Foundations of machine learning

Multi-armed bandits

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Outline

- Thus far: “Supervised machine learning” – data are given.
  Next: “Active learning” – experimentation.
- Setup: The multi-armed bandit problem.
  Adaptive experiment with exploration / exploitation trade-off.
- Two popular approximate algorithms:
  1. Thompson sampling
  2. Upper Confidence Bound algorithm
- Characterizing regret – fixed parameter asymptotics, local-to-zero asymptotics.
- Characterizing an exact solution: Gittins Index.
- Extension to settings with covariates (contextual bandits).

Takeaways for this part of class

- When experimental units arrive over time, and we can adapt our treatment choices, we can learn optimal treatment quickly.
- Treatment choice: Trade-off between
  1. choosing good treatments now (exploitation),
  2. and learning for future treatment choices (exploration).
- Optimal solutions are hard, but good heuristics are available.
- We will derive a bound on the regret of one heuristic.
  - Bounding the number of times a sub-optimal treatment is chosen,
  - using large deviations bounds (cf. testing!).
- Worst case regret occurs for intermediate effect sizes that are of order $1/\sqrt{T}$.
- We will also derive a characterization of the optimal solution in the infinite-horizon case. This relies on a separate index for each arm.

The multi-armed bandit

Setup

- Treatments $D_t \in 1, \dots, k$
- Experimental units come in sequentially over time.
  One unit per time period $t = 1, 2, \dots$
- Potential outcomes: i.i.d. over time, $Y_t = Y_t^{D_t}$,

$$Y_t^d \sim F^d \qquad\qquad E[Y_t^d] = \theta^d$$

- Treatment assignment can depend on past treatments and outcomes,

$$D_{t+1} = d_t(D_1, \dots, D_t, Y_1, \dots, Y_t).$$

The multi-armed bandit

Setup continued

- Optimal treatment:

$$d^* = \underset{d}{\operatorname{argmax}}\ \theta^d \qquad\qquad \theta^* = \max_d \theta^d = \theta^{d^*}$$

- Expected regret for treatment d:

$$\Delta^d = E\left[Y^{d^*} - Y^d\right] = \theta^{d^*} - \theta^d.$$

- Finite horizon objective: Average outcome,

$$U_T = \tfrac{1}{T} \sum_{1 \le t \le T} Y_t.$$

- Infinite horizon objective: Discounted average outcome,

$$U_\infty = \sum_{t \ge 1} \beta^t Y_t$$

The multi-armed bandit

Expectations of objectives

- Expected finite horizon objective:

$$E[U_T] = E\left[\frac{1}{T}\sum_{1\le t\le T}\theta^{D_t}\right]$$

- Expected infinite horizon objective:

$$E[U_\infty] = E\left[\sum_{t\ge 1}\beta^t\theta^{D_t}\right]$$

- Expected finite horizon regret:
  Compare to always assigning optimal treatment d^*.

$$R_T = E\left[\frac{1}{T}\sum_{1\le t\le T}\left(Y_t^{d^*} - Y_t\right)\right] = E\left[\frac{1}{T}\sum_{1\le t\le T}\Delta^{D_t}\right]$$

Practice problem

- Show that these equalities hold.
- Interpret these objectives.
- Relate them to our decision theory terminology.

Two popular algorithms

Upper Confidence Bound (UCB) algorithm

- Define

$$
\begin{aligned}
\bar{Y}_t^d &= \frac{1}{T_t^d} \sum_{1 \le s \le t} 1(D_s = d) \cdot Y_s, \\
T_t^d &= \sum_{1 \le s \le t} 1(D_s = d) \\
B_t^d &= B(T_t^d).
\end{aligned}
$$

- $B(\cdot)$ is a decreasing function, giving the width of the "confidence interval." We will specify this function later.

- At time $t+1$, choose

$$
D_{t+1} = \underset{d}{\operatorname{argmax}}\ \bar{Y}_t^d + B_t^d.
$$

- "Optimism in the face of uncertainty."

