2022-10-20

Policy space methods

Simplicity at the cost of variance

Cathy Wu

6.7950: Reinforcement Learning: Foundations and Methods

References

1. Matteo Pirotta. FAIR. Reinforcement Learning. 2019, Lecture 5.
2. Matteo Pirotta. Reinforcement Learning Summer School, 2019. Policy Search: Actor-Critic Methods.

Outline

1. From Policy Iteration to Policy Search
2. Policy gradient methods
3. Actor-critic

Outline

1. **From Policy Iteration to Policy Search**
2. Policy gradient methods
3. Actor-critic

Function approximation

Last time: adding function approximation to value iteration

This time: adding function approximation to policy iteration. Sorta.

Policy Iteration: Recap

```
Let $\pi_0$ be an arbitrary stationary policy.
while $k = 1, \dots, K$ do
    Policy Evaluation: given $\pi_k$ compute $V_k = V^{\pi_k}$
    Policy Improvement: find $\pi_{k+1}$ that is better than $\pi_k$
        - e.g. compute the greedy policy:
        $$\pi_{k+1}(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{y} p(y|s,a)V^{\pi_k}(y) \right\}$$
    return the last policy $\pi_K$
end
```

- Convergence is finite and monotonic [Bertsekas, 2007] (in exact settings)

? Issues: Function approximation for V^{π_k} $\Longrightarrow$ Is it still converging?
Continuous Actions?

Approximate Policy Iteration with Q Functions

Recall the state-action cost-to-go function: $Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) Q_\pi(s',\pi(s'))$

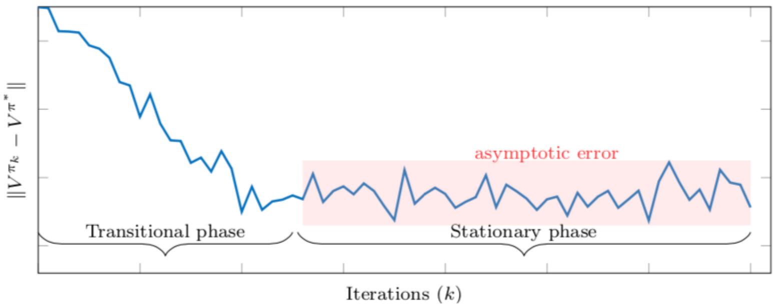
Approximate PI:

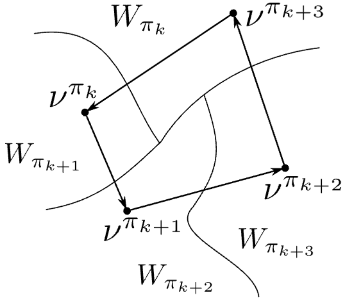
- For $k = 0, 1, 2, \ldots$
  1. Approximate the value under π_k: $Q_{\theta_k} \approx Q_{\pi_k}$
  2. Solve for an improved policy

$$\pi_{k+1}(s) \in \underset{a \in A(s)}{\operatorname{argmin}}\, Q_{\theta_k}(s,a) \qquad \forall s \in \mathcal{S}$$

Q_{π_k} can be approximated by either TD or Monte Carlo methods.

Same story as fitted Q-iteration. No longer guaranteed to converge.





From Policy Iteration to Policy Search

- Approximate a stochastic policy directly using function approximation

$$\pi_\theta : S \rightarrow \mathcal{P}(\mathcal{A}) \text{ with } \theta \in \mathbb{R}^d$$

- Let $V(\pi_\theta)$ denote the policy performance of policy π_θ

➢ Policy optimization problem

$$\max_{\pi_\theta} V(\pi_\theta)$$

Solution 1: **Policy Search/Blackbox optimization**:

Use global optimizers or gradient by finite-difference methods

Policy π_θ can also be not differentiable w.r.t. θ

Solution 2: **Policy gradient optimization**:

Compute the gradient $\nabla_\theta V(\theta)$ and follow the ascent direction

$\nabla_\theta \pi_\theta(s, a)$ should exist

Policy Gradient as Policy Update

Approximate Policy Iteration

$$\pi_{\theta_{k+1}} = \arg\max_{\pi_\theta} Q^{\pi_\theta}\big(s, \pi_\theta(s)\big)$$

Unstable (fast)

No convergence

Policy Gradient

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta V(\theta_k)$$

Smooth, fine control (slow)

Convergence to local optima

1. **How do we compute $\boldsymbol{\nabla_\theta V(\theta)}$?**
2. How quickly do we update (i.e. α_k)?

Outline

1. From Policy Iteration to Policy Search

2. **Policy gradient methods**
   a. REINFORCE
   b. Representing a policy (discrete and continuous!)
   c. Variance reduction (temporal structure and baselines)

3. Actor-critic

Assume: finite-horizon setting

Discount γ excluded to simplify notation.

Policy Gradient (Finite-Horizon)

Given an MDP $M = (\mathcal{S}, \mathcal{A}, p, r, T, \mu)$ and a policy π_{θ_0}. For k = 1,2,…

1. Use π_{θ_k} to collect data τ.
2. Use τ to approximate gradient of:

Maximizing this is ultimately what we desire

$$V(\pi_{\theta_k}) = \mathbb{E}\left[\sum_{t=0}^{T-1} r_t | \pi_{\theta_k}, M\right] = \mathbb{E}_{\tau \sim \mathbb{P}(\tau | \pi_{\theta_k}, M)}[\mathcal{R}(\tau)]$$

where

- μ is an initial state distribution
- $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$ (includes terminal reward) is a trajectory
- $\mathcal{R}(\tau)$ its return (sum of rewards).

3. Update $\theta_{k+1} = \theta_k + \alpha_k \widehat{\nabla_\theta V}(\pi_{\theta_k})$

How?

