EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

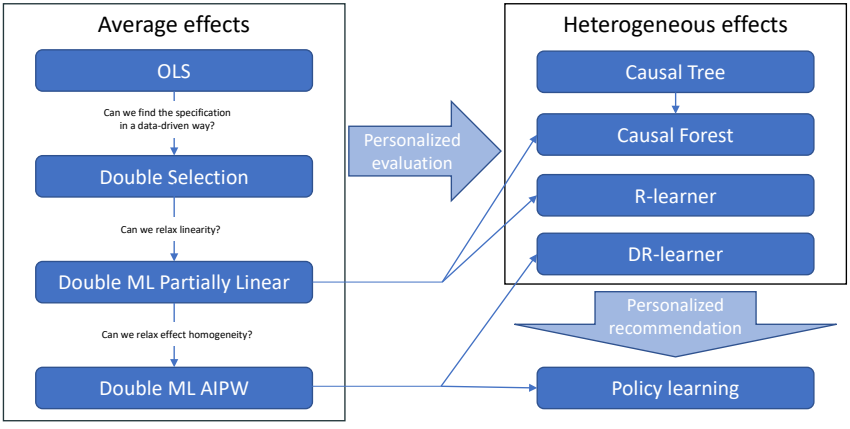
Causal Machine Learning

Multi-armed bandits (online policy learning)

Michael Knaus

WiSe 25/26

State of the journey



Last week we uncovered policy learning in the familiar context with historic data

Plan of this morning

How to use Causal ML for decision making (policy recommendation) if we can sequentially choose the treatment?

1. Online policy learning (bandits)

2. Wrapping-up

3. Outlook

Online policy learning (bandits)

A dynamic setting

Our policy learning lived so far in the familiar world where we evaluate historic data and derive policy recommendations for the future (offline policy learning)

So-called online policy learning lives in a different setting

Instead of learning policies after treatment assignments took place, we learn what works while assigning treatments

This involves algorithms that are able to actively interact/experiment with their environment

For me this is the first method we cover that might deserve the AI label

Tech companies vs. social science

Amazon, Facebook, Google, Spotify, ... use online PL to figure out which ads, playlists or other features of the user experience maximize their returns

Several features of their setting are excellent for online learning:

- Millions of users per day
- Real time data
- Real time outcomes (did the user click or not?)

These features are rarely present in social sciences and the literature of adapting the methods to productive use for us is in the beginning

Still (or for this reason) I would like to give you the idea behind these methods

Let's look at the intuition from the nice illustration in "Practitioner's Guide: Designing Adaptive Experiments" by Hadad, Rosenzweig, Athey, and Karlan (2021)

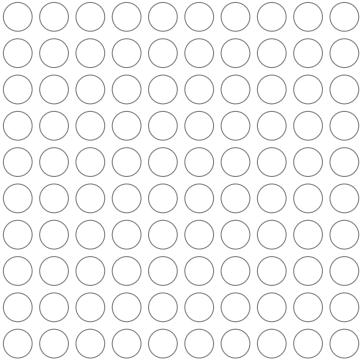
Multi-armed bandits - visual (1/4)

Figure 1

Experiment Setup

Treatments or Arms

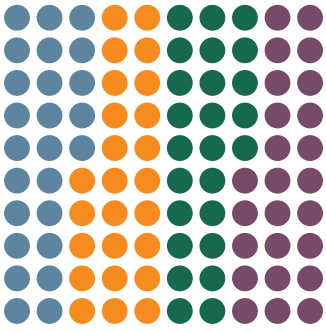
Experimental Budget (n=100)

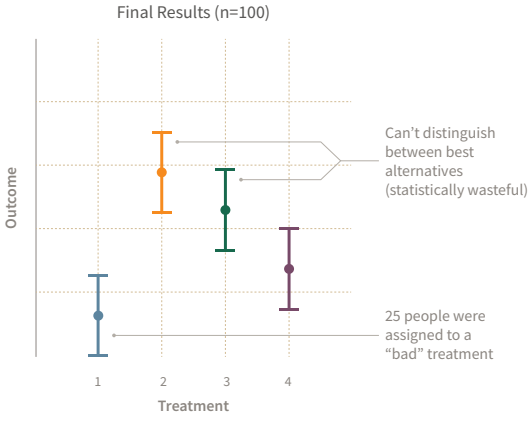


Multi-armed bandits - visual (2/4)

Figure 2

Illustration of Nonadaptive Experiment Results

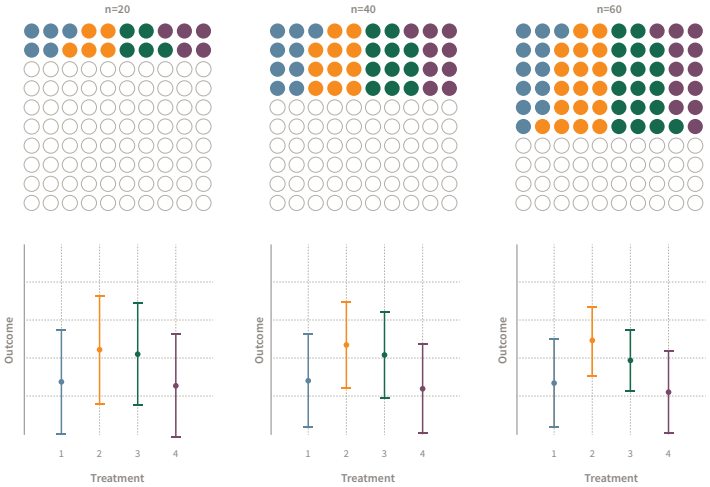




Multi-armed bandits - visual (3/4)

