# Analyzing data from RL

**Lucas Janson**

**CS/Stat 184(0): Introduction to Reinforcement Learning**

**Fall 2024**

# Today

- Feedback from last lecture
- Recap
- Motivation: analyzing data from RL
- Hypothesis testing
- Randomization testing

# Feedback from feedback forms

1. Thank you to everyone who filled out the forms!

# Today

- ✓ Feedback from last lecture
- Recap
- Motivation: analyzing data from RL
- Hypothesis testing
- Randomization testing

# Adaptively collected data

One of the main things that distinguishes most of RL from other types of machine learning is its interactive nature: learning is interlaced with data collection

This “online” RL setting is the most like how humans learn, and as we’ve learned in this class about exploration/exploitation, it is critical for the best performance

Offline methods exist but come with serious challenges unless the data collection policy already happens to be essentially optimal (e.g., imitation learning)

Online RL is focused on how we can learn *while* interacting with the environment

Today we’ll talk about how to draw probabilistic conclusions about the environment based on the thus-far adaptively collected data

# Contextual bandits

Primarily today we'll focus on contextual bandits, so a reminder:

**Formally, a contextual bandit is the following interactive learning process:**

For $t = 0 \to T - 1$

1. Learner sees context $x_t \sim \nu_x$ Independent of any previous data
2. Learner pulls arm $a_t = \pi_t(x_t) \in \{1,\ldots,K\}$ $\pi_t$ policy learned from all data seen so far
3. Learner observes reward $r_t \sim \nu^{(a_t)}(x_t)$ from arm $a_t$ in context $x_t$

# Today

- ✓ Feedback from last lecture
- ✓ Recap
- Motivation: analyzing data from RL
- Hypothesis testing
- Randomization testing

# Motivation: clinical trials

Consider a clinical trial (modeled as contextual bandit):

- $T$ patients total, arriving one at a time
- Patient $t$ has context $x_t$ drawn from $\nu_x$
  - demographics, medical history, etc.
- Receives treatment $a_t \in \{0,1\}$ according to current policy $\pi_t(x_t)$
  - $1$ = treatment, $0$ = control
- We observe reward $r_t \sim \nu^{(a_t)}(x_t)$
  - $1$ = recovery, $0$ = not recovery
  - Assume condition being treated is acute so reward is immediate

Typical clinical trial: $\pi_t$ is coin flip for all $t$, i.e., patients/treatments are i.i.d.

Ethical reasons to run bandit: maximize outcomes of patients in trial

Challenge: statistically rigorous test of whether treatment worked (e.g., for FDA)

# Motivation: online advertising

Consider online advertising (modeled as contextual bandit):

- Viewer $t$ has context $x_t$ drawn from $\nu_x$
    - Browsing cookies, site being viewed, properties of ad space, etc.
- Sees ad $a_t$ according to current policy $\pi_t(x_t)$
    - Choosing among some carefully curated set of ads (maybe 5-10)
- We observe reward $r_t \sim \nu^{(a_t)}(x_t)$
    - $1 =$ click, $0 =$ not click

Clear we want to maximize clicks—this is the whole point of placing ads!

But when we want to design a new set of ads, want to know what worked

Challenge: bandit learns good policy, but won't say what about good ads works

E.g., maybe ads with red click buttons worked better than those with blue buttons

How can we learn this? Standard statistics question! "Between-study" learning

# Unified question in bandit

Assume 2 arms, and for now no context

**Question: is there any difference between these two arms?**

In clinical trial: is the treatment making any difference?

In online advertising: does it matter what ad I show?

Idea: look at how often arms are pulled by bandit algorithm

- If one arm pulled more than another, conclude there is a difference?
- If one arm pulled a lot more than another? How much is "a lot"?
- By their nature, RL algorithms are "streaky", even when $\nu^{(0)} = \nu^{(1)}$
- E.g., Gittins index (optimal alg for Bayesian Bernoulli bandit) will never pull other arm if first arm it pulls always returns a 1

Want to focus on rewards for different arms, rather than which arms are pulled

# Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ Motivation: analyzing data from RL
- Hypothesis testing
- Randomization testing

# Hypothesis testing

**Question: is there any difference between these two arms?**

This kind of question is asked all the time in statistics, but usually for i.i.d. data

