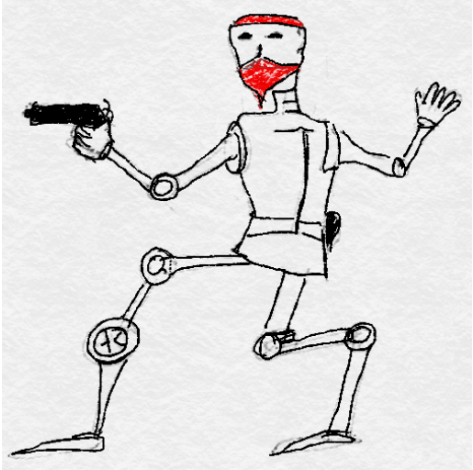
Bandits: Part I
Stochastic, Finite-Armed Bandits

Csaba Szepesvári

Department of Computing Science & AICML
University of Alberta & (@ Deepmind since August)
csaba.szepesvari@ualberta.ca

August 31, 2017

Summer School @ DS3



Outline

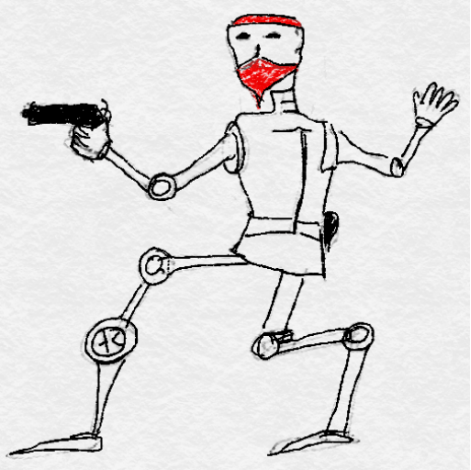
Outline

Overview of Talks

- Talk 1: Basics
  - What, why, applications
  - Bandits: Problem definition
  - Stochastic finite-armed bandits: Basics
  - Measure concentration. Subgaussianity
  - Explore-then-commit
  - UCB, Optimism, Optimality
- Talk 2: Adversarial and linear bandits
  - Adversarial finite-armed bandits
  - Contextual bandits
  - Exp4
  - Stochastic linear bandits
  - Adversarial linear bandits
- Talk 3: Bandits in the wild

Learning Material

- This lecture (and more): `http://banditalgs.com`
  - Joint effort with Tor Lattimore;
  - Book to be published by early next year: Stay tuned!
  - Tor's lightweight C++ bandit library
- I will share some exercises later
- Sebastien Bubeck's tutorial
  - Blog post 1
  - Blog post 2
- Bubeck and Cesa-Bianchi's book; (Bubeck and Cesa-Bianchi, 2012)



banditalgs.com

Outline

Learning Objectives: Knowledge

The goal is to gain knowledge about:

- Bandit problems: What are they?
- Types of bandit problems: How do they differ?
- Key ideas: Explore vs. exploit; why significant? What to do?
- Basic solution techniques
- Core results: How far did we get? How far can we go?
- Peak into contemporary research

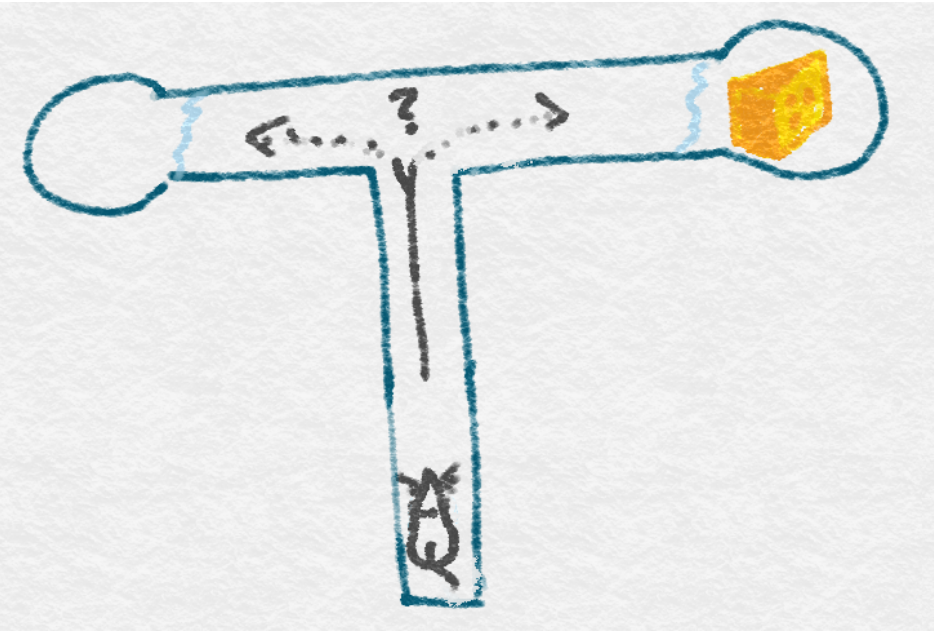
Learning Objectives: Skills

Skills to be acquired. Ability to ...

- ... recognize bandit problems;
- ... recognize types/variants of bandits;
- ... write code for bandits (algorithms, environment, ...);
- ... recognize insurmountable tradeoffs; limits of what is possible;
- ... get around in the literature.

Outline

Origin of the Name



Mouse learning in a T-maze
1953: Frederick Mosteller and Robert Bush, psychogolists
"A Stochastic Model with Applications to Learning"

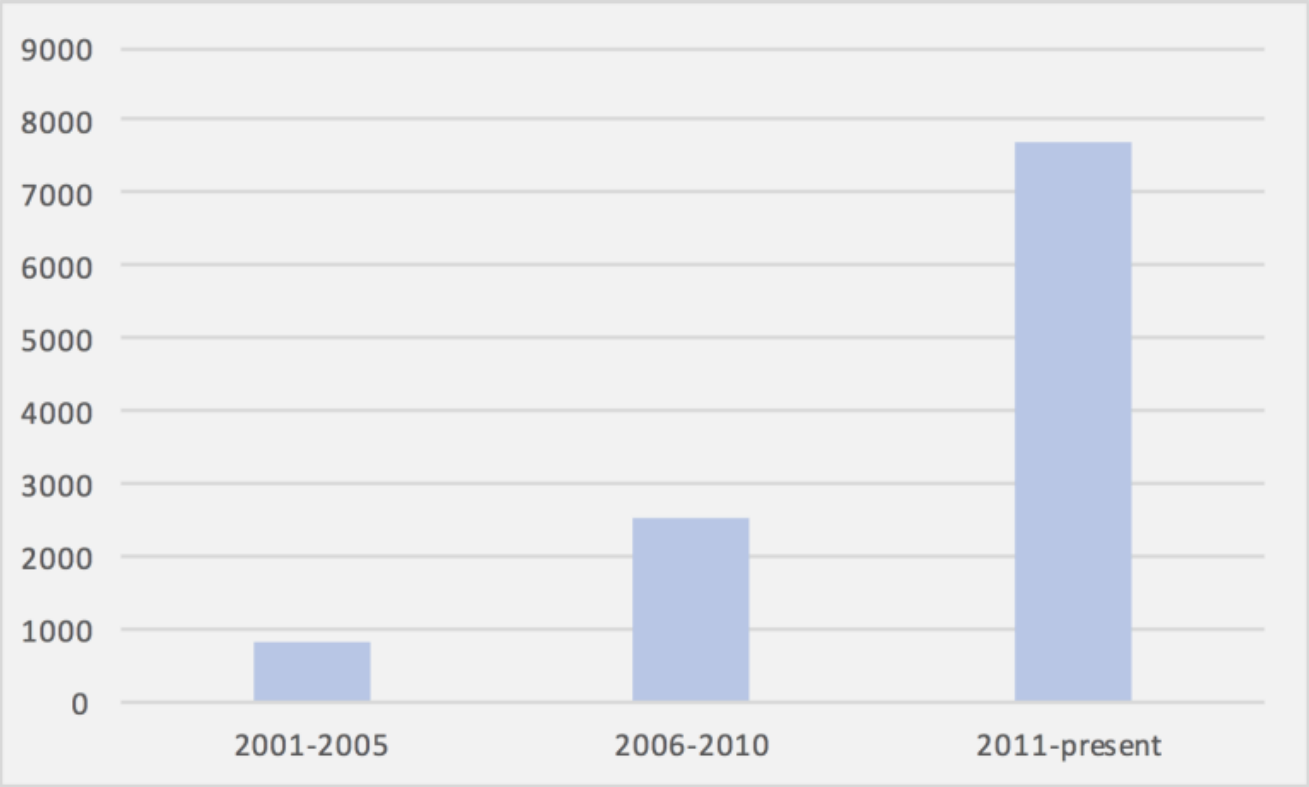
Generalization to humans:
"Two-armed bandits".