Policy Gradient (Finite-Horizon)

Policy Gradient Theorem [Williams, 1992; Sutton et al., 2000]

For any finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, p, r, T, \mu)$ and differentiable policy π_θ

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}_{\tau \sim \mathbb{P}(\cdot|\pi, M)} \left[R(\tau) \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s_t, a_t) \right]$$

- Model-free! Why?
- Compare: taking gradient through trajectory-space is difficult

$$\nabla_\theta V(\pi_\theta) = \nabla_\theta \mathbb{E}_\tau[R(\tau)] = \nabla_\theta \int \mathbb{P}(\tau|\pi_\theta, M) R(\tau) d\tau$$

Proof

- The objective is an expectation. Want to compute the gradient w.r.t. θ (simplify notation from: $V(\pi_\theta)$ to $V(\theta)$). First, bring the gradient to the inside.

$$\nabla_\theta V(\theta) = \nabla_\theta \mathbb{E}_\tau[R(\tau)] = \nabla_\theta \int \mathbb{P}(\tau|\pi_\theta, M) R(\tau) d\tau$$

Log trick

$$\nabla_\theta \log \mathbb{P}(\tau|\pi_\theta, M) = \frac{\nabla_\theta \mathbb{P}(\tau|\pi_\theta, M)}{\mathbb{P}(\tau|\pi_\theta, M)}$$

$$= \int \nabla_\theta \mathbb{P}(\tau|\pi_\theta, M) R(\tau) d\tau$$

$$= \int \mathbb{P}(\tau|\pi_\theta, M) \nabla_\theta \log \mathbb{P}(\tau|\pi_\theta, M)\, R(\tau) d\tau$$

$$= \mathbb{E}_\tau[R(\tau) \nabla_\theta \log \mathbb{P}(\tau|\pi_\theta, M)]$$

- Last expression is an unbiased gradient estimator
Just sample $\tau_t \sim \mathbb{P}(\tau|\pi_\theta, M)$, and compute $\hat{g}_t = R(\tau_t) \nabla_\theta \log \mathbb{P}(\tau_t|\pi_\theta, M)$
- Issue: Need to be able to compute & differentiate the density $\mathbb{P}(\tau|\pi_\theta, M)$ w.r.t θ

Proof

Likelihood (with stochastic policies)

$$\mathbb{P}(\tau|\pi_\theta, M) = \mu(s_0)\prod_{t=0}^{T-1}\pi_\theta(a_t|s_t)p(s_{t+1}|s_t,a_t)$$

$$\log\mathbb{P}(\tau|\pi_\theta, M) = \log\mu(s_0) + \sum_{t=0}^{T-1}\log\pi_\theta(a_t|s_t) + \log p(s_{t+1}|s_t,a_t)$$

$$\nabla_\theta\log\mathbb{P}(\tau|\pi_\theta, M) = \cancelto{0}{\nabla_\theta\log\mu(s_0)} + \sum_{t=0}^{T-1}\nabla_\theta\log\pi_\theta(a_t|s_t) + \cancelto{0}{\nabla_\theta\log p(s_{t+1}|s_t,a_t)}$$

→ model free

Alternative proof: likelihood rescaling

- Interested in policy gradient: $\nabla_\Delta V(\theta + \Delta)|_{\nabla=0}$
- Likelihood rescaling

$$V(\theta + \Delta) = \mathbb{E}_{\tau(\theta)}\left[R(\tau(\theta))\frac{\prod_t \pi_{\theta+\Delta}(a_t|s_t)}{\prod_t \pi_\theta(a_t|s_t)}\right]$$

- Apply chain rule to get

$$\nabla_\Delta V(\theta + \Delta)\Big|_{\nabla=0} = \mathbb{E}_{\tau(\theta)}\left[R(\tau(\theta))\sum_t \frac{\nabla\, \pi_\theta(a_t|s_t)}{\pi_\theta(a_t|s_t)}\right]$$

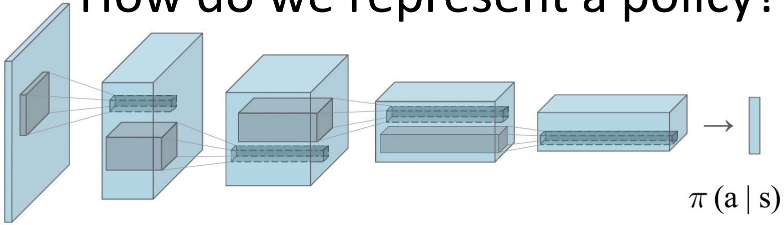
$$= \mathbb{E}_\tau[R(\tau) \textstyle\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t)]$$

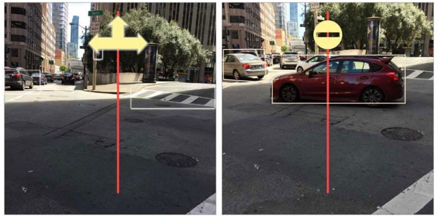
REINFORCE [Williams, 1992]

1. Let π_{θ_1} be an arbitrary policy.
2. At each iteration $k = 1, \dots, K$
   - Sample m trajectories $\tau_i = (s_0, a_0, r_0, s_1, \dots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$ following π_k
   - Compute unbiased gradient estimate:
     $$\widehat{\nabla_\theta V}\left(\pi_{\theta_k}\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\sum_{t=0}^{T-1} r_t^i\right)\left(\sum_{t=0}^{T-1} \nabla_\theta \log \pi_{\theta_k}\left(a_t^i \middle| s_t^i\right)\right)$$
   - Update parameters:
     $$\theta_{k+1} = \theta_k + \alpha_k \widehat{\nabla_\theta V}\left(\pi_{\theta_k}\right)$$
3. Return last policy π_{θ_K}

Policy Gradient: Example

How do we represent a policy?





