+1})] \quad \text{compute recursively from } t = T \text{ to } t = 1$$

$$\beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t | \mathbf{s}_t)} [\beta_t(\mathbf{s}_t, \mathbf{a}_t)]$$

$$\text{let } V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$$

$$\text{let } Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$$

log of  $\beta_t$  is “Q-function-like”

# The action prior

remember this?

$$p(\mathcal{O}_{t+1:T}|\mathbf{s}_{t+1}) = \int \underbrace{p(\mathcal{O}_{t+1:T}|\mathbf{s}_{t+1}, \mathbf{a}_{t+1})}_{\beta_t(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})} \cancel{p(\mathbf{a}_{t+1}|\mathbf{s}_{t+1})} d\mathbf{a}_{t+1}$$

(“soft max”)

$$V(\mathbf{s}_t) = \log \int \exp(Q(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t)) \mathbf{a}_t$$

what if the action prior is not uniform?

$$Q(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log E[\exp(V(\mathbf{s}_{t+1}))]$$

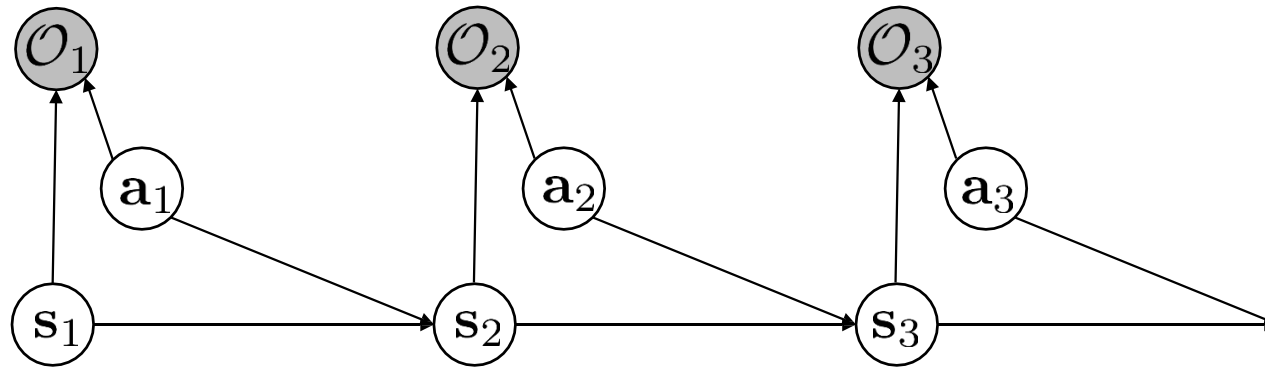
$$\text{let } \tilde{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t) + \log E[\exp(V(\mathbf{s}_{t+1}))]$$

$$V(\mathbf{s}_t) = \log \int \exp(\tilde{Q}(\mathbf{s}_t, \mathbf{a}_t)) \mathbf{a}_t \quad \Leftrightarrow \quad V(\mathbf{s}_t) = \log \int \exp(Q(\mathbf{s}_t, \mathbf{a}_t) + \log p(\mathbf{a}_t|\mathbf{s}_t)) \mathbf{a}_t$$

can **always** fold the action prior into the reward!

uniform action prior can be assumed without loss of generality

# Policy computation



$$p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) \propto \exp(r(\mathbf{s}_t, \mathbf{a}_t))$$

$$p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$$

■■■■■■■

2. compute policy  $p(\mathbf{a}_t | \mathbf{s}_t, \mathcal{O}_{1:T})$

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t, \mathbf{a}_t)$$

$$\beta_t(\mathbf{s}_t) = p(\mathcal{O}_{t:T} | \mathbf{s}_t)$$

$$p(\mathbf{a}_t | \mathbf{s}_t, \mathcal{O}_{1:T}) = \pi(\mathbf{a}_t | \mathbf{s}_t)$$

$$= p(\mathbf{a}_t | \mathbf{s}_t, \mathcal{O}_{t:T})$$

$$= \frac{p(\mathbf{a}_t, \mathbf{s}_t | \mathcal{O}_{t:T})}{p(\mathbf{s}_t | \mathcal{O}_{t:T})}$$

$$= \frac{p(\mathcal{O}_{t:T} | \mathbf{a}_t, \mathbf{s}_t) p(\mathbf{a}_t, \mathbf{s}_t) / \cancel{p(\mathcal{O}_{t:T})}}{p(\mathcal{O}_{t:T} | \mathbf{s}_t) p(\mathbf{s}_t) / \cancel{p(\mathcal{O}_{t:T})}}$$

$$= \frac{p(\mathcal{O}_{t:T} | \mathbf{a}_t, \mathbf{s}_t)}{p(\mathcal{O}_{t:T} | \mathbf{s}_t)} \frac{p(\mathbf{a}_t, \mathbf{s}_t)}{p(\mathbf{s}_t)} = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)} \cancel{p(\mathbf{a}_t | \mathbf{s}_t)}$$

$$\pi(\mathbf{a}_t | \mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)}$$

# Policy computation with value

for  $t = T - 1$  to 1:

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]$$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) \mathbf{a}_t$$

$$\pi(\mathbf{a}_t | \mathbf{s}_t) = \frac{\beta_t(\mathbf{s}_t, \mathbf{a}_t)}{\beta_t(\mathbf{s}_t)} \quad \begin{array}{l} V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t) \\ Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t) \end{array}$$

$$\pi(\mathbf{a}_t | \mathbf{s}_t) = \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t) - V_t(\mathbf{s}_t)) = \exp(A_t(\mathbf{s}_t, \mathbf{a}_t))$$