Figure 3

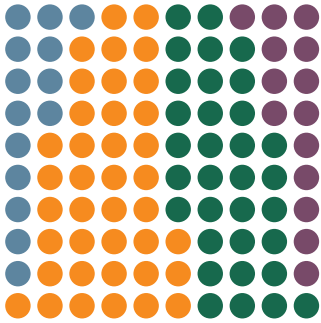
Illustration of an Adaptive Experiment

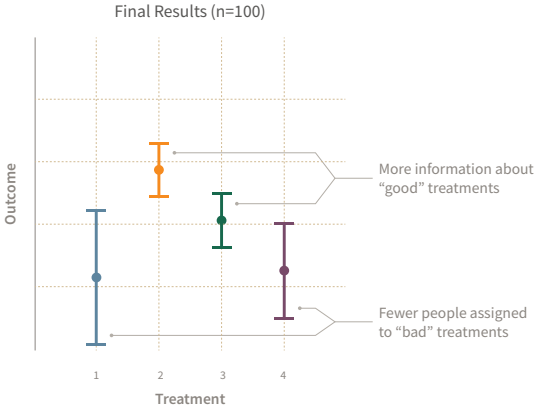


Multi-armed bandits - visual (4/4)

Figure 4

Illustrative Adaptive Experiment Result





Multi-armed bandits - formal

Multiple treatments $W \in \{0, ..., T\}$ (same framework as at the end of last week)

But units $i = 1, ..., N$ arrive sequentially to be assigned to treatment

Question: Which treatment to assign to unit $i + 1$? (no covariates for now)

Optimal treatment would be to assign those with largest average potential outcome $\gamma_w := \mathbb{E}[Y(w)] \Rightarrow \pi^*_{i+1} := \arg\max_w \mathbb{E}[Y(w)] = \arg\max_w \gamma_w$

Goal: Minimize regret for all units $\min \frac{1}{N} \sum_i (\mathbb{E}[\underbrace{Y_i(\pi^*_i)}_{\text{optimal}}] - \mathbb{E}[\underbrace{Y_i(W_i)}_{\text{actual}}])$

Remark: The name is derived from the problem gamblers face when deciding which of the one-armed bandits in the casino to play (nice cartoon explanation)

Two conflicting goals

We start out with randomly assigning each treatment

But if every treatment was assigned at least twice (i.e. we can estimate a variance), we have the chance to balance two conflicting dimensions:

- **Exploration:** we can assign units to treatment arms we are uncertain about $\Rightarrow$ use unit $i + 1$ to explore what works best (reduce uncertainty)
- **Exploitation:** we can likely decrease regret by assigning unit $i + 1$ to the treatment that is currently viewed as the best $\Rightarrow$ exploit what we know so far in an optimistic way

Two common strategies

Denote by $\hat{\gamma}_{i,w}$ the estimated average PO using all units until i, by $\hat{\sigma}^2_{i,w}$ the variance of this estimate, and by $\alpha > 0$ an appropriately chosen hyperparameter

Upper Confidence Bound (UCB) method: Calculate a confidence interval with critical value α and select the treatment with the highest upper confidence bound

$$W_{i+1} = \arg\max_w (\hat{\gamma}_{i,w} + \alpha \hat{\sigma}_{i,w}) \tag{1}$$

Thompson sampling: Draw for each treatment from a normal distribution $\tilde{\gamma}_{i,w} \sim N(\hat{\gamma}_{i,w}, \alpha^2 \hat{\sigma}^2_{i,w})$ and pick the one with the highest draw

