# Fundamentals of Reinforcement Learning

Master IASD, Université PSL

https://www.di.ens.fr/olivier.cappe/Courses/IASD-FoRL/

February 2022

PSL UNIVERSITÉ PARIS

# Roadmap

# Importance Sampling

In theory, one could estimate the value of a policy $\nu$ from a different exploration policy $\pi$ by importance sampling based on

$$v_\nu(s) = \mathbb{E}_\nu\left(\sum_{t=0}^{\infty}\gamma^t R_{t+1}\middle| S_0 = s\right) = \mathbb{E}_\pi\left(\prod_{i=0}^{\infty}\frac{\nu(A_i|S_i)}{\pi(A_i|S_i)}\sum_{t=0}^{\infty}\gamma^t R_{t+1}\middle| S_0 = s\right)$$

$$= \mathbb{E}_\pi\left(\sum_{t=0}^{\infty}\gamma^t \prod_{i=0}^{t}\frac{\nu(A_i|S_i)}{\pi(A_i|S_i)} R_{t+1}\middle| S_0 = s\right)$$

But with high variability, as the variance under $\mathbb{P}_\pi$ of the importance weigths $W_t = \prod_{i=0}^{t}\nu(A_i|S_i)/\pi(A_i|S_i)$ typically diverges exponentially in $t$.

# Parameterized Policies

A more robust idea, which can be traced back to the likelihood ratio method of [Glynn, 1990], consists in using importance sampling to estimate the gradient of the value function.

This requires considering parameterized policies.

## Softmax policy

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})}$$

Outside of the finite state (or "tabular") case this requires to use features to restrict the space of investigated policies:

$$\pi_\theta(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'} \exp(f_\theta(s,a'))}$$

where $f_\theta(s,a) = \langle \theta, \phi_{s,a} \rangle$ corresponds to log-linear policies.

# Policy Gradient

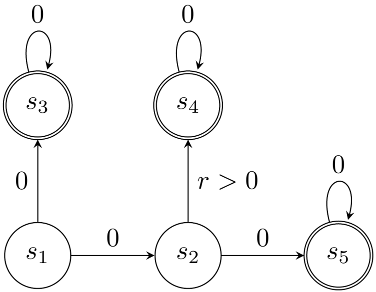
$$\nabla_\theta v_\theta(s) = \mathbb{E}_\theta\left[\left(\sum_{i=0}^{\infty}\gamma^i R_{i+1}\right)\left(\sum_{t=0}^{\infty}\nabla_\theta \log \pi_\theta(A_t|S_t)\right)\middle| S_0 = s\right] \qquad \text{(REINFORCE)}$$

$$\begin{aligned}\nabla_\theta v_\theta(s) &= \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\gamma^t\left(\sum_{i=0}^{\infty}\gamma^i R_{t+i+1}\right)\nabla_\theta \log \pi_\theta(A_t|S_t)\middle| S_0 = s\right] \\ &= \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\gamma^t q_\theta(S_t, A_t)\nabla_\theta \log \pi_\theta(A_t|S_t)\middle| S_0 = s\right] \\ &= \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\gamma^t (q_\theta(S_t, A_t) - v_\theta(S_t))\nabla_\theta \log \pi_\theta(A_t|S_t)\middle| S_0 = s\right] \qquad \text{(Advantage)}\end{aligned}$$

For softmax policies the score $\nabla_\theta \log \pi_\theta(a|s)$ is easy to compute; the first two expressions yield direct Monte-Carlo approximations while the latter requires some approximation of $v_\theta$.

# Non Concavity

Previous ideas lead to stochastic gradient schemes, but the value function is in general non concave.




From [Agarwal *et al.*, 2021]

Hint: Consider $\theta^{(1)}$ such that $\theta^{(1)}_{s_1,U} = \log 1, \theta^{(1)}_{s_1,R} = \log 3, \theta^{(1)}_{s_2,U} = \log 3, \theta^{(1)}_{s_2,R} = \log 1$ and $\theta^{(2)} = -\theta^{(1)}$ and check that $v_{\theta^{(1)}}(s_1) + v_{\theta^{(2)}}(s_1) > 2v_{(\theta^{(1)}+\theta^{(2)})/2}(s_1)$: note that it also holds in direct parameterization.

# Roadmap

# Bandits?

In the context of this course might be more accurately described as a single-state MDP!

# A Short History of Bandits

- Thompson (1933) *On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples*. Biometrika
- Robbins (1952) *Some aspects of the sequential design of experiments*. Bull. Amer. Math. Soc.
- Gittins (1979) *Bandit processes and dynamic allocation indices*. J. R. Stat. Soc. Ser. B Stat. Methodol.
- Lai & Robbins (1985) *Asymptotically efficient adaptive allocation rules*. Adv. in Appl. Math.
- Auer, Cesa-Bianchi & Fischer (2002) *Finite-time analysis of the multi-armed bandit problem*. Machine Learning Journal

And many papers since then in the machine learning literature...