Normal Policy

$$\pi(a|s) = \frac{1}{\sigma_\omega(s)\sqrt{2\pi}} e^{-\frac{(a-\mu_\theta(s))^2}{2\sigma_\omega^2(s)}}$$

Then:

$$\nabla_\theta \log \pi(a|s) = \frac{(a-\mu_\theta(s))}{\sigma_\omega^2(s)} \nabla_\theta \mu_\theta(s)$$

$$\nabla_\omega \log \pi(a|s) = \frac{(a-\mu_\theta(s))^2 - \sigma_\omega^2(s)}{\sigma_\omega^3(s)} \nabla_\omega \mu_\omega(s)$$

Gibbs (softmax) Policy

$$\pi(a|s) = \frac{e^{\mathcal{K} Q_\theta(s,a)}}{\sum_{a' \in \mathcal{A}} e^{\mathcal{K} Q_\theta(s,a')}}$$

Then:

$$\nabla_\theta \log \pi(a|s) = \mathcal{K} \nabla_\theta Q_\theta(s,a) \\ -\mathcal{K} \sum_{a' \in \mathcal{A}} \pi(a'|s) \nabla_\theta Q_\theta(s,a')$$

Policy Gradient via Automatic Differentiation

- Manually coding the derivative can be tedious
  $\Longrightarrow$ use auto diff
- Define a graph parameterized by θ such that its gradient is the policy gradient

“Pseudo loss”: weighted maximum likelihood

$$\tilde{V} = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{T-1} \log \pi_\theta(s_{i,t}, a_{i,t}) \hat{q}_{i,t}$$

Where:

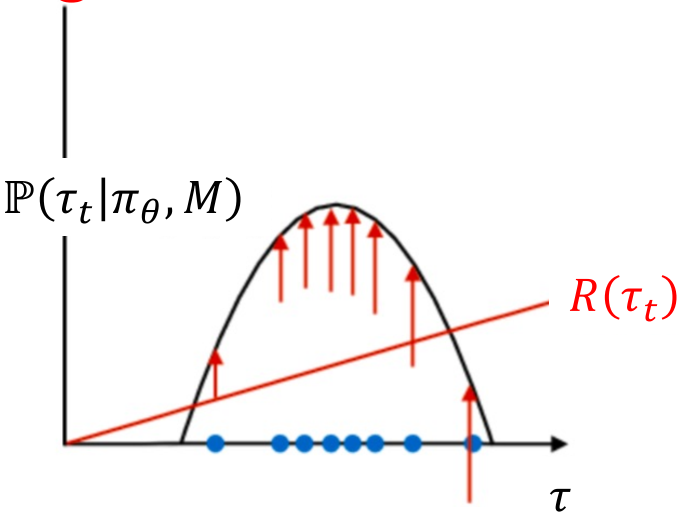
- $\hat{q}_{i,t} = \sum_{k=0}^{T_i} r_k^i$ for REINFORCE and
- $\hat{q}_{i,t} = \sum_{k=t}^{T_i} r_k^i$ for G(PO)MDP.

Note that $\mathbb{E}[\nabla_\theta \tilde{V}] = \nabla_\theta V(\pi_\theta)$.

REINFORCE as Supervised Learning

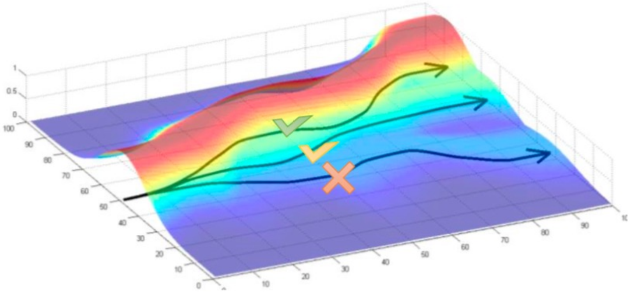
$$\hat{g}_t = R(\tau_t)\nabla_\theta \log \mathbb{P}(\tau_t|\pi_\theta, M)$$

- $R(\tau_t)$ measures how good is sample τ_t
- Moving in the direction of $\hat{g}_t$ pushes up the log probability of the sample in proportion to how good it is.



Interpretation: uses good trajectories as supervised examples

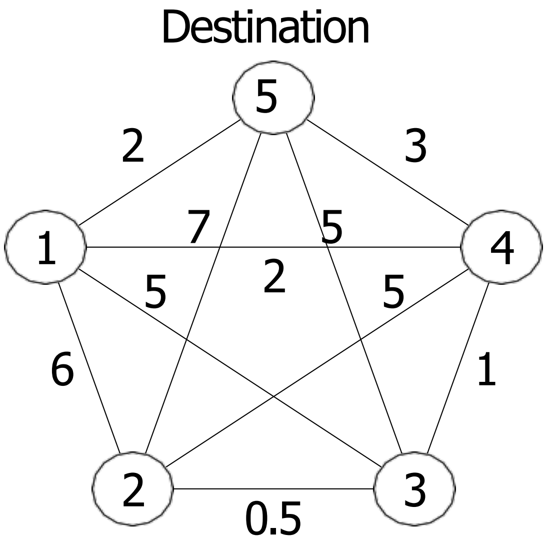
- Like maximum likelihood in supervised learning
- Good stuff are made more likely while bad less
- Trial and Error approach



From “CS 294-112: Deep Reinforcement Learning” slides by S. Levine

Dynamic programming vs policy gradient

How would policy gradient solve shortest path?



Destination is node 5.

