# From Bandits to PPO

## I. Bandit

* action value

$$q_\pi(a) = \mathbb{E}(R_t \mid A_t = a)$$

Hedge UCB

* softmax policy

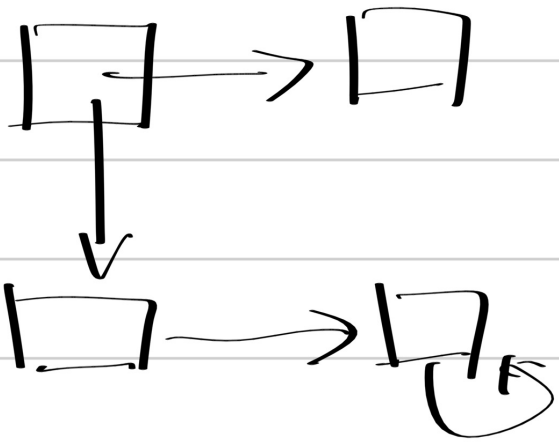
$$\pi_{t+1}(a) = \frac{\exp(H_t(a))}{\sum_b \exp(H_t(b))}$$

* Preference Update Rule

$$J(\pi) = \mathbb{E}[R_t] = \sum \pi_t(a) q_\pi(a)$$

? contexual bandit

$\pi$ state
$*$ action
policy



(immediate) reward

{ discounted return
  average return

$$V(s) \leftarrow V_t + \sum_{s',r} p(s',r|s,a)(\gamma v(s') + r)$$

$$\square = \left[\ \right]\left[\ \right] + \left[\ \right]$$

$$* \quad v = \gamma P v + r$$

$$(I_n - \gamma P) v = r$$

inverse -

* contraction mapping

III policy iteration
value iteration

* Bellman Optimality Equation

$$v^*(s) = \max \sum_{s', r} p(s', r | s, a)(\gamma v^*(s') + r)$$

X Essence : DP

+

Numerical Linear Algebra

You have to know "probability"

Model - Based.

Curse of Dimensionality

↳ To update "s", we have to [Full Sweep] state space.

# IV Monte Carlo & TD learning

— Model free

— Finite states

"Tabular"

How to evaluate

## * Monte Carlo

sample a complete trajectory

$$V(S_t) \leftarrow V(S_t) + \alpha [G_t - V(S_t)]$$

## * Temperal Difference learning

$$V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$$

✓ Q-Learning.

$Q(s,a) \leftarrow Q(s,a) + \alpha[R_{t+1} + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

$\gamma \max_{a'} Q(s',a')$

↑

so have to be tabular

✗ DQN

Couple "Sample" with "Function Approximation"

$$Loss(\theta) = \mathbb{E}_{s,a,r,s'}\left[r_t + \gamma Q_{\theta^-}(s',a') - Q_\theta(s,a)\right]^2$$

$\theta^-$ For stability

# VI. Policy Gradient

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{s_t, a_t \sim \pi_\theta} Q^\pi(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t)$$

$$J_\theta = \int_\tau R(\tau) P(\tau|\theta) d\tau$$

$$P(\tau|\theta) = \rho(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \pi_\theta(a_t|s_t)$$

$$\nabla_\theta J(\theta) \overset{\nabla_\theta}{=} \int_\tau R(\tau) P(\tau|\theta) d\tau = \int_\tau \nabla_\theta R(\tau) P(\tau|\theta) d\tau$$

$$= \int_\tau R(\tau) \left( \nabla_\theta P(\tau|\theta) \right) d\tau$$

$$= \int_\tau R(\tau) \left( \nabla_\theta \ln P(\tau|\theta) \right) P(\tau|\theta) \, d\tau$$

$$= \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau) \nabla_\theta \left( \ln P(\tau|\theta) \right) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} R(\tau) \nabla_\theta \left( \sum_{t=0}^{\infty} P(s_{t+1}|s_t, a_t) \pi(a_t|s_t) \right) P(s_0) +$$

What to optimize

→ value based.

→ policy - based

# VII.1 Reinforce.

use Monte Carlo to estimate $Q$

Monte Carlo Return