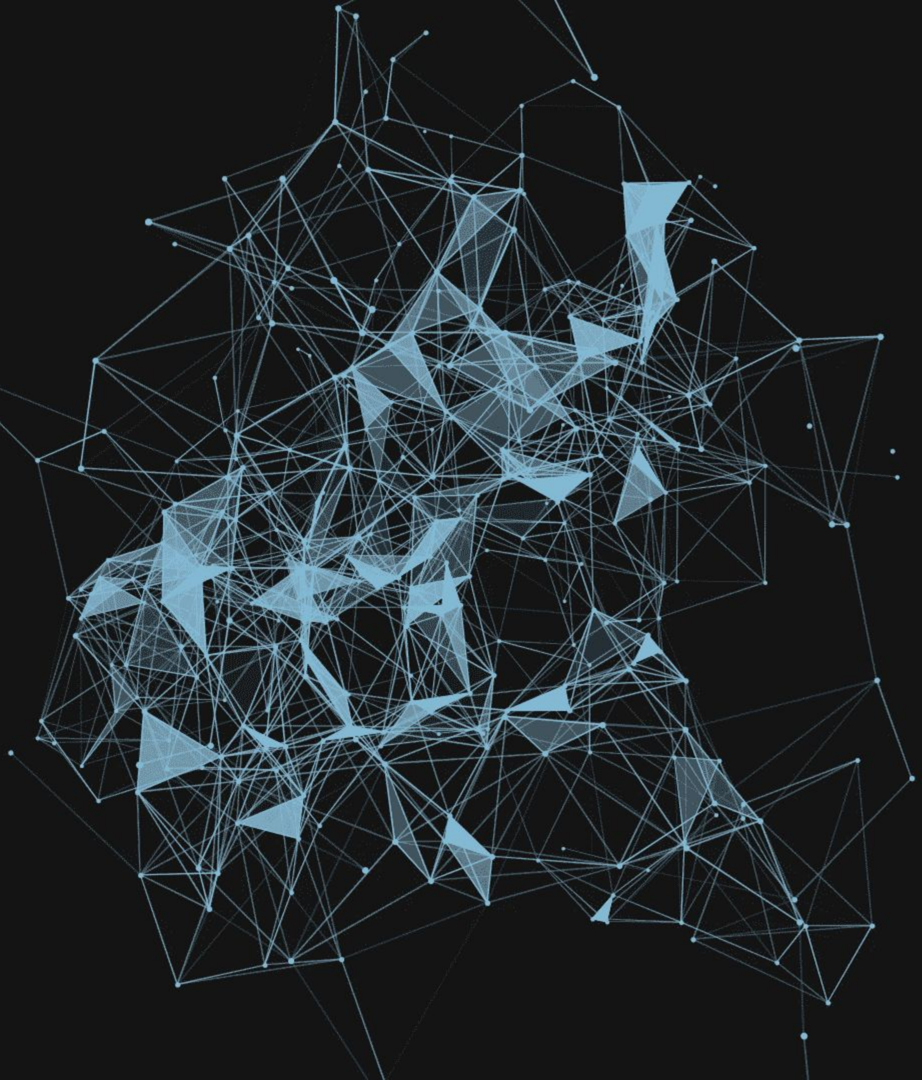
Reinforcement Learning Lesson - 6



Value-based and Policy-based RL

Value Based

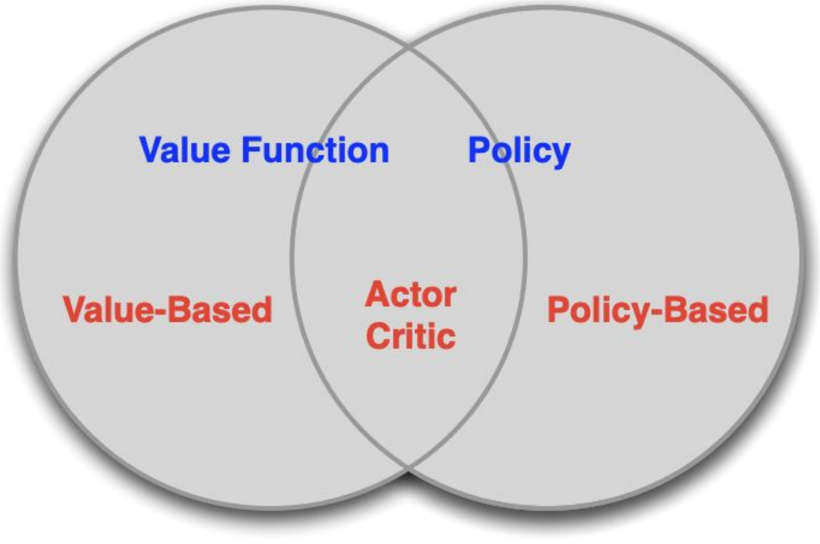
- Learnt Value Function
- Implicit policy (e.g. -greedy)

