Multi-armed Contextual Bandits

Machine Learning and Causal Inference

Professor Susan Athey, Stanford

A/B Testing and Randomized Field Experiments

- Central to innovation in major tech companies, businesses, and (future) governments
- Used in economic evaluations, particularly development
  Future opportunities
- Many alternative treatments (phrasing of text message, variations of online training, etc.)
- Personalized treatment assignment

Schizophrenia

At the same time we use:

- Complex, sophisticated algorithms, econometric methods
- Fixed, preset experimentation among small number of alternatives

Cutting edge in tech companies today (Multi-world testing (MSFT), Google Optimize 360, Facebook):

- Adaptive, online experimentation
- For personalized policies

Bringing into economics

- Unlike most ML, this literature has explicit causal model from the start
- The setup is "good economics": minimizing regret, balancing exploration and exploitation
- But almost no attention in econometrics or field experiments
- Sprawling literature is an impenetrable morass of mix and match heuristics and approaches

What do we need?

- Be able to understand the disparate literatures and jargon (contextual bandits, Gaussian processes, etc.)
- Justify the many choices in some sort of coherent way
- Efficiency in estimation, confidence intervals for evaluating final policy

1. Contextual Multi-armed Bandits

Treatments $w \in \mathbb{W} = \{1, 2, \ldots, M\}$,
potential outcomes $Y_i(1), \ldots, Y_i(M)$.
Expected outcome:

$$\mu(w, x) = \mathbb{E}[Y_i(w)|X_i = x]$$

Optimal rule:

$$\pi^*(x) = \arg\max_{w \in \mathbb{W}} \mu(w, x)$$

Unit i receives W_i, possibly different from optimal $W^*(X_i)$.
Expected average regret:

$$\mathbb{E}[\mathcal{R}_n] = \frac{1}{n} \sum_{i=1}^{n} \left(\mu(\pi^*(X_i), X_i) - \mu(W_i, X_i) \right)$$

We would like to choose a rule that assigns a new unit, say unit $n+1$, for $n = 0, 1, 2, \ldots, N$, optimally to a treatment, in order to minimize expected average regret, given the covariate/feature values, and given the outcomes, treatment, and covariate values for prior units:

$$\pi_n : \mathbf{W} \times \mathbf{X} \times \mathbf{W}^n \times \mathbf{Y}^n \times \mathbf{X}^n \mapsto [0,1]^{|\mathbf{W}|},$$

with $\sum_{w \in \mathbf{W}} \pi_n(w, x, W_1, \ldots, W_n, Y_1, \ldots, Y_n, X_1, \ldots, X_n) = 1$,
Challenge: how to balance **exploration** (information gained from assigning units to treatments that we are uncertain about) and **exploitation** (improvement in regret from assigning incoming units to the treatment that is currently viewed as the best).

Bandit problem choice:

- What heuristic to balance exploration and exploitation, when primitives of problem unknown? (UCB v. Thompson)

Contextual bandit choices

- Fixed set of policies, update weights on each using data (analog of non-contextual bandit where policy=arm) VS Estimate a more structural model, derive optimal policy
- How/whether to account for non-random assignment as data accumulates
- Parametric versus non-parametric models, Bayesian v. sort-of Bayesian v. Frequentist
- This is a problem where it is crucial to efficiently make use of available data. Efficiency theory may be insightful, and small sample properties are crucial.

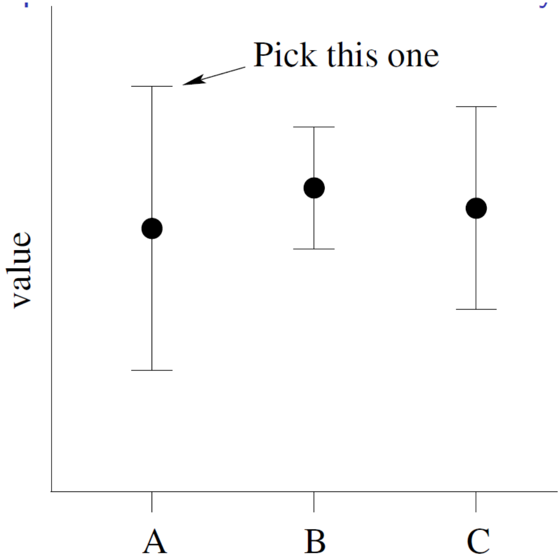
2. UCB Methods and Thompson Sampling without Covariates

Two general approaches to mult-armed bandit problems: UCB (Upper Confidence Bound) methods and Thompson sampling.
UCB methods: Develop estimator $\hat{\mu}_n(w)$ for $\mu(w)$, with measure of uncertainty, $\sigma_n(w)$, given first n units.
Then assign unit $n+1$ to treatment that solves

$$W_{n+1} = \arg\max_w \left\{ \hat{\mu}_n(w) + \sigma_n(w) \right\}.$$

$\sigma_n(w)$ goes to zero as more information about treatment level w accumulates.