$$W_{i+1} = \arg\max_w \tilde{\gamma}_{i,w} \tag{2}$$

Illustration UCB

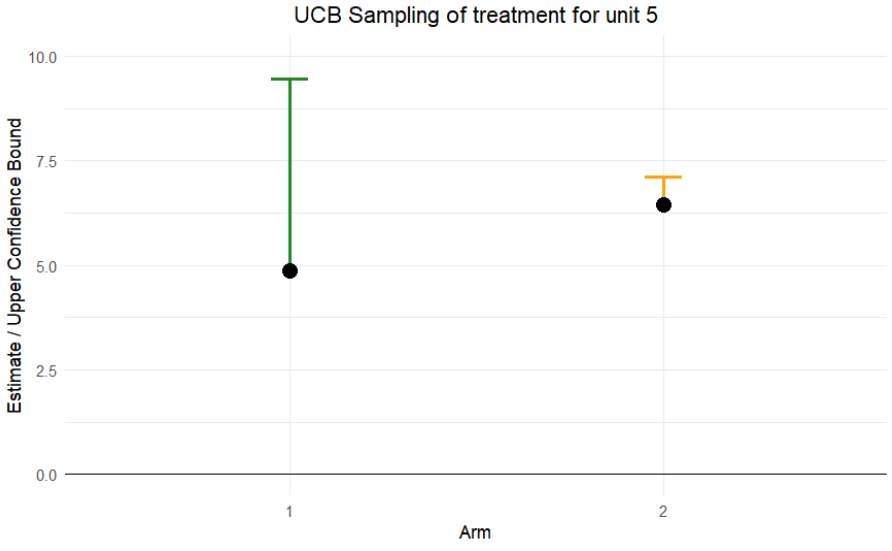


Illustration UCB

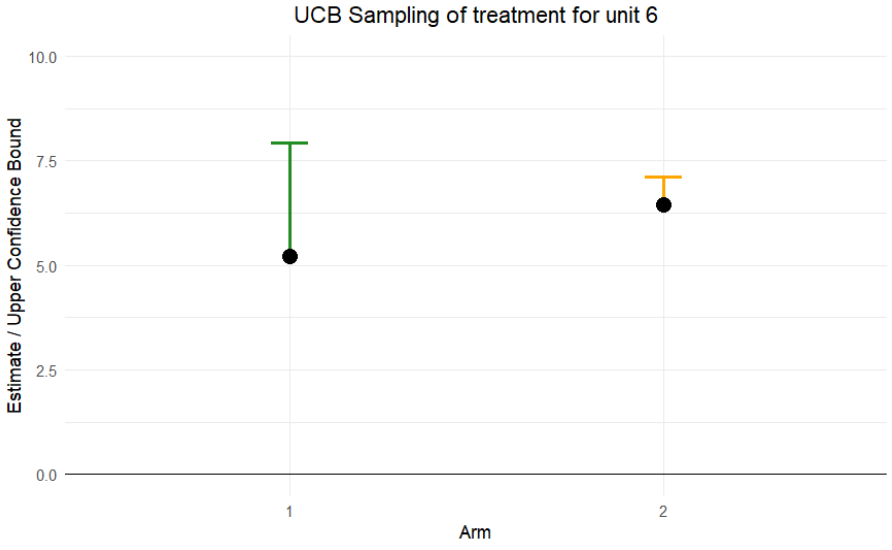


Illustration UCB

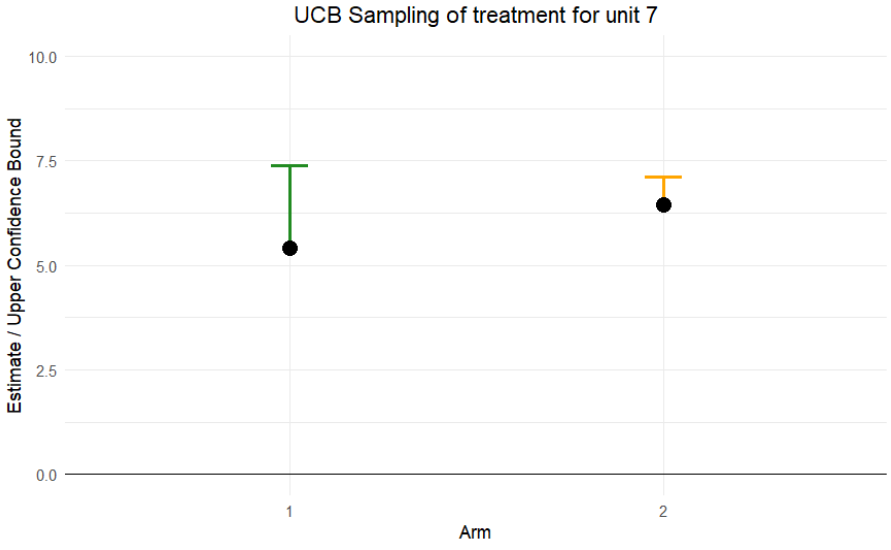


Illustration UCB

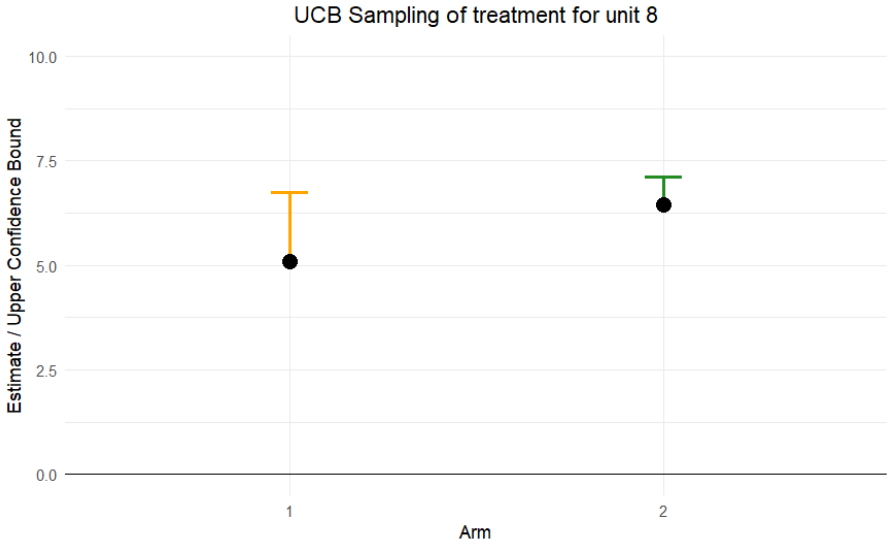


Illustration UCB

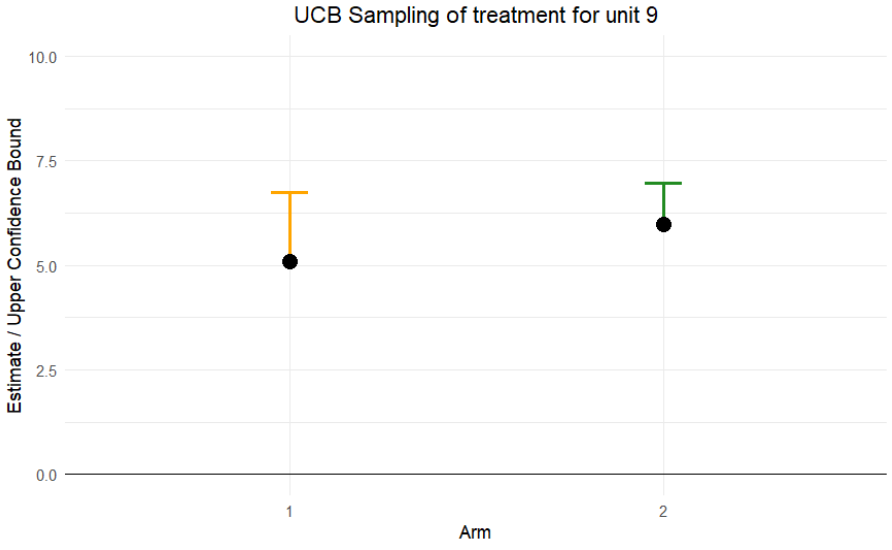


Illustration UCB

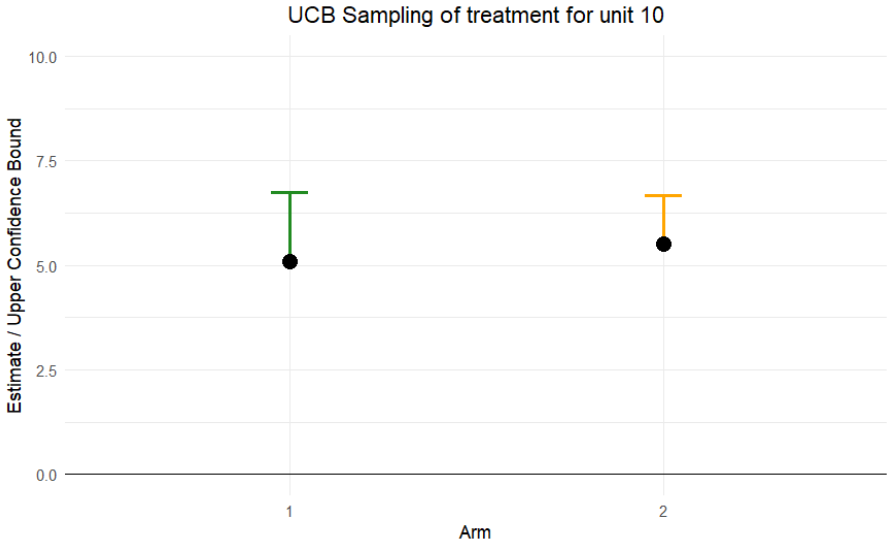


Illustration UCB

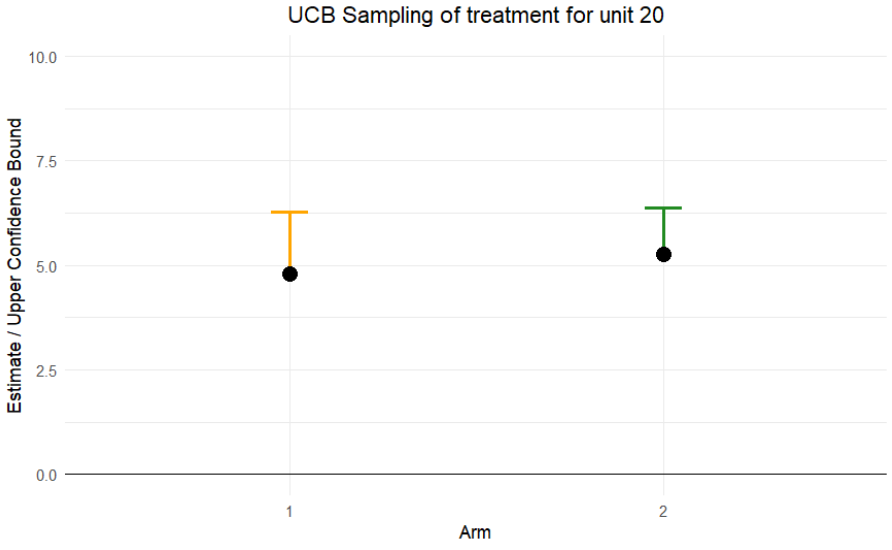


Illustration UCB

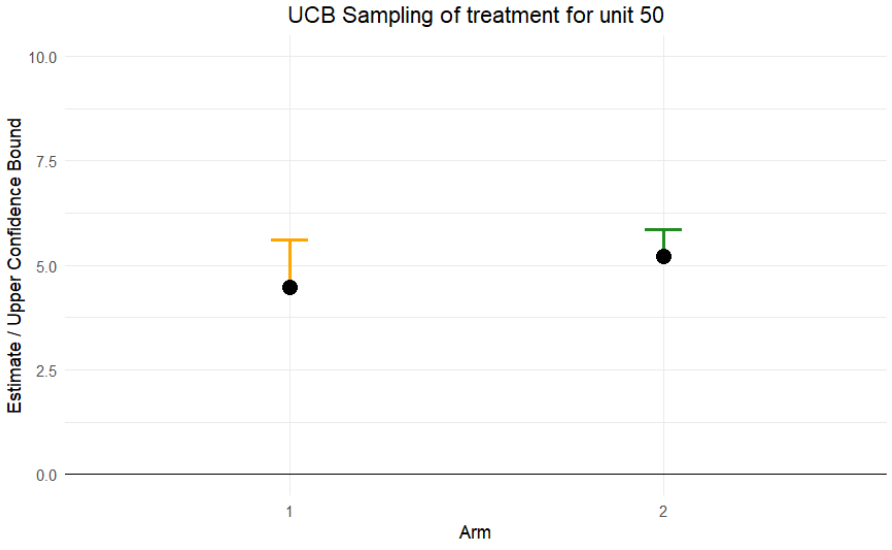


Illustration UCB