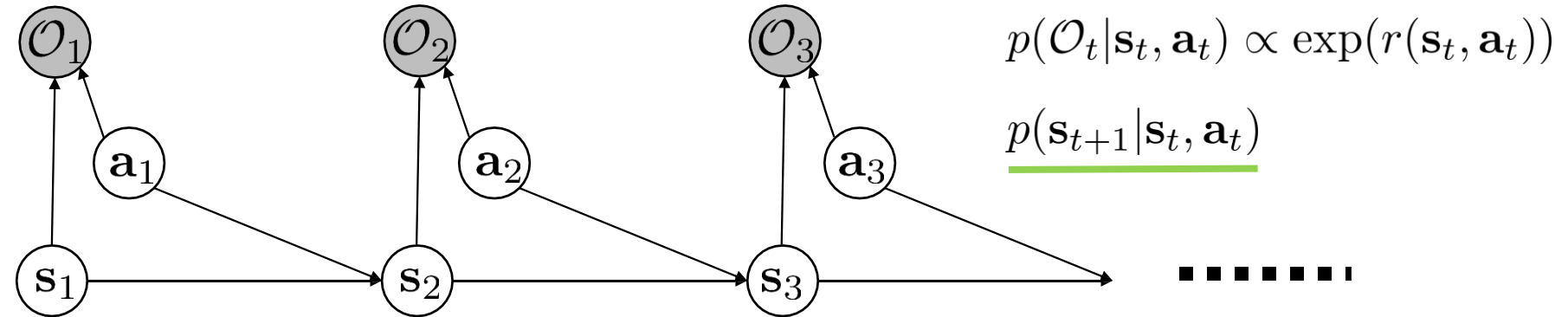
# Policy computation summary

$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t) - V_t(\mathbf{s}_t)) = \exp(A_t(\mathbf{s}_t, \mathbf{a}_t))$$

with temperature:  $\pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(\frac{1}{\alpha}Q_t(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{\alpha}V_t(\mathbf{s}_t)) = \exp(\frac{1}{\alpha}A_t(\mathbf{s}_t, \mathbf{a}_t))$

- Natural interpretation: better actions are more probable
- Random tie-breaking
- Analogous to Boltzmann exploration
- Approaches greedy policy as temperature decreases

# Forward messages



$$\alpha_1(\mathbf{s}_1) = p(\mathbf{s}_1) \text{ (usually known)}$$

$$\alpha_t(\mathbf{s}_t) = p(\mathbf{s}_t | \mathcal{O}_{1:t-1})$$

$$= \int p(\mathbf{s}_t, \mathbf{s}_{t-1}, \mathbf{a}_{t-1} | \mathcal{O}_{1:t-1}) d\mathbf{s}_{t-1} d\mathbf{a}_{t-1} = \int p(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1}, \cancel{\mathcal{O}_{1:t-1}}) p(\mathbf{a}_{t-1} | \mathbf{s}_{t-1}, \mathcal{O}_{1:t-1}) p(\mathbf{s}_{t-1} | \mathcal{O}_{1:t-1}) d\mathbf{s}_{t-1} d\mathbf{a}_{t-1}$$

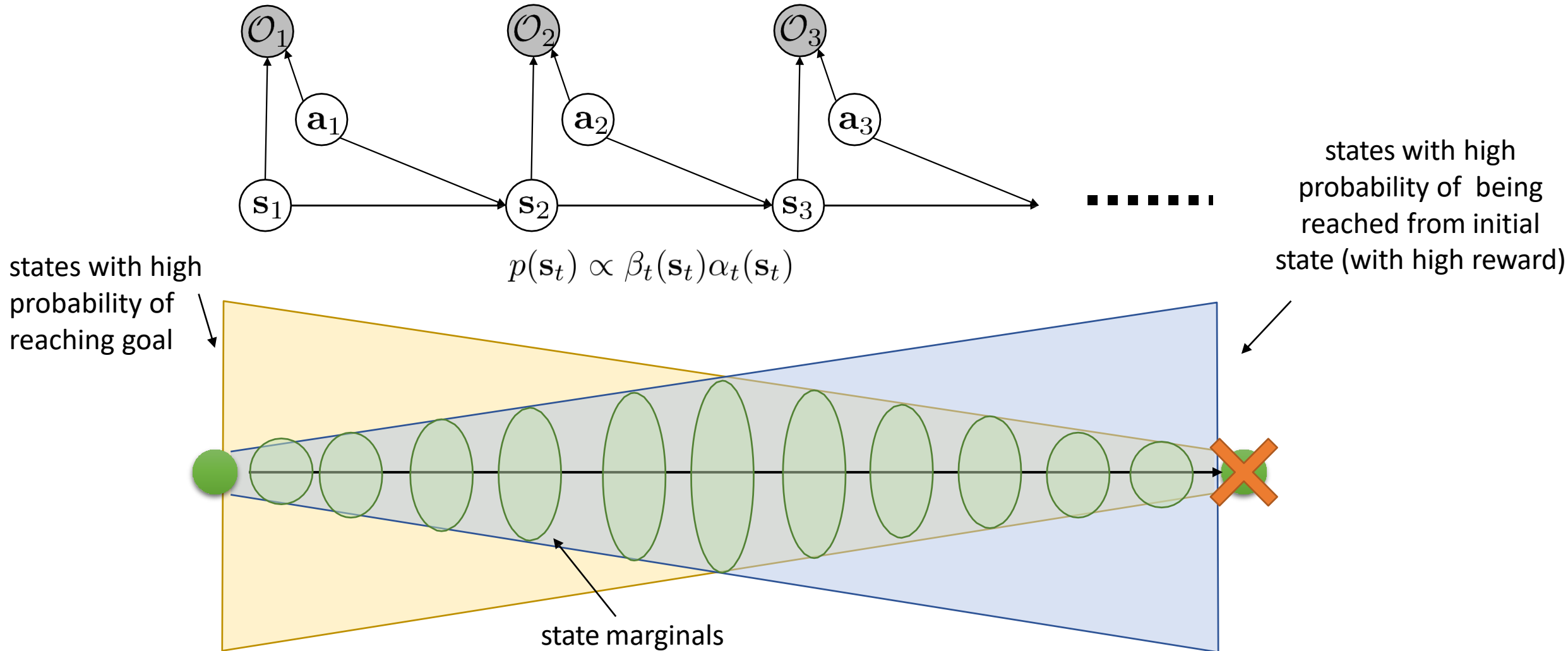
$$= \int \underline{p(\mathbf{s}_t | \mathbf{s}_{t-1}, \mathbf{a}_{t-1})} p(\mathbf{a}_{t-1} | \mathbf{s}_{t-1}, \mathcal{O}_{t-1}) p(\mathbf{s}_{t-1} | \mathcal{O}_{1:t-1}) d\mathbf{s}_{t-1} d\mathbf{a}_{t-1}$$

