Causal Machine Learning

Lecture 11: Response Adaptation

Drew Dimmery
2025-11-25

Schedule

1. Design-based causal inference & Monte Carlo simulation
2. Covariate adjustment
3. Balancing weights
4. Double robustness, Double ML and TMLE
5. Heterogeneous Treatment Effects I: Basics
6. HTE II: Neural Networks
7. Off-policy evaluation and optimization
8. Experimental design
9. Panel data and the new diff-in-diff
10. Partial Identification
11. **Response Adaptation**
12. Interference

Roadmap

- **Foundation**: (Robbins 1952) - The original formulation of sequential experiments
- (Russo et al. 2018) - Bayesian approach to exploration-exploitation
- (Sutton and Barto 2018) - Core RL concepts: value functions, policy gradients, etc
- (Schulman et al. 2017) - Stable policy optimization via clipped surrogate objectives
- (Haarnoja et al. 2018) - Entropy-regularized RL for robust exploration

Optional:

- (Hadad et al. 2021) - Valid inference when using adaptive treatment assignment
- (Ouyang et al. 2022) - RLHF: Aligning language models with human preferences
- (Kasy and Sautmann 2021) - Optimal adaptive assignment for policy decisions
- (Ariu et al. 2021) - Connections between bandits and best-arm identification

The sequential experiment problem

Robbins (1952): A doctor must choose between two treatments for their patients

The setup:

- Treatment A and Treatment B have unknown success probabilities p_A and p_B
- Each patient receives exactly one treatment
- After treating a patient, observe success or failure
- Goal: maximize total successes across all patients

The tension:

- To learn which treatment is better, must try both
- But trying the inferior treatment harms patients
- How should the doctor allocate treatments?

Robbins' key insight: This is fundamentally different from classical statistics

- Not about estimating $p_A - p_B$ with minimal variance
- About making *decisions* that balance learning and doing

Why not just randomize?

The classical approach: Randomize 50-50, then analyze
Problems with fixed randomization:

1. **Ignores accumulating evidence:** If A looks much better after 1000 patients, why keep assigning B to 50% of patients?
2. **Ethical costs:** Knowingly assigning inferior treatments
3. **Efficiency costs:** Could learn the same with fewer patients on the worse arm

A concrete example:

- After 1000 patients, recovery: $\hat{p}_A = 0.7$, $\hat{p}_B = 0.3$
- Fixed randomization: $\frac{1}{2}\hat{p}_A + \frac{1}{2}\hat{p}_B = 0.5$
- Only assigning A: $\hat{p}_A = 0.7 \implies \approx 20$pp fewer recoveries

The alternative: Adapt treatment assignment based on observed outcomes
But this creates new problems…

Regret as a performance measure

Definition: Regret is the expected loss relative to always playing the best arm

$$R_T(\pi) = T \cdot \mu^* - \sum_{t=1}^{T} \mathbb{E}_\pi[\mu_{A_t}]$$

where $\mu^* = \max_a \mu_a$ is the mean reward of the best arm
Intuition: How much reward did we lose through our assignment process?
Why regret?

- Measures cumulative cost of learning
- Accounts for both exploration (learning) and exploitation (earning)
- A policy with low regret learns quickly *and* acts on that knowledge

Key result from Robbins (1952): Optimal regret grows *logarithmically* in T

$$R_T = \mathcal{O}(\log T)$$

This is remarkable: The cost of learning is vanishingly small relative to the horizon

The exploration-exploitation tradeoff

The core dilemma:

Exploitation

Choose arm with best estimated reward

Why pure exploitation fails:

- Early estimates may be wrong via noise
- Get stuck on suboptimal arm forever
- Linear regret: $R_T = \mathcal{O}(T)$

Exploration

Choose arms to reduce uncertainty

Why pure exploration fails:

- Never capitalize on what you've learned
- Also linear regret

The solution: Explore *strategically*

- Explore arms that might be optimal
- Stop exploring arms that are clearly suboptimal
- Shift toward exploitation as certainty grows

Key insight: The optimal amount of exploration *decreases* over time

- Early: high uncertainty $\implies$ explore more
- Late: low uncertainty $\implies$ exploit more

The multi-armed bandit formalization

The model:

- K arms (treatments), each with unknown reward distribution
- Arm a has mean reward μ_a; let $\mu^* = \max_a \mu_a$
- At each round $t = 1, \ldots, T$:
  - Algorithm selects arm $A_t \in \{1, \ldots, K\}$
  - Observe reward $Y_t \sim P_{A_t}$ (typically assume bounded in $[0, 1]$)