Upper Confidence Bounds



[9]

Thompson Sampling

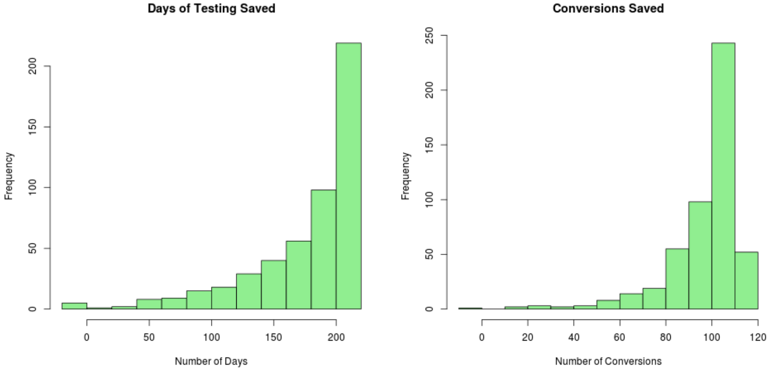
- Specify parametric joint distribution of $(Y_i(1), \ldots, Y_i(M))$, given parameter θ, e.g., $Y_i(w) \sim \mathcal{N}(\beta(w), \sigma^2(w))$, with $\theta = (\beta(1), \sigma^2(1), \ldots, \beta(M), \sigma^2(M))$.
- Specify prior distribution for θ.
- Calculate posterior distribution for θ given information for units 1 through n, and implied posterior for $\mu(1), \ldots, \mu(M)$.
- Assign unit $n+1$ to treatment w with probability equal to the posterior probability that treatment w is the best one given current information, $\mathrm{pr}(\mu(w) = \max_{w' \in \mathbf{W}} \mu(w'))$.

Bayesian way of balancing exploration and exploitation: if $\hat{\mu}(1)$ is less than $\hat{\mu}(2)$, it may still be choosen with substantial probability if we are uncertain about $\mu(2) - \mu(1)$.

Bandits use data more efficiently than A/B test

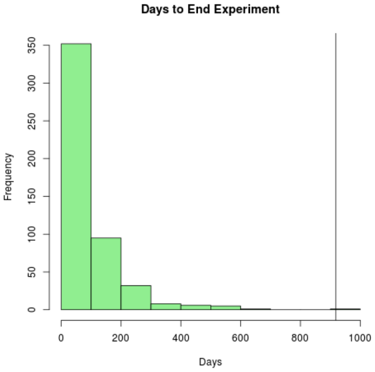
- A/B test: can do power calculation to design experiment in advance, compare to bandit with stopping rule
- Stop when "value remaining in experiment" (optimal choice versus best draw by draw choice when drawing from posterior) small enough, 95th percentile
- Example: Experiment to find ad that maximizes conversions. 100 people exposed per day. Arm 1 has conversion rate .04, arm 2 has .05.
- A/B test takes 220 days to reach 22,000 exposures

Comparison against pre-planned A/B test with correct power calculation (2 arms):



Source: https://support.google.com/analytics/answer/2844870?hl=en

Comparison against pre-planned A/B test with correct power calculation (6 arms requires more than 2 years with 100 exposures per day):



What to do with covariates?

- Run separate bandits for covariate values.
- Build parametric model for potential outcomes given covariates.

What to do with many covariates?

- Specify set of policy/assignment rules and run bandits to choose between them (Beygelzimer et al, 2011, Agarwal et al 2016)
- Use Ridge regression to model outcomes, UCB/Thompson sampling for each x (lin-UCB)
- Langford et al (2016): update policies/add to mix after batches, using weighted classifier to estimate new policies
- Gaussian process approaches: Eytan Bakshy et al (Facebook)

Proposed Approach (Athey, Dimakopoulou, Du, and Imbens (in progress)

- Select combination of modeling choices satisfying desiderata: interpretable, flexible, simple
- Model outcomes directly, flexibly (Bayesian forests) after each batch of observations
- Incorporate (known) propensity score into forest estimation to improve estimates of treatment effects and assignment.
- Use Bayesian posterior to construct probability each treatment is highest together with Thompson sampling
- Thompson sampling: in proportion to probability of being highest

Bayesian Additive Regression Trees

Model conditional expectation as sum of trees:

$$\mathbb{E}[Y_i(w)|X_i = x] = \sum_{m=1}^{M} g(x; \mathcal{T}_m, B_m)$$

where $\mathcal{T}_m$ is the m-th tree, with parameters B_m.
Within leaf l of tree m, the potential outcome is modeled as normal with leaf-tree specific mean $\mu_{m,l}$ and common variance σ^2.
We use standard prior distributions.
The prior for the tree involves a splitting probability for a node η that depends on the depth d_η of the node:

$$\mathrm{pr}_{\mathrm{split}} = \alpha \cdot (1 + d_\eta)^{-\gamma}$$

Extensions to dynamic trees possible (Taddy et al, 2012)
VS: Empirical Bayes interpretation of Random Forests

4. Other Estimation Issues

- We process new units in batches.
- In the first batch units are randomly assigned to treatments.
- After the first batch we estimate the probability that a particular treatment level is better for a given value of the covariates using the first batch of data, using BART.
- After the second batch we re-estimate the trees. Now the treatments were **not** randomly assigned.
- The assignment probabilities depend on the covariates and the batch - but they are known.
- This generates **systematic** biases in within-leaf estimates unless we account for the assignment weightings, and this is more general than trees/forests
- In general, if today's estimated outcome model does not account for previous assignment probabilities, can have bias; this generates a benefit to start simple with assignment models; see Dimakopoulou, Athey and Imbens (2017)

Within-leaf estimates of the average of potential outcomes are now in general biased upward:

- Within a leaf, units assigned to treatment w are likely to have a higher expected potential outcome for treatment w than units assigned to treatment w'.

- We address this by re-weighting units within a leaf using the (known) assignment probability.

Conclusion

- Much more to do on causality than simply estimating average treatment effects!
- Many implications and questions arising from treatment effect heterogeneity.
- Optimal assignment in settings with unknown dependence of potential outcomes on covariates is complicated.
- Flexible Bayesian methods with Thompson sampling appear promising.
- Dimakopoulou, Athey and Imbens (2017) analyze a number of issues relating to the fact that contextual bandits involve an estimation problem, including benefits to simpler assignment models, Thompson sampling v. UCB, and propensity score weighting.