$$p(\mathbf{a}_{t-1} | \mathbf{s}_{t-1}, \mathcal{O}_{t-1}) p(\mathbf{s}_{t-1} | \mathcal{O}_{1:t-1}) = \frac{p(\mathcal{O}_{t-1} | \mathbf{s}_{t-1}, \mathbf{a}_{t-1}) p(\mathbf{a}_{t-1} | \mathbf{s}_{t-1})}{p(\cancel{\mathcal{O}_{t-1}} | \mathbf{s}_{t-1})} \frac{\alpha_{t-1}(\mathbf{s}_{t-1}) p(\cancel{\mathcal{O}_{t-1}} | \mathbf{s}_{t-1})}{\underline{p(\mathcal{O}_{t-1} | \mathcal{O}_{1:t-2})}}$$

what if we want  $p(\mathbf{s}_t | \mathcal{O}_{1:T})$ ?

$$p(\mathbf{s}_t | \mathcal{O}_{1:T}) \propto p(\mathcal{O}_{t:T} | \mathbf{s}_t) p(\mathbf{s}_t | \mathcal{O}_{1:t-1}) = \beta_t(\mathbf{s}_t) \alpha_t(\mathbf{s}_t)$$

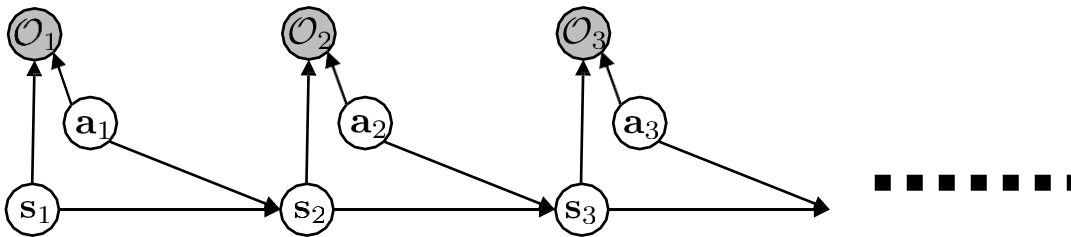
# Forward/backward message intersection





# Summary

1. Probabilistic graphical model for optimal control



2. Control = inference (similar to HMM, EKF, etc.)

3. Very similar to dynamic programming, value iteration, etc. (but “soft”)

# Control as Variational Inference

# The optimism problem

for  $t = T - 1$  to 1:

$$\beta_t(\mathbf{s}_t, \mathbf{a}_t) = p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} [\beta_{t+1}(\mathbf{s}_{t+1})]$$

“optimistic” transition  
(not a good idea!)

$$\beta_t(\mathbf{s}_t) = E_{\mathbf{a}_t \sim p(\mathbf{a}_t | \mathbf{s}_t)} [\beta_t(\mathbf{s}_t, \mathbf{a}_t)]$$

$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \overbrace{\log E[\exp(V_{t+1}(\mathbf{s}_{t+1}))]}$$

let  $V_t(\mathbf{s}_t) = \log \beta_t(\mathbf{s}_t)$

why did this happen?

let  $Q_t(\mathbf{s}_t, \mathbf{a}_t) = \log \beta_t(\mathbf{s}_t, \mathbf{a}_t)$

the inference problem:  $p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T} | \mathcal{O}_{1:T})$

marginalizing and conditioning, we get:  $p(\mathbf{a}_t | \mathbf{s}_t, \mathcal{O}_{1:T})$  (the policy)

“given that you obtained high reward, what was your action probability?”

marginalizing and conditioning, we get:  $p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T}) \neq p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$

“given that you obtained high reward, what was your transition probability?”

# Addressing the optimism problem

marginalizing and conditioning, we get:  $p(\mathbf{a}_t|\mathbf{s}_t, \mathcal{O}_{1:T})$  (the policy)  $\longleftarrow$  we want this

“given that you obtained high reward, what was your action probability?”

marginalizing and conditioning, we get:  $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t, \mathcal{O}_{1:T}) \neq p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$   $\longleftarrow$  but not this!