Goal: Minimize cumulative regret $R_T = \sum_{t=1}^{T} (\mu^* - \mu_{A_t})$

Key quantities:

- $N_a(t)$: number of times arm a pulled through round t
- $\hat{\mu}_a(t)$: sample mean of arm a through round t
- $\Delta_a = \mu^* - \mu_a$: suboptimality gap for arm a

Regret decomposition: $\mathbb{E}[R_T] = \sum_{a:\Delta_a > 0} \Delta_a \cdot \mathbb{E}[N_a(T)]$

Minimize regret $\iff$ pull suboptimal arms as few times as possible

Why standard estimators fail with adaptive data

The problem: Selection bias in sample means

Example: Two arms, $\mu_A = \mu_B = 0.5$

- Round 1: Pull A, observe $Y_1 = 1$ (lucky)
- Greedy policy: Keep pulling A because $\hat{\mu}_A = 1 > \hat{\mu}_B = 0$
- Result: $\hat{\mu}_A$ stays high, $\hat{\mu}_B$ undefined or based on few samples

The bias: $\mathbb{E}[\hat{\mu}_a \mid \text{arm } a \text{ pulled often}] > \mu_a$

Arms are pulled more when they *look* good, not when they *are* good

This is not just variance—it's systematic bias

Consequence for inference:

- Confidence intervals have wrong coverage
- Hypothesis tests have wrong size
- Cannot trust standard errors

The culprit: Treatment assignment depends on past outcomes

The martingale problem

In usual RCTs: Treatment assignment A_i is independent of potential outcomes

$$A_i \perp\!\!\!\perp (Y_i(0), Y_i(1))$$

In adaptive experiments: Assignment depends on *past* outcomes

$$A_t \not\perp\!\!\!\perp Y_t \quad \text{because} \quad A_t = f(Y_1, \ldots, Y_{t-1})$$

This is bad: Standard estimators rely on:

$$\mathbb{E}[\hat{\mu}_a] = \mathbb{E}\left[\frac{1}{N_a} \sum_{t:A_t=a} Y_t\right] = \mu_a$$

But with adaptive assignment, **the sums are not over independent terms**:

$$\mathbb{E}\left[\frac{1}{N_a} \sum_{t:A_t=a} Y_t\right] \neq \mu_a$$

Upper Confidence Bound algorithms

The UCB principle: "Optimism in the face of uncertainty"
Idea: Construct upper confidence bounds for each arm's mean reward

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \mathrm{Bonus}_a(t)$$

- Large Bonus when arm a has been pulled few times (high uncertainty)
- Small Bonus when arm a has been pulled many times (low uncertainty)

The algorithm: At each round, pull the arm with highest UCB

$$A_t = \arg\max_a \mathrm{UCB}_a(t)$$

- If an arm's UCB is low, it's either:
  - Truly bad (low $\hat{\mu}_a$), or
  - Well-estimated as mediocre (small bonus)
- Either way, safe to skip it
- Arms with high UCB are either good or uncertain—worth exploring

UCB1: The algorithm

The UCB1 formula (Auer, Cesa-Bianchi, and Fischer 2002):

$$A_t = \arg\max_a \left\{ \hat{\mu}_a(t-1) + \sqrt{\frac{2 \ln t}{N_a(t-1)}} \right\}$$

Components:

- $\hat{\mu}_a(t-1)$: sample mean of arm a (exploitation term)
- $\sqrt{\frac{2 \ln t}{N_a(t-1)}}$: exploration bonus
  - Derived from Hoeffding's inequality
  - With probability $\geq 1 - t^{-4}$: true mean lies below UCB
  - Grows with $\ln t$: ensures continued exploration
  - Shrinks with N_a: less bonus for well-sampled arms

Initialization: Pull each arm once (rounds $1, \ldots, K$)
Computational cost: $\mathcal{O}(K)$ per round—just compute K UCB values
No tuning parameters: The $\sqrt{2 \ln t}$ factor is derived, not chosen

UCB1: Regret guarantees

Theorem: For UCB1 with K arms and rewards in $[0, 1]$:

$$\mathbb{E}[R_T] \leq 8 \sum_{a:\Delta_a > 0} \frac{\ln T}{\Delta_a} + \left(1 + \frac{\pi^2}{3}\right) \sum_{a=1}^{K} \Delta_a$$

