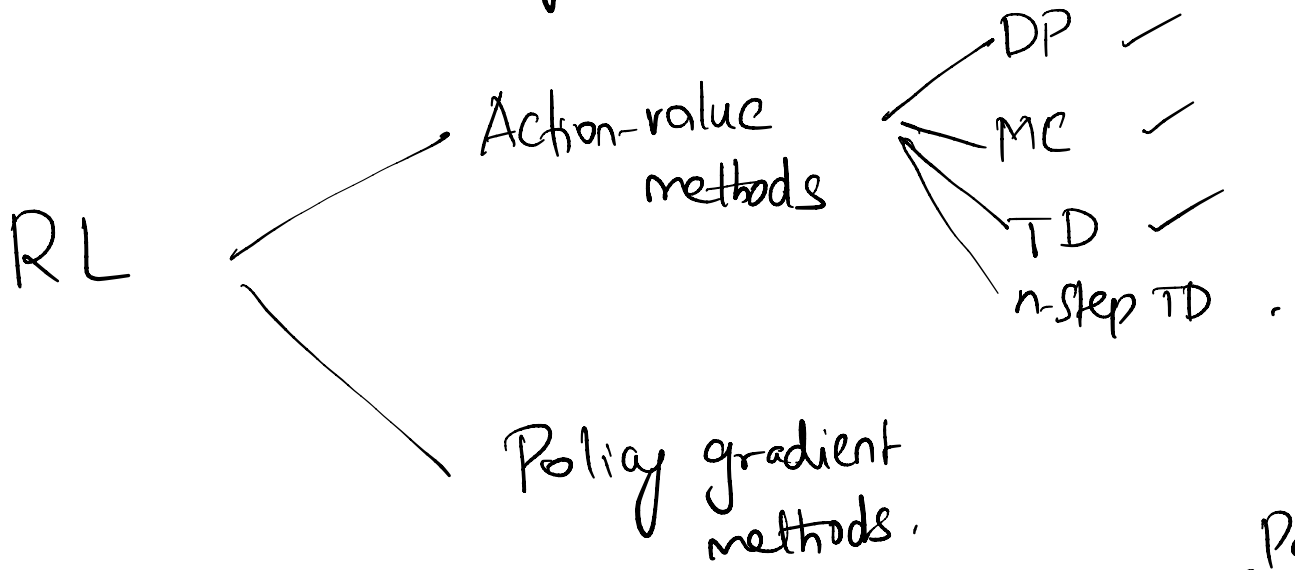
Lecture -08

Policy Gradient Methods



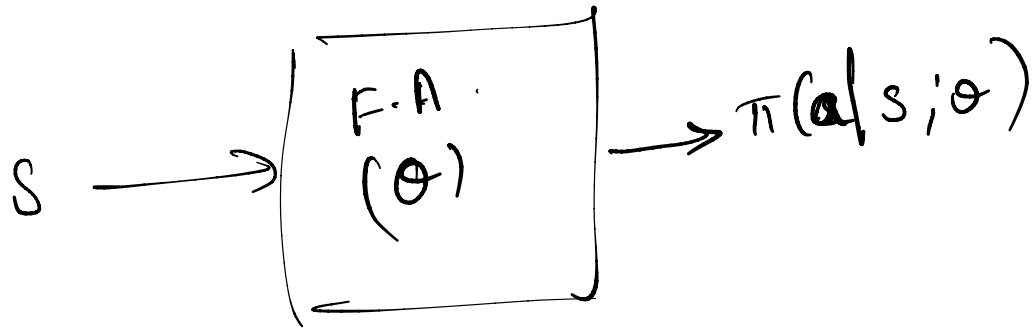
Policy gradient methods :-

$\hat{v}(s; w)$ — w: Parameter of the FA

$$\pi(a|s;\theta) = Pr\{A_t = a \mid S_t = s, \theta_t = \theta\}$$

θ - Parameter for the Policy approximator

w - Param. for the value-fn approx.



Performance measure : $J(\theta) = V_{\pi_\theta}(S_0)$

goal: maximize the perf.

$$\theta_{t+1} = \theta_t + \alpha \widehat{\nabla J(\theta_t)}$$

→ Actor-critic methods

↑ policy ↑ value fn.

Advantages of PG methods:-

numerical pref $h(s,a;\theta) \in \mathbb{R}$

$$\pi(a|s;\theta) = \frac{e^{h(s,a;\theta)}}{\sum_b e^{h(s,b;\theta)}}$$

Softmax in action pref.