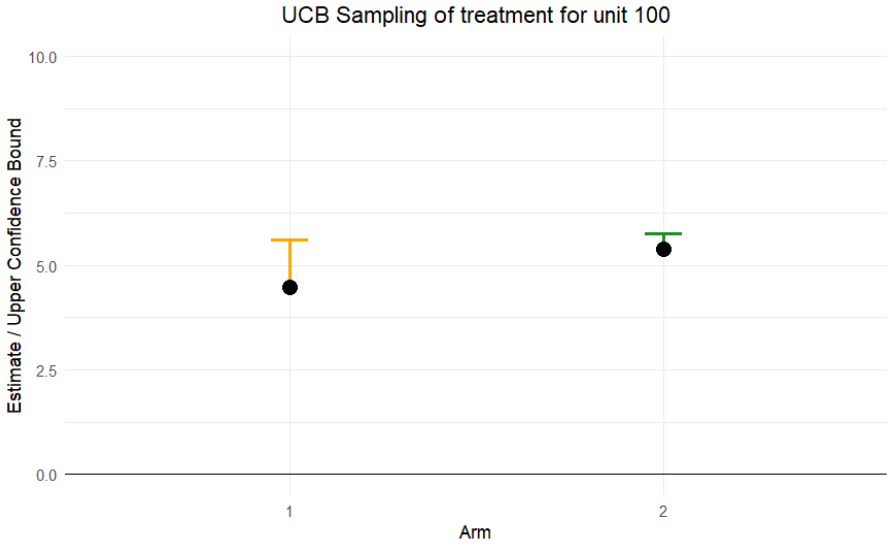


Illustration Thompson sampling

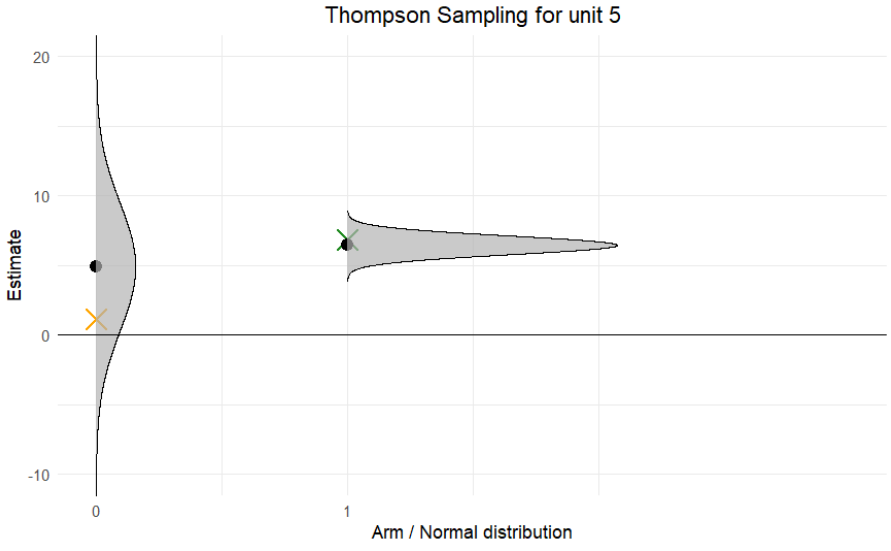


Illustration Thompson sampling

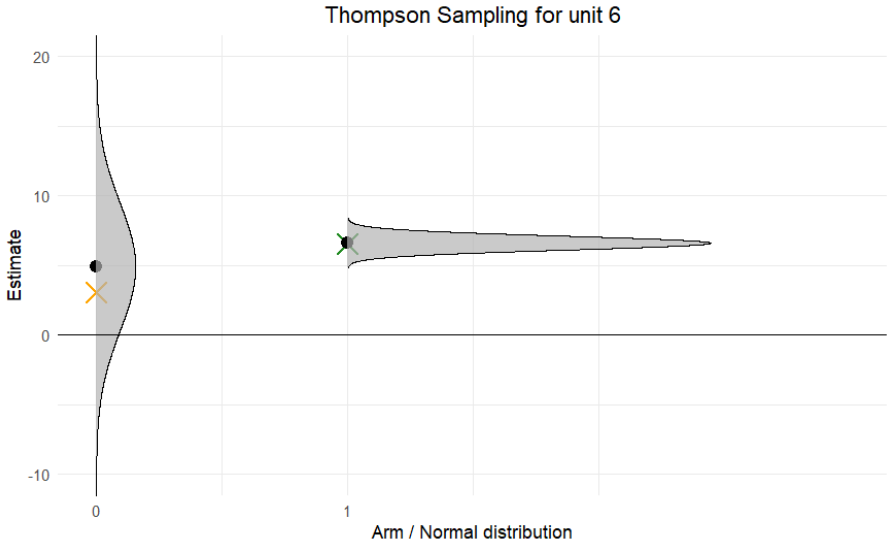


Illustration Thompson sampling

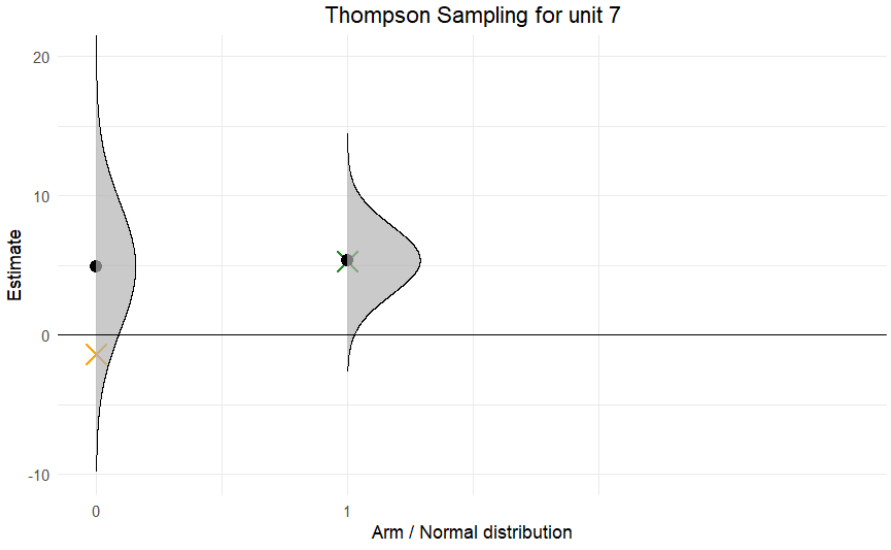


Illustration Thompson sampling

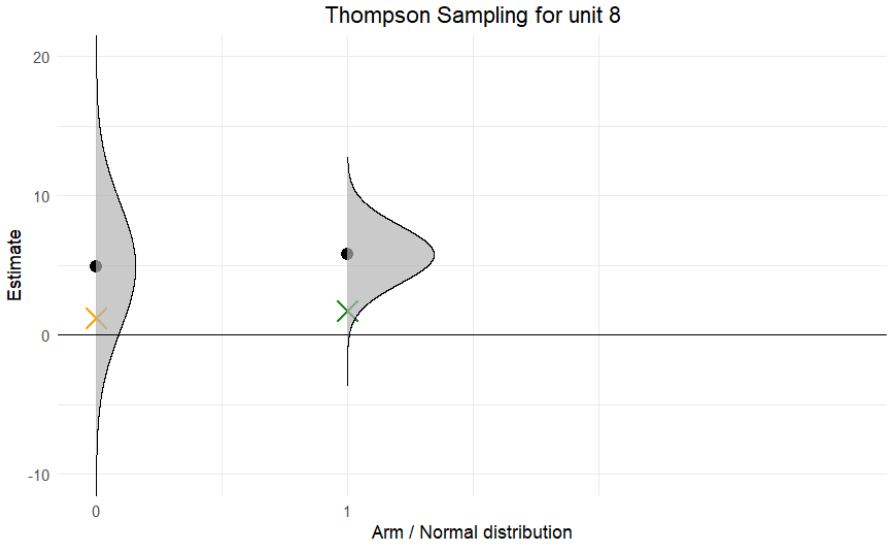


Illustration Thompson sampling

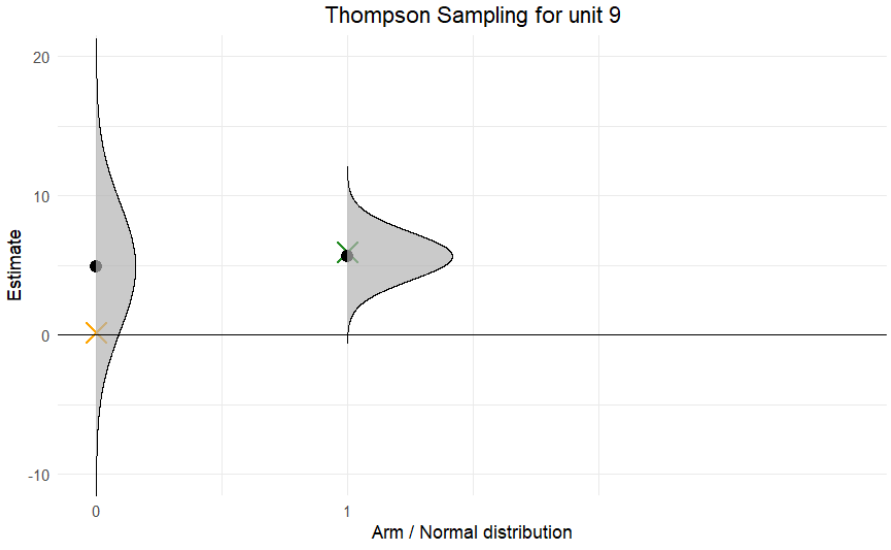


Illustration Thompson sampling

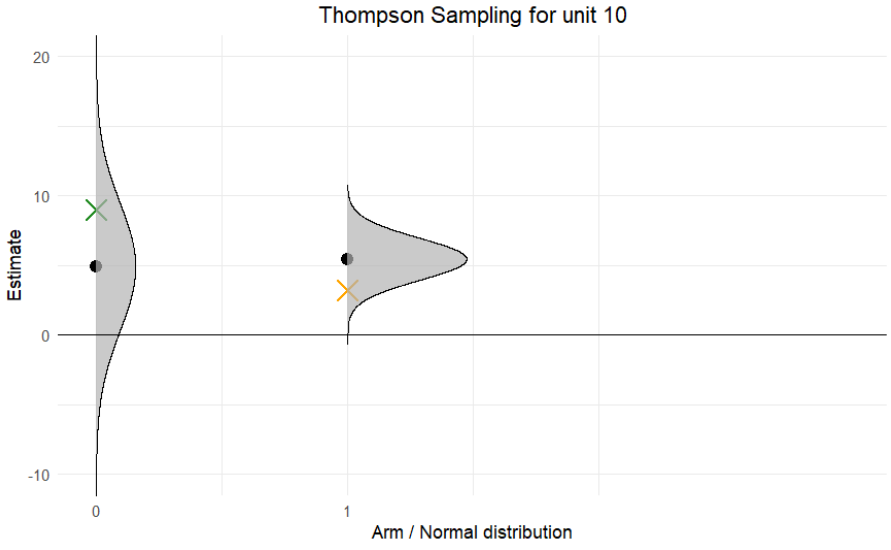


Illustration Thompson sampling