Classics

- First paper on bandits: Thompson (1933)
- Major effort by Herbert Robbins in the 50s; earliest is (Robbins, 1952)
- Another pioneer: Herman Chernoff, e.g., (Bather and Chernoff, 1967)
- Breakthrough in Bayesian bandits (computation): (Gittins, 1979)
- A breakthrough in stochastic bandits: (Lai and Robbins, 1985), asymptotic regret-optimality in a frequentist setting
- UCB as we know it: (Auer et al., 2002), finite-time optimality
- Exp3, adversarial bandits: (Auer et al., 1995)
- Early books (sequential design, Bayesian setting): (Chernoff, 1959; Berry and Fristedt, 1985; Gittins et al., 2011).

Outline

The Number Game



Number of papers published in 5-year periods (present=2016) on bandits as reported by google scholar.

Why Do People Care?

- Decision making uncertainty is a significant challenge
- Bandits capture key aspects of these challenge: The exploration-exploitation dilemma

Some examples

- Which drugs should a patient receive?
- How should I allocate my study time between courses?
- Which version of a website will generate the most revenue?
- What move should be considered next when playing chess/go?
- ...

Outline

The Exploration-Exploitation Dilemma

Payoffs after 5 pulls of each arm:

Left arm:	$0, 10, 0, 0, 10$
Right arm:	$10, 0, 0, 0, 0$

Left arm average payoff: 4 dollars per round;
Right arm average payoff: 2 dollars per round.
Budget: 20 more trials (pulls).

- Shall we pull left only? "Exploit"?
- Shall we pull the right arm at all? "Explore"?
- Why?
- How many times to pull each arm?
- "Good luck/bad luck?"

Play time!

Outline

Applications, applications..

1. A/B testing
2. Drug testing
3. Advert placement
4. Network routing (packets, planes, cars)
5. Tree search (MCTS)
6. Recommendation services (for example, news or movies)
7. Ranking (for example, search)
8. Educational games
9. Resource allocation (memory, bandwidth, manufacturing space)
10. Waiting problems (hard-disk shutdown, auto logout, waiting for a bus)
11. Dynamic pricing (for example, on Amazon)
12. A core component of RL

Questions?

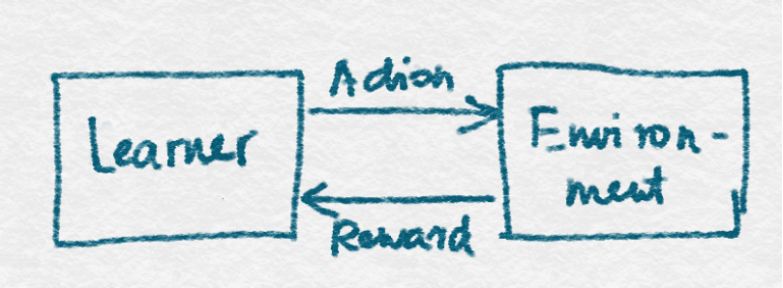
Outline

Bandits: Interaction Protocol and Goal

For **rounds** $t = 1, 2, ..., n$:

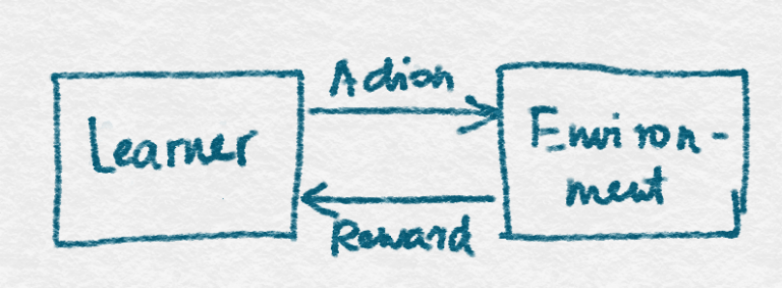
1. Learner chooses an **action** A_t from a set $\mathcal{A}$ of available actions. The chosen action is sent to the environment;
2. The environment ($\mathcal{E}$) generates a response in the form of a real-valued **reward** $X_t \in \mathbb{R}$, which is sent back to the learner.

The **goal** of the learner is to maximize the sum of rewards that it receives, $\sum_{t=1}^{n} X_t$.



Bandits: Interaction Protocol and Goal: Concepts

- History at the end of round t: $H_t = (A_1, X_1, \ldots, A_t, X_t)$.
- Learner can use history to base its action A_{t+1} on in round $t+1$.
- Learner "uses" a "policy", a map of all possible histories to actions.
- The learner is also allowed to randomize.



Bandits and Regret

Definition (Regret – Informal definition)

The **regret** of learner relative to action a =
Total reward gained when a is used for all n rounds –
Total reward gained by the specific learner in n rounds according to its chosen actions.

- Maximize reward $\Leftrightarrow$ minimize regret.
- Advantage? Normalizes the scale so that zero has special meaning (kills meaningless shifts).

Questions:

- What does it mean that a learner has positive regret?
- (What does it mean that a learner has positive reward?)
- What does it mean that a learner has zero regret?

Regret – II.

Definition (Zero-regret learner)

"Zero-regret learner": If R_n is regret at time n, $R_n/n \to 0$; a.k.a. "vanishing regret", sublinear regret; $R_n = o(n)$.

In general we care about how fast $R_n/n \to 0$ happens $\Leftrightarrow$ How slowly R_n grows.

Examples:

- $R_n = O(\sqrt{n})$ $\quad$ $(R_n/n = O(1/\sqrt{n}))$.
- $R_n = O(\log(n))$ $\quad$ $(R_n/n = O(\log(n)/n))$.

Discussion

- Why compare with fixed actions? Is this limiting? What does this capture?
- Stationarity.
- Does the policy know the "horizon"? Is it "anytime"?
- Good learner: "Small" regret (small worst-case regret!?) over a large class of environments.
- Instance-dependent regret.
- Regret lower and upper bounds: "Worst-case" vs. instance dependent.

Types of Environments

- Stochastic
- Adversarial ("unconstrained")
- Unstructured vs. structured payoff
- Contextual
- ...

Outline

Stochastic Bandits

Definition

A K**-armed stochastic bandit environment** is a tuple of distributions $\nu = (P_1, P_2, \ldots, P_K)$, where P_i is a distribution over the reals for each $i \in [K] \doteq \{1, 2, \ldots, K\}$.

Interaction

For rounds $t = 1, 2, 3, \ldots$

1. Based on its past observations (if any), the learner chooses an action $A_t \in [K]$ following some policy π: $A_t \sim \pi(\cdot | A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. The chosen action is sent to the environment.
2. The environment generates a random reward X_t whose distribution is P_{A_t} (in notation: $X_t \sim P_{A_t}$). The generated reward is sent back to the learner.

A note

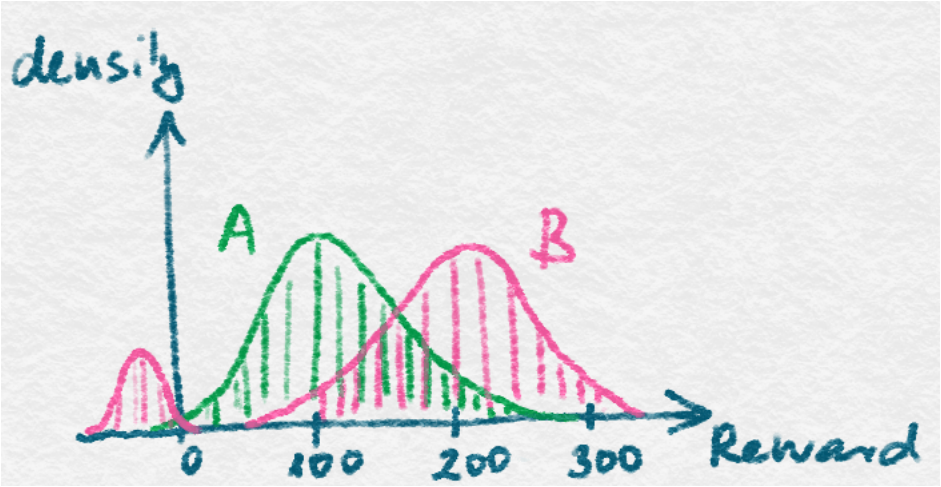
- In the environment, no joint (over all arms) is specified. Why?
  - The learner sees only X_t, not rewards from other arms. Even if those exist, they are latent, the learner cannot access them, hence, the properties of joint (even if it was specified) do not matter.
  - Precise answer: All information about an interconnected environment-policy pair (ν, π) is in the joint distribution of action-reward sequences.