(0.999, 0.0001, 0.0001)

$$h(s,a;\theta) = \theta^T x(s,a)$$

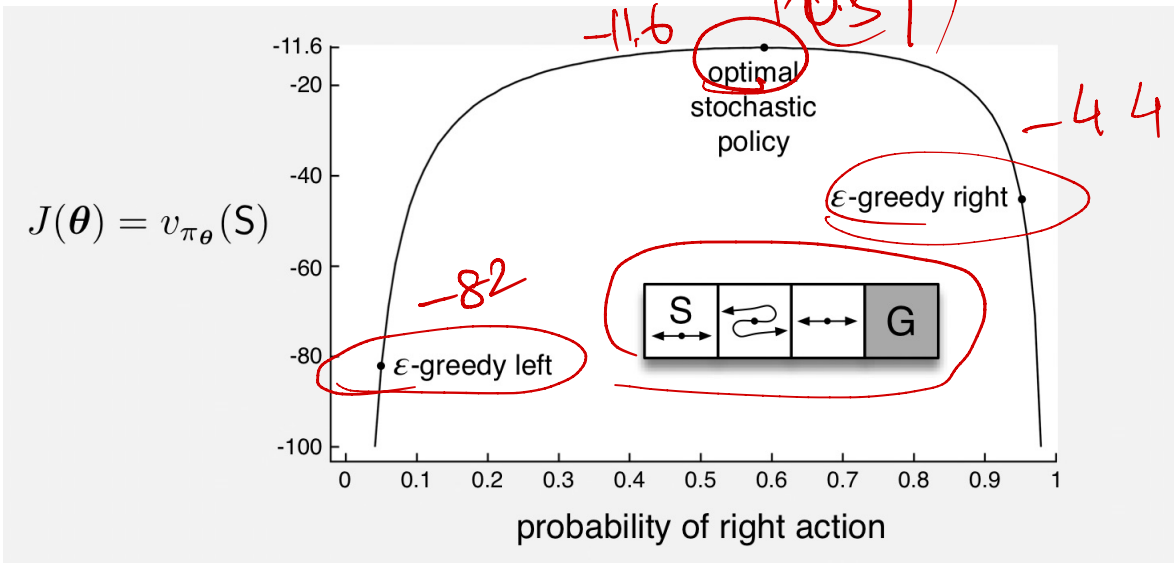
↑ feature vector for s.

① Stochastic policy can approach deterministic policy if needed.

$q(s,a)$

② → enable the selection of actions with arbitrary probs

Reward = -1 all the time



-11.6 (0.59)

-44

-82

$x(S, \text{right}) = [1, 0]^T$

$x(S, \text{left}) = [0, 1]^T$

(3) Policy may be simpler to approximate than value.

(4) Policy parameterization → injecting prior knowledge.

Policy Gradient Theorem :-

(a_1, a_2, a_3, a_4)

$\epsilon = 0.01$

0.99 + ε/4 | ε/4

ε/4 | 0.99 + ε/4

Value based method

Value fn space

Policy grad methods

policy space

$$J(\theta) = V_{\pi_\theta}(s_0)$$

action selection

distribution of states

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s;\theta)$$

$$\propto \mathbb{E}_\pi \left[\sum_a q_\pi(s,a) \nabla \pi(a|s;\theta) \right]$$

Proof: $\gamma = 1$

$$\nabla V_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla \left(\pi(a|s) q_\pi(s,a) \right) \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a)(r + V_\pi(s')) \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s',r} p(s',r|s,a) \nabla V_\pi(s') \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V_\pi(s') \right]$$

$$\nabla V^{\pi}(s) = \nabla \phi(s)$$

$$= \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V_\pi(s') \Big]$$

$p^{\pi}(s \to x, k)$ Prob of reaching state x from s, after k steps.

$k=0$, $p^{\pi}(s \to s, k=0) = 1$

$k=1$, $p^{\pi}(s \to s', k=1) = \sum_a \pi(a|s) p(s'|s,a)$

$s \xrightarrow{k+1} x$ after $k+1$ steps.

$s \xrightarrow{k} s' \to x$

$$p^{\pi}(s \to x, k+1) = \sum_{s'} p^{\pi}(s \to s', k) \cdot p^{\pi}(s' \to x, 1)$$

$$\nabla V^{\pi}(s) = \phi(s)$$

$$\sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V_\pi(s') \right]$$

$$= \phi(s) + \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V_\pi(s')$$

$$= \phi(s) + \sum_{s'} \sum_a \pi(a|s) p(s'|s,a) \nabla V_\pi(s')$$

$$\nabla V_\pi(s) = \phi(s) + \sum_{s'} p^\pi(s \to s', 1) \nabla V_\pi(s')$$

$$= \phi(s) + \sum_{s'} p^\pi(s \to s', 1) \left[\phi(s') + \sum_{s''} p^\pi(s' \to s'', 1) \nabla V_\pi(s'') \right]$$

$$= \phi(s) + \sum_{s'} p^\pi(s \to s', 1) \phi(s') + \sum_{s''} p^\pi(s \to s'', 2) \nabla V_\pi(s'')$$

$$= \phi(s) + \sum_{s'} \rho^{\pi}(s \to s', 1)\phi(s') + \sum_{s''} \rho^{\pi}(s \to s'', 2) \nabla V_{\pi}(s'')$$

$$= \phi(s) + \sum_{s'} \rho^{\pi}(s \to s', 1)\phi(s') + \sum_{s''} \rho^{\pi}(s \to s'', 2)\phi(s'') + \sum_{s'''} \rho^{\pi}(s \to s''', 3) \nabla V_{\pi}(s''')$$