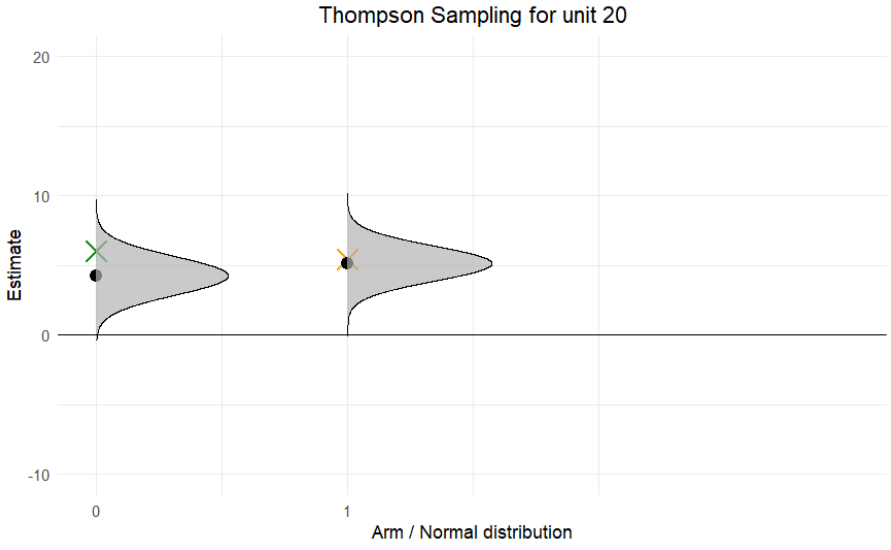


Illustration Thompson sampling

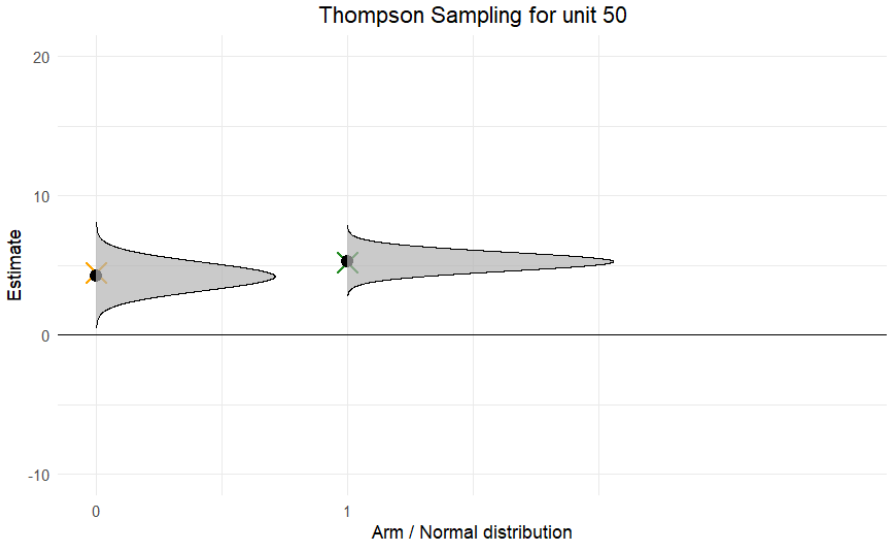
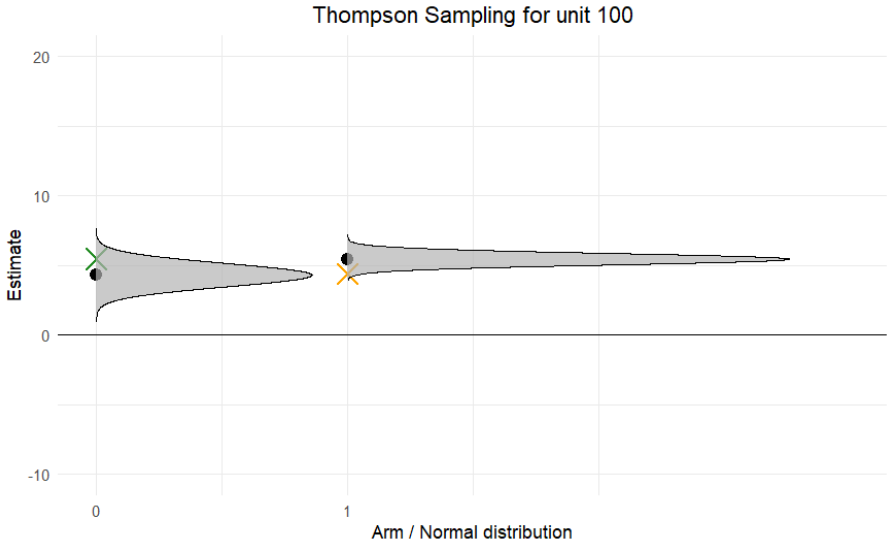


Illustration Thompson sampling



Some comments

As the number of units increases and thus variance decreases, both methods assign more often the good treatments and ignore the bad ones

Assignment of treatments is not related to potential outcomes $\Rightarrow$ no identification issues

Larger α means more exploration and less exploitation (lower risk of missing a good treatment by chance), $\alpha \to \infty$ resembles classic RCT

