Modern Sampling Methods

Class 7: Bandits

January 11, 2022

Outline

- Introduction
- Stochastic Bandit Framework
- Explore-then-commit algorithm
- UCB algorithm
- Thompson Sampling
- Issues and Extensions

Introduction: Sequential Treatment Choice

Suppose we have a sequence of individuals $i = 1, 2, \ldots$ and we can choose treatment for each individual based on observations of the prior subjects' outcomes.

We might be interested not in estimating the ATE precisely, but in allocating treatment so that the individuals in the study benefit as much as possible.

This type of problem is studied in the literature on multi-armed stochastic bandits.

Stochastic Bandits: Setup

Treatment Arms: $\mathcal{T}$ (e.g. $\mathcal{T} = \{0, 1\}$).

Potential Outcome Distributions: $F = \{F_t : t \in \mathcal{T}\}$

(e.g. $F = \{F_0, F_1\}$ with $Y(0) \sim F_0$, $Y(1) \sim F_1$).

Rounds/Subjects: $i = 1, \ldots, n$, arriving sequentially.

At each i:

1. Choose $T_i \in \mathcal{T}$,
2. Realize

$$Y_i \mid T_1, Y_1, \ldots, T_{i-1}, Y_{i-1}, T_i = t \quad \sim \quad F_t.$$

Dynamic Treatment Rule: $\pi = (\pi_1, \ldots, \pi_n)$ where

$$T_i \sim \pi_i(\cdot \mid T_1, Y_1, \ldots, T_{i-1}, Y_{i-1}).$$

The pair (π, F) generates a joint distribution for

$$(T_1, Y_1, \ldots, T_n, Y_n).$$

We will mostly focus on the case where $\mathcal{T}$ is finite, but the general framework also allows more general treatment/action spaces.

Regret

Focus on expected outcomes.

Let

$$\mu_t = \mu_t(F) = \int y dF_t(y) \quad \text{(assume exists } \forall t\text{)}.$$

Let $\mu^* = \mu^*(F) = \max_{t \in \mathcal{T}} \mu_t(F)$.

Regret:

$$R_n(\pi, F) = n\mu^* - E\left[\sum_{i=1}^{n} Y_i\right].$$

Note that $R_n(\pi, F) \geq 0$, with equality if $T_i = \arg\max \mu_t$ for all i.

For example, suppose $\mathcal{T} = \{0, 1\}$ and

$$T_i \sim \text{Bernoulli}(1/2).$$

Then half the time, we assign the inferior treatment, with shortfall

$$\Delta = |\mu_1 - \mu_0| .$$

$$\Rightarrow \quad R_n(\pi, F) = n \cdot \frac{\Delta}{2}.$$

(About the same if we simply alternate $T = 0, 1, 0, 1, \ldots$)

Can we find a policy π s.t.

$$\frac{R_n(\pi, F)}{n} \to 0 \quad \text{as } n \to \infty \quad ?$$

Remark: relation to dynamic programming

The bandit problem can be viewed as a dynamic programming problem with uncertainty/ambiguity about F:

- If we put a prior distribution on the unknown F, then a Bayesian solution is possible in principle.
- Except in special cases, full Bayes solution is extremely difficult to compute. (But see Gittins (1989) for some powerful results and see Adusumilli (2021) for recent progress on approximate Bayes solutions.)
- Bayes solution may also be sensitive to prior specification.
- We will focus on heuristic algorithms and discuss their frequentist performance uniformly over classes of outcome distributions.

Useful Lemma

For $t \in \mathcal{T}$, let the suboptimality gap of arm t be

$$\Delta_t = \Delta_t(F) = \mu^*(F) - \mu_t(F).$$

For $j = 1, \ldots, n$, let the count of arm j be

$$N_t(j) = \sum_{i=1}^{j} 1(T_i = t).$$

Lemma: For any π, F with $\mathcal{T}$ finite,

$$R_n(\pi, F) = \sum_{t \in \mathcal{T}} \Delta_t E\left[N_t(n)\right].$$

Explore-Then-Commit Algorithm

Let

$$\hat{\mu}_t(j) = \frac{1}{N_t(j)} \sum_{i=1}^{j} 1(T_i = t) Y_i$$

be the empirical average of arm t at round j.

