Policy Gradient Method.

Deep-Q-Learning

$$\underline{\hat{Q}_\theta(s,a)} = \cancel{\sum W_i f_i(s,a)}$$

$$\underline{\mathbf{F}(\theta, s, a)}$$

$$\mathcal{L}(\underline{Q} - \underline{\hat{Q}}) =$$

$$\theta \leftarrow \theta - \alpha \frac{\partial \mathcal{L}}{\partial \theta}$$

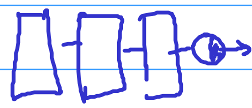
Optimal policy $G_t = \sum_{t}^{T} \gamma^{T-t} \cdot r_t$

$E[G_t]$

$Q \Rightarrow (s, a)$

$\pi_\theta = \underline{P(a|s)}$

π_θ



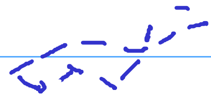
Softmax $\{a_1, a_2 \ldots a_3\}$

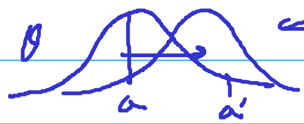
$\sim P(y=c)$

$c = \{c_1 \cdots c_N\}$

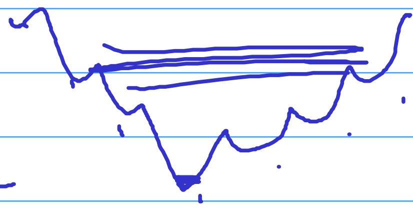
$P(a|s)$

ε-greedy $\pi \to$ random





PG : stronger "convergence" gauranteed



$$\pi_\theta = F_\theta(s)$$

$$\rightarrow \underline{J_\theta} = \underset{\tau \sim \pi_\theta}{E}[r(\tau)]$$

, τ : trajectory, rollout in an episode.

$\int P \cdot r(\tau) d\tau$ $\quad P(a|s)$

$$= \int_\tau \pi_\theta(\tau) \cdot r(\tau) d\tau$$

$$\nabla_\theta J_\theta = \int_\tau \underline{\nabla_\theta \pi_\theta} \, r(\tau) d\tau$$

$$x \, d\log x = dx$$

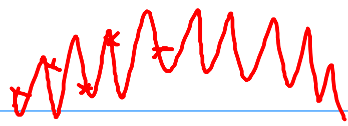
$$= \int \pi_\theta \boxed{(\nabla_\theta \log \pi_\theta) \cdot r(\tau)} d\tau = \underset{\tau \sim \pi_\theta}{E}[\nabla_\theta \log \pi_\theta \cdot r(\tau)]$$

$$= \frac{1}{N} \sum_i^N \left(\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta (a_{it} | s_{it}) \right) \cdot \left(\sum_{t=0}^{T} r(s_{it}, a_{it}) \right)$$

MC PG

REINFORCE

- gradient is noisy
- high variance



For e in Sample N episodes following π_θ

For t in $(0 \cdots T-1)$

$\boxed{G_t} \leftarrow \sum_{k=0}^{T} \gamma^{k} R_k$ t

$\theta \leftarrow \theta + \alpha \underline{\nabla_\theta J}$ $\cancel{(x)}$

$J_\theta = \mathbb{E}_{\tau\sim\pi}[r(\tau)]$

$\boxed{\nabla_\theta J} = G \cdot \nabla_\theta \log \pi_\theta$

$\mathbb{E}_{\tau\sim\pi}[\underbrace{\nabla_\theta (\log \pi_\theta)}_{g(\tau)} \cdot r(\tau)]$

$\boxed{\underbrace{= \mathbb{E}[g^2(\tau) r^2(\tau)] - \mathbb{E}[g \cdot r]^2}_{\mathrm{Var}(g(\tau) r(\tau))}}$

1) Causality

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 $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\sum_{t=0}^{T} r = \cancel{\sum_{t=0}^{t} r} + \sum_{t}^{T} r$ reward to go (G_t)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha\gamma^t G \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$

$\nabla_\theta J_\theta \propto G_t$

2) baseline

$J = \mathbb{E}_{\tau\sim\pi_\theta}[r(\tau)] \rightarrow \mathbb{E}_{\tau\sim\pi_\theta}[r(\tau) - b]$

$\nabla_\theta J \rightarrow$ Policy gradient $\nabla_\theta \log \pi_\theta$ or $\nabla_\theta \pi_\theta$.

$\nabla_\theta J = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta \cdot (r - b)] = \mathbb{E}_{\tau\sim\pi_\theta}[\nabla_\theta \log \pi_\theta \cdot r]$

$= \mathbb{E}_{\tau\sim\pi}[\nabla_\theta \log \pi_\theta \, r] - b \underbrace{\mathbb{E}_{\tau\sim\pi}[\nabla_\theta \log \pi_\theta]}_{\to 0}$

$b \hookrightarrow \frac{1}{N}\sum_{i}^{N} r_i(\tau)$ practical

$$\mathrm{Var}\left[\underbrace{\nabla_\theta \log \pi_\theta}_{g(\tau)(r(\tau)-b)} (r-b)\right]$$

$$= \underset{\tau\sim\pi}{E}\left[g^2 (r-b)^2\right] - \underset{\tau\sim\pi}{E}\left[g\cdot(r-b)\right]^2$$

$$E[g(r-b)] = E[g\,r]$$

$$\mathrm{Var} = \cancel{E[g^2 r^2]} - E[2\cdot g\cdot r\cdot b] - \cancel{E[g\cdot r]^2} + E[g^2 b^2]$$

$$b \quad \frac{\partial \mathrm{Var}}{\partial b} = 0 = -\frac{\partial}{\partial b} E[2\cdot g\cdot r\cdot b] + \frac{\partial}{\partial b} E[g^2 b^2] = 0$$

$$-E[\cancel{2}\cdot g\cdot r] + 2\cancel{b}\, E[g^2] = 0$$

$$b = \frac{E[g(\tau)^2\cdot r(\tau)]}{E[g(\tau)^2]}$$