Maximum Entropy RL

10/28/20

Miscellaneous

- HW3 due on Friday 10/30
- Quiz 2 on next *Friday 11/6* (Different from original date)
- HW4 not released until after Quiz 2

Overview for Today

1. Intuition: When is acting (slightly) randomly a good idea?
2. Algorithms for Maximum Entropy
3. Why is MaxEnt RL so appealing?

All RL Problems have a Deterministic Solution

$$\max_{\pi} E_{\pi} \left[\sum_{t} \gamma^{t} r(s_t, a_t) \right]$$

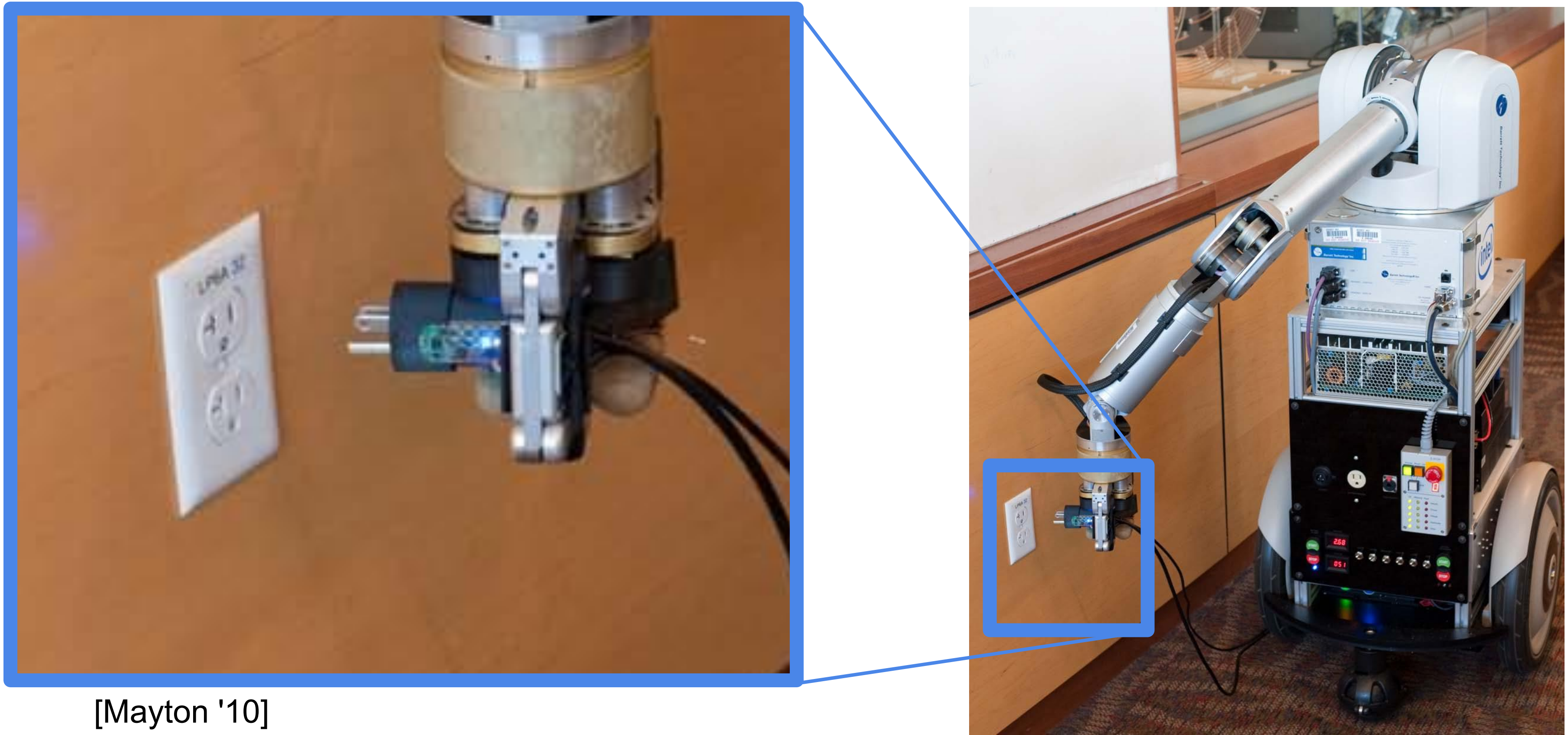
Proof?

Hint: Think about policy improvement...

$$\pi(a \mid s) = \arg\max_{a} Q(s, a)$$

When is acting (slightly) random a good idea?