“given that you obtained high reward, what was your transition probability?”

“given that you obtained high reward, what was your action probability,  
*given that your transition probability did not change?*”

can we find another distribution  $q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$  that is close to  $p(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}|\mathcal{O}_{1:T})$  but has dynamics  $p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$

where have we seen this before?

let  $\mathbf{x} = \mathcal{O}_{1:T}$  and  $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$  find  $q(\mathbf{z})$  to approximate  $p(\mathbf{z}|\mathbf{x})$

let's try variational inference!

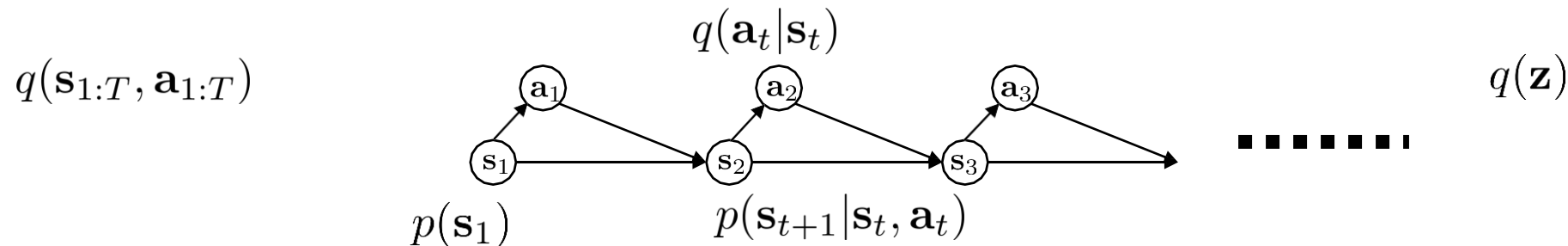
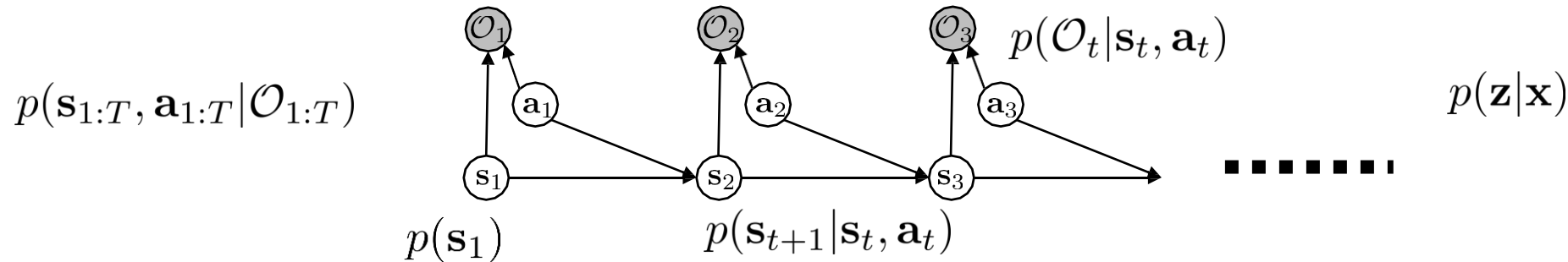
# Control via variational inference

$$\text{let } q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t | \mathbf{s}_t)$$

same dynamics and  
initial state as  $p$


only new thing

let  $\mathbf{x} = \mathcal{O}_{1:T}$  and  $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$



# The variational lower bound

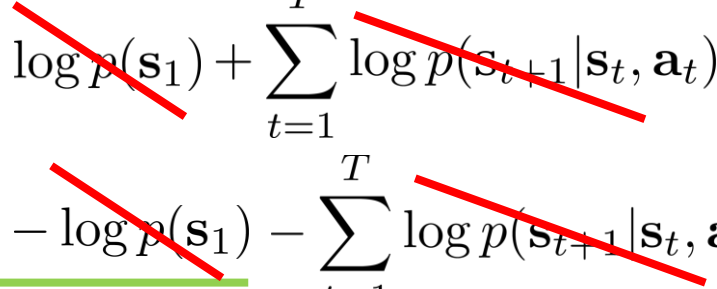
$$\log p(\mathbf{x}) \geq E_{\mathbf{z} \sim q(\mathbf{z})} [\log p(\mathbf{x}, \mathbf{z}) - \log q(\mathbf{z})]$$


 the entropy  $\mathcal{H}(q)$

let  $\mathbf{x} = \mathcal{O}_{1:T}$  and  $\mathbf{z} = (\mathbf{s}_{1:T}, \mathbf{a}_{1:T})$

$$\text{let } q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = \underbrace{p(\mathbf{s}_1)}_{\text{green underline}} \prod_t \underbrace{p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)}_{\text{orange underline}} q(\mathbf{a}_t | \mathbf{s}_t)$$

$$\log p(\mathcal{O}_{1:T}) \geq E_{(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \sim q} \left[ \cancel{\log p(\mathbf{s}_1)} + \sum_{t=1}^T \cancel{\log p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} + \sum_{t=1}^T \log p(\mathcal{O}_t | \mathbf{s}_t, \mathbf{a}_t) \right. \\ \left. - \cancel{\log p(\mathbf{s}_1)} - \sum_{t=1}^T \cancel{\log p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} - \sum_{t=1}^T \log q(\mathbf{a}_t | \mathbf{s}_t) \right]$$



$$= E_{(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) \sim q} \left[ \sum_t r(\mathbf{s}_t, \mathbf{a}_t) - \log q(\mathbf{a}_t | \mathbf{s}_t) \right]$$

$$= \sum_t E_{(\mathbf{s}_t, \mathbf{a}_t) \sim q} [r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t))]$$

maximize reward and maximize action entropy!