REINFORCE

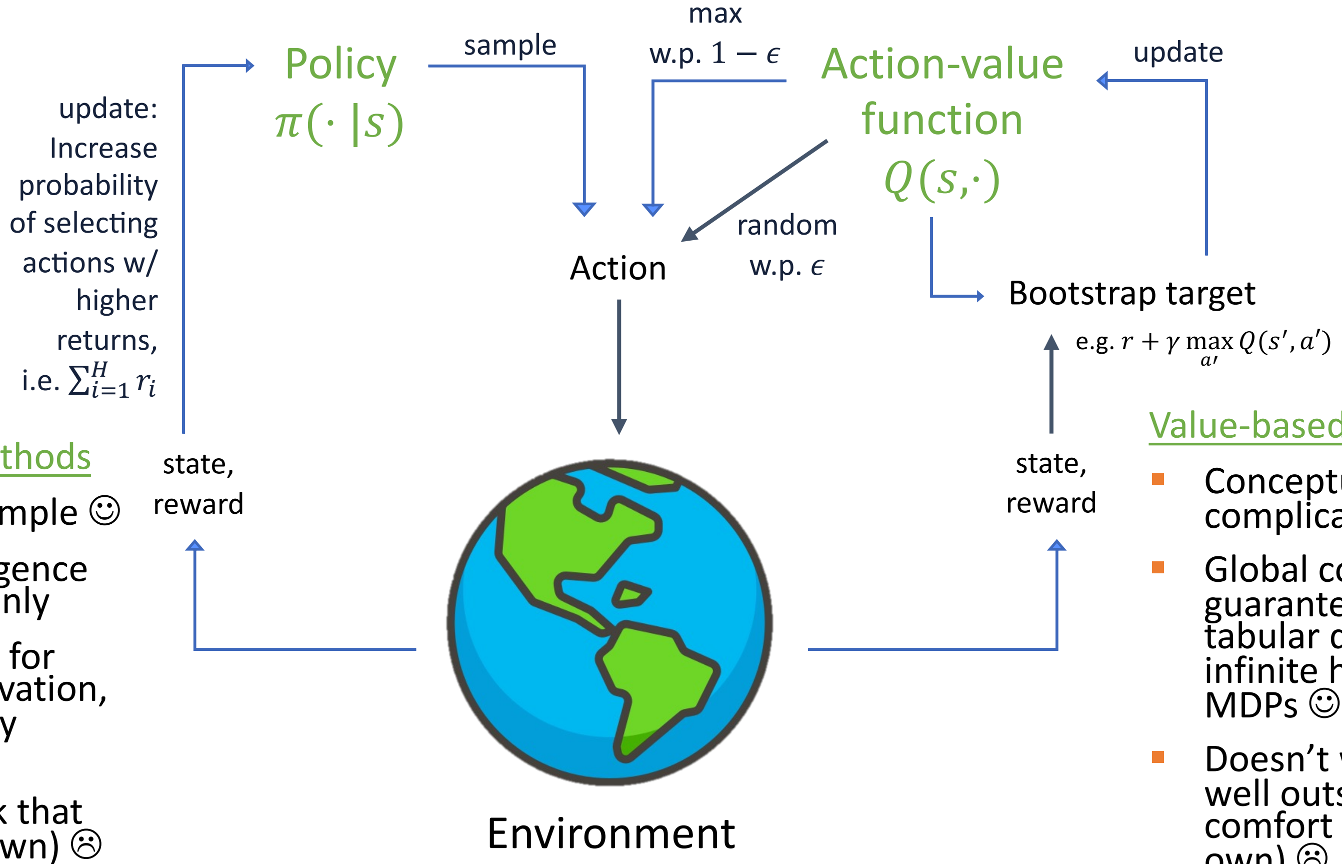
Pros

- Easy to compute
- Does not use Markov property!
- Can be used in partially observable MDPs without modification

Issues

- Use an MC estimate of $Q(s, a)$
- It has possibly a very large variance
- Needs many samples to converge

Policy-based vs value-based methods



update: Increase probability of selecting actions w/ higher returns, i.e. $\sum_{i=1}^{H} r_i$

Policy $\pi(\cdot|s)$

sample

max w.p. $1-\epsilon$

Action-value function $Q(s,\cdot)$

update

random w.p. ϵ

Action

Bootstrap target

e.g. $r+\gamma \max_{a'} Q(s',a')$

state, reward

state, reward

Environment

Policy-based methods

- Sensible & simple ☺
- Local convergence guarantees only
- Also "works" for partial observation, nonstationary settings
- Doesn't work that well (on its own) ☹

Value-based methods

- Conceptually more complicated
- Global convergence guarantees for tabular discounted infinite horizon MDPs ☺
- Doesn't work that well outside of its comfort zone (on its own) ☹

Wu

Policy Gradient: Temporal Structure

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{T-1} r_{t'}\right]$$

Discuss: Why is this better?

Because $\forall t$:

$$\mathbb{E}_{a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s_t) \sum_{t'=0}^{t-1} r_i \,|\tau_{0:t-1}\right] = \left(\sum_{t'=0}^{t-1} r_i\right) \int \pi_\theta(s_t, a) \nabla_\theta \log \pi_\theta(a|s_t) da$$

$$= \left(\sum_{t'=0}^{t-1} r_i\right) \int \nabla_\theta \pi_\theta(a|s_t) da$$

$$= \left(\sum_{t'=0}^{t-1} r_i\right) \nabla_\theta \underbrace{\int \pi_\theta(a|s_t) da}_{:= 1} = 0$$

In literature known as G(PO)MDP [Peters and Schaal, 2008b].

Policy Gradient: Baseline

Discuss: Why does this help with variance?

- Further reduce the variance by introducing a baseline $b(s)$

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(s_t, a_t) \left(\sum_{t'=t}^{T-1} r_{t'} - b(s_t)\right)\right]$$

- The gradient estimate is unbiased.
- “Near optimal choice” that minimize the variance is the expected sum of returns:

$$b^\star(s) \approx \mathbb{E}\left[\sum_{t=0}^{T-1} r_t | s_0 = s, \pi_\theta, M\right] = V^{\pi_\theta}(s)$$

Interpretation: increase the log probability of an action a_t proportionally to how much returns are better than expected (relative values).

Variance reduction via baseline?

Intuition (variance reduction):

$$\mathrm{Var}(x-y)=\mathrm{Var}(x)-2\mathrm{Cov}(x,y)+\mathrm{Var}(y)$$

baseline

$$\nabla_\theta V(\pi_\theta)=\mathbb{E}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(s_t,a_t)\left(\sum_{t'=t}^{T-1} r_{t'}-b(s_t)\right)\right]$$

Optimal Baseline Derivation

Rough Idea

$$\nabla_{\theta_i} V(\pi_\theta) = \mathbb{E}_\tau\left[\underbrace{\nabla_{\theta_i} \log \mathbb{P}(\tau|\pi_\theta)}_{:= g(\tau)}(R(\tau) - b)\right]$$

$$\mathrm{Var} = \mathbb{E}_\tau[(g(\tau)(R(\tau) - b))^2] - (\mathbb{E}_\tau[g(\tau)(R(\tau) - b)])^2$$

$$\Longrightarrow \mathbb{E}_\tau[g(\tau)R(\tau)]^2$$

[Baseline is unbiased in expectation]

$$\frac{\partial}{\partial \mathrm{b}} \mathrm{Var} = \frac{\partial}{\partial \mathrm{b}} \mathbb{E}_\tau[g(\tau)^2(R(\tau) - b)^2]$$

$$= \cancelto{0}{\frac{\partial}{\partial \mathrm{b}} \mathbb{E}_\tau[g(\tau)^2 R(\tau)^2]} - 2\frac{\partial}{\partial \mathrm{b}} \mathbb{E}_\tau[g(\tau)^2 R(\tau) b] + \frac{\partial}{\partial \mathrm{b}} \mathbb{E}_\tau[b^2 g(\tau)^2]$$

$$\Longrightarrow b^\star(\tau) = \frac{\mathbb{E}_\tau[g(\tau)^2 R(\tau)]}{\mathbb{E}_\tau[g(\tau)^2]}$$

Expected return weighted by the magnitude of the gradient.

