Talk for Computer Laboratory

Machines Regret Their Actions Too:

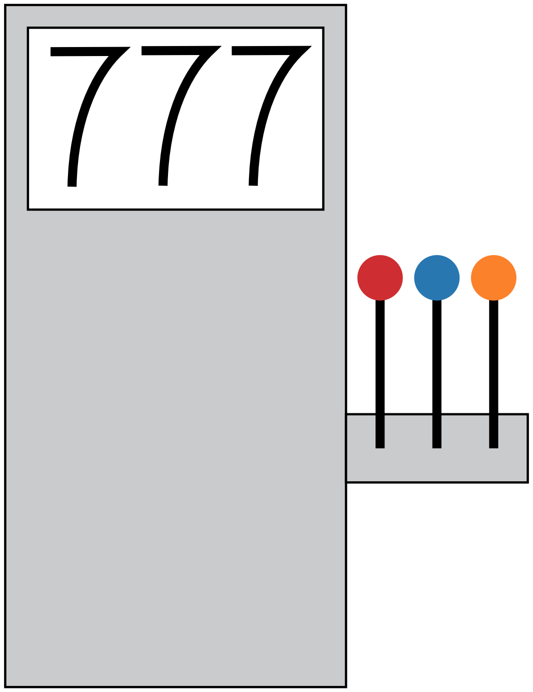
A Brief Tutorial on Multi-armed Bandits

UNIVERSITY OF CAMBRIDGE

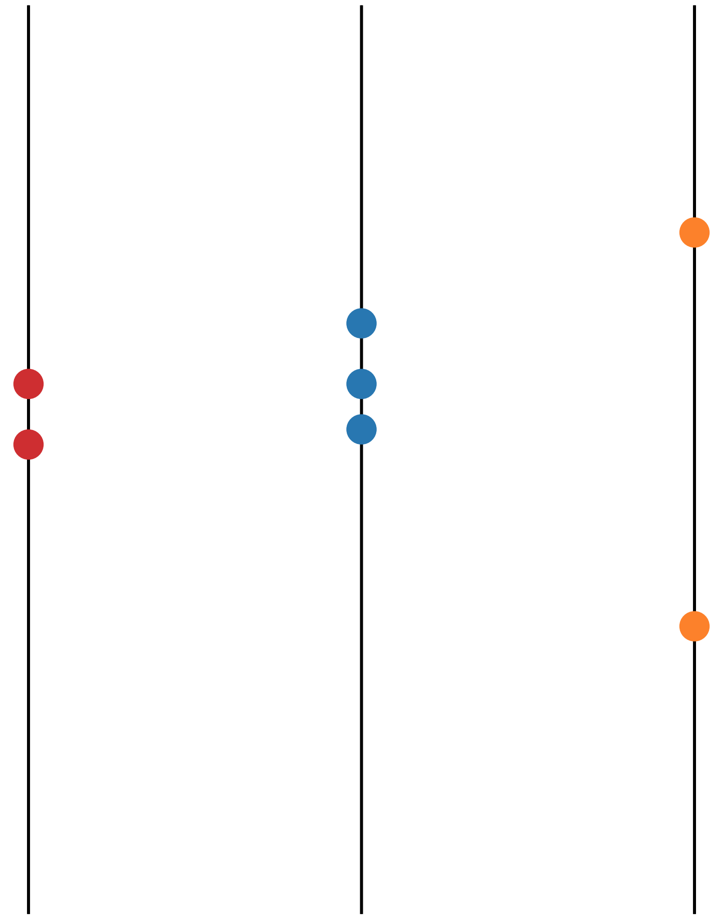
Alexander Terenin

https://avt.im/ · @avt_im

Multi-armed Bandits



Multi-armed Bandits



Formalism

Goal: for a function $f : X \to [0,1]$ where $|X| = K < \infty$ find

$$\max_{x \in X} f(x)$$

based on noisy observations $f(x_t) + \varepsilon_{x_t}$ where ε_{x_t} is random.

Given observations up to time t, how should one choose x_{t+1}?

