Fundamentals of Reinforcement Learning

Master IASD, Université PSL

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

February 2022
Roadmap

1. Temporal Difference Learning
   Reminder (MDP, Value Functions, Bellman Equations)
   Stochastic Approximation

2. Policy Gradient
   Importance Sampling
   Policy Gradient

2. The Multi-Armed Bandit Model

1. Bayesian Algorithms

2. Analysis of the Explore-then-Commit Algorithm
   Deviation Inequalities
   Regret bounds
Bayesian Optimal Algorithms

In this part we assume a prior distribution $\lambda$ on the set of bandit problems, and consider the Bayesian regret

$$\mathbb{E}_{(\nu_1,\ldots,\nu_K) \sim \lambda} \left( \mathbb{E}[R_T|\nu_1,\ldots,\nu_K] \right)$$

averaged over all possible models (under $\lambda$).

Interestingly, the Bayesian framework makes it possible to define optimal bandit algorithms (which are however not practical).
The Bayesian Approach

By specifying a prior distribution $\lambda$ on an unknown parameter $\theta$, the knowledge on $\theta$ gained from observing $X_1, \ldots, X_t$ is fully summarized by the posterior distribution

$$\Lambda_t(\theta) = p(\theta|X_1, \ldots, X_t) = \frac{p(X_1, \ldots, X_t|\theta)\lambda(\theta)}{\int p(X_1, \ldots, X_t|\theta')\lambda(\theta')d\theta'}$$

which defines

Posterior mean estimator $\int \theta \Lambda_t(\theta)d\theta$

Predictive distribution $\int p(x_{t+1}|X_1, \ldots, X_t, \theta)\Lambda_t(\theta)d\theta$

Posterior probability of hypothesis $\theta \in \mathcal{R} \int_{\mathcal{R}} \Lambda_t(\theta)d\theta$

Sequential update $\Lambda_{t+1}(\theta) \propto p(X_{t+1}|X_1, \ldots, X_t, \theta)\Lambda_t(\theta)$

Bayesian computation are usually not available in closed-form, except when using conjugate priors.
Example: Beta – Binomial Bayesian Experiment

Prior $\theta \sim \text{Beta}(\alpha, \beta)$, Likelihood $X_i | \theta \sim \text{Bernoulli}(\theta)$

Definition (Beta Distribution)

PDF $f(x | \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$

Expectation $m = \frac{\alpha}{\alpha + \beta}$

Variance $\frac{m(1-m)}{\alpha + \beta + 1} \leq \frac{1}{4(\alpha + \beta + 1)}$

- Posterior $\theta | X_1, \ldots, X_n \sim \text{Beta}(\alpha + S_n, \beta + n - S_n)$, where $S_n = \sum_{i=1}^{n} X_i$

- Predictive distribution $X_{n+1} | X_1, \ldots, X_n \sim \text{Bernoulli} \left( \frac{\alpha + S_n}{\alpha + \beta + n} \right)$
Bayesian Optimal Bandit as an MDP Planning Problem

The optimal algorithm solves the planning problem for an MDP whose state at time $t$ is the history $H_{t-1}$ of the observations and actions up to time $t - 1$. It can be solved (for fixed horizon $T$) by backward dynamic programming using the Bellman equation

$$v_{t:T}^*(H_{t-1}) = \max_k \mathbb{E} \left( X_t + v_{t+1:T}^*(H_t) \mid H_{t-1}, A_t = k \right)$$

initializing with the final greedy action

$$v_{T:T}^*(H_{T-1}) = \max_k \mathbb{E} \left( X_T \mid H_{T-1}, A_T = k \right)$$

so as to obtain $v_{1:T}^* = \mathbb{E}(\sum_{t=1}^{T} X_t)$ for the optimal policy $\pi^*$.

The optimal bandit algorithm is given by $A_t = \arg\max_k q_{t:T}^*(H_{t-1}, k)$ (note that it is non-stochastic).
A Toy Example

Let’s see how it works in

- Two armed bandit ($K = 2$)
- With Bernoulli distributions ($\mathbb{P}(X_{k,i} = 1) = \mu_k, \mathbb{P}(X_{k,i} = 0) = 1 - \mu_k$)
- For horizon $T = 2$ (homework: do it for $T = 3$ at home...)
- Using independent $\text{Beta}(\alpha_k, \beta_k)$ priors on $\mu_k$
\begin{align*}
q^*_1(\mathcal{A}_1 = 1) &= \frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0}} + \frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0}} \max \left( \frac{\alpha_{1,0} + 1}{\alpha_{1,0} + \beta_{1,0} + 1}, \frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}} \right) \\
&\quad + \frac{\beta_{1,0}}{\alpha_{1,0} + \beta_{1,0}} \max \left( \frac{\alpha_{1,0}}{\alpha_{1,0} + \beta_{1,0} + 1}, \frac{\alpha_{2,0}}{\alpha_{2,0} + \beta_{2,0}} \right)
\end{align*}
Gittins Indices

In the infinite horizon, $\gamma$-discounted case, Gittins (1979) showed that the optimal policy is an index policy where, if $\Lambda_{k,t-1}$ denotes the posterior on arm $k$ at time $t-1$

$$A_t = \arg \max_{k \in \{1,...,K\}} g_\gamma(\Lambda_{k,t-1})$$

**Definition (Gittins index)**

$$g_\gamma(\lambda) = \inf \left\{ \rho : \sup_{\tau \geq 0} \mathbb{E}_\lambda \left[ \sum_{t=1}^{\tau} \gamma^{t-1} X_t + \frac{\gamma^\tau \rho}{1-\gamma} \right] = \frac{\rho}{1-\gamma} \right\}$$

where the supremum is taken over all random stopping times $\tau$.

$g_\gamma(\lambda)$ can be interpreted as the exploration threshold in the one-armed bandit model with retirement (with prior $\lambda$ on the unknown arm).
**Thompson Sampling**

A Bayesian-inspired randomized algorithm that is successfully used in practice, has been proposed by Thompson (in 1933!) but was only analyzed very recently.

**Thompson Sampling**

- Draw $I_{k,t}$ from each posterior distribution $\Lambda_{k,t-1}$, for $k = 1, \ldots, K$
- Select
  \[ A_t = \arg \max_{k \in \{1, \ldots, K\}} I_{k,t} \]
- Observe $X_t$ and update the posterior $\Lambda_{A_t,t-1}$ to obtain $\Lambda_{A_t,t}$

A key observation is that Thompson sampling selects arm $k$ according to the posterior probability $P(K^* = k | H_{t-1})$ that it is actually optimal.