Typical Environments

Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^K$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^K\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^K$	$\{(\mathcal{U}(a_i, b_i))_i : a, b \in \mathbb{R}^K, a \leq b\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^K(\sigma^2)$	$\{(\mathcal{N}(\mu_i, \sigma^2))_i : \mu \in \mathbb{R}^K\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{N}}^K$	$\{(\mathcal{N}(\mu_i, \sigma_i^2))_i : \mu \in \mathbb{R}^K, \sigma^2 \in \mathbb{R}_+^K\}$
Finite variance	$\mathcal{E}_{\mathbb{V}}^K(\sigma^2)$	$\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \leq \sigma^2 \text{ for all } i\}$
Finite kurtosis	$\mathcal{E}_{\mathrm{Kurt}}^K(\kappa)$	$\{(P_i)_i : \mathrm{Kurt}_{X\sim P_i}[X] \leq \kappa \text{ for all } i\}$
Bounded support	$\mathcal{E}_{[a,b]}^K$	$\{(P_i)_i : \mathrm{Supp}(P_i) \subseteq [a,b]\}$
Subgaussian	$\mathcal{E}_{\mathrm{SG}}^K(\sigma^2)$	$\{(P_i)_i : P_i \text{ is } \sigma\text{-subgaussian for all } i\}$
$\mathrm{Supp}(P)$ is the support of distribution P; $\mathrm{Kurt}(X) = \frac{\mathbb{E}[(X-\mathbb{E}[X])^4]}{\mathbb{V}[X]^2}$.		

Table: Typical environment classes for stochastic bandits

Expected Reward/Regret



- $S_n = \sum_{t=1}^{n} X_t$: Total reward. Random!!!
- Possible goal: Maximize $\mathbb{E}[S_n]$, the expected reward.
- OK?
- Same as minimizing R_n, the **(expected) regret**, where

$$R_n \doteq n\mu^* - \mathbb{E}[S_n] \ ,$$

$\mu^* = \max_{i \in [K]} \mu_i$, $\mu_i = \int_{-\infty}^{+\infty} x \, P_i(dx)$.

Outline

Basic Properties of Regret – I.

- Let $R_n(\pi, \nu)$ be the regret (expected!) of **policy** π on **environment** ν.

Lemma (Warmup – Exercise 0)

(a) $R_n(\pi, \nu) \geq 0$ *for all policies* π.

(b) *The policy* π *choosing* $A_t \in \operatorname{argmax}_i \mu_i$ *for all* t *satisfies*

$$R_n(\pi, \nu) = 0\,.$$

(c) *If* $R_n(\pi, \nu) = 0$ *for some policy* π *then for all* t, $A_t \in [K]$ *is optimal with probability one:*

$$\mathbb{P}\left(\mu_{A_t} = \mu^*\right) = 1\,.$$

Regret Decomposition – Important!!!

- Let $\nu = (P_1, \ldots, P_K)$ be a bandit environment.
- Let $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$: **sub-optimality gap** or **action gap** or **immediate regret** of action i.
- Usage count for action i:

$$T_i(t) = \sum_{s=1}^{t} \mathbb{I}_{\{A_s = i\}} \,.$$

- Note: $\Delta_i \doteq \Delta_i(\nu)$ is non-random, $T_i(t)$ is random! (Why?)

Lemma (Regret Decomposition Lemma)

For any policy π and K-armed stochastic bandit environment ν and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies

$$R_n = \textstyle\sum_{i=1}^{K} \Delta_i \, \mathbb{E}\left[T_i(n)\right] \,.$$

Proof of the Regret Decomposition Lemma

Proof.

For any fixed t we have $\sum_k \mathbb{1}_{\{A_t=k\}} = 1$.
Hence, $S_n = \sum_t X_t = \sum_t \sum_k X_t \mathbb{1}_{\{A_t=k\}}$ and thus

$$R_n = n\mu^* - \mathbb{E}[S_n] = \sum_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}\left[(\mu^* - X_t)\mathbb{1}_{\{A_t=k\}}\right].$$

Now, knowing A_t, the expected reward is μ_{A_t}. Thus we have

$$\begin{aligned}
\mathbb{E}\left[(\mu^* - X_t)\mathbb{1}_{\{A_t=k\}} \mid A_t\right] &= \mathbb{1}_{\{A_t=k\}}\mathbb{E}\left[\mu^* - X_t \mid A_t\right] \\
&= \mathbb{1}_{\{A_t=k\}}(\mu^* - \mu_{A_t}) \\
&= \mathbb{1}_{\{A_t=k\}}(\mu^* - \mu_k).
\end{aligned}$$

Take expectations, sum both sides over $t = 1, \ldots, n$. $\square$

Questions?

Exercise 1: (get code ↗)

Implement a Bernoulli bandit environment in Python using the code snippet below:

```
class BernoulliBandit:
  # accepts a list of K >= 2 floats, each lying in
      [0,1]
  def __init__(self, means):
    pass

  # Function should return the number of arms
  def K(self):
    pass

  # Accepts a parameter 0 <= a <= K-1 and returns the
  # realisation of random variable X with P(X = 1)
      being
  # the mean of the (a+1)th arm.
  def pull(self, a):
    pass

  # Returns the regret incurred so far.
  def regret(self):
    pass
```

Exercise 2: Follow-the-Leader (get code ↗)

Implement the following simple algorithm called 'Follow-the-Leader' (FTL), which chooses each action once and subsequently chooses the action with the largest average observed so far. Ties should be broken randomly.

```
def FollowTheLeader(bandit, n):
  # implement the Follow-the-Leader algorithm by
      replacing
  # the code below that just plays the first arm in
      every round
  for t in range(n):
    bandit.pull(0)
```

Note: Depending on the literature you are reading, Follow-the-Leader may be called 'stay with the winner' or the 'greedy algorithm'.

Exercise 3: Distribution of (random) regret

Consider a Bernoulli bandit with two arms and means $\mu_1 = 0.5$ and $\mu_2 = 0.6$.

(a) Using a horizon of $n = 100$, run 1000 simulations of your implementation of Follow-the-Leader on the Bernoulli bandit above and record the (random) regret, $n\mu^* - S_n$, in each simulation.

(b) Plot the results using a histogram (see fig. on the right).

(c) Explain the results in the figure.

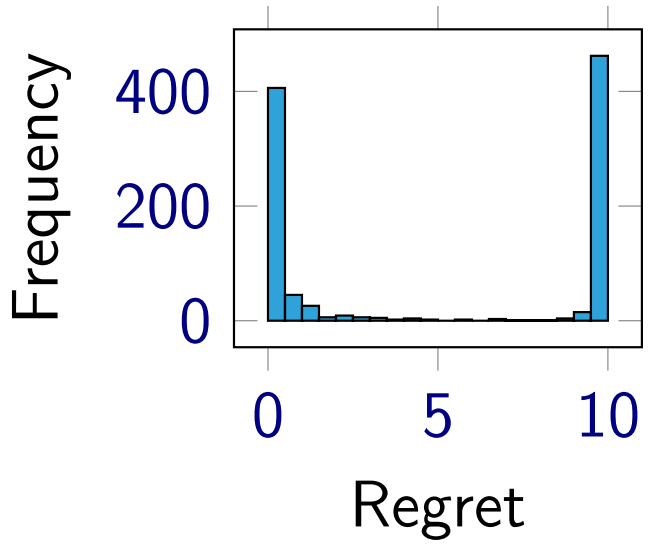


Figure: Histogram of regret for FTL over 1000 trials on Bernoulli bandit with means $\mu_1 = 0.5, \mu_2 = 0.6$

Exercise 4: Regret Over Time

Consider the same Bernoulli bandit as used in the previous question.

(a) Run 1000 simulations of your implementation of FTL for each horizon $n \in \{100, 200, 300, \ldots, 1000\}$.

(b) Plot the average regret obtained as a function of n (see the fig. on the right). Include error bars.

(c) Explain the plot. Do you think FTL is a good algorithm? Why/why not?