Contextual bandits use "context" information X and the CAPO instead of average PO for assignment (see e.g Dimakopoulou et al., 2019)

Many open questions regarding implementation in social science context (only few exploratory studies) $\Rightarrow$ **exciting area for research**

Simulation notebook: Multi-armed bandit

Further material on bandits - intuition and applications

Introduction videos: Multi-armed bandits, UCB method, Thompson sampling, reinforcement learning

Applications: Covid testing (Bastania et al., 2021), refugee employment assistance (Caria et al., 2021), charitable giving (Athey et al., 2022), job search support (Hoffman et al., 2023), Hiring (Li et al. (2025)

Shiny app to play with

Practitioner's guide

Further material on bandits - technical

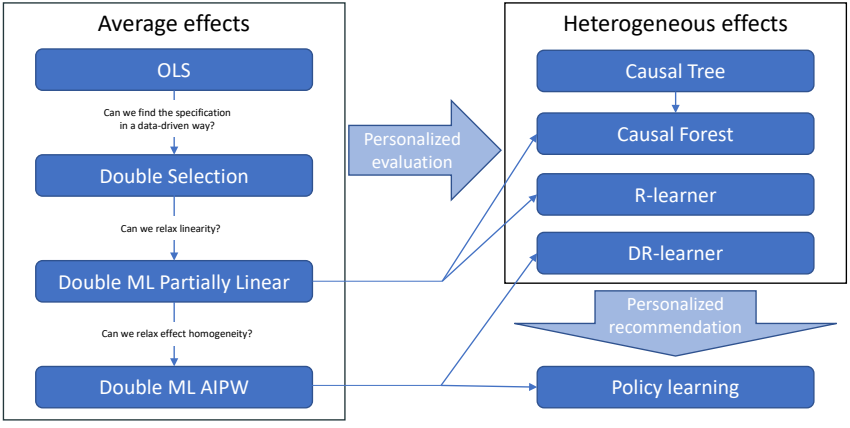
With links to causal inference:

- Max Kasy's bandit and RL teaching slides
- An Introduction to Causal Reinforcement Learning by Barenboim, Zhan and Lee
- Reinforcement Learning in Modern Biostatistics: Constructing Optimal Adaptive Interventions by Deliu et al.