Two popular algorithms

Thompson sampling

- Start with a Bayesian prior for θ.
- Assign each treatment with probability equal to the posterior probability that it is optimal.
- Put differently, obtain one draw $\hat{\theta}_{t+1}$ from the posterior given $(D_1, \ldots, D_t, Y_1, \ldots, Y_t)$, and choose

$$D_{t+1} = \underset{d}{\operatorname{argmax}} \ \hat{\theta}^d_{t+1}.$$

- Easily extendable to more complicated dynamic decision problems, complicated priors, etc.!

Two popular algorithms

Thompson sampling - the binomial case

- Assume that $Y \in \{0,1\}$, $Y_t^d \sim Ber(\theta^d)$.
- Start with a uniform prior for θ on $[0,1]^k$.
- Then the posterior for θ^d at time $t+1$ is a $Beta$ distribution with parameters

$$\alpha_t^d = 1 + T_t^d \cdot \bar{Y}_t^d,$$
$$\beta_t^d = 1 + T_t^d \cdot (1 - \bar{Y}_t^d).$$

- Thus

$$D_t = \underset{d}{\operatorname{argmax}}\ \hat{\theta}_t.$$

where

$$\hat{\theta}_t^d \sim Beta(\alpha_t^d, \beta_t^d)$$

is a random draw from the posterior.

Regret bounds

- Back to the general case.
- Recall expected finite horizon regret,

$$R_T = E\left[\frac{1}{T}\sum_{1\le t\le T}\left(Y_t^{d^*} - Y_t\right)\right] = E\left[\frac{1}{T}\sum_{1\le t\le T}\Delta^{D_t}\right].$$

- Thus,

$$T \cdot R_T = \sum_d E[T_T^d] \cdot \Delta^d.$$

- Good algorithms will have $E[T_T^d]$ small when $\Delta^d > 0$.
- We will next derive upper bounds on $E[T_T^d]$ for the UCB algorithm.
- We will then state that for large T similar upper bounds hold for Thompson sampling.
- There is also a lower bound on regret across all possible algorithms which is the same, up to a constant.

Reminder: Large deviations inequality

- Let $\bar{Y}_T = \frac{1}{T}\sum_{1\leq t\leq T} Y_t$ for i.i.d. Y_t.
- Suppose that
$$E[\exp(\lambda \cdot (Y - E[Y]))] \leq \exp(\psi(\lambda)).$$
- Define the Legendre-transformation of ψ as
$$\psi^*(\varepsilon) = \sup_{\lambda \geq 0} [\lambda \cdot \varepsilon - \psi(\lambda)].$$
  - For distributions bounded by $[0,1]$:
    $\psi(\lambda) = \lambda^2/8$ and $\psi^*(\varepsilon) = 2\varepsilon^2$.
  - For normal distributions:
    $\psi(\lambda) = \lambda^2\sigma^2/2$ and $\psi^*(\varepsilon) = \varepsilon^2/(2\sigma^2)$.
- Then
$$P(\bar{Y}_T - E[Y] > \varepsilon) \leq \exp(-T \cdot \psi^*(\varepsilon)).$$

Applied to the Bandit setting

- Suppose that for all d

$$E[\exp(\lambda \cdot (Y^d - \theta^d))] \leq \exp(\psi(\lambda))$$
$$E[\exp(-\lambda \cdot (Y^d - \theta^d))] \leq \exp(\psi(\lambda)).$$

- Recall / define

$$\bar{Y}_t^d = \tfrac{1}{T_t^d} \sum_{1 \leq s \leq t} 1(D_s = d) \cdot Y_s, \qquad B_t^d = (\psi^*)^{-1} \left(\frac{\alpha \log(t)}{T_t^d} \right).$$

- Then we get

$$\begin{aligned} P(\bar{Y}_t^d - \theta^d > B_t^d) &\leq \exp(-T_t^d \cdot \psi^*(B_t^d)) \\ &= \exp(-\alpha \log(t)) = t^{-\alpha} \\ P(\bar{Y}_t^d - \theta^d < -B_t^d) &\leq t^{-\alpha}. \end{aligned}$$

Why this choice of $B(\cdot)$?