# Optimizing the variational lower bound

$$\text{let } q(\mathbf{s}_{1:T}, \mathbf{a}_{1:T}) = p(\mathbf{s}_1) \prod_t p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t) q(\mathbf{a}_t | \mathbf{s}_t) \quad \log p(\mathcal{O}_{1:T}) \geq \sum_t E_{(\mathbf{s}_t, \mathbf{a}_t) \sim q} [r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t))]$$

base case: solve for  $q(\mathbf{a}_T | \mathbf{s}_T)$ :

$$\begin{aligned} q(\mathbf{a}_T | \mathbf{s}_T) &= \arg \max E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T)] + \mathcal{H}(q(\mathbf{a}_T | \mathbf{s}_T))] \\ &= \arg \max E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T) - \log q(\mathbf{a}_T | \mathbf{s}_T)]] \end{aligned}$$

optimized when  $q(\mathbf{a}_T | \mathbf{s}_T) \propto \exp(r(\mathbf{s}_T, \mathbf{a}_T))$

$$q(\mathbf{a}_T | \mathbf{s}_T) = \frac{\exp(r(\mathbf{s}_T, \mathbf{a}_T))}{\int \exp(r(\mathbf{s}_T, \mathbf{a})) d\mathbf{a}} = \exp(Q(\mathbf{s}_T, \mathbf{a}_T) - V(\mathbf{s}_T))$$

$$V(\mathbf{s}_T) = \log \int \exp(Q(\mathbf{s}_T, \mathbf{a}_T)) d\mathbf{a}_T$$

$$E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T) - \log q(\mathbf{a}_T | \mathbf{s}_T)]] = E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [V(\mathbf{s}_T)]]$$

# Optimizing the variational lower bound

$$\log p(\mathcal{O}_{1:T}) \geq \sum_t E_{(\mathbf{s}_t, \mathbf{a}_t) \sim q} [r(\mathbf{s}_t, \mathbf{a}_t) + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t))]$$

$$q(\mathbf{a}_T | \mathbf{s}_T) = \frac{\exp(r(\mathbf{s}_T, \mathbf{a}_T))}{\int \exp(r(\mathbf{s}_T, \mathbf{a})) d\mathbf{a}} = \exp(Q(\mathbf{s}_T, \mathbf{a}_T) - V(\mathbf{s}_T))$$

$$E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [r(\mathbf{s}_T, \mathbf{a}_T) - \log q(\mathbf{a}_T | \mathbf{s}_T)]] = E_{\mathbf{s}_T \sim q(\mathbf{s}_T)} [E_{\mathbf{a}_T \sim q(\mathbf{a}_T | \mathbf{s}_T)} [V(\mathbf{s}_T)]]$$

$$\begin{aligned} q(\mathbf{a}_t | \mathbf{s}_t) &= \arg \max E_{\mathbf{s}_t \sim q(\mathbf{s}_t)} [E_{\mathbf{a}_t \sim q(\mathbf{a}_t | \mathbf{s}_t)} [\boxed{r(\mathbf{s}_t, \mathbf{a}_t) + E_{\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)} [V(\mathbf{s}_{t+1})]}] + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t))] \\ &= \arg \max E_{\mathbf{s}_t \sim q(\mathbf{s}_t)} [E_{\mathbf{a}_t \sim q(\mathbf{a}_t | \mathbf{s}_t)} [Q(\mathbf{s}_t, \mathbf{a}_t)] + \mathcal{H}(q(\mathbf{a}_t | \mathbf{s}_t))] \\ &= \arg \max E_{\mathbf{s}_t \sim q(\mathbf{s}_t)} [E_{\mathbf{a}_t \sim q(\mathbf{a}_t | \mathbf{s}_t)} [Q(\mathbf{s}_t, \mathbf{a}_t) - \log q(\mathbf{a}_t | \mathbf{s}_t)]] \end{aligned}$$

*regular Bellman backup*  
*not optimistic*

$$\boxed{Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + E[V_{t+1}(\mathbf{s}_{t+1})]}$$

optimized when  $q(\mathbf{a}_t | \mathbf{s}_t) \propto \exp(Q(\mathbf{s}_t, \mathbf{a}_t))$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t$$

$$q(\mathbf{a}_t | \mathbf{s}_t) = \exp(Q(\mathbf{s}_t, \mathbf{a}_t) - V(\mathbf{s}_t))$$




# Backward pass summary - variational

for  $t = T - 1$  to 1:


$$Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + E[V_{t+1}(\mathbf{s}_{t+1})]$$