## Definition (Multi-Armed Bandit Model)

- "Arms" from $1$ to $K$;
- Each arm $k \in \{1, \dots, K\}$ is associated with an infinite sequence of undisclosed "rewards" $(X_{k,i})_{i \geq 1} \in [0, 1]$;
- At each round $t = 1, \dots$
  - "play" arm $A_t \in \{1, \dots, K\}$,
  - obtain $X_t = X_{A_t, N_{a_t}(t)}$, where

$$N_k(t) = \sum_{s=1}^{t} \mathbb{1}\{A_s = k\}$$

The Regret is defined as

$$R_T = \max_{k \in \{1,\dots,K\}} \sum_{t=1}^{T} X_{k,t} - \sum_{t=1}^{T} X_t$$

# A First Intuitive Approach

## Algorithm (Explore-then-Commit (ETC))

- *For "rounds"* $i = 1, \ldots, m$, *play arm* $k = 1, \ldots, K$ *such that* $N_k(mK) = m$ *for each* $k \in \{1, \ldots, K\}$.
- *For* $t \geq 1 + mK$ *play*

$$A_t = \underset{k \in \{1,\ldots,K\}}{\arg\max}\ \overline{X}_k(mK)$$

*where*

$$\overline{X}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} X_s \mathbb{1}\{A_s = k\}$$

Note that by construction in ETC

$$\overline{X}_k(mK) = \overline{X}_{k,m} = 1/m \sum_{i=1}^{m} X_{k,i}$$

# Exploration/Exploitation Tradeoff

How to balance knowledge acquisition and reward maximization?

- The major question in bandit models
- In ETC, instanciates as the choice of $m$
- Needs a model of the world
  - Deterministic environment
  - Stochastic environment
  - Adversarial environment

✎

## Definition (Stochastic MAB)

$(X_{k,t})_{t\geq 1}$ are mutally independent i.i.d. sequences such that

$$X_{k,t} \sim \nu_k$$

We denote by

- $\mu_k = \mathbb{E}[X_{k,t}]$
- $k^* = \arg\max_{k\in\{1,\dots,K\}} \mu_k$
- $\mu^* = \max_{k\in\{1,\dots,K\}} \mu_k$
- $\Delta_k = \mu^* - \mu_k$

and assume that $\Delta_k > 0$ for $k \neq k^*$.

Efiicent algorithms are invariant w.r.t. arm indexing, and we can assume w.l.o.g. for analysis that $k^* = 1$.

## Definition (Bandit Algorithm)

A sequential allocation rule such that $A_t$ is $\mathcal{H}_{t-1}$ – measurable, where $\mathcal{H}_{t-1} = \sigma(X_1, \ldots, X_{t-1})$ .

A *ramdomized* bandit algorithm is $\mathcal{H}_{t-1} \vee \mathcal{G}$ – measurable, where $\mathcal{G}$ is independent of $\mathcal{H}_\infty$.

$\longrightarrow$ We are interested in bandit algorithms that are optimal w.r.t. a performance criterion.

# Reward Maximization – Regret Minimization

**Goal** Make sure that $1/T \sum_{t=1}^{T} X_t \xrightarrow{\mathbb{L}_1} \mu^*$ as fast as possible by minimizing the regret.

## Definition ((Stochastic) Regret or Pseudo-Regret)

$$R_T = \max_{k \in \{1,\ldots,K\}} \mathbb{E}\left[\sum_{t=1}^{T} X_{k,t}\right] - \sum_{t=1}^{T} X_t$$

# Expected Regret Decomposition

## Proposition

$$\mathbb{E}[R_T] = \mu^* T - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_{\substack{k=1 \\ k \neq k^*}}^{K} \Delta_k \mathbb{E}[N_k(T)]$$

$\hookrightarrow$ A sequential decision task that is not equivalent to estimating the values of the arm means $(\mu_k)$.

# Alternative Objective: Best Arm Identification

Goal: Find which of the $K$ hypotheses $\mathcal{H}_k : \mu_k = \mu^*$ is true

**Definition (Fixed Confidence Setting)**

Given a probability $\delta$, design an allocation rule and a stopping time $\tau$ such that

1. $\mathbb{P}(A_{\tau+1} \neq k^*) < \delta$
2. $\mathbb{E}[\tau]$ is minimal

- related to classical sequential hypothesis testing, with active added allocation
- requires more exploration than the reward maximization objective
- will not be addressed in the rest of this course