- A smaller $B(\cdot)$ is better for exploitation.
- A larger $B(\cdot)$ is better for exploration.
- Special cases:
  - Distributions bounded by $[0, 1]$:

$$B_t^d = \sqrt{\frac{\alpha \log(t)}{2T_t^d}}.$$

  - Normal distributions:

$$B_t^d = \sqrt{2\sigma^2 \frac{\alpha \log(t)}{T_t^d}}.$$

- The $\alpha \log(t)$ term ensures that coverage goes to 1, but slow enough to not waste too much in terms of exploitation.

When d is chosen by the UCB algorithm

- By definition of UCB, at least one of these three events has to hold when d is chosen at time $t+1$:

$$\bar{Y}_t^{d^*} + B_t^{d^*} \leq \theta^* \tag{1}$$
$$\bar{Y}_t^{d} - B_t^{d} > \theta^d \tag{2}$$
$$2B_t^d > \Delta^d. \tag{3}$$

- 1 and 2 have low probability. By previous slide,

$$P\left(\bar{Y}_t^{d^*} + B_t^{d^*} \leq \theta^*\right) \leq t^{-\alpha}, \qquad P\left(\bar{Y}_t^{d} - B_t^{d} > \theta^d\right) \leq t^{-\alpha}.$$

- 3 only happens when T_t^d is small. By definition of B_t^d, 3 happens iff

$$T_t^d < \frac{\alpha \log(t)}{\psi^*(\Delta^d/2)}.$$

Practice problem

Show that at least one of the statements 1, 2, or 3 has to be true whenever $D_{t+1} = d$, for the UCB algorithm.

Bounding $E[T_t^d]$

- Let
$$\tilde{T}_T^d = \left\lfloor \frac{\alpha \log(T)}{\psi^*(\Delta^d/2)} \right\rfloor.$$

- Forcing the algorithm to pick d the first $\tilde{T}_T^d$ periods can only increase T_T^d.

- We can collect our results to get
$$\begin{aligned}
E[T_T^d] = \sum_{1\le t\le T} 1(D_t = d) &\le \tilde{T}_T^d + \sum_{\tilde{T}_T^d < t \le T} E[1(D_t = d)] \\
&\le \tilde{T}_T^d + \sum_{\tilde{T}_T^d < t \le T} E[1((1) \text{ or } (2) \text{ is true at } t)] \\
&\le \tilde{T}_T^d + \sum_{\tilde{T}_T^d < t \le T} E[1((1)\text{is true at } t)] + E[1((2) \text{ is true at } t)] \\
&\le \tilde{T}_T^d + \sum_{\tilde{T}_T^d < t \le T} 2t^{-\alpha+1} \le \tilde{T}_T^d + \frac{\alpha}{\alpha - 2}.
\end{aligned}$$

Upper bound on expected regret for UCB

- We thus get:

$$
\begin{aligned}
E[T_T^d] &\leq \frac{\alpha \log(T)}{\psi^*(\Delta^d/2)} + \frac{\alpha}{\alpha - 2}, \\
R_T &\leq \tfrac{1}{T} \sum_d \left(\frac{\alpha \log(T)}{\psi^*(\Delta^d/2)} + \frac{\alpha}{\alpha - 2} \right) \cdot \Delta^d.
\end{aligned}
$$

- Expected regret (difference to optimal policy) goes to 0 at a rate of $O(\log(T)/T)$ – pretty fast!
- While the cost of "getting treatment wrong" is Δ^d, the difficulty of figuring out the right treatment is of order $1/\psi^*(\Delta^d/2)$. Typically, this is of order $(1/\Delta^d)^2$.

Related bounds - rate optimality

- **Lower bound**: Consider the Bandit problem with binary outcomes and any algorithm such that $E[T_t^d] = o(t^a)$ for all $a > 0$. Then

$$\liminf_{t\to\infty} \frac{T}{\log(T)} \bar{R}_T \geq \sum_d \frac{\Delta^d}{kl(\theta^d, \theta^*)},$$

  where $kl(p,q) = p \cdot \log(p/q) + (1-p) \cdot \log((1-p)/(1-q))$.