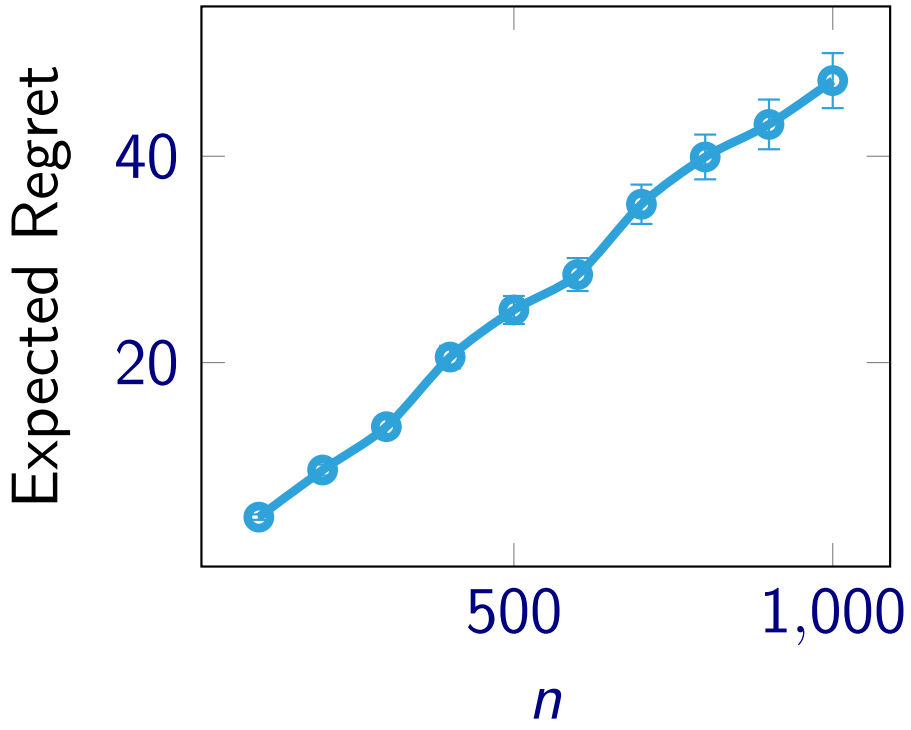


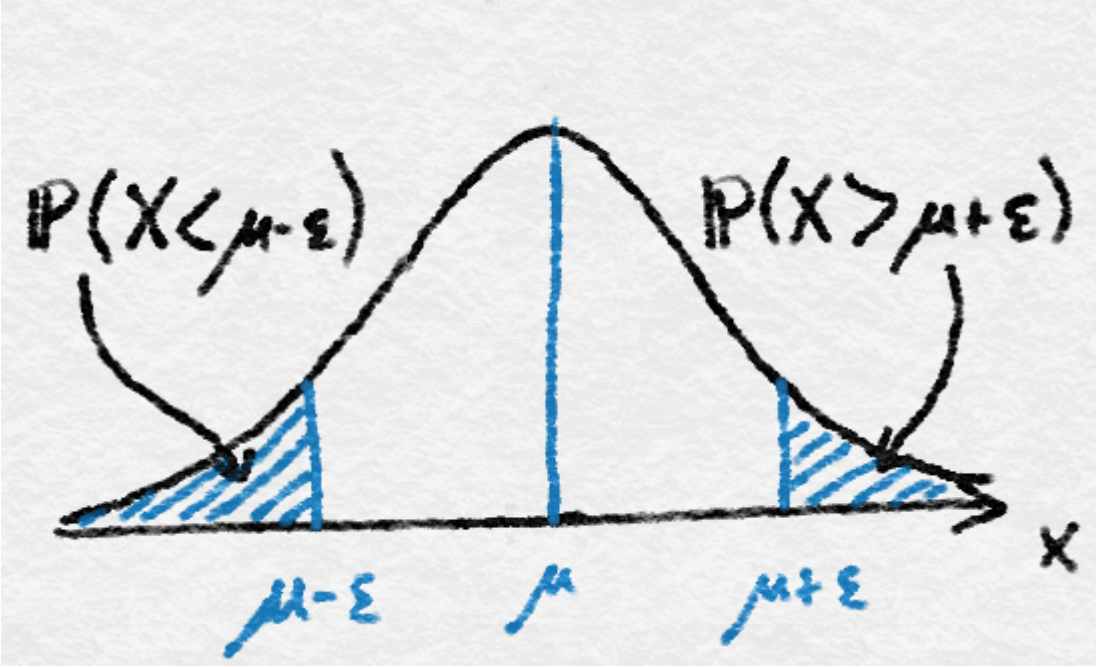
Figure: Histogram of regret for FTL over 1000 trials.

Outline

Statistics 101

- $X_1, \ldots, X_n$ independent, identically distributed ($\sim P$), real-valued random variables (think rewards).
- What is $\mu \doteq \mathbb{E}[X_1] \, (= \mathbb{E}[X_t])$?
- Estimate: $\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t$
- "Statistics" of data: any function of $X_1, \ldots, X_n$.
- Notice that $\hat{\mu}$ is random, but $\hat{\mu} \approx \mu$.
- The distribution of $\hat{\mu}$ depends on P.
- "How close"?: Characterize the distribution of $|\hat{\mu} - \mu|$.
- A priori characterization: The characterization depends on $\mathcal{P}$ where all we know that $P \in \mathcal{P}$.
- A posteriori characterization: The characterization depends on $X_1, \ldots, X_n$.

Tail probabilities



Imagine $X \doteq \hat{\mu}$.

We care about the probability masses $\mathbb{P}\left(\hat{\mu} < \mu - \varepsilon\right)$, $\mathbb{P}\left(\hat{\mu} > \mu - \varepsilon\right)$.

Either, given ε, give lower or upper bounds on this ("probability bound"). Or for a given probability mass, give upper and/or lower bounds on ε ("deviation bound").

Markov and Chebyshev

Lemma

For any random variable X with finite mean and $\varepsilon > 0$ it holds that:

(a) *(Markov):* $\mathbb{P}(|X| \geq \varepsilon) \leq \dfrac{\mathbb{E}[|X|]}{\varepsilon}$.

(b) *(Chebyshev):* $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \dfrac{\mathbb{V}[X]}{\varepsilon^2}$.

- Exercise 1: Prove Markov. Hint: Prove it for nonnegative r.v.s., use $\mathbb{E}[X] = \int_0^\infty x\, P(dx)$ and split the integral.
- Exercise 2: Prove Chebyshev. Hint: Apply Markov.
- Chebyshev applied to $\hat{\mu}$: $\mathbb{V}[\hat{\mu}] = \frac{\sigma^2}{n}$. Hence, $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$.
- Note: Chebyshev is more precise. You can further increase precision by applying Markov to $|X - \mathbb{E}[X]|^k$, k big.

Central Limit Theorem (CLT)

Theorem (CLT)

Let X_t iid, $\sigma^2 = \mathbb{V}[X_1] < \infty$, $S_n = \sum_{t=1}^{n}(X_t - \mu)$, $Z_n = S_n/\sqrt{\sigma^2 n}$. Then $F_{Z_n} \to F_Z$ as $n \to \infty$ where $Z \sim \mathcal{N}(0,1)$.

Note: $F_Z(u) = \mathbb{P}(Z \geq u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$.

Bounding $F_Z(u)$:

$$\int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \leq \frac{1}{u\sqrt{2\pi}} \int_u^\infty x \exp\left(-\frac{x^2}{2}\right) dx$$

$$= \sqrt{\frac{1}{2\pi u^2}} \exp\left(-\frac{u^2}{2}\right) .$$

Hence

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) = \mathbb{P}\left(\frac{S_n}{\sqrt{\sigma^2 n}} \geq \varepsilon\sqrt{\frac{n}{\sigma^2}}\right) \approx \mathbb{P}\left(Z \geq \varepsilon\sqrt{\sigma^2 n}\right)$$

$$\leq \sqrt{\frac{\sigma^2}{2\pi n \varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) .$$

How good is the CLT?

$$\mathbb{P}\left(\hat{\mu}_n \geq \mu + \varepsilon\right) \lessapprox \sqrt{\frac{\sigma^2}{2\pi n \varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) .$$

Question: Can we safely swap $\lessapprox$ to $\leq$, e.g., when $X_t \sim P$ and P is supported on $[0, 1]$.

(a) Yes, when $n \geq 30$, the error will be very small for these distributions.

(b) Yes, when $n \geq 1000$, the error will be very small for these distributions.

(c) No: No n will make the error uniformly small when P ranges over all distributions with support in $[0, 1]$.

Subgaussianity

Definition (Subgaussianity)

A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$ it holds that $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\lambda^2\sigma^2/2\right)$.

Moment/cumulant generating function of X, $M_X, \psi_X : \mathbb{R} \to \mathbb{R}$: $M_X(\lambda) = \mathbb{E}\left[\exp(\lambda X)\right]$, $\psi_X(\lambda) = \log M_X(\lambda)$, $\lambda \in \mathbb{R}$.

Lemma

X σ-subgaussian iff $\psi_X(\lambda) \leq \frac{1}{2}\lambda^2\sigma^2$ for all $\lambda \in \mathbb{R}$.

Example: $Z \sim N(0, \sigma^2)$. Then, $M_X(\lambda) = \exp(\lambda^2\sigma^2/2)$.

Does M_X (or ψ_X) exist always? No, e.g., for $X \sim \mathrm{Exp}$, $M_X(\lambda) = \infty$ for $\lambda \geq 1$.

Why the Name?

Theorem

If X is σ-subgaussian, then for any $\varepsilon \geq 0$,

$$\mathbb{P}\left(X \geq \varepsilon\right) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right) . \tag{1}$$

Proof

Proof.