. . . .

$$\nabla J(\theta) = \nabla V^{\pi}(s_0) = \sum_{x \in S} \sum_{k=0}^{\infty} \rho^{\pi}(s \to x, k)\phi(x)$$

$$= \sum_{x \in S} \eta(s)\phi(s)$$

$$= \left(\sum_{s} \eta(s)\right) \sum_{s} \frac{\eta(s)}{\sum_{s} \eta(s)} \phi(s)$$

$$\propto \sum_{s} \frac{\eta(s)}{\sum_{s} \eta(s)} \phi(s)$$

$$\nabla J(\theta) \propto \sum_s \mu(s) \phi(s)$$

$$= \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a)$$

Monte Carlo Policy Gradient :-

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s;\theta)$$

$$= \mathbb{E}_\pi \left[\sum_a q_\pi(s,a) \nabla \pi(a|s;\theta) \right]$$

① $$\theta_{t+1} = \theta_t + \alpha \sum_a \hat{q}(s_t, a; w) \nabla \pi(a|s_t;\theta)$$

② REINFORCE (Williams, 1992)

REINFORCE:

$$\nabla J(\theta) \propto \mathbb{E}_\pi \left[\sum_a q_\pi(S_t, a) \nabla \pi(a|S_t, \theta) \right]$$

$$= \mathbb{E}_\pi \left[\sum_a \pi(a|S_t, \theta) \cdot q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \theta)}{\pi(a|S_t, \theta)} \right]$$

$$= \mathbb{E}_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t; \theta)}{\pi(A_t|S_t, \theta)} \right]$$

$$= \mathbb{E}_\pi \left[G_t \frac{\nabla \pi(A_t|S_t; \theta)}{\pi(A_t|S_t, \theta)} \right]$$

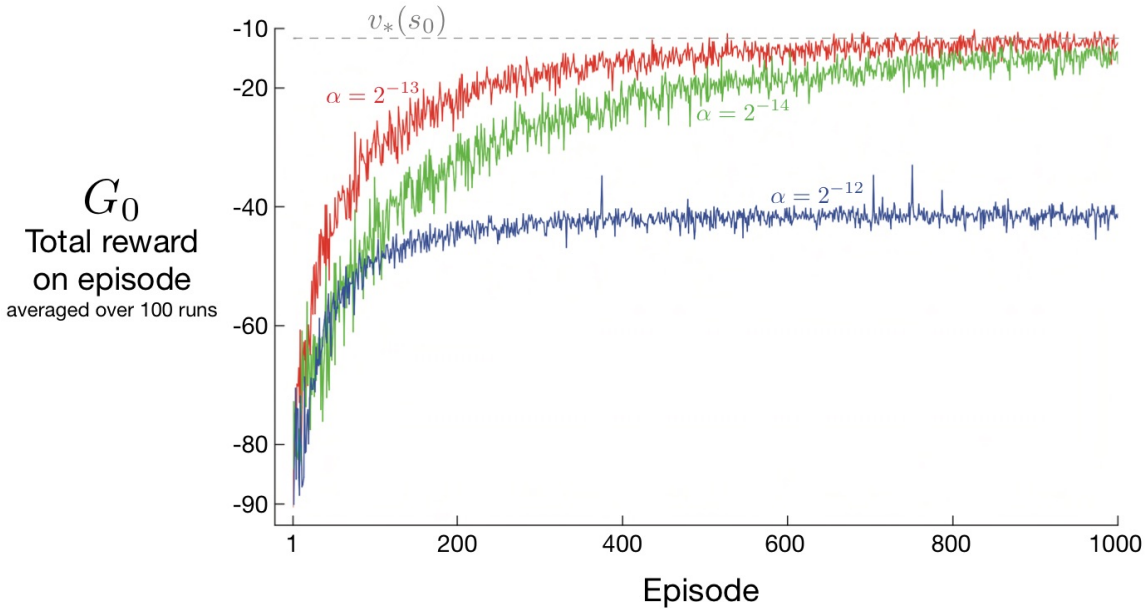
$$\theta_{t+1} = \theta_t + \alpha G_t \frac{\nabla \pi(A_t|S_t; \theta)}{\pi(A_t|S_t; \theta)}$$