State-Action baseline (side note)

Several recent methods [Gu et al., 2017, Thomas and Brunskill, 2017, Grathwohl et al., 2018, Liu et al., 2018, Wu et al., 2018] have extended to state-action baselines

$$b(s) \to b(s, a)$$

Going Beyond the Finite-Horizon Case

Theorem

For an infinite horizon MDP (average or discounted), the policy gradient is:

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

- $d_\mu^{\pi_\theta}$ is the stationary distribution
- $\hat{Q}^{\pi_\theta}$ is the state-action value estimate $R(\tau_t)$

Preview to actor-critic methods. See HW to formalize the connection.

Gradient estimate based on m trajectories $(\tau_i)_{i=1}^m$:

$$\overline{\nabla_\theta V}(\pi_\theta) := \frac{1}{m}\sum_{i=1}^{m}\sum_{t=0}^{T_i-1}\gamma^t\nabla_\theta \log \pi_\theta\left(s_t^i, a_t^i\right)\sum_{t'=t}^{T_i}\gamma^{t'-t}r_{t'}^i$$

Where T_i is the length of trajectory τ_i.

Convergence Results

- Policy gradient is stochastic gradient

$$\theta_{k+1} = \theta_k + \alpha_k(\nabla V(\theta_k) + \text{noise})$$

- V is non-convex
- $\Longrightarrow$ converge asymptotically to a stationary point or a local minimum (under standard technical assumptions)

What is the quality of this point?

Dynamics are linear (LQ systems) $\Longrightarrow$ global convergence [Fazel et al., 2018].

- Surprising since $\min_\pi V_{\mathrm{LQ}}(\pi)$ may be not convex, and V_{LQ} is not smooth but is “almost” smooth (far from un/stable boundaries).
- *Hint*: use properties of functions that are gradient dominated.

Convergence Results

Issues

- Non-convexity of the loss function
- Unnatural policy parameterization: parameters that are far in Euclidean distance may describe the same policy (we will talk about this later)
- Insufficient exploration: naïve stochastic exploration
- Large variance of stochastic gradients: generally increases with the length of the horizon

Solution:

$\Longrightarrow$ similar to LQ, we need structural assumptions [Bhandari and Russo, 2019]

See also [Zhang et al., 2019] for convergence results.

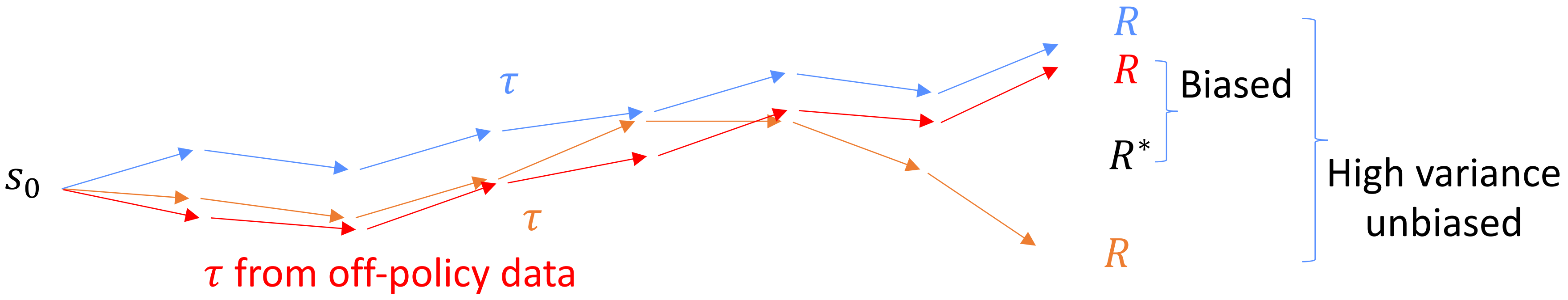
Outline

1. From Policy Iteration to Policy Search

2. Policy gradient methods

3. **Actor-critic**
   a. Compatible function approximation
   b. Advantages and Advantage Actor-Critic (A2C)
   c. Asynchronous A2C (A3C)
   d. Deep Deterministic Policy Gradient (DDPG)
   e. Soft Actor-Critic (SAC)

Policy gradients & high variance: the saga continues



- Monte-Carlo policy gradient is unbiased but still has high variance

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{T-1} r_{t'}\right]$$

- Policy gradient is on-policy (doesn’t re-use data → inefficient!)

