Fundamentals of Reinforcement Learning

Master IASD, Université PSL

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

February 2022
Roadmap

1. Analysis of the Explore-then-Commit Algorithm
   - Deviation Inequalities
   - Regret bounds
   - Adaptive ETC

2. The Lai and Robbins Lower Bound
   - Kullback-Leibler Divergence
   - Lower Bound

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

2. Policy Gradient
   - Importance Sampling
   - Policy Gradient

3. The Multi-Armed Bandit Model

4. Bayesian Algorithms
Why do We Need Deviation Inequalities?

Contrary to deterministic or purely randomized allocations, bandit allocation does not preserve distributions: neither $\overline{X}_k(t)$, nor $\overline{X}_k(t) | N_k(t) = n$ are distributed as any of the $(\overline{X}_{k,m})_{k \geq 1}$. 
Facts About Bandit Allocation

The following is true:

- $X_t | \mathcal{H}_{t-1} \sim \nu_{A_t}$;
- In the Bayesian approach, if a prior distribution $\lambda$ is specified on $(\nu_k)$, the posterior distribution, given $\mathcal{H}_{t-1}$ is also available in close-form (as seen in the previous course);
- Denoting $S_k(t) = \sum_{s=1}^{t} X_s 1\{A_s = k\}$,
  
  $S_k(t) - \mu_k N_k(t)$ is a $(\mathcal{H}_t)$ martingale increment

implying, in particular, that

$$E[S_k(t)] = \mu_k E[N_k(t)]$$

Which is true as well if $t$ is replaced by a stopping time $\tau$, due to Doob’s optimal stopping theorem.
Typical Use of Deviation Inequalities

But, the distribution of \((S_k(t), N_k(t))\) is not fixed as it depends on the learning algorithm.

- Cannot rely on distribution-dependent or asymptotic statistical results.
- Resort to (maximal) deviation inequalities, e.g.,

\[
\mathbb{P} \left( \sqrt{N_k(t)} (\bar{X}_k(t) - \mu_k) > \delta \right) \leq \mathbb{P} \left( \max_{1 \leq m \leq t} \sqrt{m} (\bar{X}_{k,m} - \mu_k) > \delta \right)
\]

\[
= \mathbb{P} \left( \exists m, 1 \leq m \leq t : \sqrt{m} (\bar{X}_{k,m} - \mu_k) > \delta \right) \leq \sum_{m=1}^{t} \mathbb{P} \left( \sqrt{m} (\bar{X}_{k,m} - \mu_k) > \delta \right)
\]

(union bound)

There exist finer bounds that will not be discussed in this course, see, e. g., [Garivier & Cappé, 2011].
Lemma (Cramér-Chernoff Method)

Assume \((X_i)_{i \geq 1}\) i.i.d. \(\sim \nu\), with \(\mathbb{E}[e^{\lambda X_1}] < \infty\), \(\forall \lambda \in \mathbb{R}\). Let \(\mu = \mathbb{E}[X_1]\), 
\[ \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \phi(\lambda) = \log \mathbb{E}[e^{\lambda X_1}] \text{ and } I(x) = \phi^*(x) = \sup_{\lambda \in \mathbb{R}} \lambda x - \phi(\lambda). \]
For \(x > \mu\),
\[ \mathbb{P}(\bar{X}_n > x) \leq e^{-nI(x)} \]

Lemma (Underestimation Bound)

Under the same conditions, for \(x < \mu\),
\[ \mathbb{P}(\bar{X}_n < x) \leq e^{-nI(x)} \]
These results are non improvable “in rate”, in the sense of the following Large Deviation Theorem.

Theorem (Cramér Theorem)

Under the same conditions,

\[
\lim_{n \to \infty} \frac{1}{n} \log P(\overline{X}_n > x) = -I(x) \quad (x > \mu)
\]

\[
\lim_{n \to \infty} \frac{1}{n} \log P(\overline{X}_n < x) = -I(x) \quad (x < \mu)
\]
Lemma (Gaussian Concentration Bound (Underestimation))

If $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, $\phi(\lambda) = \lambda^2 \sigma^2 / 2 + \mu \lambda$, $I(x) = (x - \mu)^2 / (2\sigma^2)$.