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla \ln \pi(A_t|S_t; \theta)$$

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta})$



REINFORCE

$$\theta_{t+1} = \theta_t + \alpha G_t \nabla \ln \pi(A_t | S_t; \theta_t)$$

REINFORCE with Baseline :-

$$\nabla J(\theta) \propto \sum_s \mu(s) \sum_a \left(q_\pi(s,a) - b(s) \right) \nabla \pi(a|s;\theta)$$

$$\sum_a b(s) \nabla \pi(a|s;\theta) = b(s) \sum_a \nabla \pi(a|s;\theta)$$

$$= b(s) \nabla 1 = 0.$$

$$\theta_{t+1} = \theta_t + \alpha \left(G_t - b(S_t) \right) \ln \pi(A_t|S_t,\theta_t)$$

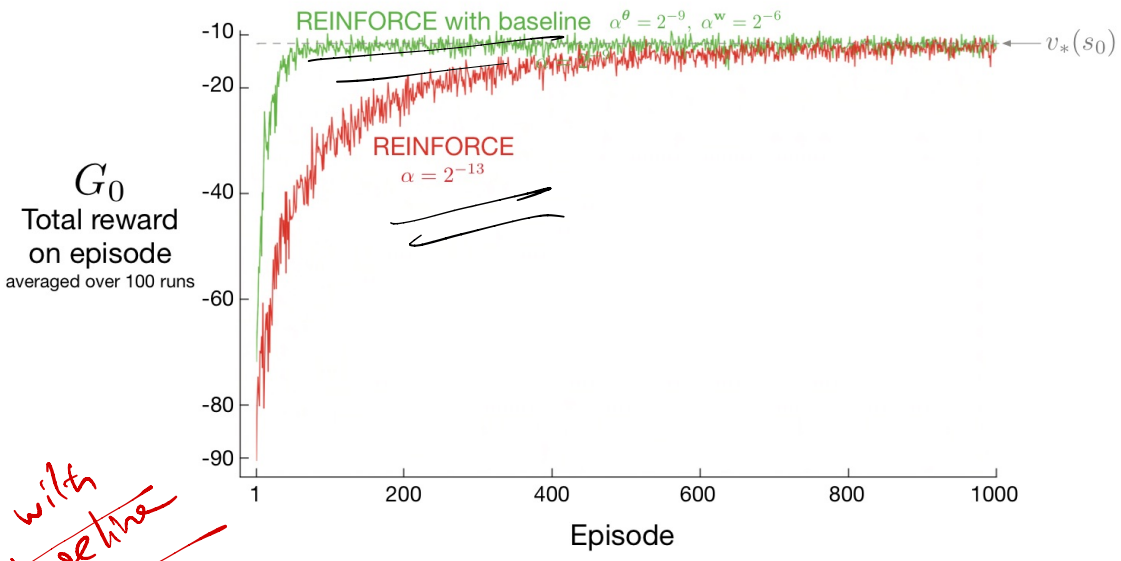
$$\theta_{t+1} = \theta_t + \alpha \left(G_t - b(S_t) \right) \ln \pi \left(A_t | S_t, \theta_t \right)$$

$$= \theta_t + \alpha \left(G_t - \hat{v}(S_t; w) \right) \ln \pi \left(A_t | S_t, \theta_t \right)$$

REINFORCE with Baseline (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^d$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
    $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$



REINFORCE with baseline

$$\theta_{t+1} = \theta_t + \alpha \left(G_t - \hat{v}(S_t, w) \right) \nabla \ln \pi(A_t | S_t, \theta_t)$$

Actor-Critic Method:-

$$\theta_{t+1} = \theta_t + \alpha \left(G_{t:t+1} - \hat{v}(S_t, w) \right) \nabla \ln \pi(A_t | S_t, \theta_t)$$

$$= \theta_t + \alpha \left(R_{t+1} + \gamma \hat{v}(S_{t+1}, w) - \hat{v}(S_t, w) \right) \nabla \ln \pi$$

$$\theta_{t+1} = \theta_t + \alpha \delta_t \nabla \ln \pi(A_t | S_t, \theta_t)$$

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
  Initialize S (first state of episode)
  $I \leftarrow 1$
  Loop while S is not terminal (for each time step):
    $A \sim \pi(\cdot|S,\boldsymbol{\theta})$
    Take action A, observe S', R
    $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S,\boldsymbol{\theta})$
    $I \leftarrow \gamma I$
    $S \leftarrow S'$