Formalism

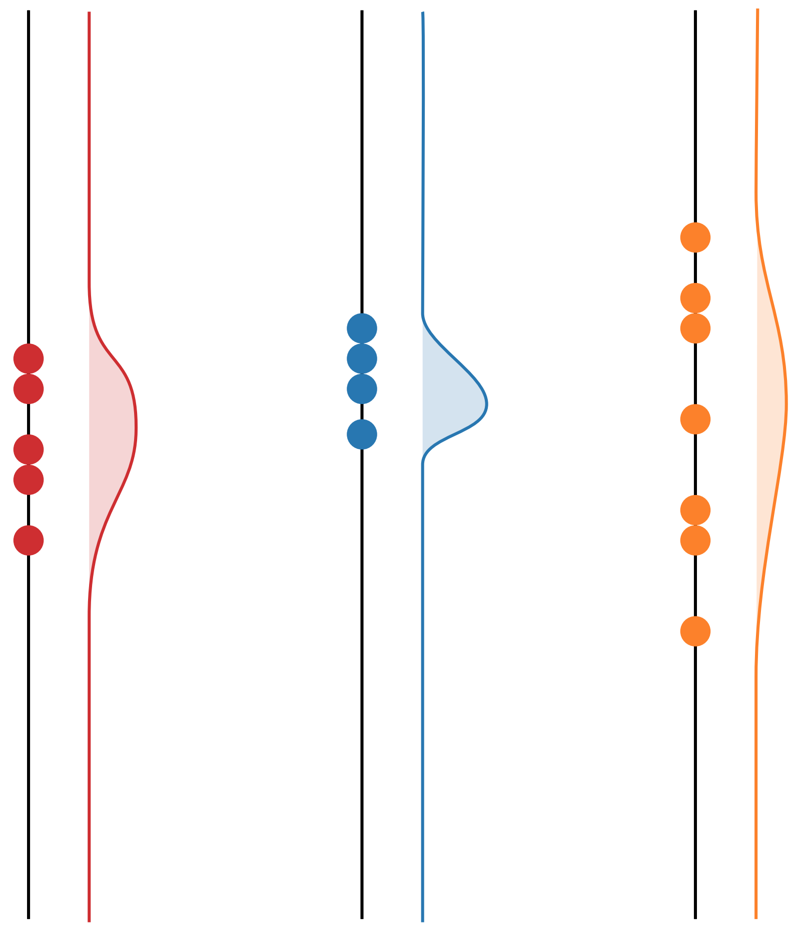
Regret:

$$R(T) = \sum_{t=1}^{T} f(x^*) - f(x_t)$$

where $x^* = \arg\max_{x \in X} f(x)$.

Different strategies yield different regret asymptotics

Balancing Explore-Exploit Tradeoffs



A Regret Lower Bound

Theorem. For any algorithm there is an f such that

$$\mathbb{E}[R(T)] \geq \Omega(\sqrt{KT}).$$

Some regret is always incurred in order to learn f

Error Bars

Result (Hoeffding). Let $y_1, .., y_t$ be an IID sequence of random variables with values in $[0, 1]$. Let $\bar{y}_t$ be their sample mean. Then