Example: Inserting a Plug

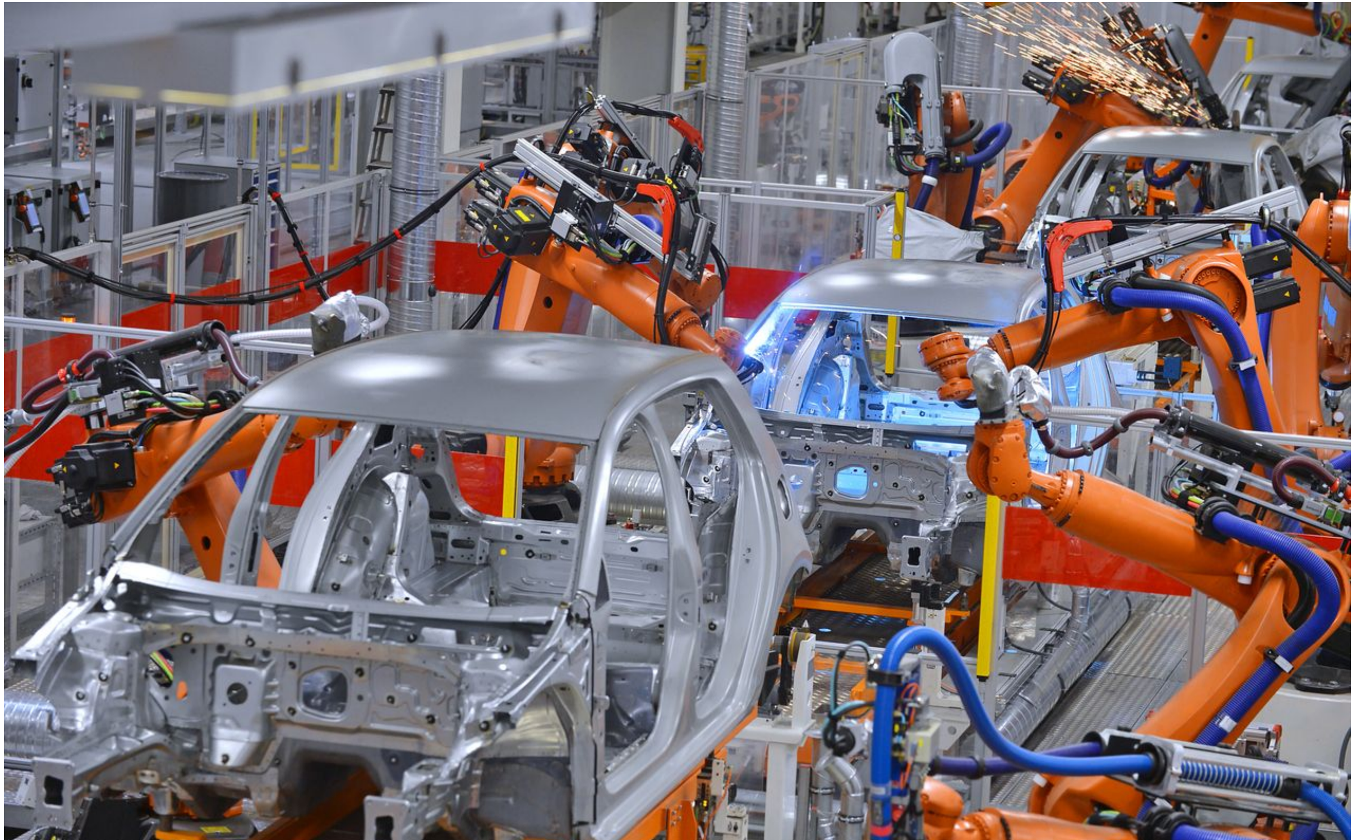


[Mayton '10]

Example: Looking for sugar in a new kitchen?

Example: Handling adversarial hockey players

Counterexample: Precise Manufacturing?



What objective results in stochastic ("random") policies?

What is Maximum Entropy RL?

$$\max_{\pi} E_{\pi}\left[\sum_{t}\gamma^{t}\left(r(s_t,a_t)+\mathcal{H}_{\pi}[a_t \mid s_t]\right)\right]$$

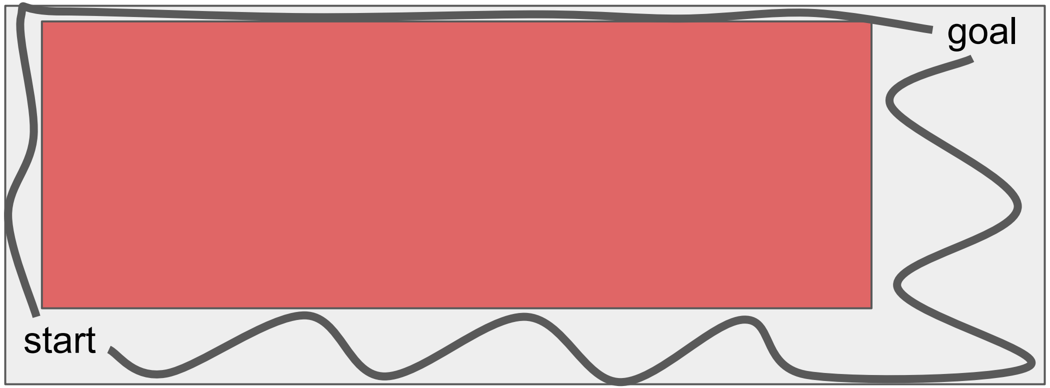
$$\mathcal{H}_{\pi}[a_t \mid s_t] \triangleq E_{\pi}[-\log \pi(a_t \mid s_t)]$$

Intuition

- Want to maximize expected *future* reward and action entropy
- Take actions that lead to high reward, and allow us to act randomly in the future
- If there are many ways to solve the task, try all of them!
- If there are many paths to a goal, try all possible paths, but more frequently use short paths.