ETC Algorithm with k arms and tuning parameter m:

1. Explore: in first mk rounds, alternate between the arms.
2. Commit: after round mk, always choose the arm with highest value of $\hat{\mu}_t(mk)$.

Basic tradeoff:

- If m is too large, the algorithm spends too many rounds sampling inferior arms in the "explore" phase.
- If m is too small, there is a higher probability of selecting an inferior arm for the "commit" phase.

The (infeasible) optimal choice for m depends on the distributions F_t and on n, k.

With the infeasible optimal m, can show that (under some conditions) $R_n \approx C\sqrt{n}$.

However, feasible versions of ETC converge more slowly.

Upper Confidence Bound Algorithm

Due to Lai & Robbins (1985).

Idea is to choose treatment arm based on an "optimistic" estimate of μ_t.

For each arm t, let

$$\text{UCB}_t(j-1) = \hat{\mu}_t(j-1) + \hat{u}(\cdot),$$

where $\hat{u}(\cdot)$ can depend on n, $N_t(j-1)$, etc.

Think of $\text{UCB}_t(j-1)$ as the upper endpoint of a confidence interval for μ_t based on data up to time $j-1$.

UCB Algorithm:

1. Initialize by alternating treatment arms once.
2. For subsequent times j, choose:

$$T_j = \arg\max_{t \in \mathcal{T}} \text{UCB}_t(j-1).$$

If we choose $\hat{u}(\cdot)$ appropriately, then under some conditions can show

$$R_n \approx \log(n),$$

which means that $R_n/n \to 0$ fairly quickly as $n \to \infty$.

Details of UCB

Subgaussianity: a random variable X is σ-subgaussian if, for all $\lambda \in \mathbb{R}$, $E[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$.

Subgaussianity means that the tails of the distribution behave approximately like a $N(0, \sigma^2)$ random variable. This lets us control the probability of extreme values of, e.g., the sample average.

In our bandit problem, suppose that every potential outcome distribution F_t is 1-subgaussian. For UCB use:

$$\text{UCB}_t(j-1) = \hat{\mu}_t(j-1) + \sqrt{\frac{2\log f(j)}{N_t(j-1)}},$$

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

Theorem (see Lattimore and Szepesvári, Ch. 8):

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq \sum_{t:\Delta_t>0} \frac{2}{\Delta_t}.$$

(This turns out to be asymptotically optimal, in the sense of giving the best possible rate.)

Thompson Sampling

From Thompson (1933), the first paper on bandits.

Based on a "naive" application of Bayesian updating:

- Initialize with independent prior distributions for each F_t.
- At every stage j, let $F_t(j-1)$ be the posterior updated distribution for arm t based on information available, and let $p_t(j-1)$ be the posterior probability that arm t has highest payoff.
- Choose T_j from $\mathcal{T}$ with probabilities $\{p_t(j-1) : t \in \mathcal{T}\}$.

Thompson Sampling: Example

Suppose $\mathcal{T} = \{0, 1\}$ and Gaussian model for potential outcomes:

$$Y(0) \sim N(\mu_0, 1), \quad Y(1) \sim N(\mu_1, 1).$$

Initialize prior, say:

$$\mu_0 \sim N(0, 1/\gamma), \quad \mu_1 \sim N(0, 1/\gamma).$$

Under this prior, $\Pr(\mu_1 > \mu_0) = 1/2$, so:

Step 1: Draw T_1 with

$$\Pr(T_1 = 1) = 1/2,$$

and observe

$$Y_1 = T_1 Y_1(1) + (1 - T_1) Y_1(0).$$

Suppose $T_1 = 1$, then the posterior for μ_1 becomes

$$\mu_1 \sim N\left(Y_1/(1+\gamma), 1/(1+\gamma)\right),$$

while the posterior for μ_0 does not change.

Step 2:
Now $\Pr(T_2 = 1) =$ probability that $\mu_1 > \mu_0$ based on the updated distributions. Can implement this by:

- Draw $\mu_0(2) \sim N(0, 1/\gamma)$
- Draw $\mu_1(2) \sim N(Y_1/(1+\gamma), 1/(1+\gamma))$.
- Set $T_2 = 1(\mu_1(2) > \mu_0(2))$.