$$\mathbb{P}(\mathbb{E}(y_1) > \bar{y}_t + \delta) \leq e^{-2t\delta^2}.$$

Concentration inequalities enable us to construct error bars for f

Error Bars

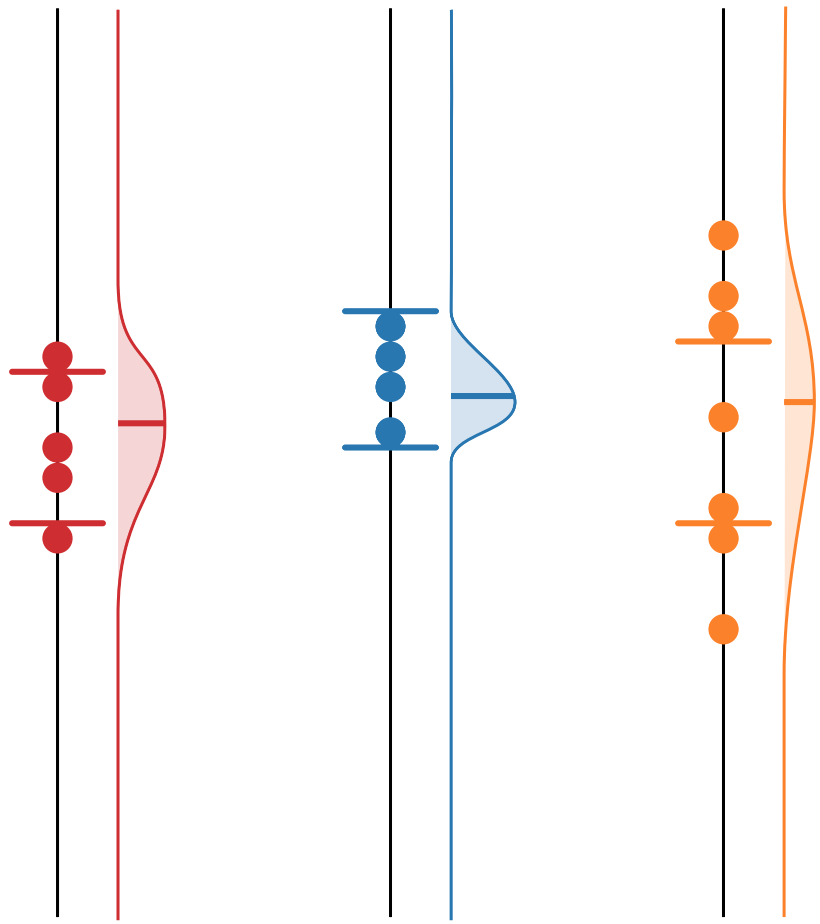
Assume $f(x_t) + \varepsilon_{x_t}$ satisfy Hoeffding's inequality. Choose δ, η so that

$$\mathbb{P}\left(|\bar{y}_t(x) - f(x)| \leq \underbrace{\sqrt{\frac{2 \ln T}{n_t(x)}}}_{\sigma_t(x)} \right) \geq 1 - \eta.$$

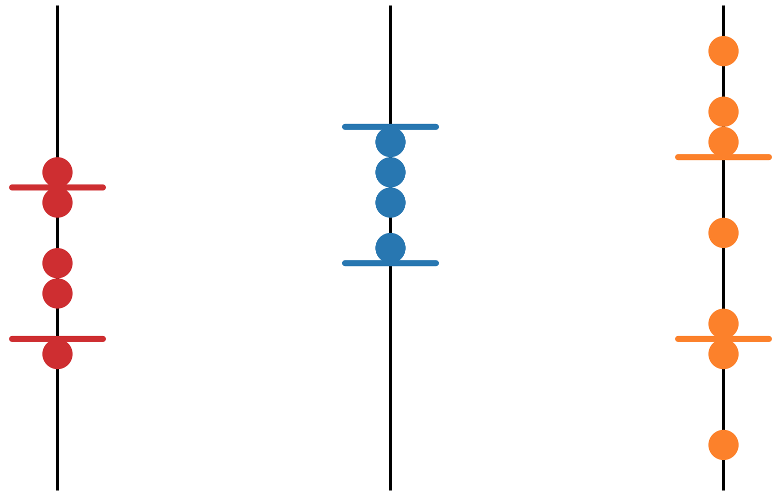
where

- $n_t(x)$ is the expected number of times x is selected by time t, and
- $\bar{y}_t(x_t)$ is the empirical mean of $f(x_t) + \varepsilon_{x_t}$ up to time t.

Error Bars



The Upper Confidence Bound Algorithm



$$x_{t+1} = \arg\max_{x \in X} f_t^+(x_t)$$

$$f_t^{\pm}(x) = \bar{y}_t(x) \pm \sigma_t(x)$$

$\bar{y}_t$: empirical mean

σ_t: error bar width

The Upper Confidence Bound Algorithm

Theorem. Hoeffding–UCB's regret satisfies

$$\mathbb{E}[R(T)] \leq \widetilde{\mathcal{O}}(\sqrt{KT})$$

uniformly for all f.