Policy Based

- Value Function Optional
- Learnt Policy

Actor-Critic

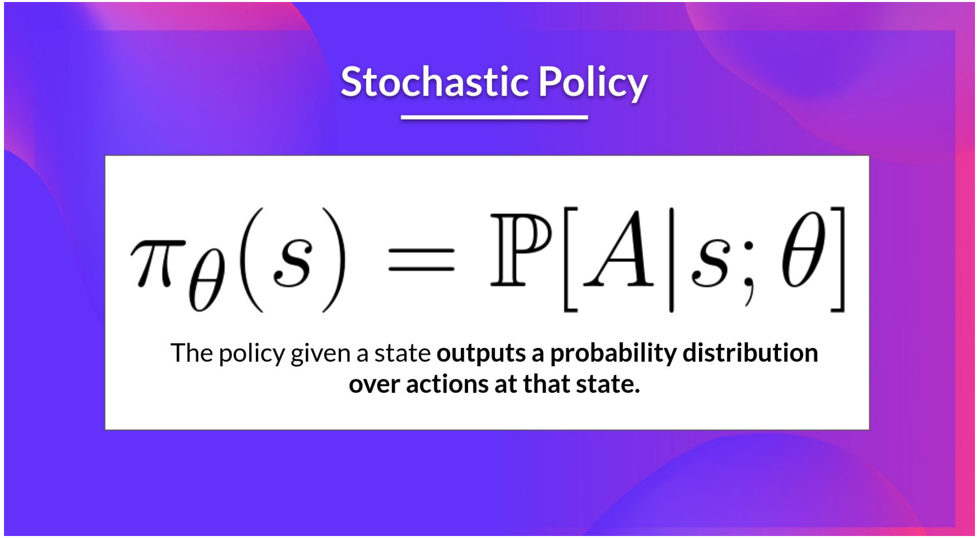
- Learnt Value Function
- Learnt Policy



https://www.davidsilver.uk/wp-content/uploads/2020/03/pg.pdf

Policy-based RL

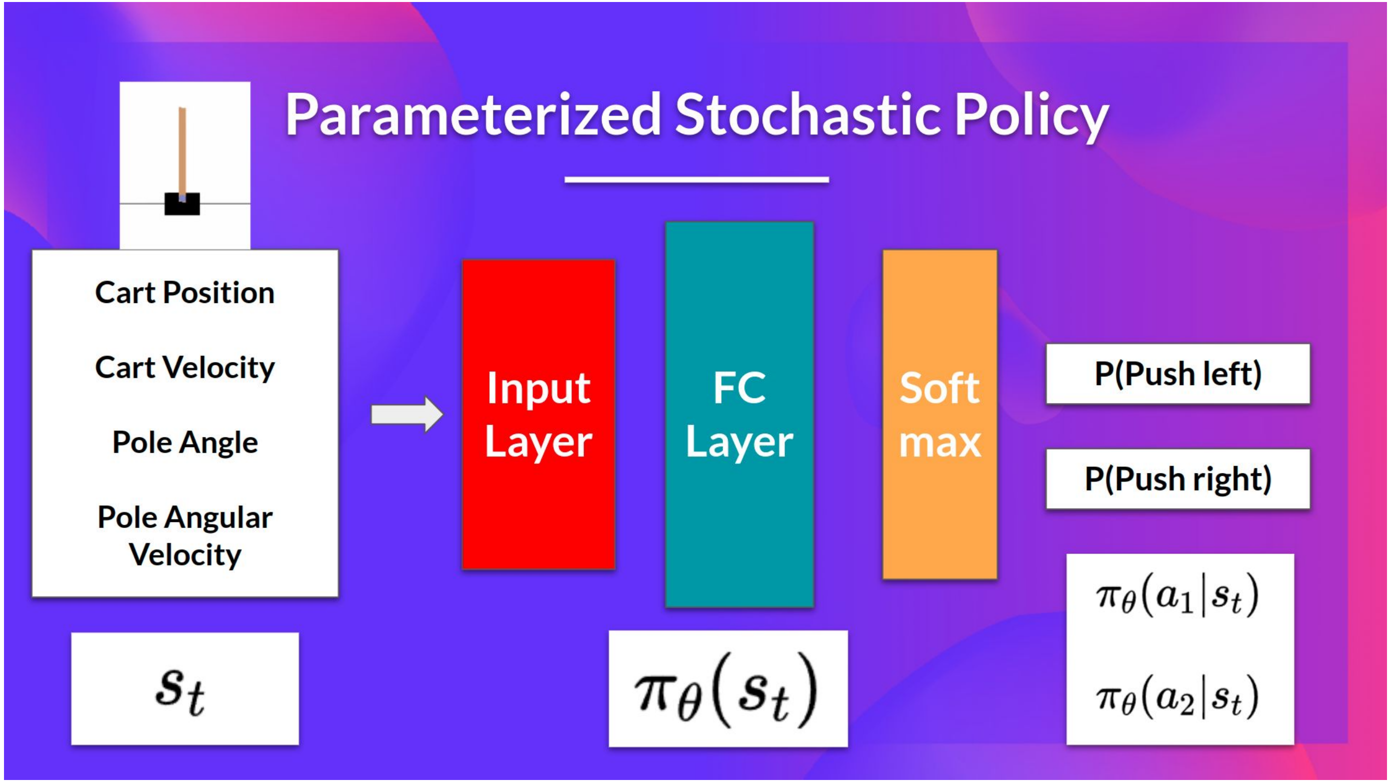
- The idea is to parameterize the policy.
- Policy will output a probability distribution over actions (stochastic policy).