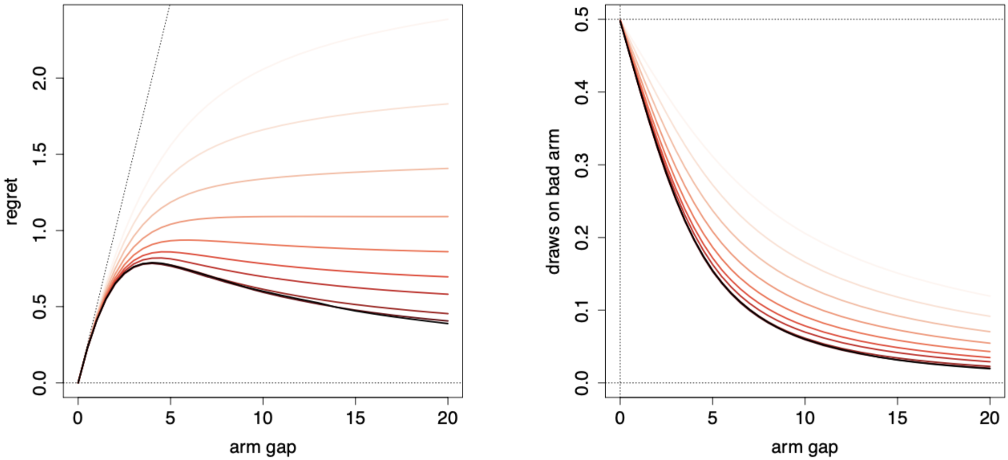
- **Upper bound for Thompson sampling**: In the binary outcome setting, Thompson sampling achieves this bound, i.e.,

$$\liminf_{t\to\infty} \frac{T}{\log(T)} \bar{R}_T = \sum_d \frac{\Delta^d}{kl(\theta^d, \theta^*)}.$$

Local-to-zero asymptotics

- The regret rate we just derived holds θ constant, as $T \to \infty$.
- This provides a good characterization in the "high-powered" case, where it is easy to detect the best treatment quickly.
- What about the low-powered case?
- Here is a heuristic calculation, for two arms, normal outcomes, variance 1:
  1. The probability of correctly identifying the best arm, after $T/2$ observations on each arm, is $\Phi\left(2\sqrt{T}\Delta\right)$.
  2. The regret if we get the arm wrong equals Δ.
  3. Thus the expected average regret is on the order of $\Delta \cdot \Phi\left(-2\sqrt{T}\Delta\right)$.
  4. This vanishes for $\Delta \to 0$ and for $\Delta \to \infty$, and peaks in between, for $\Delta = O(1/\sqrt{T})$, yielding a worst-case average regret of order $1/\sqrt{T}$. (Not $\log(T)/T$, as in the fixed parameter case!)

Limiting regret of two-arm Thompson sampling



From Wager and Xu (2021).
Darker hues indicate a higher prior variance.

Formalizing local-to-zero asymptotics

- Consider a set of sequential experiments, indexed by their sample size T.
- Suppose $\theta^d = \theta_1^d/\sqrt{T}$, and $\sigma^{2d} = \mathrm{Var}(Y^d)$ is the same for all T.
- Denote

$$\tilde{Y}_t^d = \frac{1}{\sqrt{T}} \sum_{s=1}^{t} 1(D_s = d) \cdot Y_s$$
$$\tilde{T}_t^d = \frac{1}{T} \sum_{s=1}^{t} 1(D_s = d).$$

- Assume that the assignment probability for treatment d, p_t^d, is given by a function

$$p_t^d = \psi^d(\tilde{Y}_t, \tilde{T}_t)$$

- This covers, for instance, Thompson sampling for normal outcomes.

Practice problem

Suppose that $Y_t^d \sim N(\theta^d, \sigma^d)$.

- What is the distribution of the stochastic process $\frac{1}{\sqrt{T}} \sum_{s=1}^{t} Y_s^d$?
  What is the limit of this stochastic process?
- Given $\tilde{Y}_t^d$, what is the expectation of $\tilde{T}_{t+1}^d - \tilde{T}_t^d$?
- Given $(\tilde{T}_t^d, \tilde{Y}_t^d)_{d=1}^k$, what is the expectation and variance of $\tilde{Y}_{t+1}^d - \tilde{Y}_t^d$?