Policy- and value-based methods → actor-critic

- Monte-Carlo policy gradient is unbiased but still has high variance

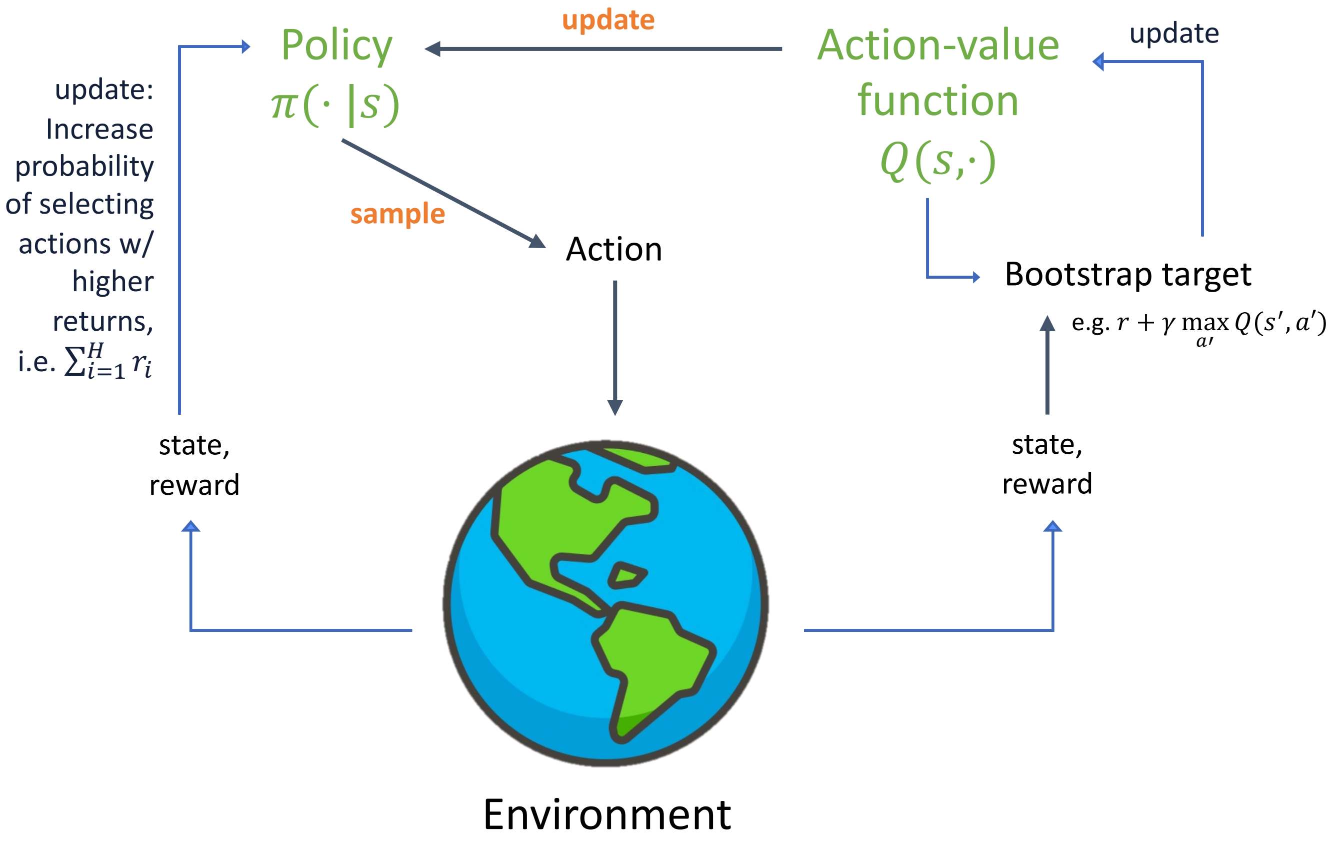
$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \sum_{t'=t}^{T-1} r_{t'}\right]$$

- Incorporate an estimate of $Q^\pi(s,a) \Longrightarrow$ actor-critic
  - Critic: estimate the value function
  - Actor: update the policy in the direction suggested by the critic
- Actor-critic

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t)\right]$$

- These are equivalent (see HW).

Actor-critic methods



Actor-Critic

- Algorithm maintains two sets of parameters: $\theta \longmapsto \pi_\theta, \omega \longmapsto Q_\omega$
- Critic can use $TD(0)$

```
for $t = 0, \ldots, T-1$ do
    $a_t \sim \pi_\theta(s_t, \cdot)$ and observe $r_t$ and $s_{t+1}$
    Compute temporal difference
        $\delta_t = r_t + \gamma Q_\omega(s_{t+1}, a_{t+1}) - Q_\omega(s_t, a_t)$
    Update $Q$ estimate
        $\omega = \omega + \beta\delta_t \nabla_\omega Q_\omega(s_t, a_t)$
    Update policy
        $\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) Q_\omega(s_t, a_t)$
end
```

Actor-Critic

Issues:

- $Q_\omega(s,a)$ is a biased estimate of $Q^{\pi_\theta}(s,a)$
- The update of θ may not follow the gradient of $\nabla_\theta V(\pi_\theta)$

Solution:

- Choose the approximation space $Q_\omega(s,a)$ carefully
  $\Longrightarrow$ compatible function approximation between Q_ω and π_θ

Compatible Function Approximation

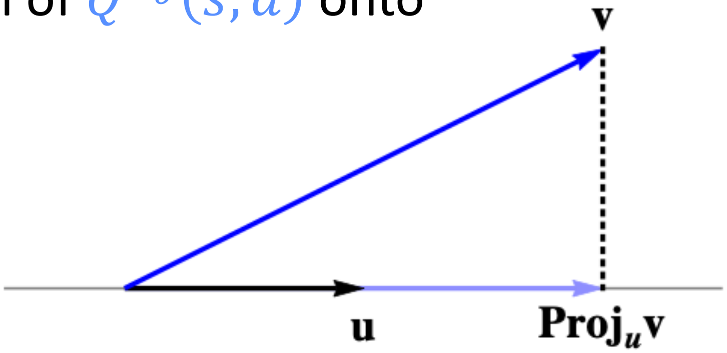
- Actor-critic

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t)\right]$$

- Re-write using occupancy measures

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}} E_{a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s)\, Q^{\pi_\theta}(s, a)]$$

- Interpretation (inner product): projection of $Q^{\pi_\theta}(s,a)$ onto subspace spanned by $\nabla_\theta \log \pi_\theta(a|s)$
- Let $Q_\omega(s,a) = \sum_i \alpha_i [\nabla_\theta \log \pi_\theta(s,a)]_i$ where $\omega = (\alpha_i)_{|\theta|}$



Compatible Function Approximation

Theorem

An action value function space Q_ω is compatible with a policy space π_θ if:

1. $\nabla_\omega Q_\omega(s,a) = \nabla_\theta \log \pi_\theta(s,a)$
2. And if ω minimizes the squared error

$$\omega = \arg\min_\omega \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_a \pi_\theta(a|s)\left(Q^{\pi_\theta}(s,a) - Q_\omega(s,a)\right)^2\right]$$

Then:

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}} E_{a\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q_\omega(s,a)]$$

- Remark 1: conditions for which the policy gradient is exact.
- Remark 2: approximately satisfied by linear function approximation.