We take a generic approach called **Cramer-Chernoff's method**. Let $\lambda > 0$ be some constant to be tuned later. Then

$$\begin{aligned}
\mathbb{P}(X \geq \varepsilon) &= \mathbb{P}(\exp(\lambda X) \geq \exp(\lambda\varepsilon)) \\
&\leq \mathbb{E}[\exp(\lambda X)] \exp(-\lambda\varepsilon) && \text{(Markov's inequality)} \\
&\leq \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right). && \text{(Def. of subgaussianity)}
\end{aligned}$$

Now λ was any positive constant, and in particular may be chosen to minimize the bound above, which is achieved by $\lambda = \varepsilon/\sigma^2$. $\square$

Variations

Union bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

Corollary: $\mathbb{P}(|X| \geq \varepsilon) \leq 2\exp(-\varepsilon^2/(2\sigma^2))$.

Equivalent "deviation" forms:

$$\mathbb{P}\left(X \geq \sqrt{2\sigma^2 \log(1/\delta)}\right) \leq \delta \qquad \mathbb{P}\left(|X| \geq \sqrt{2\sigma^2 \log(2/\delta)}\right) \leq \delta\,,$$

or w.p. $1-\delta$,

$$\left(-\sqrt{2\sigma^2 \log(2/\delta)},\ \sqrt{2\sigma^2 \log(2/\delta)}\right)\ .$$

Algebra with Subgaussian Random Variables

Lemma

Suppose that X is σ-subgaussian and X_1 and X_2 are independent and σ_1 and σ_2-subgaussian respectively, then:

(a) $\mathbb{E}[X] = 0$ *and* $\mathrm{Var}[X] \leq \sigma^2$.

(b) cX *is* $|c|\sigma$*-subgaussian for all* $c \in \mathbb{R}$.

(c) $X_1 + X_2$ *is* $\sqrt{\sigma_1^2 + \sigma_2^2}$*-subgaussian.*

Note

No matter what, $X_1 + X_2$ *is* $(\sigma_1 + \sigma_2)$*-subgaussian.*
Independence improves this to $\sqrt{\sigma_1^2 + \sigma_2^2}$.

Concentration of the Mean

Corollary

Assume that $X_i - \mu$ *are independent,* σ*-subgaussian random variables. Then, for any* $\varepsilon \geq 0$,

$$\mathbb{P}\left(\hat{\mu} \geq \mu + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right) \text{ and } \mathbb{P}\left(\hat{\mu} \leq \mu - \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \tag{2}$$

where $\hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t$.

Exercise 5

Using $\exp(-x) \leq 1/(ex)$ (which holds for all $x \geq 0$), show that except for a very small ε the above inequality is strictly stronger than what we obtained via Chebyshev's inequality and exponentially smaller (tighter) if $n\varepsilon^2$ is large relative to σ^2.

Deviation Form

Corollary

For any $\delta \in [0, 1]$*, with probability at least* $1 - \delta$*,*

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\,. \tag{3}$$

Symmetrically, it also follows that with probability at least $1 - \delta$*,*

$$\mu \geq \hat{\mu} - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\,. \tag{4}$$

Further Examples

- If X is distributed like a Gaussian with zero mean and variance σ^2, then X is σ-subgaussian.
- If X is bounded, zero-mean (i.e., $\mathbb{E}[X] = 0$ and $|X| \leq B$ almost surely for some $B \geq 0$) then X is B-subgaussian.
- Specifically, if X is a shifted Bernoulli with $\mathbb{P}(X = 1 - p) = p$ and $\mathbb{P}(X = -p) = 1 - p$, it also holds that X is $1/2$-subgaussian.

Extension of the Definition

- X is σ-subgaussian if the **noise** $X - \mathbb{E}[X]$ is σ-subgaussian.
- A distribution is called σ-subgaussian if a random variable drawn from that distribution is σ-subgaussian.

Bibliography

- Basic reference: Boucheron et al. (2013) (concentration under independence).
- Matrix versions of many standard results: Tropp (2015).
- Survey of classical results: McDiarmid (1998).
- Self-normalization (by standard deviation/variance): Peña et al. (2008).
- Empirical process theory: van de Geer (2000) or Dudley (2014).

Questions?

Outline

Standing Assumption (until further notice)

Assumptions

All bandit instances are in $\mathcal{E}_{\mathrm{SG}}^{K}(1)$, i.e., the reward distribution for all arms is 1-subgaussian.

Is this restrictive?

1. All the algorithms that follow rely on the knowledge of σ.
2. Unequal subgaussianity constant across arms.

Notation

Empirical mean of arm i:

$$\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathbb{I}_{\{A_s = i\}} X_s \,,$$

where

$$T_i(t) = \sum_{s=1}^{t} \mathbb{I}_{\{A_s = i\}}$$

and $A_s \in [K] = \{1, \ldots, K\}$ is the index of arm chosen in round s.

Explore-then-Commit (ETC)

- Explore all the K arms m times.
- Go with the winner for the remaining rounds.

Input $m \in \mathbb{N}$.

In round t choose action

$$A_t = \begin{cases} i\,, & \text{if } (t \bmod K) + 1 = i \text{ and } t \leq mK\,; \\ \operatorname{argmax}_i \hat{\mu}_i(mK)\,, & t > mK\,. \end{cases}$$

(ties in the argmax are broken arbitrarily)

History and Related Algorithms

- ε-greedy and friends: Choose winner in every round with probability $1 - \varepsilon$, explore uniformly at random all arms with probability ε (origin lost in history);
- "Certainty equivalence with forcing": Robbins (1952);
- In more complex bandits:
  - "epoch-greedy": Langford and Zhang (2008);
  - "Forced Exploration": Abbasi-Yadkori et al. (2009); Abbasi-Yadkori (2009);
  - "Phased exploration and greedy exploitation" (PEGE) Rusmevichientong and Tsitsiklis (2010).

Notes

- Looks silly: Explore for m steps, then exploit? Why should we care?
- Simplicity is great! Educational!
- ε-greedy and friends: Choose winner in every round with probability $1 - \varepsilon$, explore uniformly at random all arms with probability ε.
  - In n rounds, a particular arm i will be chosen on the average about $n\varepsilon/K$ times. So, $m \approx n\varepsilon/K$, or $\varepsilon = mK/n$.
  - Will the additional randomness help ε greedy (for the environments considered)?
  - Does it make sense to intermix exploration and exploitation steps, rather than exploring first and then exploiting?
- How to choose m? (How to choose ε?)

Outline

How Big Can the Regret Be?

Let $a \wedge b = \min(a, b)$ and $(x)^+ = \max(x, 0)$, for $a, b, x \in \mathbb{R}$.

Basic regret decomposition identity:

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(n)\right] .$$

Thus, enough to bound $\mathbb{E}\left[T_i(n)\right]$:

$$\begin{aligned}\mathbb{E}\left[T_i(n)\right] &\leq m \wedge \left\lceil \frac{n}{K} \right\rceil + (n - mK)^+ \, \mathbb{P}\left(i = A_{mK+1}\right) \\ &\leq m \wedge \left\lceil \frac{n}{K} \right\rceil + (n - mK)^+ \, \mathbb{P}\left(\hat{\mu}_i(mK) \geq \max_{j \neq i} \hat{\mu}_j(mK)\right) .\end{aligned}$$

Bounding the Probability

We have

$$\mathbb{P}\left(\hat{\mu}_i(mK) \geq \max_{j \neq i} \hat{\mu}_j(mK)\right) \leq \mathbb{P}\left(\hat{\mu}_i(mK) \geq \hat{\mu}_1(mK)\right)$$
$$= \mathbb{P}\left(\,\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1) \geq \Delta_i\,\right) .$$

Claim: $\hat{\mu}_i(mK) - \mu_i - (\hat{\mu}_1(mK) - \mu_1)$ is $\sqrt{2/m}$-subgaussian.

Hence, by (2) we have

$$\mathbb{P}\left(\hat{\mu}_i(mK) - \mu_i - \hat{\mu}_1(mK) + \mu_1 \geq \Delta_i\right) \leq \exp\left(-\frac{m\Delta_i^2}{4}\right) .$$

ETC Regret Upper Bound

Theorem (Instance-Dependent Bound)

After n rounds, the expected regret R_n of the ETC policy satisfies

$$R_n \leq \left(m \wedge \left\lceil \frac{n}{K} \right\rceil\right) \sum_{i=1}^{K} \Delta_i + (n - mK)^+ \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right) . \quad (5)$$

Instance dependent: Because the bound depends on Δ_i (properties of the bandit environment instances).
AKA: Gap-dependent, problem-dependent.

How to choose m?!

Outline

Optimal choice of m.