Maximize the performance of the parameterized policy using **gradient ascent**.

https://huggingface.co/learn/deep-rl-course/unit4/what-are-policy-based-methods

Policy-based RL



https://huggingface.co/learn/deep-rl-course/unit4/what-are-policy-based-methods

Policy Gradient

Training Loop:

Collect an **episode with the π** (policy).

Calculate the return (sum of rewards).

Update the weights of the π:

If positive return → **increase** the probability of each (state, action) pairs taken during the episode.

If negative return → **decrease** the probability of each (state, action) taken during the episode

https://huggingface.co/learn/deep-rl-course/unit4/what-are-policy-based-methods

Likelihood Ratio Policy Gradient

$$U(\theta) = \sum_{\tau} P(\tau;\theta) R(\tau)$$

Taking the gradient w.r.t. θ gives

$$\begin{aligned}
\nabla_\theta U(\theta) &= \nabla_\theta \sum_{\tau} P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} \nabla_\theta P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta) R(\tau) \\
&= \sum_{\tau} P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \\
&= \sum_{\tau} P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau)
\end{aligned}$$

Approximate with the empirical estimate for m sample paths under policy π_θ:

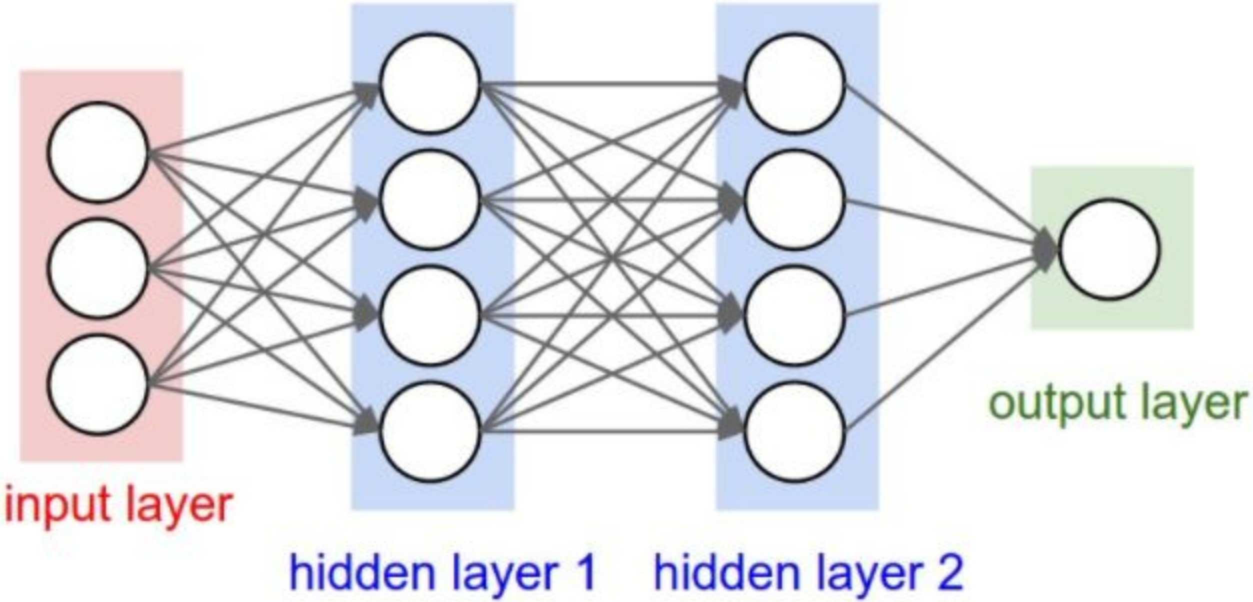
$$\nabla_\theta U(\theta) \approx \hat{g} = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log P(\tau^{(i)};\theta) R(\tau^{(i)})$$

[Aleksandrov, Sysoyev, & Shemeneva, 1968]
[Rubinstein, 1969]
[Glynn, 1986]
[Reinforce, Williams 1992]
[GPOMDP, Baxter & Bartlett, 2001]

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


input layer
hidden layer 1
hidden layer 2
output layer