Hence, for $x < \mu$,

$$P \left( \bar{X}_n < x \right) \leq e^{-n \frac{(x-\mu)^2}{2\sigma^2}}$$

Corollary (Gaussian Upper Confidence Bound)

For any probability $\delta \in (0, 1)$,

$$P \left( \bar{X}_n + \sqrt{\frac{2\sigma^2}{n} \log \frac{1}{\delta}} < \mu \right) \leq \delta$$
Lemma (Hoeffding Lemma)

If \( X_1 \in [0, 1], \phi(\lambda) \leq \lambda^2/8 + \mu \lambda, \) \( i.e., \) \( \nu \) \( is \ 1/2 \ — \) sub-Gaussian

Thus for \( X_1 \in [0, 1], \) the previous bounds hold with \( \sigma^2 = 1/4, \) in particular,

\[
\mathbb{P} \left( \overline{X}_n + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} < \mu \right) \leq \delta
\]

Warning: Assuming that rewards are in \([0, 1]\) is the most common assumption in the bandit literature (used in this course) but others –such as, e.g., Lattimore & Szepesvári’s book– consider instead 1–sub-Gaussian rewards.
Theorem (Regret of ETC)

The regret of the Explore-Then-Commit algorithm may be bounded as

$$\mathbb{E}[R_T] \leq \sum_{k=1 \atop k \neq k^*}^K \Delta_k \left( m + Te^{-m\Delta_k^2} \right)$$

- The interesting regime occurs when $1 \ll m \ll T$
- Optimizing $m$ requires knowledge of $T$ and $\Delta_{\text{min}} \leq (\Delta_k)$ — not anytime, not adaptive!
- The latter is very conservative
Instance (or Parameter) Dependent Bound

Taking $m = \left\lceil \frac{\log T}{\Delta_{\text{min}}^2} \right\rceil$, 

\[ \mathbb{E}[R_T] \leq \sum_{k=1}^{K} \Delta_k \left( 1 + \frac{\log T}{\Delta_{\text{min}}^2} \right) \]

Minimax Bound

When $K = 2$, taking $m = \left\lceil \frac{\log(T\Delta^2)}{\Delta^2} \right\rceil$ if $\Delta > \frac{1}{\sqrt{T}}$ and anything otherwise, 

\[ \mathbb{E}[R_T] \leq \sqrt{T}(1 + \log T) \]
In simple cases, ETC can be made adaptive (but not anytime)

Algorithm (Adaptive ETC (Two Arms))

Given an horizon $T$,

- $M = 1$, play arms 1 and 2
- While $|\bar{X}_1(2M) - \bar{X}_2(2M)| \leq \sqrt{\gamma \log T/M}$:
  - Play arms 1 and 2
  - $M++$
- For $1 + 2M \leq t \leq T$, play $A_t = 1$ if $\bar{X}_1(2M) > \bar{X}_2(2M)$, or $A_t = 2$ otherwise
Proposition

For $\gamma > 2$, Adaptive ETC satisfies

$$\mathbb{E}[R_T] \leq \frac{\gamma(1 + \epsilon) \log T}{\Delta} + O_{\gamma, \epsilon}(1)$$

for all $\epsilon > 0$, where $\Delta$ denotes the gap between the two arms.