Sample Efficiency in Actor-Critic

Issues:

- Sample efficiency is pretty poor
- All samples need to be generated by the current policy (on-policy learning)
- Samples are discarded after a single update

Solutions:

- Variance reduction techniques
- Asynchronous training (A3C)
- Use samples from other policies via importance sampling (not very stable) (next time)
- Conservative approaches (next time)
- Newton for Quasi-newton methods

Actor-Critic with a Baseline

$$\nabla_\theta V(\pi_\theta) = \mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_a \nabla_\theta \pi_\theta(s,a)\big(Q^{\pi_\theta}(s,a) - b(s)\big)\right]$$

- $b(s)$ minimizes the variance
- $V^\pi(s)$ is a good choice as baseline
  - It minimizes the variance in average reward [Bhatnagar et al., 2009]
- $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is the advantage function

Actor-Critic with Advantage Function (A2C)

- It is possible to estimate V^{π} and Q^{π} independently (e.g. by $TD(0)$)
- $A^{\pi} = Q_{\omega} - V_{\nu}$ is a biased and unstable estimate

Solution:

- Consider the temporal difference error

$$\delta^{\pi_\theta} = r(s, a) + \gamma V^{\pi_\theta}(s') - V^{\pi_\theta}(s)$$

- δ^{π_θ} is an unbiased estimate of the advantage

$$\begin{aligned}\mathbb{E}[\delta^{\pi_\theta} | s, a] &= \mathbb{E}[r(s, a) + \gamma V^{\pi_\theta}(s') | s, a] - V^{\pi_\theta}(s) \\ &= Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)\end{aligned}$$

Actor-Critic with Advantage Function (A2C)

- Estimate only $V_v \longmapsto \delta_v = r + \gamma V_v(s') - V_v(s)$

☞ Convergence results with compatible function approximation [Bhatnagar et al., 2009]

```
for $t = 0, \dots, T$ do
    $a_t \sim \pi^\theta(s_t, \cdot)$ and observer $r_t$ and $s_{t+1}$
    Compute temporal difference
        $\delta_t = r_t + \gamma V_v(s_{t+1}) - V_v(s_t)$
    Update $V$ estimate
        $v = v + \beta \delta_t \nabla_v V_v(s_t)$
    Update policy
        $\theta = \theta + \alpha \delta_t \nabla_\theta \log \pi_\theta(a_t|s_t)$
end
```

Compare (actor-critic):

$$\delta_t = r_t + \gamma Q_\omega(s_{t+1}, a_{t+1}) - Q_\omega(s_t, a_t)$$
$$\omega = \omega + \beta \delta_t \nabla_\omega Q_\omega(s_t, a_t)$$
$$\theta = \theta + \alpha \nabla_\theta \log \pi_\theta(a_t|s_t) Q_\omega(s_t, a_t)$$

Asynchronous Advantage Actor-Critic (A3C)

- Multiple independent agents (networks) with their own weights, who interact with a different copy of the environment in parallel.
- The agents (or workers) train in parallel using a global network θ. They periodically update the global network with their $d\theta$.
- Improved training exploration, stability.

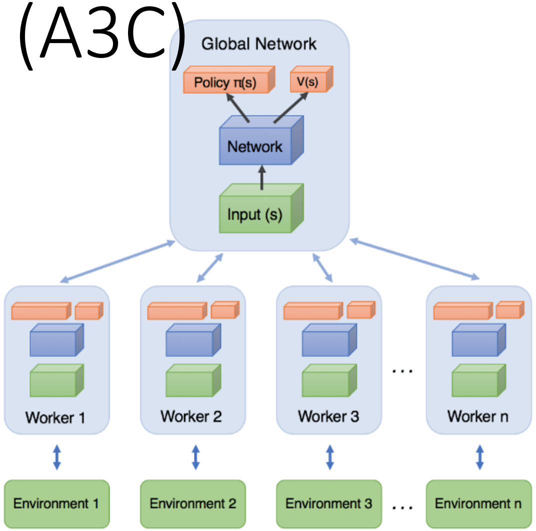
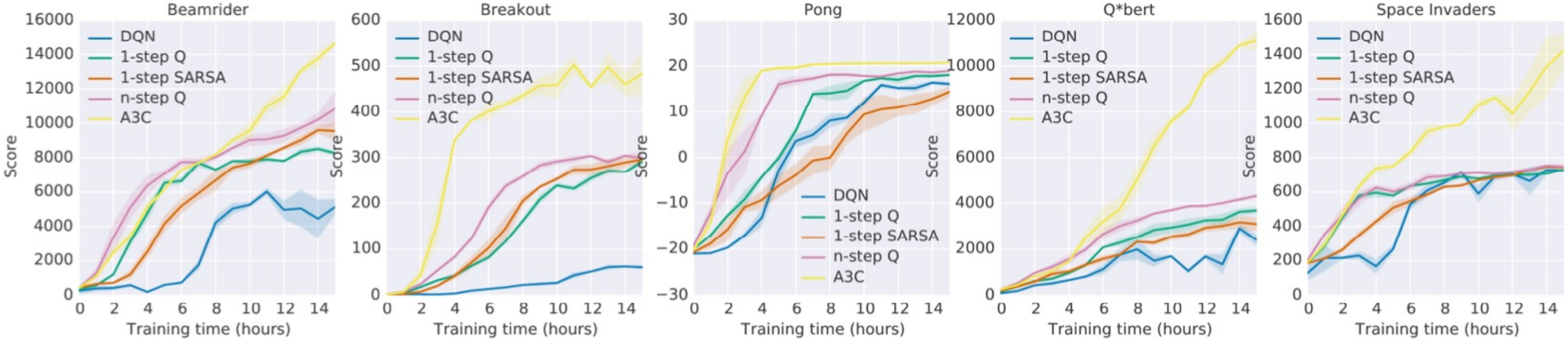