What is Maximum Entropy RL?

Intuition



- Want to maximize expected *future* reward and action entropy
- Take actions that lead to high reward, and allow us to act randomly in the future
- If there are many ways to solve the task, try all of them!
- If there are many paths to a goal, try all possible paths, but more frequently use short paths.

What is Maximum Entropy RL?

Common mistake: Don't ignore *future* entropy

$$E_\pi\left[\sum_t \gamma^t \left(r(s_t, a_t) + \mathcal{H}_\pi[a_t \mid s_t]\right)\right] \neq E_\pi\left[\sum_t \gamma^t r(s_t, a_t)\right] + \mathcal{H}_\pi[a_t \mid s_t]$$

Algorithms for Maximum Entropy RL

Solving Maximum Entropy RL

$$E_\pi\left[\sum_t \gamma^t \underbrace{\left(r(s_t, a_t) + \mathcal{H}_\pi[a_t \mid s_t]\right)}_{\tilde{r}(s,a) \triangleq r(s,a) - \log \pi(a \mid s)}\right]$$

DQN:

$$y = r(s,a) + \gamma \max_{a'} Q(s', a') \\ = r(s,a) + \gamma E_{\pi(a'|s')}[Q(s', a')]$$

$$\min_\theta \left(Q_\theta(s,a) - y\right)^2$$

$$\pi(a \mid s) = \delta(a = \arg\max_a Q(s,a))$$

Soft Q Learning

$$y = r(s,a) - \log \pi(a \mid s) + \gamma E_{\pi(a'|s')}[Q(s', a')]$$

$$\min_\theta \left(Q_\theta(s,a) - y\right)^2$$

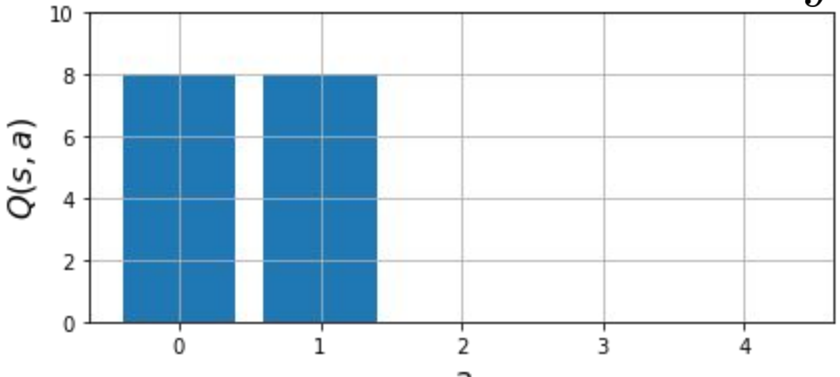
$$\max_\pi E_{\pi(a|s)}[Q(s,a) - \log \pi(a \mid s)]$$

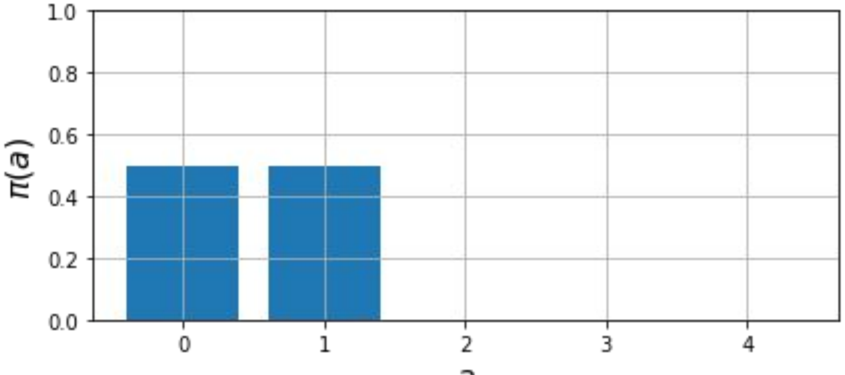
Side Note: Why is it called "soft"?

$$\max_{\pi} E_{\pi(a|s)}[Q(s,a) - \log \pi(a \mid s)]$$

Exercise: Assume Q(s, a) is given, and actions are discrete. What are the probabilities $\pi(a \mid s)$?