PG for episodic problems:-

Continuing pbm setting with no discounting

Average reward setting

$$J(\theta) = r(\pi)$$

$$= \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}\left[R_t \mid S_0, A_{0:t-1} \sim \pi \right]$$

$$= \lim_{t \to \infty} \mathbb{E}\left[R_t \mid S_0, A_{0:t-1} \sim \pi \right]$$

$$= \sum_{s} \mu(s) \sum_{a} \pi(a|s) \sum_{s' r} p(s', r | s, a)\, r$$

where μ — steady state distrib.

$$\mu(s) = \lim_{t \to \infty} Pr\{ S_{t=s} \mid A_{0:t} \sim \pi \}$$

$$\sum_{s} \mu(s) \sum_{a} \pi(a|s,\theta)\, p(s'|s,a) = \mu(s')$$

$$G_t = R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi)$$

$$+ \ldots$$

PGT (Continuing case)

$$J(\theta) = r(\pi)$$

$$\nabla V_\pi(s) = \nabla \left[\sum_a \pi(a|s) q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla q_\pi(s,a) \right]$$

$$= \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \nabla \sum_{s',r} p(s',r|s,a) \left(r - r(\theta) + V_\pi(s') \right) \right]$$

$$\nabla V_\pi(s) = \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \left[-\nabla r(\theta) + \sum_{s'} p(s'|s,a) \nabla V_\pi(s') \right] \right]$$

$$\nabla r(\theta) = \sum_a \left[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla V_\pi(s') \right]$$

$$- \nabla V_\pi(s)$$

$$\nabla r(\theta) =$$

$$\sum_s \mu(s) \sum_a \Big[\nabla \pi(a|s) q_\pi(s,a) + \pi(a|s) \sum_{s'} p(s'|s,a) \nabla v_\pi(s') \Big]$$

$$- \nabla v_\pi(s)$$

$$= \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) +$$

$$\sum_s \mu(s) \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \nabla v_\pi(s')$$

$$- \sum_s \mu(s) \nabla v_\pi(s)$$

$$= \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) +$$

$$\sum_{s'} \sum_s \mu(s) \sum_a \pi(a|s) p(s'|s,a) \nabla v_\pi(s')$$

$$- \sum_s \mu(s) \nabla v_\pi(s)$$

$$= \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) +$$

$$\sum_{s'} \mu(s') \nabla v_\pi(s')$$

$$- \sum_s \mu(s) \nabla v_\pi(s')$$

$$\nabla J(\theta) = \sum_s \mu(s) \sum_a \nabla \pi(a|s) q_\pi(s,a) \quad //$$

on-policy PG algorithms :

Off-policy PG algorithms ?

π — target policy
b — behavior policy.

$$J(\theta) = \sum_s \mu^b(s) \sum_a q_\pi(s,a)\,\pi(a|s;\theta)$$

$$= \mathbb{E}_{s\sim\mu^b(s)}\left[\sum_a \pi(a|s;\theta)\, q_\pi(s,a)\right]$$

↗

Off-Policy PGT. (Degris, White & Sutton, 2012)

$$\nabla J(\theta) = \nabla \mathbb{E}_{s\sim\mu^b}\left[\sum_a \pi(a|s;\theta)\, q_\pi(s,a)\right]$$

$$= \mathbb{E}_{s\sim\mu^b}\left[\sum_a q_\pi(s,a)\nabla\pi(a|s;\theta) + \pi(a|s;\theta)\nabla q_\pi(s,a)\right]$$

$$\approx \mathbb{E}_{s\sim\mu^b}\left[\sum_a q_\pi(s,a)\nabla\pi(a|s;\theta)\right]$$

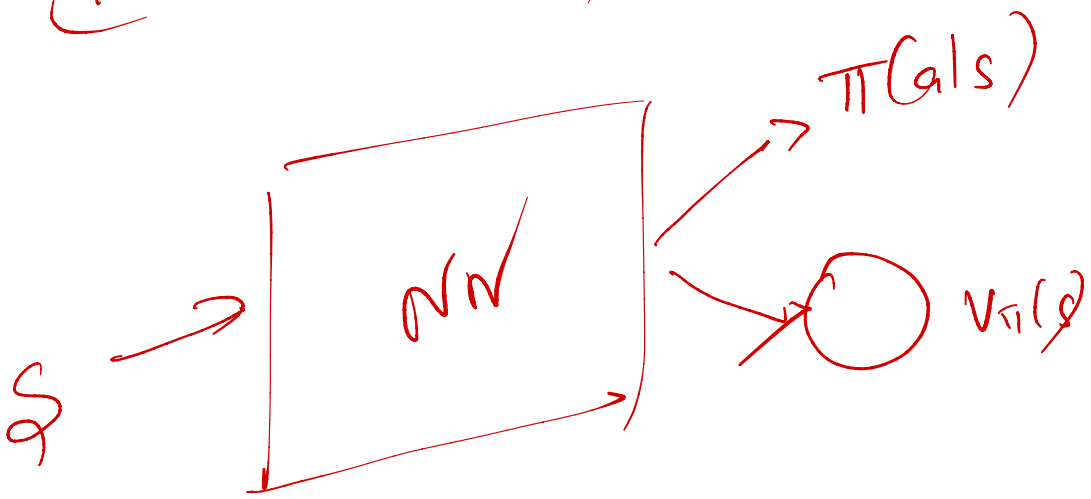
$$= \mathbb{E}_{s \sim \mu^b} \left[\sum_a b(a|s) \frac{\pi(a|s)}{b(a|s)} q_\pi(s,a) \frac{\nabla \pi(a|s)}{\pi(a|s)} \right]$$