Well-calibrated error bars lead to asymptotically efficient strategies

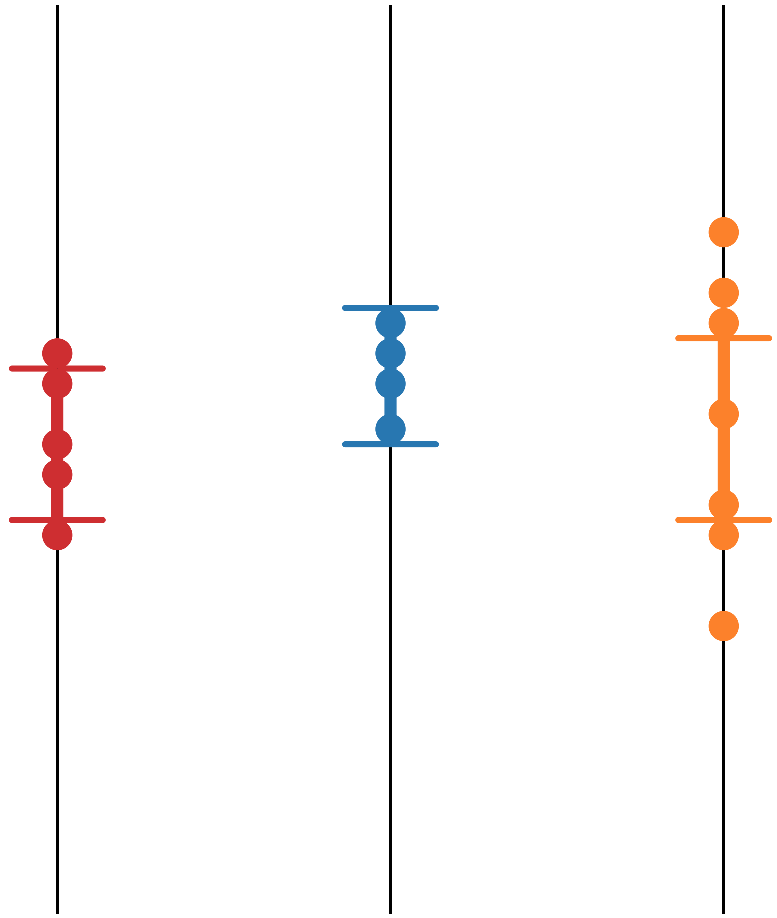
The Upper Confidence Bound Algorithm

Key idea. With sufficiently high probability, we have

$$\begin{aligned}\Delta(x_t) &= f(x^*) - f(x_t) \\ &\leq f_t^+(x^*) - f_t^-(x_t) \\ &\leq f_t^+(x_t) - f_t^-(x_t) \\ &= 2\sigma_t(x_t) = 2\sqrt{\tfrac{2\ln T}{n_t(x)}} \stackrel{t=T}{=} \widetilde{\mathcal{O}}(n_T^{-1/2})\end{aligned}$$

using $f_t^-(x) \leq f(x) \leq f_t^+(x)$, and $f_t^+(x_t) \geq f_t^+(x), \forall x, t$.

The Upper Confidence Bound Algorithm



The Upper Confidence Bound Algorithm

This gives

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{t=1}^{T} \underbrace{f(x^*) - f(x_t)}_{\Delta(x_t)} = \sum_{x \in X} \underbrace{\Delta(x)}_{\widetilde{\mathcal{O}}(n_T^{-1/2})} n_T(x) \\
&\leq \sum_{x \in X} \widetilde{\mathcal{O}}(\sqrt{n_T(x)}) \leq \widetilde{\mathcal{O}}\left(\sqrt{K \sum_{x \in X} n_T(x)}\right) \\
&= \widetilde{\mathcal{O}}(\sqrt{KT}).
\end{aligned}$$

Extensions

- $\widetilde{\mathcal{O}}(\sqrt{KT}) \rightsquigarrow \mathcal{O}(\sqrt{KT})$
- Adversarial and Contextual Bandits