$$\pi(a \mid s) = \frac{e^{Q(s,a)}}{\int e^{Q(s,a')} da'}$$



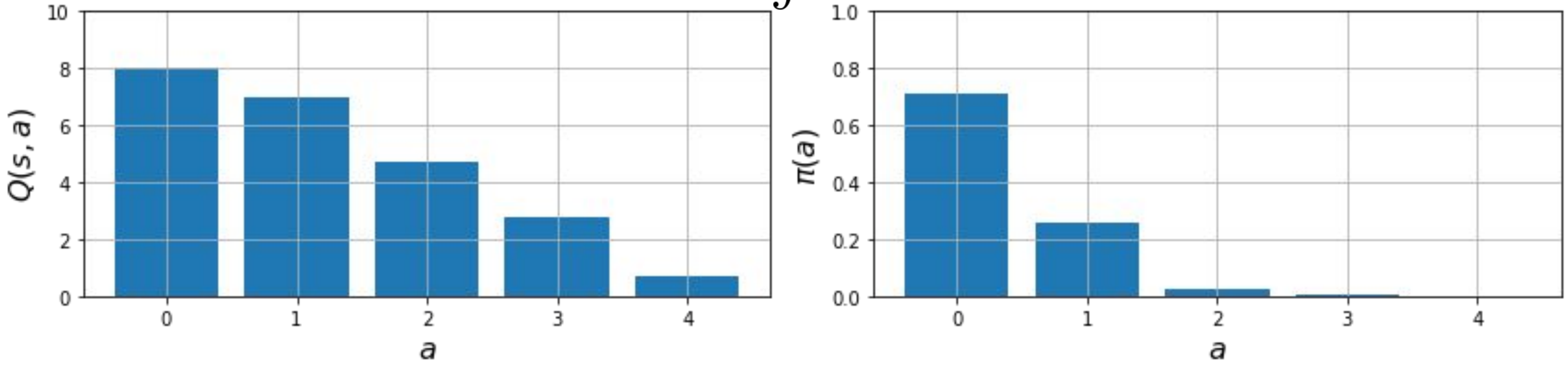


Side Note: Why is it called "soft"?

$$\max_{\pi} E_{\pi(a|s)}[Q(s,a) - \log \pi(a \mid s)]$$

Exercise: Assume Q(s, a) is given, and actions are discrete. What are the probabilities $\pi(a \mid s)$?

$$\pi(a \mid s) = \frac{e^{Q(s,a)}}{\int e^{Q(s,a')} da'}$$

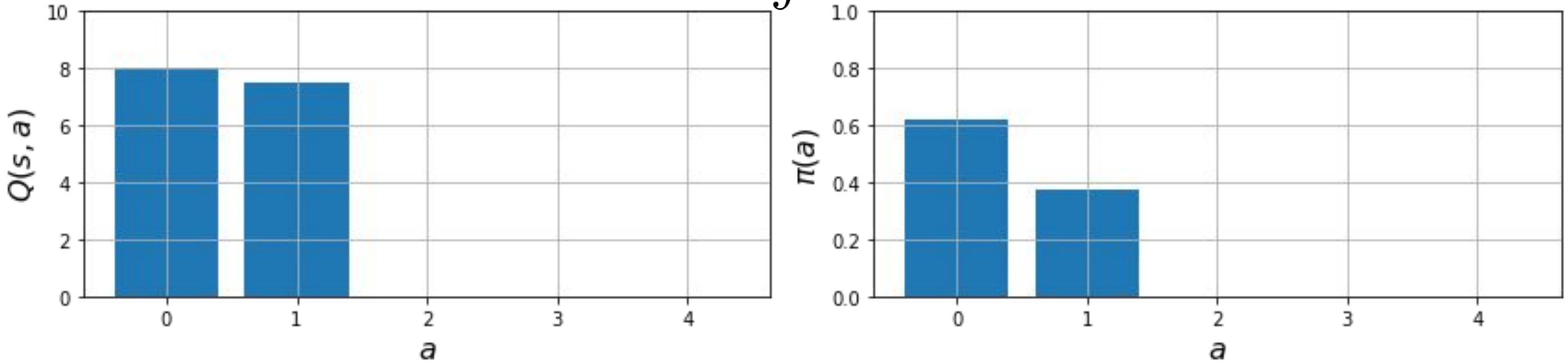


Side Note: Why is it called "soft"?

$$\max_{\pi} E_{\pi(a|s)}[Q(s,a) - \log \pi(a \mid s)]$$

Exercise: Assume Q(s, a) is given, and actions are discrete. What are the probabilities $\pi(a \mid s)$?

$$\pi(a \mid s) = \frac{e^{Q(s,a)}}{\int e^{Q(s,a')} da'}$$



Solving Maximum Entropy RL

$$E_\pi\left[\sum_t \gamma^t \underbrace{\left(r(s_t, a_t) + \mathcal{H}_\pi[a_t \mid s_t]\right)}_{\tilde{r}(s,a) \triangleq r(s,a) - \log \pi(a \mid s)}\right]$$

DDPG:

$$y = r(s,a) + \gamma E_{\pi(a'|s')}[Q(s', a')]$$

$$\min_\theta \left(Q_\theta(s,a) - y\right)^2$$

$$\max_\phi Q(s, a = \pi_\phi(s)) = E_{\pi_\phi(a|s)}[Q(s,a)]$$

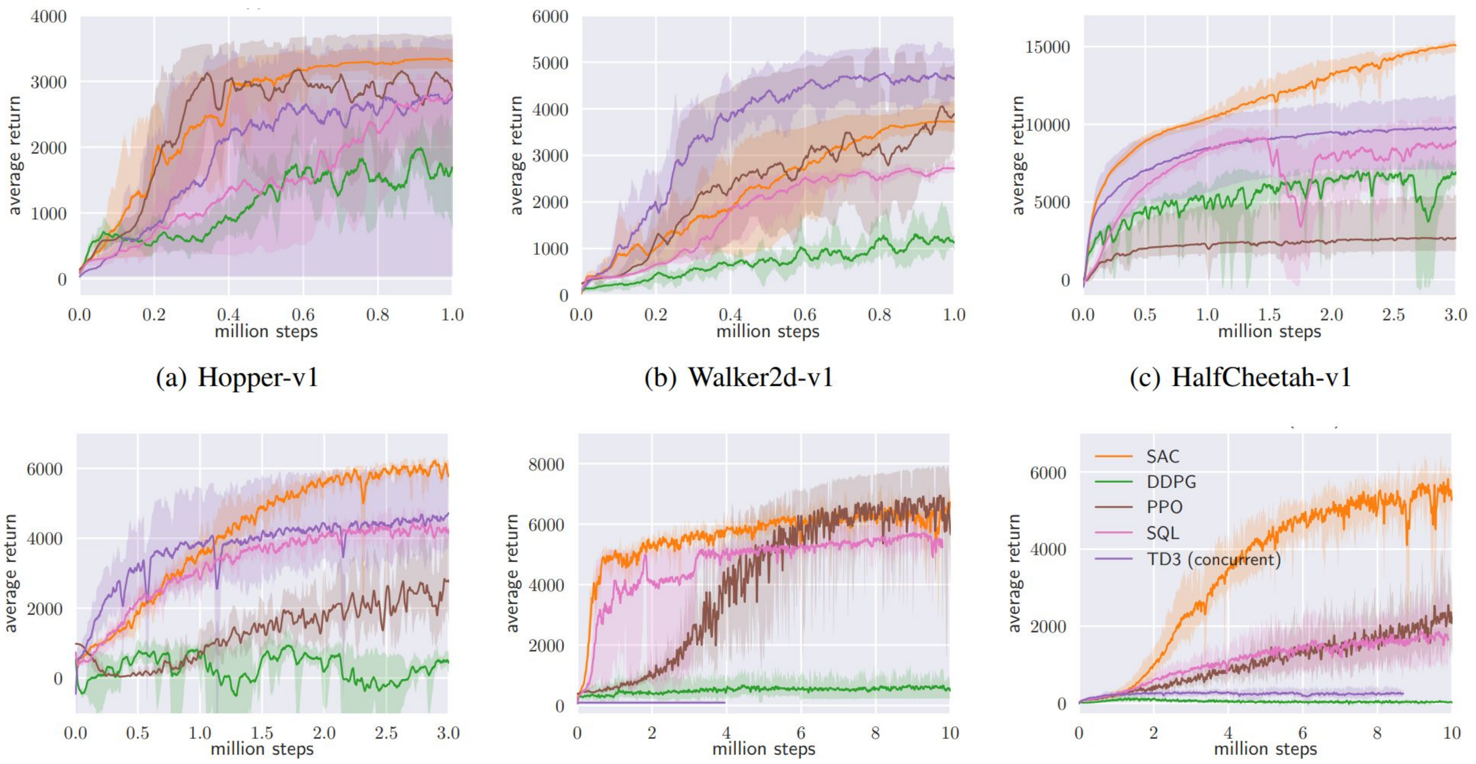
Soft Actor Critic [Haarnoja 18]