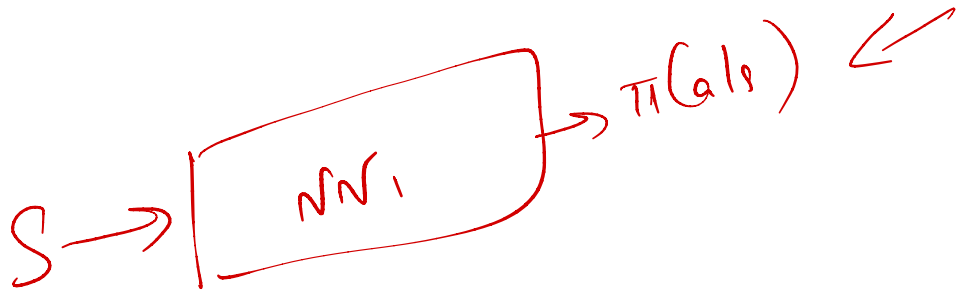
$$= \mathbb{E}_b \left[\frac{\pi(a|s;\theta)}{b(a|s)} q_\pi(s,a) \nabla \ln \pi(a|s) \right]$$

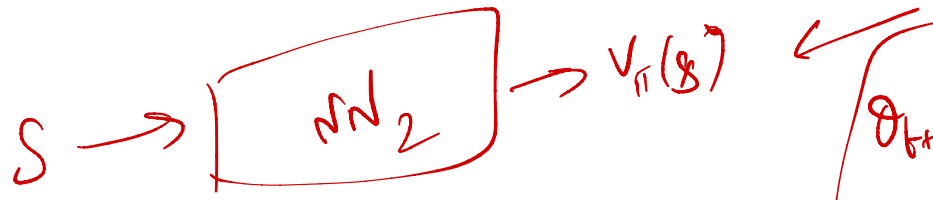
Importance weight.

A3C — Asynchronous Advantage Actor-Critic

(Mnih et al. 2016)







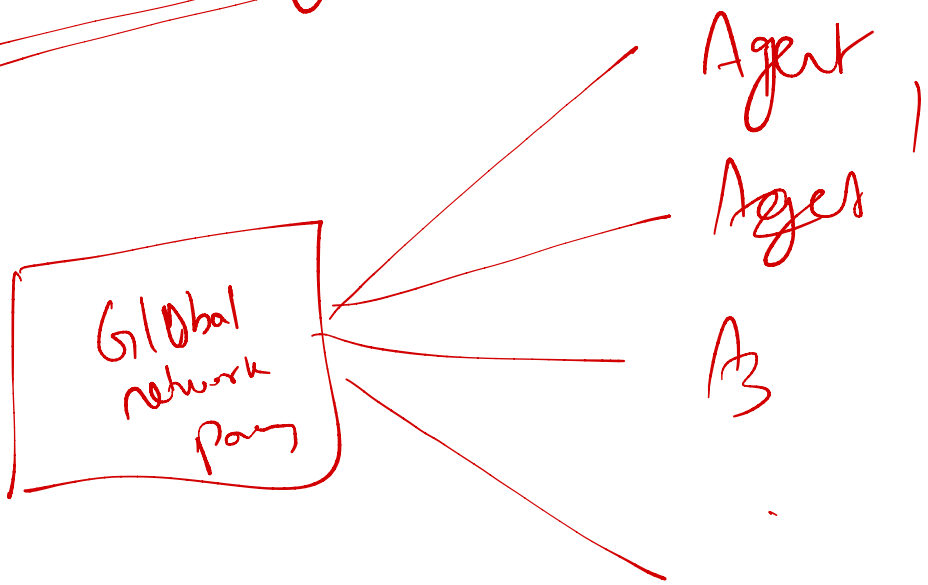
$$\theta_{t+1} = \theta_t + \alpha \left(G_{t:t+1} - \hat{V}(s;w) \right) \nabla \ln \pi$$

