Modern Sampling Methods

Class 8: Bandit Applications and Extensions

January 11, 2022

Outline

- Dynamic Pricing
- Online advertising and recommendation systems
- Development experiments

A Simple Dynamic Pricing Problem

Consider a monopolist facing an unknown demand function, who can vary prices dynamically and obtain (noisy) observations.

Unknown "true" demand: $D(p)$ where p is price.

Revenue/profit:

$$\texttt{profit}(p) = p \cdot D(p).$$

Seller can vary prices over time to learn about $D(\cdot)$.

Tradeoff between exploration/learning and in-sample revenue maximization.

Time periods (or customers): $i = 1, \ldots, n$.

At each time i:

1. Choose price from a finite menu

$$p_i \in \mathcal{P} = \{p^{(1)}, \ldots, p^{(K)}\},$$

   based on information obtained up to time $i - 1$.
2. Observe a (noisy) signal of profit Y_i with

$$E[Y_i | p_i] = \texttt{profit}(p_i).$$

Seller wants to maximize sum of profits over $i = 1, \ldots, n$.

This can be fit into the multiarmed bandit framework with

- Treatment arms: $t = p$, $\mathcal{T} = \mathcal{P}$.
- Arm means: $\mu_t = \texttt{profit}(p) = p \cdot D(p)$.
- Regret is the (undiscounted) lost profit relative to infeasible optimal monopoly pricing:

$$R_n = \sum_{i=1}^{n} \left[\texttt{profit}(p^*) - \texttt{profit}(p_i)\right],$$

where $p^* = \arg\max_{p \in \mathcal{P}} \texttt{profit}(p)$.

Idea of viewing dynamic pricing problem as a multiarmed bandit dates back to Rothschild (1974).

Dynamic Pricing in E-Commerce

- Can change prices and observe market response relatively quickly.
- At sufficiently high frequency, strategic considerations may be muted.
- Applications often have many goods with many prices, but that can be fit into the current framework, with p as a price vector, etc.

But applications often also involve:

- Need to put more structure on demand model;
- Inventory management;
- Demand dynamics.

Inventory Constraints

Besbes & Zeevi (2009): ETC

- Exploration phase: randomize uniformly over some discretized set of prices; estimate demand nonparametrically.
- Optimization phase: solve profit-maximization problem with inventory constraints

Badanidiyuru, Kleinberg, Slivkins (2013, 2017): UCB

- Bandits with knapsacks: choosing an arm consumes certain resources and generates payoffs.
- Propose variations of the upper confidence bound algorithm and analyze their properties.

Ferreira, Simchi-Levi, & Wang (2018)

Multiple goods: $g \in \{1, \ldots, G\}$.

Price p and demand $D(p)$ are G-vectors. $\mathcal{P}$ is a finite set.

Parametric modeling of demand: $Y(p) \sim F_\theta(p)$, and

$$D(p) = E_\theta[Y(p)].$$

Inventory:

- producing one unit of good g costs b_{gm} units of input $m \in \{1, \ldots, M\}$.
- there is a fixed budget of inputs $B = (B_1, \ldots, B_M)$

Thompson Sampling with Inventory Constraints:

Initialize a prior for θ.

At each time i:

1. Obtain a draw θ_i from the current posterior distribution for θ.
2. Choose $p_i \in \mathcal{P}$ to solve the revenue-maximization problem given θ_i subject to resource constraints. (A linear program.)
3. Observe Y_i, and update the posterior distribution for θ.

Note: in step 2, also allow for mixed strategies over $\mathcal{P}$ due to inventory constraints.

Misra, Schwartz, Abernethy (2019)

Dynamic pricing with robust demand estimation.

Focus on single-product case with a finite menu of prices for simplicity.

Recall that UCB involves

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

where $\hat{\mu}_p(j-1)$ is a sample-average estimate of `profit` under price p based on data up to time $(j-1)$.

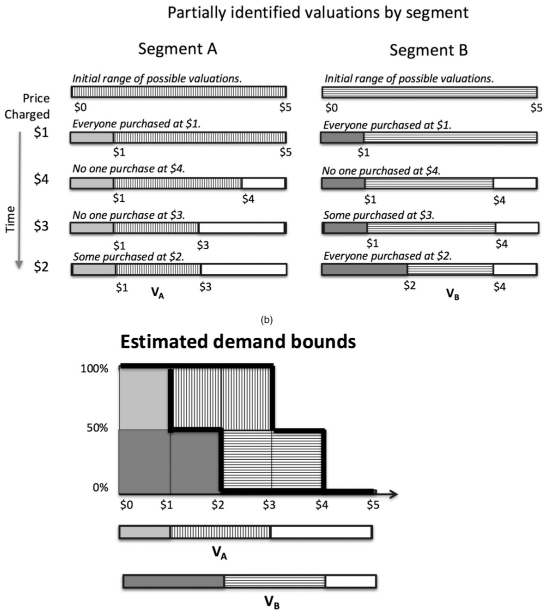
Idea is to modify $\mathrm{UCB}_p(j-1)$ to reflect nonparametric bounds on demand.