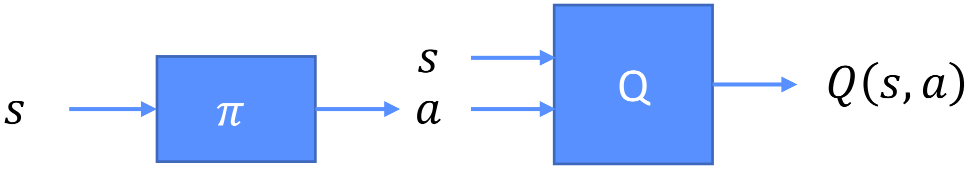
Figure from Atrisha Sarkar

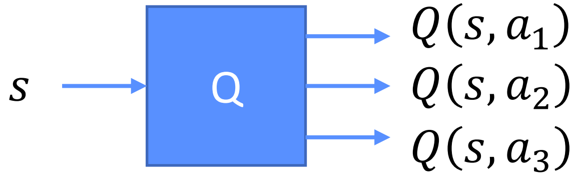


Mnih, Volodymyr, et al. "Asynchronous methods for deep reinforcement learning." *ICML*, 2016.

Bringing policies back to value-based methods

- Recall: value-based methods have trouble handling continuous actions/large action spaces
- Key idea: simplify Q using deterministic policies





Deterministic Policy Gradient (2014)

- Recall: $V_D(\pi) = \mathbb{E}_{s\sim d^\pi}\left[r\left(s,\pi(s)\right)\right]$
- $\nabla_\theta V_D(\theta) = \sum_s d^\pi(s)\nabla_\theta\pi_\theta(s)\nabla_a Q^\pi(s,a)|_{a=\pi_\theta(s)} = \mathbb{E}_{s\sim d^\pi}\left[\nabla_\theta\pi_\theta(s)\nabla_a Q^\pi(s,a)|_{a=\pi_\theta(s)}\right]$

Plug it into an actor-critic framework

- Use $TD(0)$ to update a parametric representation of Q^π

$$\delta_t = R_t + \gamma Q_w(s_{t+1},a_{t+1}) - Q_w(s_t,a_t)$$

; TD error in SARSA

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q_w(s_t,a_t)$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_a Q_w(s_t,a_t)\nabla_\theta\pi_\theta(s)\Big|_{a=\pi_\theta(s)}$$

; Deterministic policy gradient theorem

- Issue: Need to explicitly force exploration, e.g. “behavior policy” $\beta(\cdot) \sim \mathcal{N}(\theta,\sigma\beta^2)$

Soft actor-critic [Haarnoja, 2018]

1. **[Soft policy evaluation]**
   Train the action-value function Q_θ, minimizing:

$$\arg\min_\theta \mathbb{E}_{(s,a)\in H}\left[\frac{1}{2}\left(Q_\theta(s_t,a_t) - \left(r(s_t,a_t) + \gamma\mathbb{E}[V_{\bar{\psi}}(s')]\right)\right)^2\right]$$

   ! Fix the target network (e.g. DQN) → increase stability / break dependences

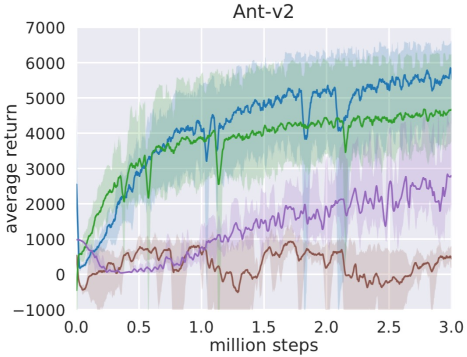
2. Train the value function V_ψ, minimizing:

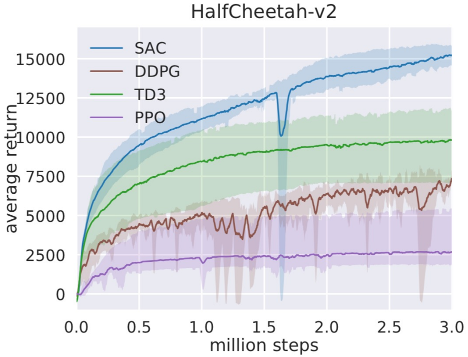
$$J_V(\psi) = \mathbb{E}_{s_t\sim D}\left[\frac{1}{2}\left(V_\psi(s_t) - \mathbb{E}_{a_t\sim\pi_\phi}[\underbrace{Q_\theta(s_t,a_t) - \overbrace{\log\pi_\phi(a_t|s_t)}^{\text{entropy regularization}}}_{\text{soft state value function}}]\right)^2\right]$$

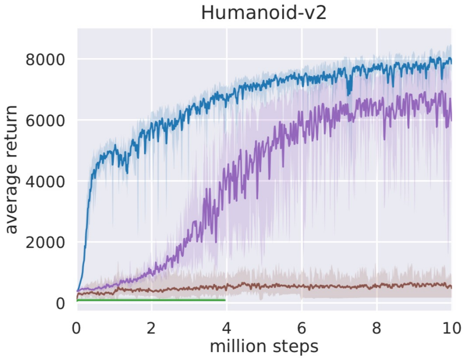
3. **[Soft policy improvement]**
   Fit the new (stochastic) policy π_ϕ:

$$\arg\min_\phi \mathbb{E}_{s\in H}\left[D_{KL}\left(\pi_\phi || \underbrace{\frac{\exp[\eta Q_\theta]}{Z}}_{\text{replace max with softmax}}\right)[s]\right]$$

Soft actor-critic (SAC) [Haarnoja, 2018]







Summary

- Policy gradient methods are an alternative and powerful class of reinforcement learning methods, based on directly optimizing the policy, rather than the value function.
- Policy gradient methods attempt to maximize the likelihood of good trajectories.
- Benefits over *value-function based methods* include not needing Markovian assumption and are often more effective for continuous action space problems.
- Disadvantages: high variance and on-policy (less sample efficient).
- Similar challenges include: exploration vs exploitation.
- A variety of approaches help to reduce variance: temporal structure, baselines, actor-critic methods.
- Core practical policy gradient methods: REINFORCE, SAC, TRPO, PPO. More on these later.