$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t$$

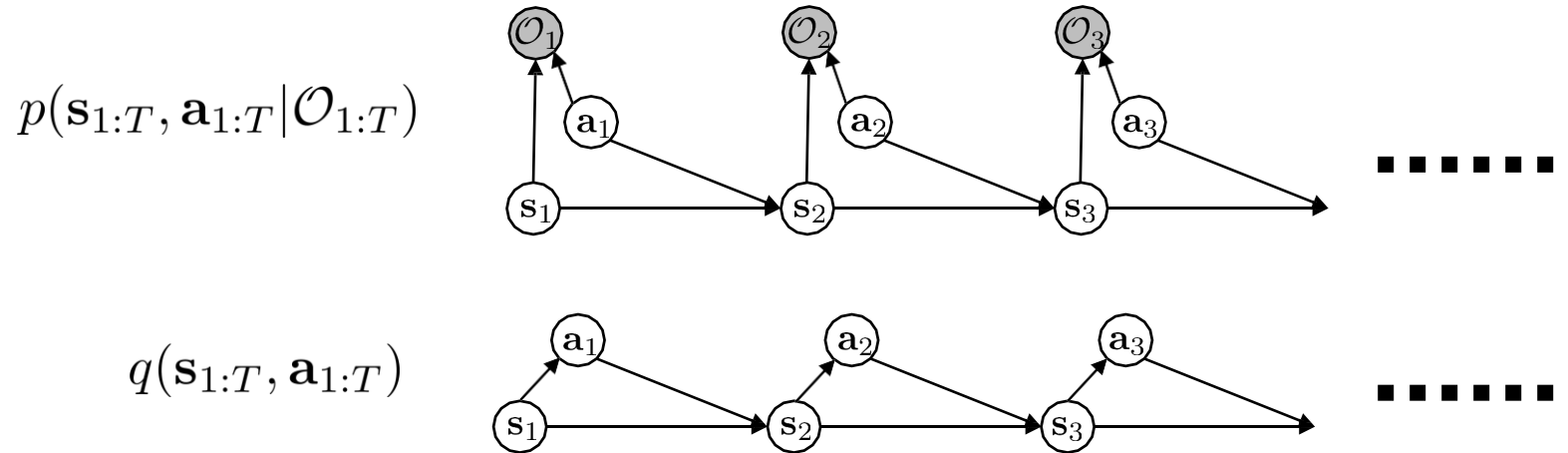
value iteration algorithm:

- 
1. set  $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')] ]$
  2. set  $V(\mathbf{s}) \leftarrow \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

*soft* value iteration algorithm:

- 
1. set  $Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \gamma E[V(\mathbf{s}')] ]$
  2. set  $V(\mathbf{s}) \leftarrow \underline{\text{soft}} \max_{\mathbf{a}} Q(\mathbf{s}, \mathbf{a})$

# Summary



$$V_t(\mathbf{s}_t) = \log \int \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t)) d\mathbf{a}_t \quad Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + E[(V_{t+1}(\mathbf{s}_{t+1}))]$$

variants:

discounted SOC:  $Q_t(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma E[V_{t+1}(\mathbf{s}_{t+1})]$

explicit temperature:  $V_t(\mathbf{s}_t) = \alpha \log \int \exp\left(\frac{1}{\alpha} Q_t(\mathbf{s}_t, \mathbf{a}_t)\right) d\mathbf{a}_t$

For more details, see: Levine. (2018). Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review.

# Algorithms for RL as Inference

# Q-learning with soft optimality


standard Q-learning:  $\phi \leftarrow \phi + \alpha \nabla_{\phi} Q_{\phi}(\mathbf{s}, \mathbf{a})(r(\mathbf{s}, \mathbf{a}) + \gamma V(\mathbf{s}') - Q_{\phi}(\mathbf{s}, \mathbf{a}))$

target value:  $V(\mathbf{s}') = \max_{\mathbf{a}'} Q_{\phi}(\mathbf{s}', \mathbf{a}')$

soft Q-learning:  $\phi \leftarrow \phi + \alpha \nabla_{\phi} Q_{\phi}(\mathbf{s}, \mathbf{a})(r(\mathbf{s}, \mathbf{a}) + \gamma V(\mathbf{s}') - Q_{\phi}(\mathbf{s}, \mathbf{a}))$

target value:  $V(\mathbf{s}') = \text{soft max}_{\mathbf{a}'} Q_{\phi}(\mathbf{s}', \mathbf{a}') = \log \int \exp(Q_{\phi}(\mathbf{s}', \mathbf{a}')) d\mathbf{a}'$

$\pi(\mathbf{a}|\mathbf{s}) = \exp(Q_{\phi}(\mathbf{s}, \mathbf{a}) - V(\mathbf{s})) = \exp(A(\mathbf{s}, \mathbf{a}))$

- 
1. take some action  $\mathbf{a}_i$  and observe  $(\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_i)$ , add it to  $\mathcal{R}$
  2. sample mini-batch  $\{\mathbf{s}_j, \mathbf{a}_j, \mathbf{s}'_j, r_j\}$  from  $\mathcal{R}$  uniformly
  3. compute  $y_j = r_j + \gamma \text{soft max}_{\mathbf{a}'_j} Q_{\phi'}(\mathbf{s}'_j, \mathbf{a}'_j)$  using *target* network  $Q_{\phi'}$
  4.  $\phi \leftarrow \phi - \alpha \sum_j \frac{dQ_{\phi}}{d\phi}(\mathbf{s}_j, \mathbf{a}_j)(Q_{\phi}(\mathbf{s}_j, \mathbf{a}_j) - y_j)$
  5. update  $\phi'$ : copy  $\phi$  every  $N$  steps, or Polyak average  $\phi' \leftarrow \tau \phi' + (1 - \tau)\phi$

