Part 8: Minimal exploration

EL2805 - Reinforcement Learning

Alexandre Proutiere

KTH, The Royal Institute of Technology

Objectives of this lecture

Discuss how to best tune the exploration of sub-optimal actions in RL.
For simplicity, we deal with *bandit optimization* (only) – RL with no state dynamics.

- Regret lower bound
- Algorithms based on the "optimism in front of uncertainty" principle
- Thompson Sampling algorithm

For more refer to the following tutorial:
`https://arxiv.org/abs/1204.5721`

References

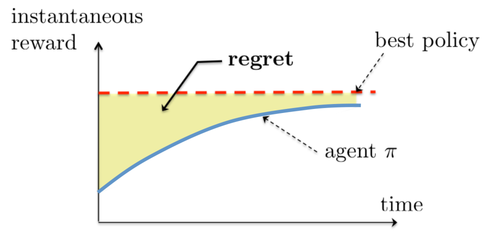
- Sutton-Barto's book chapter 2
- Bubeck-CesaBianchi survey: `https://arxiv.org/abs/1204.5721`

Bandit Optimization

- Interact with an i.i.d. or adversarial environment
- Set of available actions A with unknown sequences of rewards $r_t(a), t = 1, \ldots$
- The reward is the only feedback – **bandit feedback**
- Stochastic vs. adversarial bandits
  - i.i.d. environment: $r_t(a)$ random variable with mean θ_a
  - adversarial environment: $r_t(a)$ is arbitrary!
- Objective: develop an action selection rule π maximising the expected cumulative reward up to step T

  Remark: π must select an action depending on the entire history of observations!

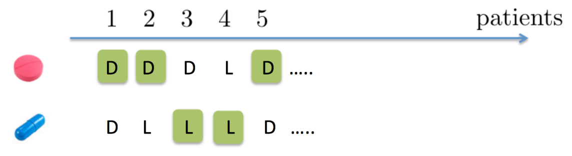
Regret



- Difference between the cumulative reward of an "Oracle" policy and that of agent π
- Regret quantifies the price to pay for learning
- Exploration vs. exploitation trade-off: we need to probe all actions to play the best later

Applications

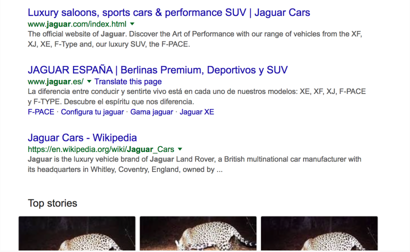
Clinical trial, Thompson 1933



- Two available treatments with unknown rewards ('Live' or 'Die')
- Bandit feedback: after administrating the treatment to a patient, we observe whether she survives or dies
- Goal: design a treatment selection scheme π maximising the number of patients cured after treatment

Applications

Search engines



- The engine should list relevant webpages depending on the request 'jaguar'
- The CTRs (Click-Through-Rate) are unknown
- Goal: design a list selection scheme that learns the list maximising its global CTRs

Unstructured Stochastic Bandits – Robbins 1952

- Finite set of actions $A = \{1, \dots, K\}$
- (Unknown) rewards of action $a \in A$: $(r_t(a), t \geq 0)$ i.i.d. Bernoulli with $\mathbb{E}[r_t(a)] = \theta_a$
- Optimal action $a^\star \in \arg\max_a \theta_a$
- Online policy π: select action a_t^π at time t depending on $a_1^\pi, r_1(a_1^\pi), \dots, a_{t-1}^\pi, r_{t-1}(a_{t-1}^\pi)$
- Regret up to time T: $R^\pi(T) = T\theta_{a^\star} - \sum_{t=1}^{T} \theta_{a_t^\pi}$

Concentration

The main tools in the analysis of stochastic bandits are **concentration-of-measure** results.
Let $X_1, X_2, \dots$ i.i.d. real-valued random variable with mean μ, and with *all* moments $G(\lambda) = \log(\mathbb{E}[e^{\lambda(X_n - \mu)}])$. $S_n = \sum_{i=1}^n X_i$.