The traditional reinforcement learning perspective:

- "Introduction to Multi-Armed Bandits" by Aleksandrs Slivkins
- "Foundations of Reinforcement Learning and Interactive Decision Making" by Dylan Foster and Alex Rakhlin

End of the journey



Congrats! 🥳

Wrapping-up

Lessons learned (on a high level)

- Naive integration of ML in causal analysis can go wrong
- We have to think about our target parameter
- We should target the underlying objective function, e.g. #CATEvsPL
- If we know what we want to do, we can leverage the power of ML ...
  - ... to make our lives easier (outsource what the machine can do better)
  - ... to get more out of the same data #nonparametricGATE #policylearning

$\Rightarrow$ More time to think about the important stuff (identification, interpretation, ...)

$\Rightarrow$ More fun in the estimation part (tuning a ML models is much more fun than specifying models yourself + (arguably) better science, at least for me)

BUT also no silver bullet: Methods help with estimation, not with identification

Key concepts

We have learned several key concepts for Causal ML:

- Nuisance parameters: Parameters (so far CEFs of observed variables) that are not of interest *per se* but help us to get our hands on the target parameter
- Neyman-orthogonal scores: Smart combination of **multiple** nuisance parameters such that these can be estimated using supervised ML (if high-quality and cross-fit)
- Pseudo-outcomes: May be used as outcomes (unbiased signals) in standard regressions to model/validate inherently unobservable causal quantities
- Modified splitting criteria: Teach regression trees and forests to model inherently unobservable causal quantities

Key concepts reoccur

The key concepts are useful beyond RCTs and VAS settings

I will point you to references that use similar ideas with other research designs such that you know where to start in case it is/becomes relevant for you

- Instrumental variable
- Regression discontinuity
- Difference-in-differences
- Mediation analysis
- Quantile treatment effects

IV assuming homogeneous effects

Belloni et al. (2012) consider a linear model for unlikely case that you have a high-dimensional set of potential instruments

Chernozhukov, Hansen & Spindler (2015) cover both high-dimensional instruments and/or controls

`hdm` can be used to implement both

Section 4.2 of Chernozhukov et al. (2018) and `DoubleML` for theory and implementation of partially linear model with IV

IV with heterogeneous effects

Several papers study estimation of conditional local average treatment effects (CLATEs):

- Stoffi & Gnecco (2019) modify the splitting criterion in the Causal Tree setup with binary Z_i and binary W_i to estimate conditional LATE
- Athey, Tibshirani & Wager (2019) adapt the splitting criterion in the Generalized Random Forest framework with one Z_i and one W_i (`grf`)
- Kugler & Biewen (2021) adapt the splitting criterion in the Generalized Random Forest framework with multiple Z_i and one W_i (2SLS)
- Takatsu et al. (2024) use a pseudo-outcome as unbiased signal for CLATE

Regression discontinuity

Noack, Olma & Rothe (2021) use relations to Double ML to improve precision of estimators in RD designs

Reguly (2021) adapts splitting criteria of Causal Trees for regression discontinuity designs to investigate heterogeneity

Difference-in-differences

Chang (2020) proposes Double ML estimator based on Neyman-orthogonal scores for 2x2 standard case, see more in the `DoubleML` documentation

Heterogeneous effects in the 2x2 standard can also be estimated for pre-specified low-dimensional heterogeneity variables (Zimmert & Zimmert, 2020) or by adapting Causal Forests (Gulyas & Pytka, 2020))

Probably I missed it, but I am not aware of suitable proposals for more time periods

More Double ML methods

Mediation effects: Farbmacher et al. (2022)

Dynamic treatment effects: Bodory et al. (2020)

Quantile treatment effects: Belloni et al. (2017) and Kallus, Mao & Uehara (2024)

...

Outlook

Topics that will keep us busy

Fill the gaps: Many methods are currently provided for relatively clean settings, but practitioners have often more complex settings ⇒ tailored estimators needed

Understand what works: You have seen only a fraction of methods and more flow in as we speak

⇒ We need to figure out what works in which settings

⇒ Requires many more applications, simulations and theoretical studies to get a clearer picture

More and better implementations: Many papers provide novel theory, but user-friendly implementations are often missing (so far)

Offline and online policy learning could be very powerful tools for social scientists, but the journey just starts

That's it from my side

I hope this course was a nice add-on to the more classic econometrics courses

We are still in the middle of understanding the fruitful integration of ML into economics/econometrics

However, the concepts you learned in this course should enable you to digest future developments in causal ML for policy evaluation/recommendation

More free resources - collections

- Collection ”Machine Learning for Economists” by Dario Sansone
- Collection ”Dive into Causal Machine Learning” of Alexander Quispe
- ”Public goods” collection of Christine Cai
- ”Must-read recent papers and resources on Causal ∩ ML”

More free resources - single resources

Loosely ordered from introductory to advanced:

- ”ML & Causal Inference: A Short Course” by Athey, Spiess and Wager
- grf package documentation
- DoubleML user guide
- ”ML-based causal inference tutorial” by Golub Capital Social Impact Lab
- ”Causal Inference for the Brave and True” Part II by Matheus Facure Alves
- Lecture notes of Stefan Wager
- Lecture notes of Christophe Gaillac and Jeremy L’Hour

Intersection of causal inference and ML is a very exciting field

I hope that you at least stay tuned