# Policy gradient with soft optimality

$\pi(\mathbf{a}|\mathbf{s}) = \exp(Q_\phi(\mathbf{s}, \mathbf{a}) - V(\mathbf{s}))$  optimizes  $\sum_t E_{\pi(\mathbf{s}_t, \mathbf{a}_t)}[r(\mathbf{s}_t, \mathbf{a}_t)] + \underbrace{E_{\pi(\mathbf{s}_t)}[\mathcal{H}(\pi(\mathbf{a}_t|\mathbf{s}_t))]}_{\text{policy entropy}}$

intuition:  $\pi(\mathbf{a}|\mathbf{s}) \propto \exp(Q_\phi(\mathbf{s}, \mathbf{a}))$  when  $\pi$  minimizes  $D_{\text{KL}}(\pi(\mathbf{a}|\mathbf{s}) \parallel \frac{1}{Z} \exp(Q(\mathbf{s}, \mathbf{a})))$   
 $\longrightarrow \pi$  minimizes  $-E_{\pi(\mathbf{a}|\mathbf{s})}[Q(\mathbf{s}, \mathbf{a})] - \mathcal{H}(\pi)$

often referred to as “entropy regularized” policy gradient  
combats premature entropy collapse

Ziebart et al. ‘10 “Modeling Interaction via the Principle of Maximum Causal Entropy”

# Soft actor-critic

## 1. Q-function update

Update Q-function to evaluate current policy:

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow r(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathbf{s}' \sim p_{\mathbf{s}}, \mathbf{a}' \sim \pi} [Q(\mathbf{s}', \mathbf{a}') - \log \pi(\mathbf{a}'|\mathbf{s}')] ]$$

This converges to  $Q^\pi$ .

## 2. Update policy

Update the policy with gradient of information projection:

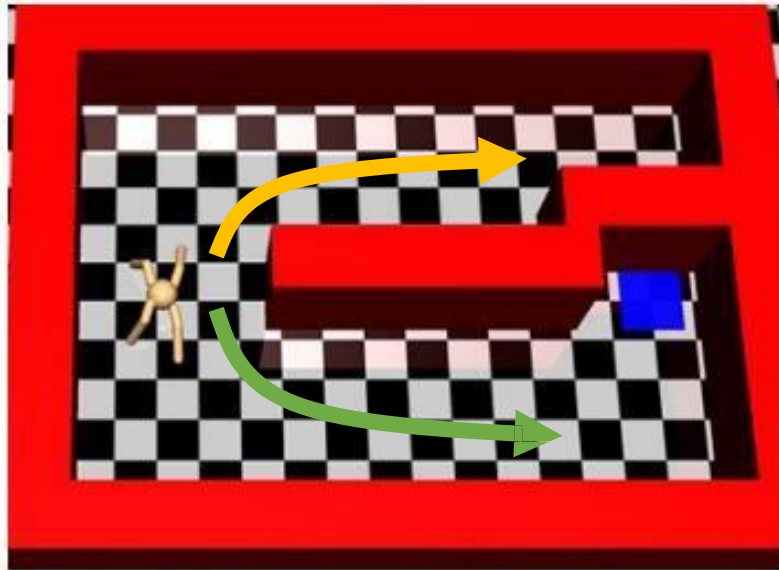
$$\pi_{\text{new}} = \arg \min_{\pi'} D_{\text{KL}} \left( \pi'(\cdot | \mathbf{s}) \parallel \frac{1}{Z} \exp Q^{\pi_{\text{old}}}(\mathbf{s}, \cdot) \right)$$

In practice, only take one gradient step on this objective

## 3. Interact with the world, collect more data

Haarnoja, et al. **Soft Actor-Critic Algorithms and Applications**. '18

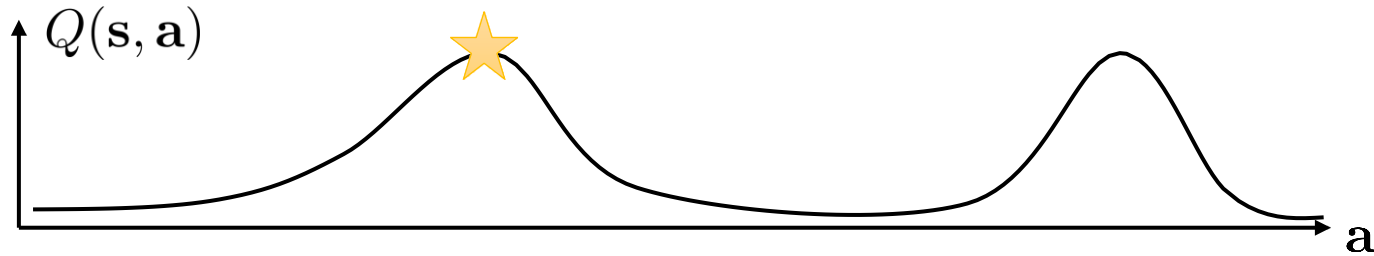
# Stochastic models for learning control



- How can we track *both* hypotheses?