Then observe $Y_2 = T_2 Y_2(1) + (1 - T_2) Y_2(0)$.

At Step j:

- Draw

$$\mu_0(j) \sim N\left(\frac{\sum_{i=1}^{j-1}(1-T_i)Y_i}{N_0(j-1)+\gamma}, \frac{1}{N_0(j-1)+\gamma}\right).$$

- Draw

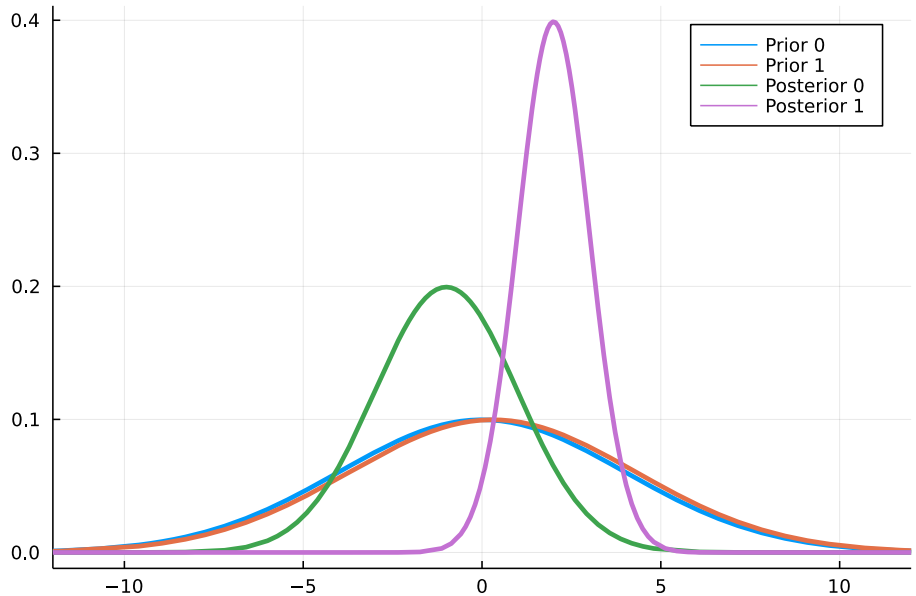
$$\mu_1(j) \sim N\left(\frac{\sum_{i=1}^{j-1} T_i Y_i}{N_1(j-1)+\gamma}, \frac{1}{N_1(j-1)+\gamma}\right).$$

- Set $T_j = 1(\mu_1(j) > \mu_0(j))$.
- Observe

$$Y_j = T_j Y_j(1) + (1-T_j) Y_j(0).$$

We are randomizing T_i at each step, but the probabilities are changing:

- As we obtain more observations on an arm, the posterior for that arm will become tighter:



- If one arm is better than the others, then probability of selecting that arm will increase.

Some Advantages of Thompson Sampling

- Thompson sampling is not a fully forward-looking Bayesian solution to the bandit problem and is not Bayesian-optimal.
- But under some conditions, it achieves the same optimal rate as UCB.
- Unlike ETC and UCB, Thompson sampling involves randomization of T_i, which can facilitate randomization-based inference (to be discussed later).
- Can do "batched" Thompson sampling, e.g.:
  - Batch 1: randomized experiment with equal arm probabilities.
  - Batch 2: randomized experiment, but arm probabilities adjusted based on results of Batch 1.

Issues and Extensions

Classic bandit algorithms target regret R_n. This may not be the most appropriate objective.

Other objectives:

- identifying the best arm with high probability
- identifying the m best arms
- balancing regret minimization with tests for significance
- etc.

So far, we have considered a basic type of bandit with finite set of arms, no covariates immediate observation of outcomes, etc.

- Specification/structural modeling of arm distributions F_t.
- Constraints on choices (e.g. capacity constraints)
- Contextual bandits: incorporate covariates/stratification
- Batched bandits; delayed observation of response
- Richer set of arms (e.g. pricing and other continuous policy spaces)
- Can also study and analyze bandits in continuous time (sometimes a useful approximation)

ETC, UCB, and Thompson Sampling can often be extended to handle these richer settings, as we'll see next.