- **Main term:** $\mathcal{O}\left(\frac{K \ln T}{\Delta_{min}}\right)$ where $\Delta_{min} = \min_{a:\Delta_a > 0} \Delta_a$
- **Regret is logarithmic in T:** The cost of learning is negligible for large T
- **Regret scales with $1/\Delta_a$:** Harder to distinguish arms with similar means

Gap-dependent vs. gap-independent bounds:

- Gap-dependent: $\mathcal{O}\left(\sum_a \frac{\ln T}{\Delta_a}\right)$—better when gaps are large
- Gap-independent: $\mathcal{O}(\sqrt{KT \ln T})$—better when gaps are small
  - Worst-case gaps $\Delta_a \approx \sqrt{K \ln T / T}$ give $K \cdot \frac{\ln T}{\Delta} = \sqrt{KT \ln T}$

Lai and Robbins (1985): Any consistent policy has $\mathbb{E}[R_T] \geq \Omega\left(\sum_a \frac{\ln T}{\Delta_a}\right)$

$\Longrightarrow$ **UCB1 is optimal up to constants!**

The confidence bound intuition

Why does optimism work?

Consider two scenarios for arm a:

1. **Arm a is optimal:** UCB is (correctly) high, we pull it often, low regret
2. **Arm a is suboptimal:**
   - If UCB is high due to uncertainty $\implies$ we explore and learn it's bad
   - Once we learn it's bad, UCB drops and we stop pulling it

The key: Optimism ensures we don't prematurely abandon good arms

Contrast with pessimism:

- Pessimistic policy: assume worst case for uncertain arms
- Would never explore sufficiently
- Could get stuck on suboptimal arm

UCB's elegance: Deterministic, simple, and optimal

From bandits to adaptive experiments

Explore/exploit is everywhere:

- **Clinical trials:** Adaptive assignment to maximize patient outcomes
- **A/B testing:** Allocate traffic to better-performing variants
- **Ad placement:** Learn which ads get clicks
- **Recommendation systems:** Learn user preferences

But there's a tension:

- **Bandit objective:** Minimize regret (maximize total reward)
- **Inference objective:** Estimate treatment effects with valid confidence intervals

These can conflict:

- Low-regret policies concentrate assignment on best arm
- Good estimation requires sufficient samples from *all* arms

The readings address this tension from different angles

How the readings extend this framework

Russo et al. (2018) - Thompson Sampling:

- Randomized alternative to UCB's deterministic optimism
- Posterior sampling naturally balances exploration-exploitation

Sutton and Barto (2018) - Reinforcement Learning:

- RL allows actions to affect future outcomes
- Same exploration-exploitation tradeoff, but incorporates long-term consequences

Schulman et al. (2017) - Proximal Policy Optimization:

- When action spaces are large/continuous, need safe new policies
- Policy gradients optimize exploration-exploitation directly

Haarnoja et al. (2018) - Soft Actor-Critic:

- Actor-critic methods combine policy learning with value estimation
- Like double robustness: use two models to improve efficiency and reliability

References

Ariu, Kaito, Masahiro Kato, Junpei Komiyama, Kenichiro McAlinn, and Chao Qin. 2021. "Policy Choice and Best Arm Identification: Asymptotic Analysis of Exploration Sampling." https://arxiv.org/abs/2109.08229.

Auer, Peter, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. "Finite-Time Analysis of the Multiarmed Bandit Problem." *Machine Learning* 47 (2-3): 235–56.

Haarnoja, Tuomas, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor." In *Proceedings of the 35th International Conference on Machine Learning*, 1861–70.

Hadad, Vitor, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. 2021. "Confidence Intervals for Policy Evaluation in Adaptive Experiments." *Proceedings of the National Academy of Sciences* 118 (15): e2014602118.

Kasy, Maximilian, and Anja Sautmann. 2021. "Adaptive Treatment Assignment in Experiments for Policy Choice." *Econometrica* 89 (1): 113–32.

Lai, Tze Leung, and Herbert Robbins. 1985. "Asymptotically Efficient Adaptive Allocation Rules." *Advances in Applied Mathematics* 6 (1): 4–22.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35: 27730–44.

Robbins, Herbert. 1952. "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society* 58 (5): 527–35.

Russo, Daniel, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. 2018. "A Tutorial on Thompson Sampling." *Foundations and Trends in Machine Learning* 11 (1): 1–96.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." https://arxiv.org/abs/1707.06347.

Sutton, Richard S, and Andrew G Barto. 2018. *Reinforcement Learning: An Introduction*. 2nd ed. MIT Press.