$$\theta_{t+1} = \theta_t + \alpha A_t \nabla \ln \pi$$

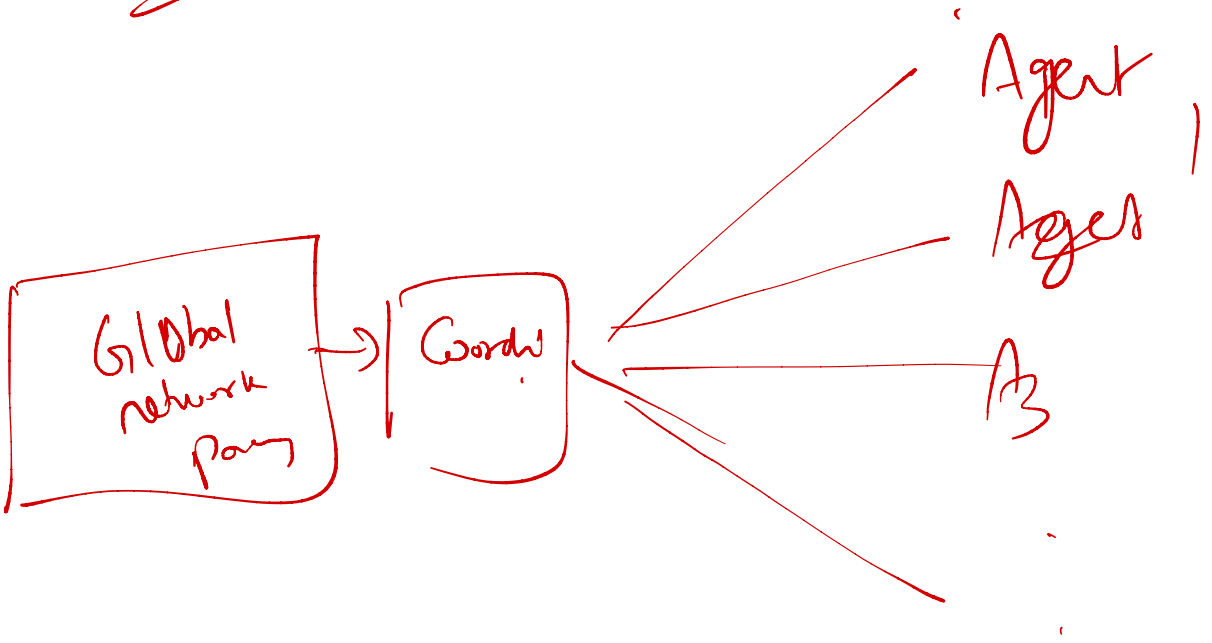
Advantage

$$\mathbb{E}(q(s,a)) - \mathbb{E}(v_\pi(s))$$

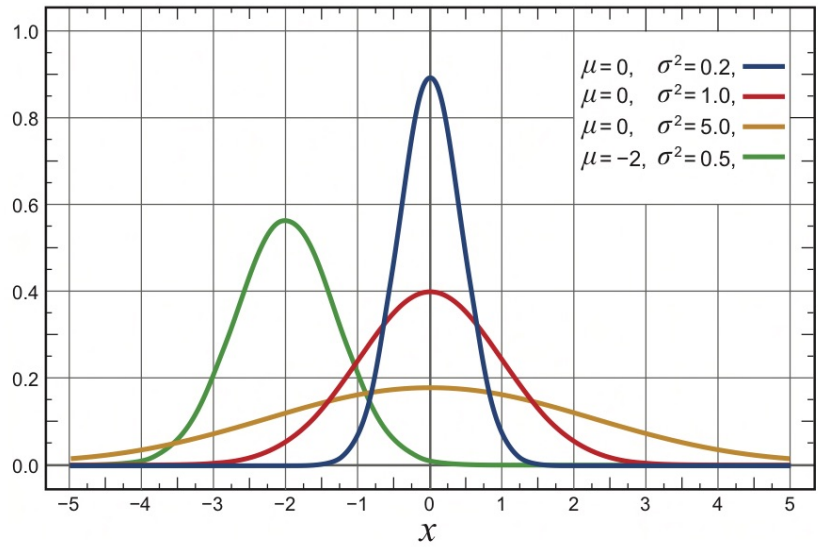
distributed training



A2C



Continuous actions:



$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

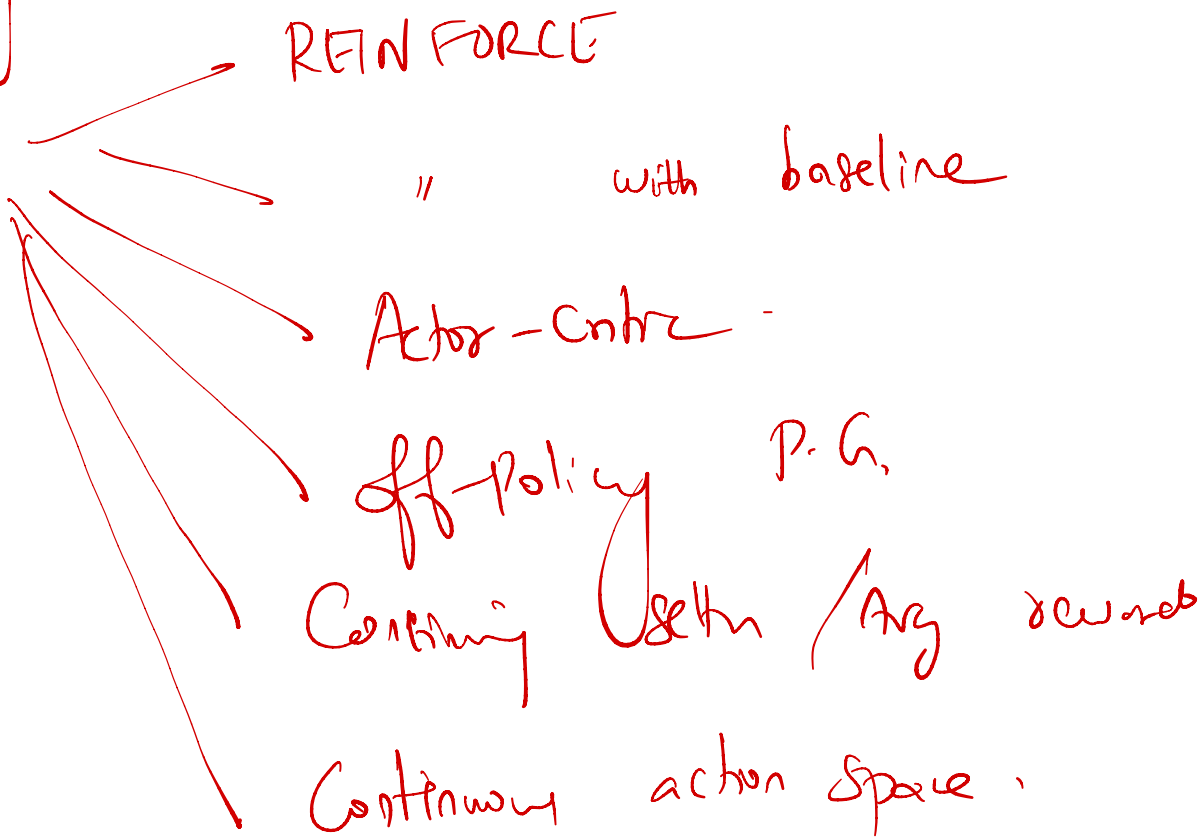
$$\pi(a|s,\theta) = \frac{1}{\sigma(s;\theta)\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s;\theta))^2}{2\sigma(s;\theta)^2}\right)$$

$$\theta = [\theta_\mu, \theta_\sigma]$$

$$\mu(s;\theta) = \theta_\mu^T x_\mu(s)$$

$$\sigma(s;\theta) = \exp\left(\theta_\sigma^T x_\sigma(s)\right)$$

Policy Gradient methods



Algorithm S3 Asynchronous advantage actor-critic - pseudocode for each actor-learner thread.

// Assume global shared parameter vectors θ *and* θ_v *and global shared counter* $T = 0$
// Assume thread-specific parameter vectors θ' *and* θ'_v
Initialize thread step counter $t \leftarrow 1$
repeat
  Reset gradients: $d\theta \leftarrow 0$ and $d\theta_v \leftarrow 0$.
  Synchronize thread-specific parameters $\theta' = \theta$ and $\theta'_v = \theta_v$
  $t_{start} = t$
  Get state s_t
  repeat
    Perform a_t according to policy $\pi(a_t|s_t;\theta')$
    Receive reward r_t and new state s_{t+1}
    $t \leftarrow t + 1$
    $T \leftarrow T + 1$
  until terminal s_t **or** $t - t_{start} == t_{max}$
  $R = \begin{cases} 0 & \text{for terminal } s_t \\ V(s_t, \theta'_v) & \text{for non-terminal } s_t \text{// Bootstrap from last state} \end{cases}$
  for $i \in \{t-1, \ldots, t_{start}\}$ **do**
    $R \leftarrow r_i + \gamma R$
    Accumulate gradients wrt θ': $d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta')(R - V(s_i;\theta'_v))$
    Accumulate gradients wrt θ'_v: $d\theta_v \leftarrow d\theta_v + \partial \left(R - V(s_i;\theta'_v)\right)^2 / \partial \theta'_v$
  end for
  Perform asynchronous update of θ using $d\theta$ and of θ_v using $d\theta_v$.
until $T > T_{max}$