Take $K = 2$. WLOG, $\Delta_1 = 0$, $\Delta \doteq \Delta_2$. Then,

$$R_n \leq m\Delta + (n - 2m)^+ \Delta \exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) . \tag{6}$$

Assume n is reasonably large. Then, optimal choice for m is

$$m^*(n, \Delta) = \max\left\{0, \left\lceil \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \right\rceil\right\} \tag{7}$$

and we get

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \max\left\{0, \log\left(\frac{n\Delta^2}{4}\right)\right\}\right) . \tag{8}$$

A Simple but Important Improvement

Regardless of a policy, $R_n = \Delta \mathbb{E}[T_2(n)] \leq \Delta n$. Combine with (8), to get

$$R_n \leq \min \left\{ n\Delta,\ \Delta + \frac{4}{\Delta} \left(1 + \log \left(\frac{n\Delta^2}{4} \right) \right) \right\} . \tag{9}$$

Corollary ((Infeasible) Worst-Case Bound)

Consider ETC with $m = \lceil n/K \rceil \wedge m^*(\Delta, n)$. Then, there exists $C > 0$ such that for any $\Delta > 0$ and $n > 0$, $R_n \leq C\sqrt{n}$.

Instance-Adaptive Algorithms

Is there a good, non-cheating choice for m (dependence on n is allowed, but not on Δ)?

<u>Claim</u>: Best such choice gives $R_n = \Theta(n^{2/3})$.

Earlier we had $R_n = O(n^{1/2})$, *much* better! Is there some *other* algorithm that achieves this *without* knowing Δ? "Adaptation to Δ". (E.g., Auer and Ortner (2010) and Garivier et al. (2016)).

Notice that a worst-case bound shows(!) whether adaptation to individual instances happens!

Outline

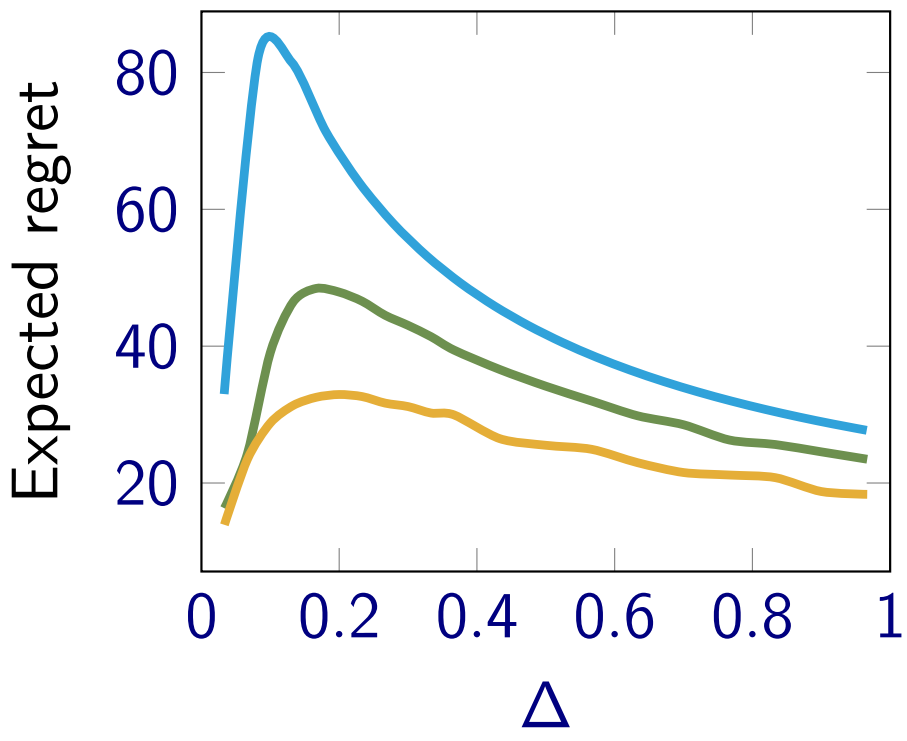
Exercise 6 (soln. ↗)

Gaussian bandit, $K = 2$, $\sigma^2 = 1$, $\mu_1 = 0$ and $\mu_2 = -\Delta$. Set $n = 1000$, repeat simulations $N = 10^4$ times and report the average. Plot, as a function of $\Delta \in [0, 1]$,

(a) The theoretical upper bound given in (9) (blue).

(b) The regret of the ETC algorithm with m set as suggested in (7) (green).

(c) The regret of the ETC algorithm with "the" optimal m (yellow: calculated numerically using that the noise is exactly Gaussian).



What can we conclude?

Exercise 7

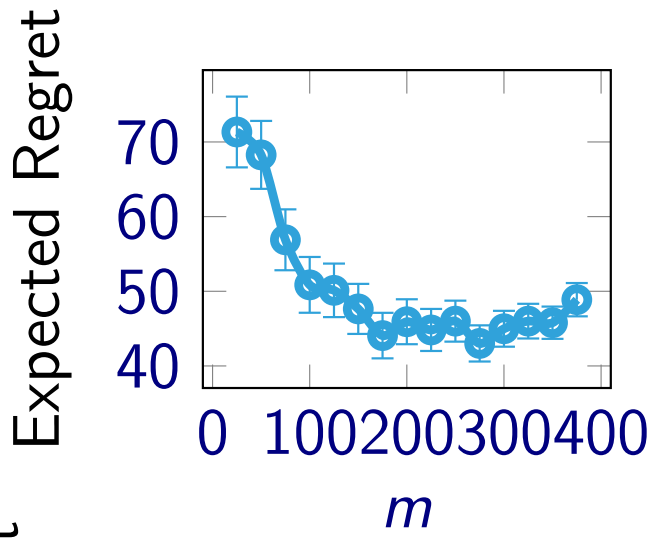
Let $\bar{R}_n = \sum_{t=1}^{n} \Delta_{A_t}$ (random, "pseudo-regret").

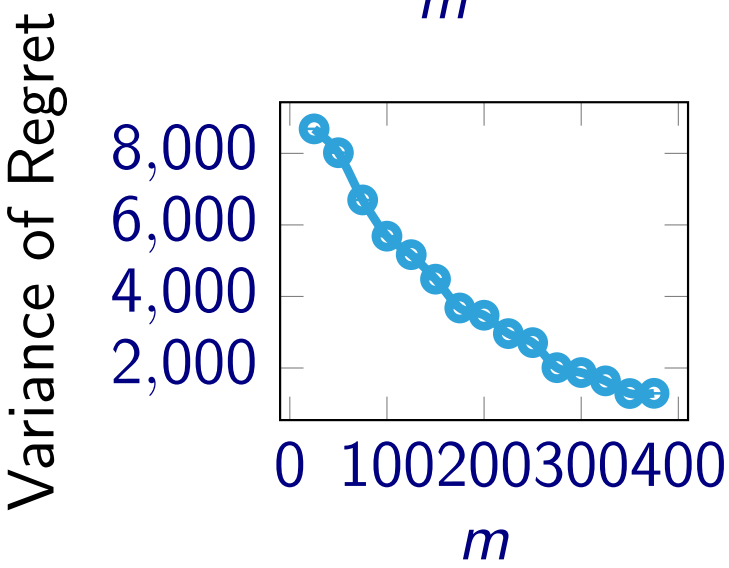
(a) Fix $\Delta = 1/10$ and plot $\bar{R}_n$ as a function of m with $n = 2000$. See upper plot on right.

(b) Plot the variance $\mathbb{V}[\bar{R}_n]$ as a function of m for the same bandit as above. See lower plot on right.

(c) Explain the curves and reconcile with theory.

(d) Did it make sense to plot $\mathbb{V}[\bar{R}_n]$? Why or why not?





Questions?

Outline

Optimism Principle

Optimism in the Face of Uncertainty (OFU) Principle

One should choose their actions as if the environment is as nice as **plausibly possible**.

Illustration

Visiting a new country.

Shall I try local cuisine/beer/. . . ?

Or stick to what I know?

Optimism: Yes.

Pessimism: No.

Optimism leads to exploration, pessimism prevents exploration.

Exploration is necessary: One can be unlucky with one's priors!
Hence, optimism is good.

How much??

Outline

On What is Plausible

Recall: if $X_1, X_2, \ldots, X_n$ are independent and 1-subgaussian with mean μ and $\hat{\mu} = \sum_{t=1}^{n} X_t/n$, then for any $\delta \in [0,1]$,

$$\mathbb{P}\left(\mu \geq \hat{\mu} + \sqrt{\frac{2\log(1/\delta)}{n}}\right) \leq \delta\,. \tag{10}$$

Round t. How big can μ_i be? Data: $\hat{\mu}_i(t-1)$ (empirical mean), based on $T_i(t-1)$ observations. Define

$$\mathrm{UCB}_i(t-1,\delta) = \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_i(t-1)}}\,. \tag{11}$$

Caveat: $T_i(t-1)$ is random, hence it is not clear whether $\mathbb{P}\left(\mathrm{UCB}_i(t-1,\delta) \geq \mu_i\right) \leq \delta$ holds.