- Strong law of large number: $\mathbb{P}[\lim_{n\to\infty} \frac{S_n}{n} = \mu] = 1$
- Concentration inequality: let $\delta, \lambda > 0$,

$$\begin{aligned}
\mathbb{P}[S_n - n\mu \geq \delta] &= \mathbb{P}[e^{\lambda(S_n - n\mu)} \geq e^{\lambda\delta}] \\
&\leq e^{-\lambda\delta}\mathbb{E}[e^{\lambda(S_n - n\mu)}] \\
&= e^{-\lambda\delta} \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mu)}] \\
&= e^{nG(\lambda) - \lambda\delta} \\
&\leq e^{-\sup_{\lambda>0}(\lambda\delta - nG(\lambda))}
\end{aligned}$$

Concentration

$$\mathbb{P}[S_n - n\mu \geq \delta] \leq e^{-\sup_{\lambda>0}(\lambda\delta - nG(\lambda))}$$

- Bounded r.v. $X_n \in [a,b]$, $G(\lambda) \leq \lambda^2 \frac{(b-a)^2}{8}$
  Hoeffding's inequality:

$$\mathbb{P}[S_n - n\mu \geq \delta] \leq e^{-\frac{2\delta^2}{n(b-a)^2}}$$

- Sub-gaussian r.v.: $G(\lambda) \leq \sigma^2\lambda^2/2$
- Bernoulli r.v.: $G(\lambda) = \log(\mu e^{\lambda(1-\mu)} - (1-\mu)e^{-\lambda\mu})$
  Chernoff's inequality:

$$\mathbb{P}[S_n - n\mu \geq \delta] \leq e^{-nKL(\mu+\delta/n,\mu)}$$

where $KL(a,b) = a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$ (KL divergence)

Regret Lower Bound

Uniformly good algorithms: An algorithm π is uniformly good if for all $\theta \in \Theta$, for any sub-optimal arm a, the number of times $n_a(t)$ arm a is selected up to round t satisfies: $\mathbb{E}[n_a(t)] = o(t^\alpha)$ for all $\alpha > 0$.

Fundamental performance limits: (Lai-Robbins1985)
For any uniformly good algorithm π:

$$\liminf_T \frac{R^\pi(T)}{\log(T)} \geq \sum_{a \neq a^\star} \frac{\theta_{a^\star} - \theta_a}{KL(\theta, \theta_{a^\star})}$$

where $KL(a,b) = a\log(\frac{a}{b}) + (1-a)\log(\frac{1-a}{1-b})$ (KL divergence)

Proof

- Change-of-measure: $\theta \to \nu$ with $\theta_j = \nu_j$ for all $j \neq a$, $\nu_a = \theta_{a^\star} + \epsilon$
- Log-likelihood ratio:
  $\mathbb{E}_\theta[L] = \sum_j \mathbb{E}_\theta[n_j(t)] KL(\theta_j, \nu_j) = \mathbb{E}_\theta[n_a(t)] KL(\theta_a, \theta_{a^\star} + \epsilon)$
- For any event A, $\mathbb{P}_\nu(A) = \mathbb{E}_\theta[\exp(-L) 1_A]$. Jensen's inequality yields:

$$\mathbb{P}_\nu(A) \geq \exp(-\mathbb{E}_\theta[L] | A) \mathbb{P}_\theta(A)$$

$$\mathbb{P}_\nu(A^c) \geq \exp(-\mathbb{E}_\theta[L] | A^c) \mathbb{P}_\theta(A^c)$$

- Hence $\mathbb{E}_\theta[L] \geq KL(\mathbb{P}_\theta(A), \mathbb{P}_\nu(A))$
- Select $A = \{n_{a^\star}(t) \leq t - \sqrt{t}\}$. We obtain:

$$\lim \inf_{t \to \infty} \frac{\mathbb{E}_\theta[n_a(t)]}{\log(t)} \geq \frac{1}{KL(\theta_a, \theta_{a^\star} + \epsilon)}$$

Algorithms

Estimating the average reward of arm a:

$$\hat{\theta}_a(t) = \frac{1}{n_a(t)} \sum_{n=1}^{t} r_n(a) 1_{a(n)=a}$$

- ϵ**-greedy.** In each round t:
  - with probability $1-\epsilon$, select the best empirical arm $a^\star(t) \in \arg\max_a \hat{\theta}_a(t)$
  - with probability ϵ, select an arm uniformly at random

  The algorithm has linear regret (not uniformly good)

Algorithms

- ϵ_t**-greedy.** In each round t:
  - with probability $1 - \epsilon_t$, select the best empirical arm $a^\star(t) \in \arg\max_a \hat{\theta}_a(t)$
  - with probability ϵ_t, select an arm uniformly at random

  The algorithm has logarithmic regret for Bernoulli rewards and $\epsilon_t = \min(1, \frac{K}{t\delta^2})$ where $\delta = \min_{a \neq a^\star}(\theta_{a^\star} - \theta_a)$

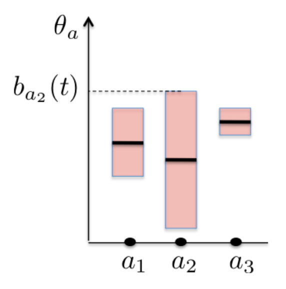
Sketch of proof. For $a \neq a^\star$ to be selected in round t, we need (most often) $\hat{\theta}_a(t) \geq \theta_a + \delta$. The probability that this occurs is less than $\exp(-2\delta^2 n_a(t))$. But $n_a(t)$ is close to $\log(t)/\delta^2$. Summing over t yields the result.