Proof Hint

$$\mathbb{E}(R_T) \leq \Delta \mathbb{E}(M) + T \sum_{m=1}^{T/2} \mathbb{P} \left( \bar{X}_{1,m} - \bar{X}_{2,m} < -\sqrt{\frac{\gamma \log T}{m}} \right)$$
Roadmap

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

2. Policy Gradient
   Importance Sampling
   Policy Gradient

3. The Multi-Armed Bandit Model

4. Bayesian Algorithms

1. Analysis of the Explore-then-Commit Algorithm
   Deviation Inequalities
   Regret bounds
   Adaptive ETC

2. The Lai and Robbins Lower Bound
   Kullback-Leibler Divergence
   Lower Bound
Theorem (Data-Processing Inequality)

Let \((\Omega, \mathcal{A})\) be a measurable space, and let \(P\) and \(Q\) be two probability measures on \((\Omega, \mathcal{A})\). Let \(X : \Omega \to (\mathcal{X}, \mathcal{B})\) be a random variable, and let \(P^X\) (resp. \(Q^X\)) be the push-forward measures, i.e., the laws of \(X\) w.r.t. \(P\) (resp. \(Q\)). Then

\[
\text{KL}(P, Q) \geq \text{KL}(P^X, Q^X)
\]

Corollary

If \(X \in [0, 1]\),

\[
\text{KL}(P, Q) \geq d(\mathbb{E}_P[X], \mathbb{E}_Q[X])
\]

where \(d\) is the Bernoulli Kullback-Leibler divergence

\[
d(p, q) = p \log(p/q) + (1 - p) \log((1 - p)/(1 - q))
\]
Two useful inequalities for $d$

**Lemma ((Basic) Pinsker Inequality)**

\[ d(p, q) \geq 2(p - q)^2 \]

**Lemma**

\[ d(p, q) \geq p \log \frac{1}{q} - \log 2 \]

and

\[ d(p, q) \geq (1 - p) \log \frac{1}{1 - q} - \log 2 \]
Lemma (Change of Distribution)

Consider two stochastic MAB models with arm distributions $\nu = (\nu_1, \ldots, \nu_k, \ldots, \nu_K)$ and $\nu' = (\nu_1, \ldots, \nu'_k, \ldots, \nu_K)$, respectively,

$$\text{KL}(P_{\nu}^{X_1, \ldots, X_T}, Q_{\nu'}^{X_1, \ldots, X_T}) = \text{KL}(\nu_k, \nu'_k) \mathbb{E}_{\nu}[N_k(T)]$$

Definition (Consistent Strategy)

A strategy is consistent if for any parameters $\nu$ of the stochastic MAB model and all $\alpha > 0$,

$$\lim_{T \to \infty} \frac{\mathbb{E}_{\nu}[R_T]}{T^\alpha} = 0$$

This implies that for all $k \neq k^*$, $\lim_{T \to \infty} \mathbb{E}_{\nu}[N_k(T)]/T^\alpha = 0$
Proposition

For any consistent strategy and $k \neq k^*$, and under regularity conditions,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu[N_k(T)]}{\log T} \geq \frac{1}{\text{KL}(\nu_k, \nu^*)}$$

Corollary (Lai and Robbins Lower Bound)

For any consistent strategy,

$$\liminf_{T \to \infty} \frac{\mathbb{E}_\nu[R_T]}{\log T} \geq \sum_{k=1}^{K} \sum_{k \neq k^*} \frac{\Delta_k}{\text{KL}(\nu_k, \nu^*)}$$
Proof hint

Assuming w.l.o.g. that \( k^* = 1 \) under model \( \nu \), consider

- \( \nu \) such that \( \nu_k \) is not the best arm, i.e., that
  \[
  \mathbb{E}_{\nu_k}[X_{k,t}] < \mathbb{E}_{\nu_1}[X_{1,t}]
  \]
- \( \nu' \) such that \( \nu'_k \) is the best arm, i.e., that
  \[
  \mathbb{E}_{\nu'_k}[X_{k,t}] > \mathbb{E}_{\nu'_1}[X_{k_1,t}]
  \]
while all other arms but \( k \) are unchanged under either \( \nu \) or \( \nu' \)

This implies in particular that, for any consistent strategy,

- \( \frac{1}{T} \mathbb{E}_{\nu}[N_k(T')] \to 0 \)
- \( \frac{1}{T} \mathbb{E}_{\nu'}[N_k(T)] \to 1 \)