$$y = r(s,a) - \log \pi(a \mid s) + \gamma E_{\pi(a'|s')}[Q(s', a')]$$

$$\min_\theta \left(Q_\theta(s,a) - y\right)^2$$

$$\max_\phi E_{\pi_\phi(a|s)}[Q(s,a) - \log \pi_\phi(a \mid s)]$$

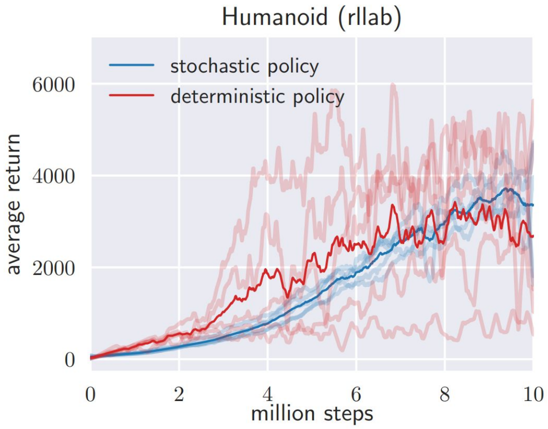
Results from Soft Actor Critic

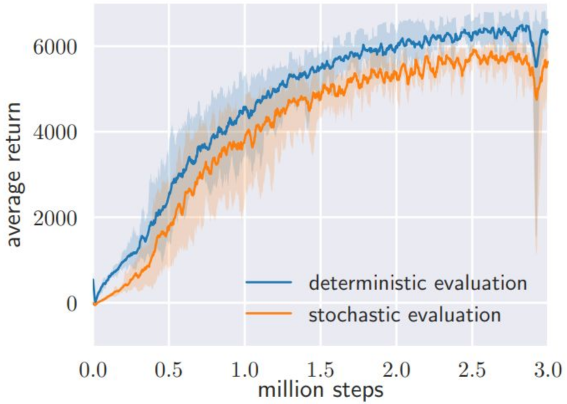


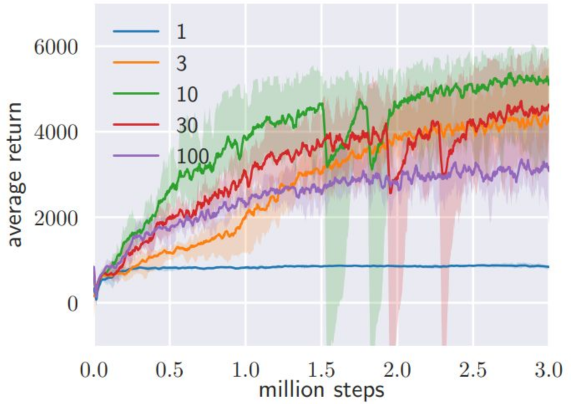
(a) Hopper-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

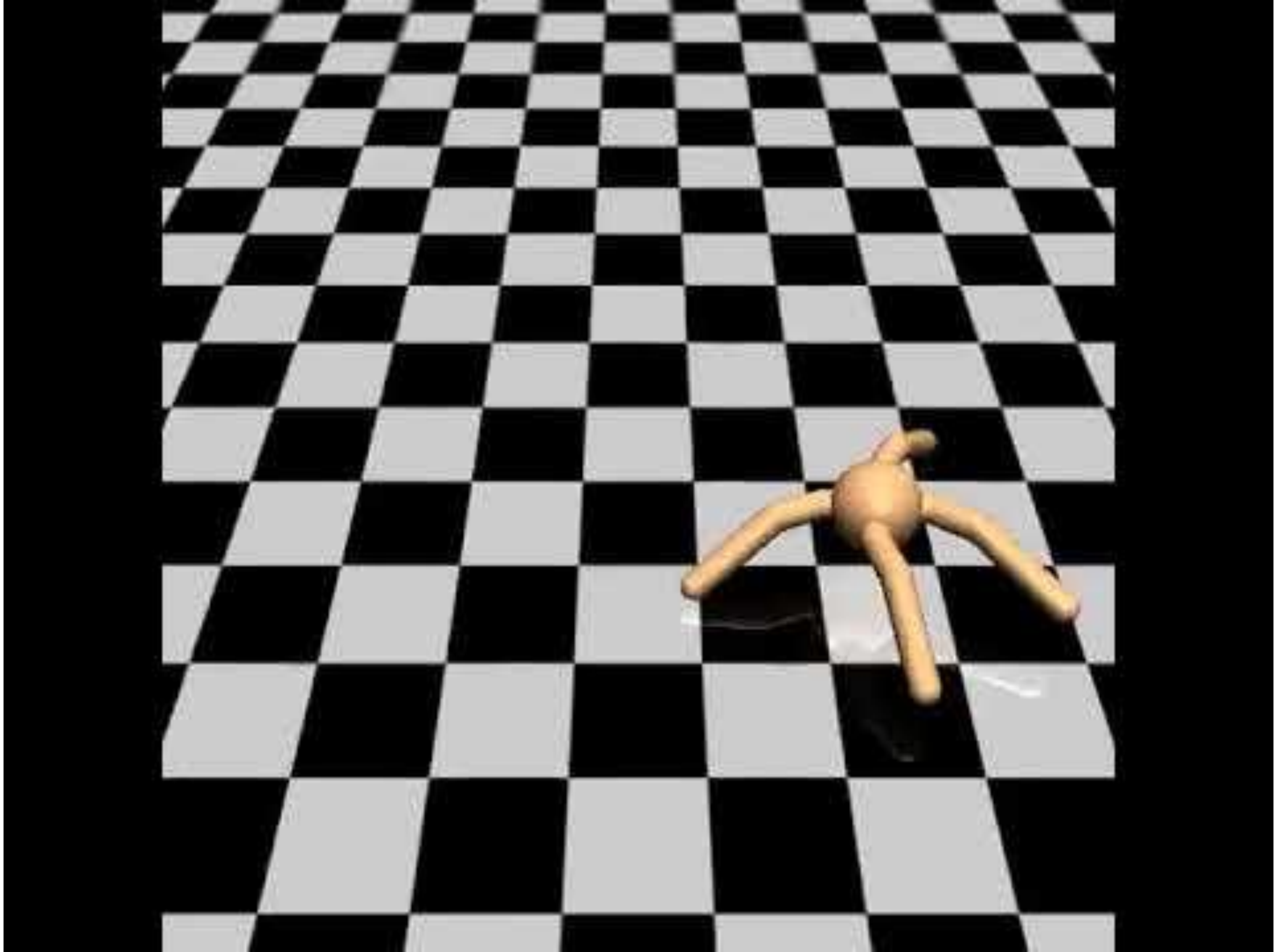
(d) Ant-v1 (e) Humanoid-v1 (f) Humanoid (rllab)