Algorithms

Optimism in front of Uncertainty

Upper Confidence Bound algorithm:

$$b_a(t) = \hat{\theta}_a(t) + \sqrt{\frac{2\log(t)}{n_a(t)}}$$



$\hat{\theta}(t)$: empirical reward of a up to t

$n_a(t)$: nb of times a played up to t

In each round t, select the arm with highest index $b_a(t)$

Under UCB, the number of times $a \neq a^\star$ is selected satisifies:

$$\mathbb{E}[n_a(T)] \leq \frac{8\log(T)}{(\theta_{a^\star} - \theta_a)^2} + \frac{\pi^2}{6}$$

Algorithms

KL-UCB algorithm:

$$b_a(t) = \max\{q \leq 1 : n_a(t) KL(\hat{\theta}_a(t), q) \leq f(t)\}$$

where $f(t) = \log(t) + 3 \log\log(t)$ is the *confidence* level.
In each round t, select the arm with highest index $b_a(t)$

Under KL-UCB, the number of times $a \neq a^\star$ is selected satisifies: for all $\delta < \theta_{a^\star} - \theta_a$,

$$\mathbb{E}[n_a(T)] \leq \frac{\log(T)}{KL(\theta_a + \delta, \theta_{a^\star})} + C \log\log(T) + \delta^{-2}$$

Algorithms

Bayesian framework, put a prior distribution on the parameters θ
Example: Bernoulli distribution with uniform prior on $[0, 1]$, we observed p successes ('1') and q failures ('0'). Then $\theta \sim \beta(p+1, q+1)$, i.e., the density is proportional to $\theta^p(1-\theta)^q$.

Thompson Sampling algorithm: Assume that at round t, arm a had $p_a(t)$ successes and $q_a(t)$ failures. Let $b_a(t) \sim \beta(p_a(t)+1, q_a(t)+1)$. The algorithm selects the arm a with the highest $b_a(t)$.

Under Thompson Sampling, for any suboptimal arm a, we have:

$$\lim \sup_{T\to\infty} \frac{\mathbb{E}[n_a(T)]}{\log(T)} = \frac{1}{KL(\theta_a, \theta_{a^\star})}$$

Illustration: UCB vs. KL-UCB

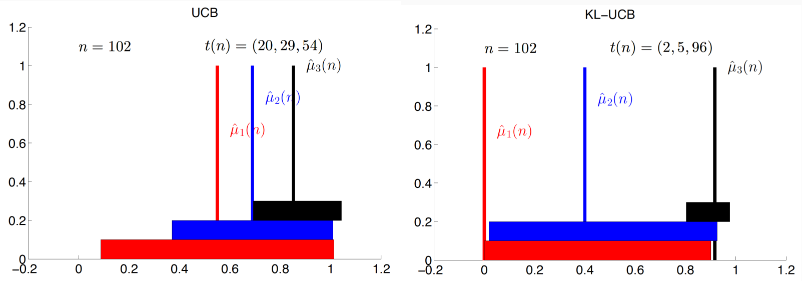
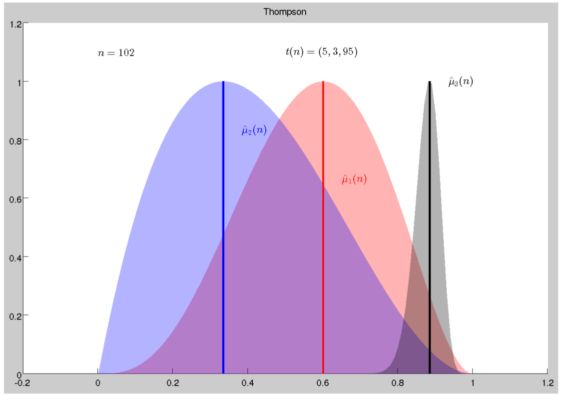


Illustration: Thompson Sampling



Performance