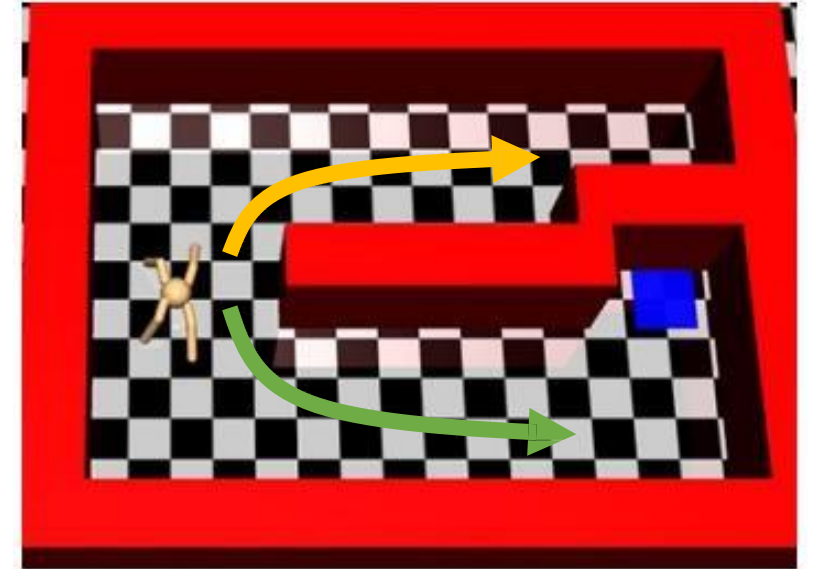
# Stochastic energy-based policies

Q-function:  $Q(\mathbf{s}, \mathbf{a}) : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$



$$\pi(\mathbf{a}|\mathbf{s}) \propto \exp(Q(\mathbf{s}, \mathbf{a}))$$

$$\pi(\mathbf{a}_t|\mathbf{s}_t) = \exp(Q_t(\mathbf{s}_t, \mathbf{a}_t) - V_t(\mathbf{s}_t)) = \exp(A_t(\mathbf{s}_t, \mathbf{a}_t))$$



Haarnoja\*, Tang\*, Abbeel, L., Reinforcement Learning with Deep Energy-Based Policies. ICML 2017

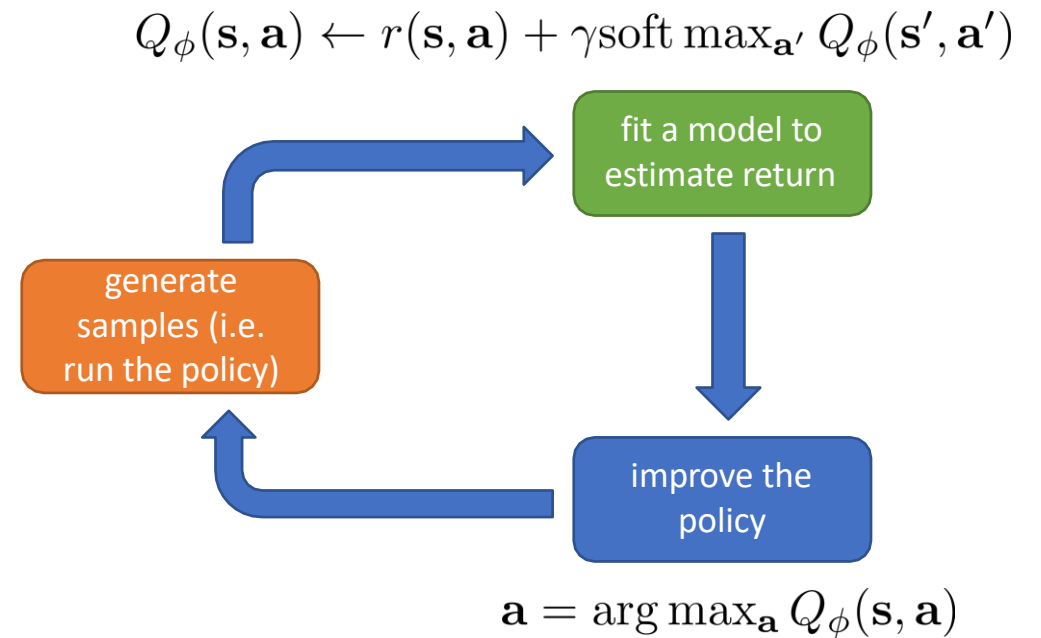


# Benefits of soft optimality

- Improve **exploration** and prevent entropy collapse
- **Easier to specialize** (finetune) policies for more specific tasks
- Principled approach to **break ties**
- **Better robustness** (due to wider coverage of states)
- **Can reduce to hard optimality** as reward magnitude increases
- Good model for **modeling human behavior**

# Review

- Reinforcement learning can be viewed as **inference** in a graphical model
  - Value function is a backward message
  - Maximize reward and entropy
  - Variational inference to remove optimism
- Soft Q-learning
- Entropy-regularized policy gradient



# Soft optimality suggested readings

- Todorov. (2006). Linearly solvable Markov decision problems: one framework for reasoning about soft optimality.
- Todorov. (2008). General duality between optimal control and estimation: primer on the equivalence between inference and control.
- Kappen. (2009). Optimal control as a graphical model inference problem: frames control as an inference problem in a graphical model.
- Ziebart. (2010). Modeling interaction via the principle of maximal causal entropy: connection between soft optimality and maximum entropy modeling.
- Rawlik, Toussaint, Vijaykumar. (2013). On stochastic optimal control and reinforcement learning by approximate inference: temporal difference style algorithm with soft optimality.
- Haarnoja\*, Tang\*, Abbeel, L. (2017). Reinforcement learning with deep energy based models:
  - soft Q-learning algorithm, deep RL with continuous actions and soft optimality
- Nachum, Norouzi, Xu, Schuurmans. (2017). Bridging the gap between value and policy based reinforcement learning.
- Schulman, Abbeel, Chen. (2017). Equivalence between policy gradients and soft Q-learning.
- Haarnoja, Zhou, Abbeel, L. (2018). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor.
- Levine. (2018). Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review