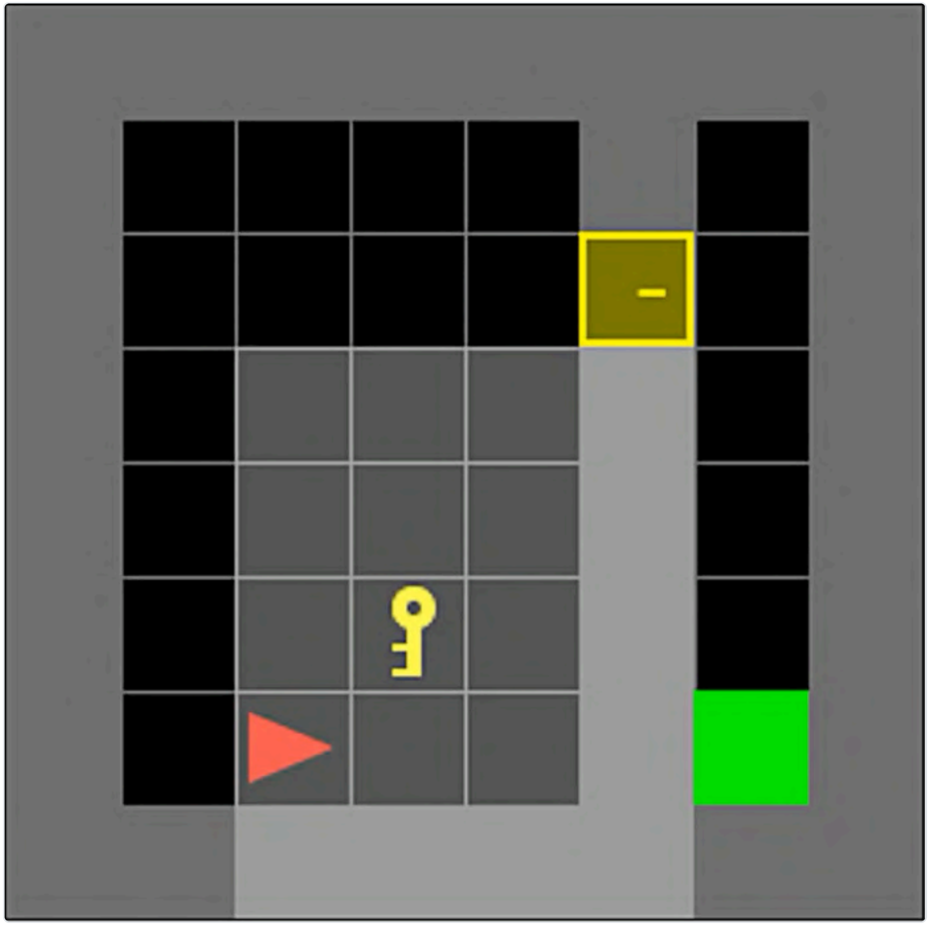
$$\min_{\varepsilon \in \mathcal{E}} \max_{x \in X} f(x)$$

- Bayesian methods and Thompson Sampling

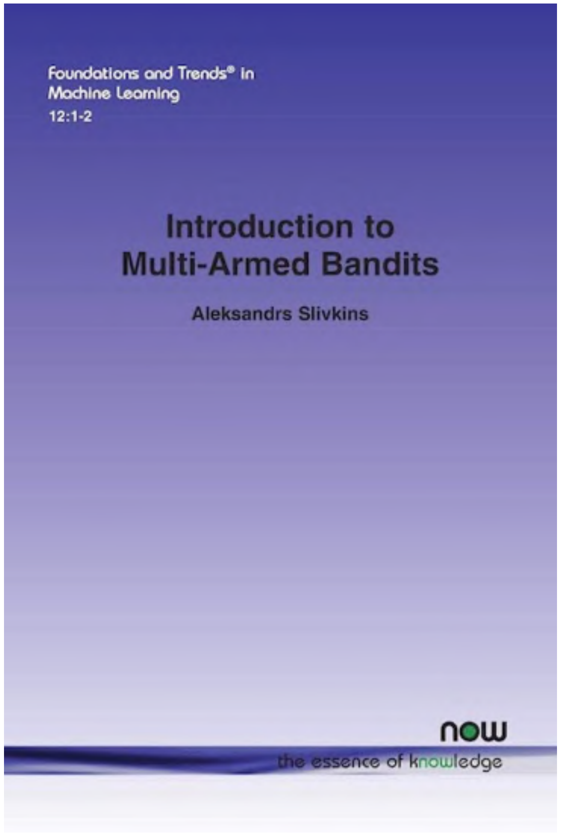
$$x_{t+1} = \arg\max_{x \in X} \phi_t(x) \qquad \phi_t \sim f \mid y_1, .., y_t$$

- Partial Monitoring and Information Directed Sampling

Reinforcement Learning



References


Foundations and Trends® in
Machine Learning
12:1-2
Introduction to
Multi-Armed Bandits
Aleksandrs Slivkins
now
the essence of knowledge


Bandit
Algorithms
TOR LATTIMORE
CSABA SZEPESVÁRI

Thank you!

HTTPS://AVT.IM/ · @AVT_IM

UNIVERSITY OF CAMBRIDGE