Building on Handel and Misra (2015) and Manski bound approach:

Market segments $1, \ldots, S$ with known sizes.

In segment s, consumer valuations $v \in [\nu_s - \delta, \nu_s + \delta]$.

At any time j, use past data to obtain bounds on ν_s, bounds on overall demand, and bounds on `profit`.



From Misra et al, 2019.

Modified UCB:

If p is not dominated by another price based on nonparametric bounds, set

$$\mathrm{UCB}_p(j-1) = \hat{\mu}_p(j-1) + \sqrt{\frac{2\log f(j)}{N_t(j-1)}}.$$

Otherwise, set

$$\mathrm{UCB}_p(j-1) = 0.$$

This rules out arms (prices) that the nonparametric bounds analysis indicates are inferior.

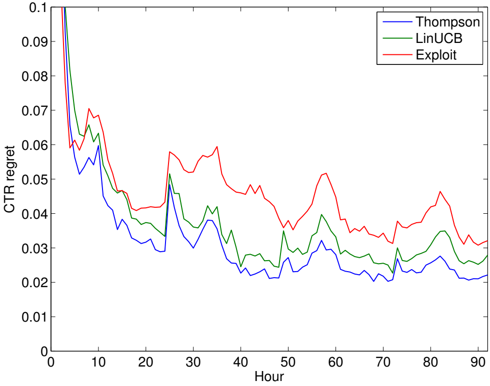
(See paper for addt'l details, including a scaling of the exploration bonus.)

Online Advertisement and Recommendation Systems

Chapelle & Li (2011)

- User visiting a web page, being served an advertisement.
- Arms: set of possible advertisements
- Outcome: ad click-through
- Mean outcome: click-through rate (CTR)
- Thompson sampling with logistic regression model for CTR
- Also applied to news article recommendation

Thompson sampling has very good small sample performance.



From Chapelle & Li (2011)

Schwartz, Bradlow, & Fader (2017):

Worked with a large retail bank.

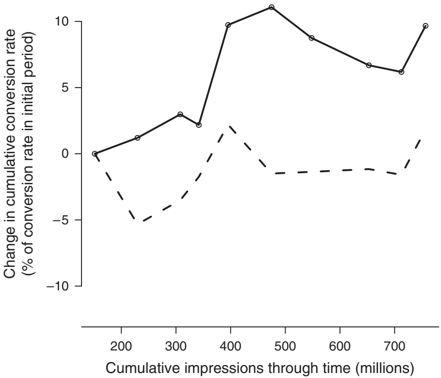
Treatments/Arms: online ad placements characterized by publisher, targeted group, ad size. 532 arms

Outcomes: impressions; clicks; conversions

Thompson Sampling:

- with a hierarchical GLM (logit with random coefficients)
- used Laplace approx. posterior for computational speed
- batched bandit – reallocations done in batches
- ran a horse-race against 'control' policy of uniform allocation across arms

Figure 3. Results Observed in the Field Experiment



Batched Thompson algorithm resulted in 8% higher conversion rate than control.

From Schwartz, Bradlow, & Fader (2017)

Remark: Alternative Objectives

So far, we have considered applications where the goal is to maximize in-sample payoffs (minimize in-sample regret).

This is different from, and can be in tension with:

- Hypothesis testing about parameters;
- Point estimation of arm means;
- Choice of policy for *future* subjects based on data from the completed experiment;
- etc.

Alternative bandit policies can be used to balance in-sample optimal allocation and other objectives

Bandits in Development Economics

Caria, Gordon, Kasy, Quinn, Shami, & Teytelboym (2021):

- Field experiment on job-finding interventions in Jordan
- Four treatment arms: cash transfer; information intervention; behavioral nudge; and control.
- 16 strata based on refugee status, gender, education, work experience.
- Outcome: employment indicator (observed with delay)
- Want to balance welfare of experimental participants with statistical inference on the treatment effects, so classic bandit algorithms may not be well suited.

Tempered Thompson Sampling: at each decision stage and stratum:

- With probability γ, choose arms with equal probability;
- With probability $1-\gamma$, choose arm based on Thompson sampling.

This ensures that probability of any arm will never fall below $\gamma/4$.

Implemented with a hierarchical model: for stratum s, individual i, arm t

$$
\begin{aligned}
Y_{si}(t) &\sim \text{Bernoulli}(\theta_{ts}), \\
\theta_{ts} &\sim \text{Beta}(\alpha_s, \beta_s),
\end{aligned}
$$

and a prior is placed on the hyperparameters α_s, β_s.

- Overall average treatment effects appear to be small, but some evidence for strata-specific gains.
- For implementation, need to observe outcome without too much delay; used employment 6 weeks later.
- Also observed employment at 2 and 4 months and used these for post-experimental analysis.
- If short-run outcome is not a good surrogate for long-run outcome, then there may be limited gains from adaptive experimentation.
- Tempered TS (see also Kasy & Sautmann 2021 for a related scheme) helps with statistical inference, but care is still needed, as will be discussed in next class.