Results from Soft Actor Critic










54 min
5x

Tips and Tricks for MaxEnt RL

TD3 trick [Fujimoto 18]

$$y = r(s,a) + \gamma \min_{i=1,2} Q_1(s', a' \sim \pi(a' \mid s'))$$

Automatic entropy tuning ("Entropy Constrained SAC")

$$E\left[\sum_t \gamma^t r(s_t, a_t) + \boxed{\alpha} \mathcal{H}_\pi[a_t \mid s_t]\right] \Rightarrow E\left[\sum_t \gamma^t r(s_t, a_t)\right]$$

$$\text{s.t.} \quad E\left[\sum_t \mathcal{H}_\pi[a_t \mid s_t]\right] \geq \epsilon$$

Side Note: Dual Gradient Ascent

How do you solve *constrained* optimization problems with SGD?

$$\max_x f(x)$$
$$\text{s.t.} \quad g(x) \geq \epsilon$$

"Lagrangian"

$$\mathcal{L}(x, \lambda) = f(x) + \lambda(g(x) - \epsilon)$$

$$\nabla_x \mathcal{L} = \nabla_x f(x) + \lambda \nabla_x g(x)$$
$$\nabla_\lambda \mathcal{L} = g(x) - \epsilon$$

$$x \leftarrow x + \eta \nabla_x \mathcal{L}$$
$$\lambda \leftarrow \lfloor \lambda + \eta(\epsilon - g(x)) \rfloor_+$$

Soft Bellman Optimality

Bellman equations

Fixed point

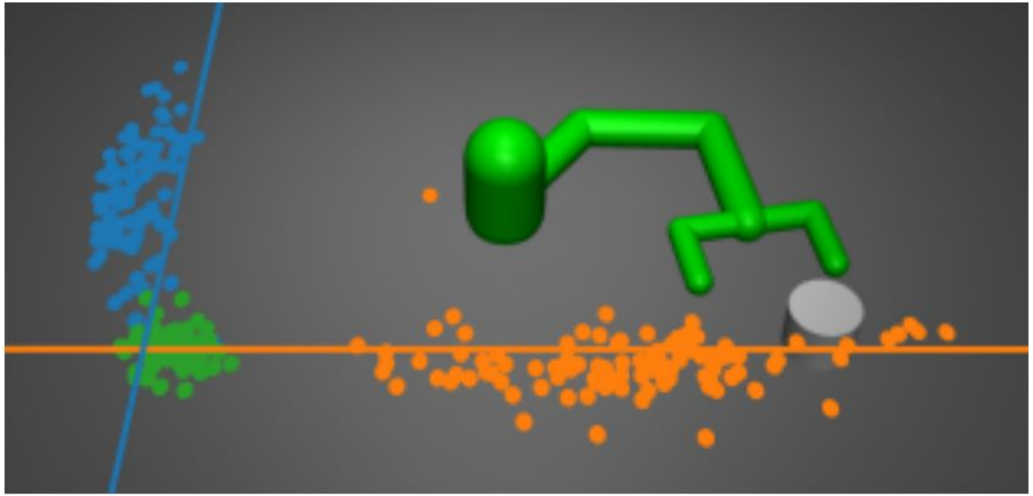
Policy improvement Thm

Regularized policy improvement

Why is MaxEnt RL so Appealing?

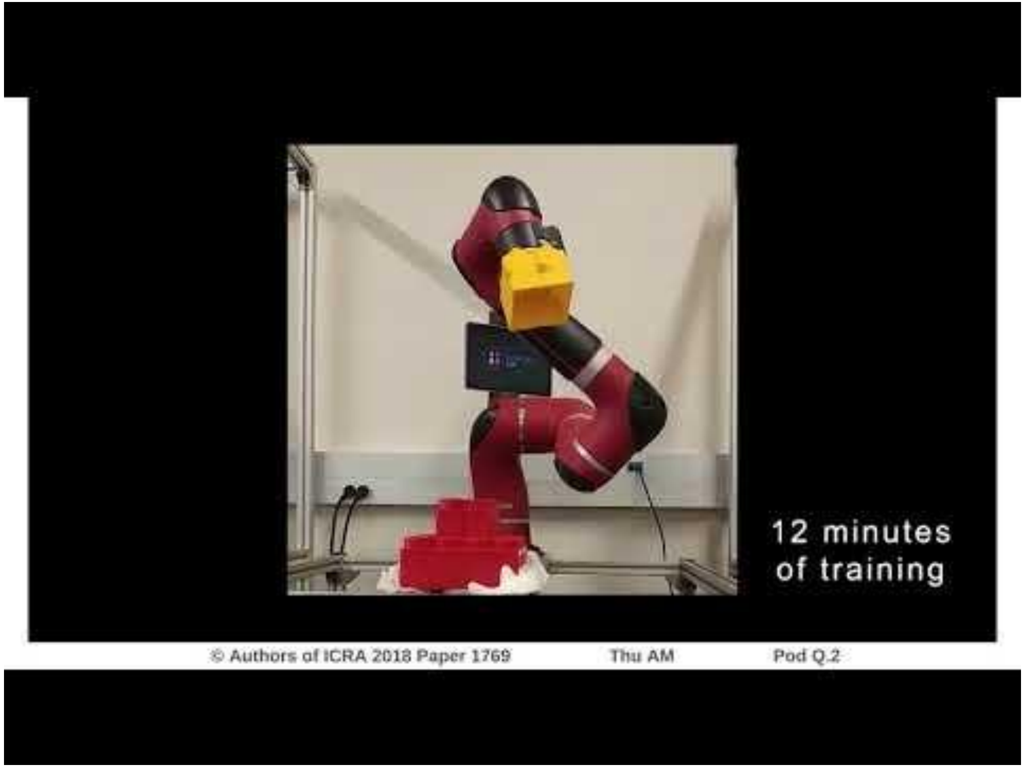
Soft Q functions are Composable

$$Q^*_{\mathcal{C}}(\mathbf{s}, \mathbf{a}) \approx Q_{\Sigma}(\mathbf{s}, \mathbf{a}) = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} Q^*_i(\mathbf{s}, \mathbf{a})$$