Practice problem

Write the expected average regret R_T as a function of $(\tilde{T}_T^d)_{d=1}^k$.

A stochastic differential equation

Theorem 1 in the paper:
Under *Assumption 1*, the stochastic process given by $(\tilde{Y}_t^d, \tilde{T}_t^d)_{d=1}^k$
(with the range of t normalized to $[0,1]$)
converges to the solution of the stochastic differential equations

$$d\tilde{T}_t^d = \psi^d(\tilde{T}_t^d, \tilde{Y}_t^d)dt,$$
$$d\tilde{Y}_t^d = \psi^d(\tilde{T}_t^d, \tilde{Y}_t^d) \cdot \theta^d dt + \sqrt{\psi^d(\tilde{T}_t^d, \tilde{Y}_t^d)}\sigma^d dB_t^d,$$

where B_t^d is a standard Brownian motion.

Gittins index

Setup

- Consider now the discounted infinite-horizon objective, $E[U_\infty] = E\left[\sum_{t\geq 1} \beta^t \theta^{D_t}\right]$, averaged over independent (!) priors across the components of θ.
- We will characterize the optimal strategy for maximizing this objective.
- To do so consider the following, simpler decision problem:
  - You can only assign treatment d.
  - You have to pay a charge of γ^d each period in order to continue playing.
  - You may stop at any time, then the game ends.
- Define γ_t^d as the charge which would make you indifferent between playing or not, given the period t posterior.

Gittins index

Formal definition

- Denote by π_t the posterior in period t, by $\tau(\cdot)$ an arbitrary stopping rule.
- Define

$$\gamma_t^d = \sup\left\{\gamma : \sup_{\tau(\cdot)} E_{\pi_t}\left[\sum_{1\leq s\leq \tau} \beta^s \left(\theta^d - \gamma\right)\right] \geq 0\right\}$$
$$= \sup_{\tau(\cdot)} \frac{E_{\pi_t}\left[\sum_{1\leq s\leq \tau} \beta^s \theta^d\right]}{E_{\pi_t}\left[\sum_{1\leq s\leq \tau} \beta^s\right]}$$

- Gittins and Jones (1974) prove: The optimal policy in the bandit problem always chooses

$$D_t = \underset{d}{\operatorname{argmax}}\ \gamma_t^d.$$

Heuristic proof (sketch)

- Imagine a per-period charge for each treatment is set initially equal to γ_1^d.
  - Start playing the arm with the highest charge, continue until it is optimal to stop.
  - At that point, the charge is reduced to γ_t^d.
  - Repeat.
- This is the optimal policy, since:
  1. It maximizes the amount of charges paid.
  2. Total expected benefits are equal to total expected charges.
  3. There is no other policy that would achieve expected benefits bigger than expected charges.

Contextual bandits

- A more general bandit problem:
  - For each unit (period), we observe covariates X_t.
  - Treatment may condition on X_t.
  - Outcomes are drawn from a distribution $F^{x,d}$, with mean $\theta^{x,d}$.
- In this setting Gittins' theorem fails when the prior distribution of $\theta^{x,d}$ is not independent across x or across d.
- But Thompson sampling is easily generalized.
  For instance to a hierarchical Bayes model:

$$\begin{aligned} Y^d|X = x, \theta, \alpha, \beta &\sim Ber(\theta^{x,d}) \\ \theta^{x,d}|\alpha, \beta &\sim Beta(\alpha^d, \beta^d) \\ (\alpha^d, \beta^d) &\sim \pi. \end{aligned}$$

- This model updates the prior for $\theta^{x,d}$ not only based on observations with $D = d, X = x$, but also based on observations with $D = d$ & different values for X.

References

- *Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Non-stochastic Multi-armed Bandit Problems.* Foundations and Trends® in Machine Learning, *5(1):1–122.*
- *Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling.* Foundations and Trends® in Machine Learning, *11(1):1–96.*
- *Wager, S. and Xu, K. (2021). Diffusion asymptotics for sequential experiments.* arXiv preprint arXiv:2101.09855.
- *Weber, R. et al. (1992). On the Gittins index for multiarmed bandits.* The Annals of Applied Probability, *2(4):1024–1033.*