The UCB(δ) Algorithm

Input K and δ
Choose each action once
For rounds $t > K$ choose action

$$A_t = \operatorname*{argmax}_i \mathrm{UCB}_i(t-1, \delta)$$

Note

Although there are many versions of the UCB algorithm, we often do not distinguish them by name and hope the context is clear. For the rest of this tutorial we'll usually call UCB(δ) just UCB.

Notes

- The algorithm first chooses each arm once, which is necessary because the term inside the square root is undefined when $T_i(t-1) = 0$.
- The value inside the argmax is called the **index** of arm i.
  - An **index algorithm** chooses the arm in each round that maximizes some value (the index), which usually only depends on current time-step and the samples from that arm.
  - In the case of UCB, the index is the sum of the empirical mean of rewards experienced so far and the **exploration bonus** (also known as the **confidence width**).
- The algorithm will work so that the UCB indices are approximately the same all the time (why?).

Demo: `http://downloads.tor-lattimore.com/bandits/`

Outline

Instance-Dependent Bound for UCB

Theorem

Consider UCB as shown earlier on a stochastic K-armed 1-subgaussian bandit problem. For any horizon n, if $\delta = 1/n^2$ then

$$R_n \leq 3 \sum_{i=1}^{K} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log(n)}{\Delta_i} .$$

Proof: Main Ideas

Regret decomposition identity:

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(n)\right] .$$

Take i so that $\Delta_i > 0$. Bound $\mathbb{E}\left[T_i(n)\right]$.

Key observation: After initialization, action i can only be chosen if its index is higher than that of an optimal arm.

This can only happen if at least one of the following is true:

(a) The index of action i is larger than the true mean of a specific optimal arm.

(b) The index of a specific optimal arm is smaller than its true mean.

Both of these have low probability. In particular, $\mathbb{E}\left[T_i(n)\right] \leq c_1 + c_2 \log(n)/\Delta_i^2$. Qu.e.d.

Proof: The Reward Consumption Model

For $i \in [K]$, let $(Z_{i,s})_s$ be an i.i.d. sequence with $Z_{i,s} \sim P_i$.

Define X_t to be the T_{A_t}th element of sequence $(Z_{A_t,s})_s$:

$$X_t = Z_{A_t, T_{A_t}(t)} \,. \tag{12}$$

Is there any loss of generality?

No: Interaction protocol $\Leftrightarrow$ constraint on the distribution of $(A_1, X_1, \ldots, A_n, X_n)$. This constraint clearly holds in this case.

Benefit: Defining

$$\hat{\mu}_{i,s} = \frac{1}{s} \sum_{u=1}^{s} Z_{i,u} \,, \qquad s \in [n]$$

(usual sample means), we have

$$\hat{\mu}_i(t) = \hat{\mu}_{i, T_i(t)} \,.$$

Proof: Some Details II.

WLOG $\mu_1 = \mu^*$. Fix i such that $\Delta_i > 0$. Let G_i be the 'good' event defined by

$$G_i = \left\{\mu_1 < \min_{t\in[n]} \mathrm{UCB}_1(t)\right\} \cap \left\{\hat{\mu}_{i,u_i} + \sqrt{\frac{2}{u_i}\log\left(\frac{1}{\delta}\right)} < \mu_1\right\},$$

where $u_i \in [n]$ is a constant to be chosen later. Then

1. If G_i occurs, then $T_i(n) \leq u_i$.
2. The complement event G_i^c occurs with low probability (governed in some way yet to be discovered by u_i).

Because $T_i(n) \leq n$ no matter what, this will mean that

$$\mathbb{E}\left[T_i(n)\right] = \mathbb{E}\left[\mathbb{I}_{\{G_i\}} T_i(n)\right] + \mathbb{E}\left[\mathbb{I}_{\{G_i^c\}} T_i(n)\right] \leq u_i + \mathbb{P}\left(G_i^c\right) n\,. \quad (13)$$

For details, see our website.

Worst-case Bound

Theorem

If $\delta = 1/n^2$, then the regret of UCB(δ) on any $\nu \in \mathcal{E}_{\mathrm{SG}}^K(1)$ environment is bounded by

$$R_n \leq 8\sqrt{nK\log(n)} + 3\sum_{i=1}^{K} \Delta_i \,.$$

Notes

Recall: For all $\nu \in \mathcal{E}_{\mathrm{SG}}^K(1)$,

$$R_n(\nu) \leq 8\sqrt{nK\log(n)} + 3\sum_{i=1}^{K} \Delta_i \,.$$

- Same tuning as in the other result! Good!
- Not anytime. Hmm..
- The additive $\sum_i \Delta_i$ term is unavoidable: all reasonable algorithms must play each arm once (exercise: what if not??).
- The bound is close to optimal: no algorithm can enjoy regret smaller than $\mathrm{const} \cdot \sqrt{nK}$ over all problems in $\mathcal{E}_{\mathrm{SG}}^K(1)$.
- A more complicated variant of UCB(δ) shaves the logarithmic term from the upper bound given above.

Proof

Previous proof actually gives:

$$\mathbb{E}[T_i(n)] \leq 3 + \frac{16\log(n)}{\Delta_i^2}\,.$$

Using the basic regret decomposition, choosing some $\Delta > 0$,

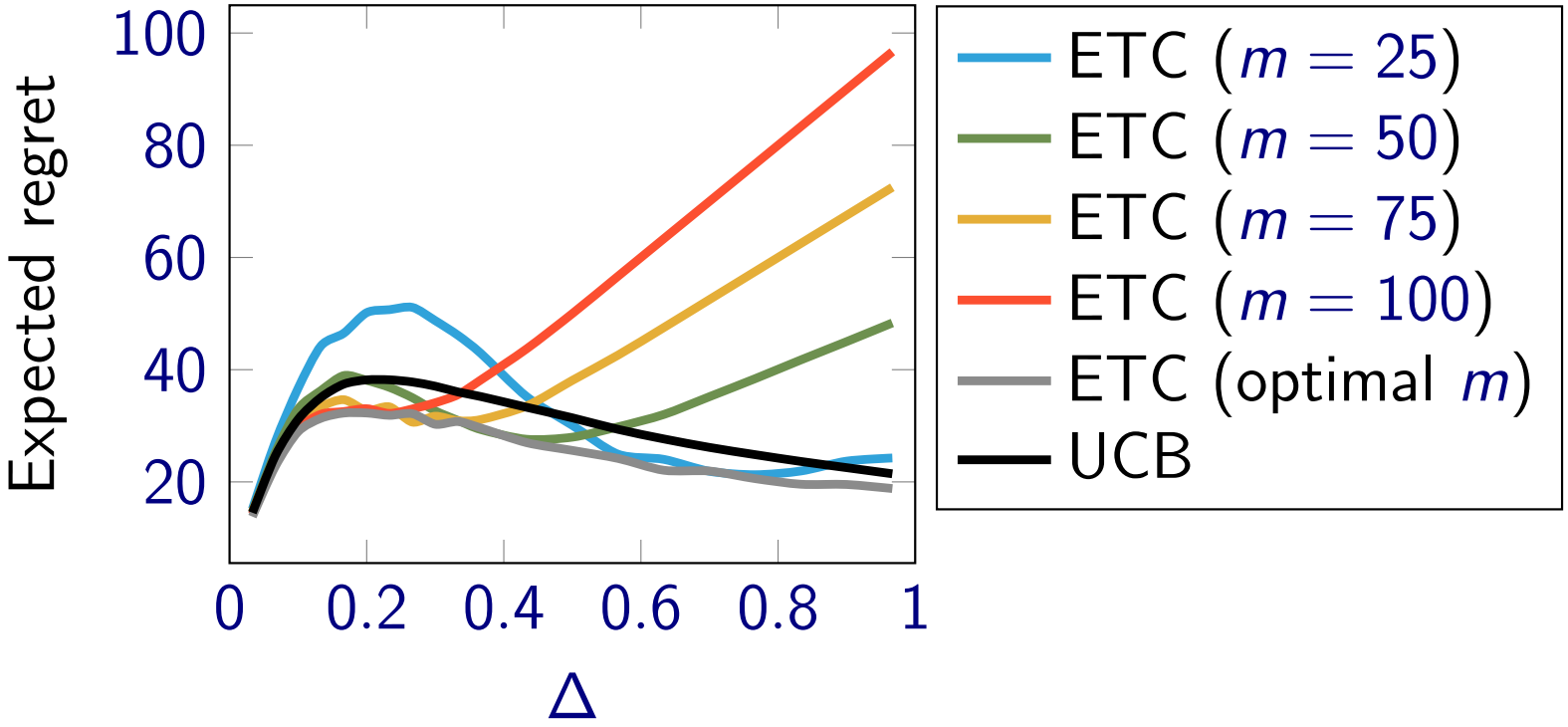
$$\begin{aligned}
R_n &= \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(n)\right] = \sum_{i:\Delta_i<\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] + \sum_{i:\Delta_i\geq\Delta} \Delta_i \mathbb{E}\left[T_i(n)\right] \\
&\leq n\Delta + \sum_{i:\Delta_i\geq\Delta} \left(3\Delta_i + \frac{16\log(n)}{\Delta_i}\right) \leq n\Delta + \frac{16K\log(n)}{\Delta} + 3\sum_i \Delta_i \\
&\leq 8\sqrt{nK\log(n)} + 3\sum_{i=1}^{K} \Delta_i\,,
\end{aligned}$$

where the first inequality follows because $\sum_{i:\Delta_i<\Delta} T_i(n) \leq n$ and the last line by choosing $\Delta = \sqrt{16K\log(n)/n}$. Qu.e.d.