Composable Deep Reinforcement Learning for Robotic Manipulation [Haarnoja 18]

Soft Q functions are Composable



Composable Deep Reinforcement Learning for Robotic Manipulation [Haarnoja
18]

Linearly Solvable MDPs

Idea: Agent to "pay" to modify the "passive dynamics" to maximize reward

$$\tilde{r}(s,a) = r(s) - KL(p(s' \mid s,a) \| p(s' \mid s, a=\emptyset))$$

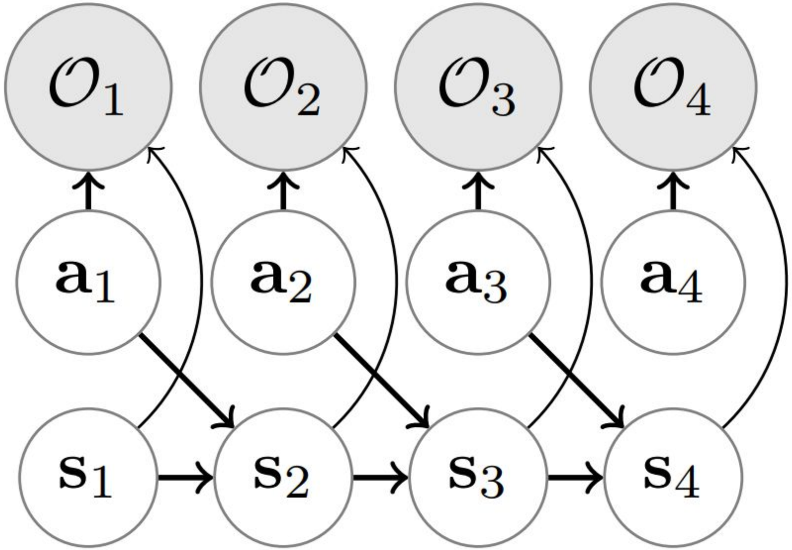
$$V(s) = r(s) + \log\left(\sum_{s'} p(s' \mid s, a=\emptyset) e^{V(s')}\right)$$

$$[e^V] = [e^r] P [e^V]$$

- Just a linear equation ("X = AX"). Can solve for "X" = e^V
- (Exponentiated) value function is an eigenvector.

Linearly Solvable MDPs [Todorov 06]

MaxEnt RL is Message Passing on a PGM



- Optimal = "not failing"
- Probability of being "optimal" in future

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

- Intuitively, choose actions to maximize
- HMM message passing = Soft Bellman Equation

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

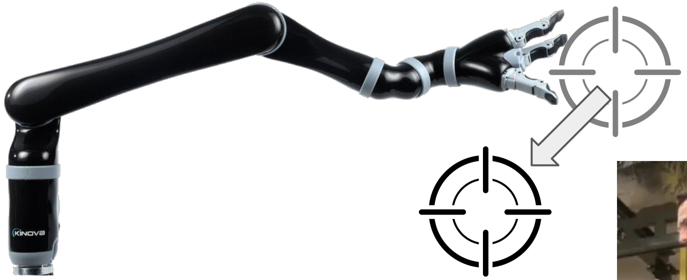
Reinforcement Learning and Control as Probabilistic Inference [Levine 18]

MaxEnt RL Policies are Robust

Theorem (informal): MaxEnt RL is optimal under disturbances to the reward function

Theorem (informal): MaxEnt RL is optimal under disturbances to the dynamics

Theorem (informal): MaxEnt RL is optimal under certain types of partial observability



If MaxEnt RL is the Answer, What is the Question? [BE 19]

Overview for Today

1. Intuition: When is acting (slightly) randomly a good idea?
2. Algorithms for Maximum Entropy
3. Why is MaxEnt RL so appealing?
    a. Compositionality
    b. Special case is just a linear system
    c. Equivalent to inference on a graphical model
    d. Robustness

What is MaxEnt RL Cool?

- Robustness [BE]
- Solves POMDPs [BE]
- Exploration [SAC]
- Easier optimization [Zaf]
- Connections with probabilistic inference [Rawlik, Levine]
  - Automatically handles uncertainty
  - Rewards can be interpretted as priors
  - Readily combined with (probabilistic) sensor tracking and fusion
- FEP [Friston]
- Compositionality [Tuomas paper]
- Path Integral Control
- Linearly Solvable MDPs [Todorov]
- Equivalence of Estimation and Control

Theorem (informal): MaxEnt RL is optimal under certain types of partial observability