Outline

Demo/Exercise 8

Setup: $n = 1000$, $K = 2$, $P_1 = \mathcal{N}(0, 1)$, $P_2 = \mathcal{N}(-\Delta, 1)$. Plot estimates of the expected regret of UCB, ETC with $m \in \{25, 50, 75, 100, \mathrm{Optimum}\}$ as $\Delta \in [0, 1]$. Your plot should resemble this:



What can we conclude?

Exercise 9/Demo

Compare the regret histogram of UCB and ETC.

Demo

Literature

- Confidence bounds, OFU principle is by Lai and Robbins (1985) (context: parametric bandits, asymptotics).
- UCB algorithm: Katehakis and Robbins (1995) (Gaussian bandits) and Agrawal (1995). Still asymptotics. Agrawal (1995)'s analysis is modular: All that is needed is appropriate UCBs (the form is irrelevant).
- Independently, Kaelbling (1993) also discovered UCB. No regret analysis, or clear advice on how to tune the confidence parameter.
- "This" UCB is most similar to UCB1 of Auer et al. (2002), except that n is t in UCB1. Finite-time regret bound, $[0, 1]$ bounded payoffs (i.e., $1/2$ subgaussian).
- Worst-case bound: Bubeck and Cesa-Bianchi (2012), which focuses on the subgaussian setup.

Questions?

Outline

Asymptotically Optimal UCB (AO-UCB)

1: **Input** K
2: Choose each arm once
3: Subsequently choose

$$A_t = \operatorname*{argmax}_i \left(\hat{\mu}_i(t-1) + \sqrt{\frac{2 \log f(t)}{T_i(t-1)}} \right)$$

where $f(t) = 1 + t \log^2(t)$

Asymptotic Regret, Asymptotic Optimality

Theorem (Upper Bound)

The regret of AO-UCB satisfies

$$\limsup_{n\to\infty} \frac{R_n(\nu)}{\log(n)} \leq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}\,. \tag{14}$$

Theorem (Lower Bound)

For any policy π that has subpolynomial regret for all 1-subgaussian environments (i.e., $R_n(\nu,\pi) = o(n^p)$ for all $p > 0$ and all ν), for any instance ν with gaps $\Delta = \Delta(\nu)$,

$$\liminf_{n\to\infty} \frac{R_n(\nu,\pi)}{\log(n)} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}\,. \tag{15}$$

Finite-time Regret for AO-UCB

Corollary

there exists some universal constant $C > 0$ *such that the regret of AO-UCB is bounded by*

$$R_n \leq C \sum_{i:\Delta_i > 0} \left(\Delta_i + \frac{\log(n)}{\Delta_i} \right) ,$$

and, in particular,

$$R_n \leq C \sum_{i=1}^{K} \Delta_i + 2\sqrt{CnK \log(n)} \, .$$

Outline

Zoo of UCBs

Strategy	Conf.	Dist. free regret	Asy. Opt.	Inst. Opt.	Anytime
UCB Auer et al. (2002)	t	$\sqrt{kn\log(n)}$	✓	✓	✓
UCB* Lai (1987)	n/T	$\sqrt{kn\log(k)}$	✓	✓	✗
UCB+ Garivier and Cappé (2011)	t/T	$\sqrt{kn\log(k)}$	✓	✓	✓
MOSS Audibert and Bubeck (2009)	$n/(kT)$	$\sqrt{kn}$	✓	✗	✗
Anytime MOSS Degenne and Perchet (2016)	$t/(kT)$	$\sqrt{kn}$	✓	✗	✓
OCUCB Lattimore (2015)	$(n/t)^{1+\varepsilon}$	$\sqrt{kn}$	✗	✓	✗
UCB† Lattimore (2017)	see ref.	$\sqrt{kn}$	✓	✓	✗
UCB‡ Lattimore (2017)	see ref.	$\sqrt{kn\log\log(n)}$	✓	✓	✓

All strategies choose $A_t = t$ for $1 \le t \le k$ and subsequently

$$A_t = \operatorname*{argmax}_i \hat{\mu}_i(t-1) + \underbrace{\sqrt{\frac{2\,\overline{\log}\,(\text{Conf.})}{T_i(t-1)}}}_{\text{exploration bonus}}, \qquad (16)$$

where $\overline{\log}(x) \sim \log(x)$ is approximately logarithmic. In the table T stands for $T_i(t-1)$.

Demo with UCB and OCUCB:
http://downloads.tor-lattimore.com/bandits/

Dealing with Risk

Setup: $K = 2$, Gaussian noise, gap Δ. If δ is the probability of missing the optimal arm by some (reasonable) algorithm then

$$R_n = O\left(n\Delta\delta + \frac{1}{\Delta}\log\left(\frac{1}{\delta}\right)\right) ,$$

Optimize δ to get $\delta = 1/(n\Delta^2)$ and

$$R_n = O\left(\frac{1}{\Delta}\left(1 + \log\left(n\Delta^2\right)\right)\right) .$$

How about $\mathbb{E}\left[R_n^2\right]$? We get: $\mathbb{E}\left[R_n^2\right] \approx \delta(n\Delta)^2 = n$. Too big!! Choose $\delta = (n\Delta)^{-2}$ to get

$$R_n = O\left(\frac{1}{\Delta}\left(\frac{1}{n} + \log\left(n^2\Delta^2\right)\right)\right)$$

and $\mathbb{E}\left[R_n^2\right] \approx \log^2(n)$!! For further info, see, Audibert et al. (2007).

Outline

Summary

- Bandits: The drosophila of "Explore-Exploit" problems.
- Learner interacts with its environment, which is initially unknown.
- We care about the total reward/regret.
  - Exploration is necessary to avoid large loss due to (plausible) unlucky start
  - Exploitation is necessary to avoid large loss due to being just curious all the time
  - What is the right amount?
- Stochastic, finite-armed bandits
  - Explore-then-commit: Ideal, but infeasible tuning shows what can be achieved.
  - Optimism does it: UCB.
  - Simultaneously satisfying many optimality criteria is possible with a carefully tuned UCB.
- We left out lower bound proofs and many other topics.

Questions?

References I

Abbasi-Yadkori, Y. (2009). Forced-exploration based algorithms for playing in bandits with large action sets. Master's thesis, University of Alberta, Department of Computing Science.

Abbasi-Yadkori, Y., Antos, A., and Szepesvári, C. (2009). Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*.

Agrawal, R. (1995). Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, pages 1054–1078.

Audibert, J.-Y. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of Conference on Learning Theory (COLT)*, pages 217–226.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory (ALT)*, pages 150–165. Springer.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE.

Auer, P. and Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

References II

Bather, J. and Chernoff, H. (1967). Sequential decisions in the control of a spaceship. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, pages 181–207.

Berry, D. and Fristedt, B. (1985). *Bandit problems : sequential allocation of experiments*. Chapman and Hall, London ; New York :.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford.

Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Degenne, R. and Perchet, V. (2016). Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*.

Dudley, R. M. (2014). *Uniform central limit theorems*, volume 142. Cambridge university press.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of Conference on Learning Theory (COLT)*.

Garivier, A., Kaufmann, E., and Lattimore, T. (2016). On explore-then-commit strategies. In *Advances in Neural Information Processing Systems (NIPS)*.

Gittins, J. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177.

References III

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Kaelbling, L. P. (1993). *Learning in embedded systems*. MIT press.

Katehakis, M. N. and Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, pages 817–824.

Lattimore, T. (2015). Optimally confident UCB: Improved regret for finite-armed bandits. *arXiv preprint arXiv:1507.07880*.

Lattimore, T. (2017). Auto-tuning the confidence level for optimistic bandit strategies. *technical report*.

McDiarmid, C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

References IV

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230.

van de Geer, S. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.