Standard framework: null hypothesis $H_0 : \nu^{(0)} = \nu^{(1)}$

A hypothesis test is a binary function $\Phi$ of the data $D$ such that:

- $\mathbb{P}(\Phi(D) = 1) \leq 5\,\%$ when $H_0$ is true
- $\mathbb{P}(\Phi(D) = 1)$ as high as possible when $H_0$ is false

Conclude there is a difference between the two arms if $\Phi(D) = 1$

# If the data were i.i.d…

A standard approach for testing $H_0$ is the t-test

Central limit theorem: $\sqrt{T/2}(\hat{\mu}_T^{(k)} - \mu^{(k)}) \to \mathcal{N}(0,\sigma^2)$ for $k = 0,1$

Use sample standard deviations to estimate $\sigma^2$ via $\hat{\sigma}_T^2$ such that $\hat{\sigma}_T^2 \to \sigma^2$

Then when $H_0$ is true and $T$ is large ( $\gtrapprox 20$), $\quad Z_T := \sqrt{T}\dfrac{\hat{\mu}_T^{(1)} - \hat{\mu}_T^{(0)}}{2\hat{\sigma}_T} \approx \mathcal{N}(0,1)$

Let $\Phi(D) = 1\{\,|Z_T| > z_{0.975}\}$, where $z_{0.975}$ is the 97.5th percentile of $\mathcal{N}(0,1)$

Then when $H_0$ is true, $\mathbb{P}(\Phi(D) = 1) = \mathbb{P}(\,|Z_T| > z_{0.975}) = 2\mathbb{P}(Z_T > z_{0.975}) = 5\,\%$

And when $\mu^{(0)} \neq \mu^{(1)}$, $|Z_T| \to \infty$ and hence $\mathbb{P}(\Phi(D) = 1) \to 1$

# Concentration inequality approach

As we've discussed before, data is not i.i.d., and CLT doesn't hold

Our solution in the past was Hoeffding's inequality:

$$\mathbb{P}\left(\forall k = 1,2 : |\hat{\mu}_T^{(k)} - \mu^{(k)}| \leq \sqrt{\ln(4T/\delta)/2N_T^{(k)}}\right) \geq 1 - \delta$$

Set $\delta = 5\,\%$ : $\Phi(D) = 1$ when $|\hat{\mu}_T^{(1)} - \hat{\mu}_T^{(0)}| > \sqrt{\ln(80T)}\left(\frac{1}{\sqrt{2N_T^{(0)}}} + \frac{1}{\sqrt{2N_T^{(1)}}}\right)$

Then when $H_0$ true: $\mathbb{P}(\Phi(D) = 1) \leq 5\,\%$

When $\mu^{(0)} \neq \mu^{(1)}$, if RL never stops exploring, $|\hat{\mu}_T^{(1)} - \hat{\mu}_T^{(0)}| \to |\mu^{(1)} - \mu^{(0)}| > 0$ while threshold $\to 0$, so $\mathbb{P}(\Phi(D) = 1) \to 1$

# Limitations of concentration inequalities

Concentration inequalities like Hoeffding have two main limitations:

1. **Assumptions**: e.g., bounded rewards (in clinical trial reward could be unbounded if it's cholesterol level or survival time)

2. **Conservative**: e.g., Hoeffding's proof upper-bounds the reward by 1 (in online advertising where reward is binary, this bound is very loose since vast majority of ad viewers don't click)

Can we test $H_0$ with adaptive data non-conservatively, without assumptions on $\nu^{(k)}$?

# Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ Motivation: analyzing data from RL
- ✓ Hypothesis testing
- Randomization testing

# Notation

Let $r := (r_0, \ldots, r_{T-1})$ w/ mean $\bar{r}$ and $a := (a_0, \ldots, a_{T-1})$ w/ mean $\bar{a}$; note $D = (r, a)$

Consider the empirical correlation $\rho := \dfrac{(r - \bar{r})^\top (a - \bar{a})}{\|r - \bar{r}\| \, \|a - \bar{a}\|}$

When $H_0$ is true, we expect $\rho \approx 0$ since the rewards when $a_t = 0$ look no different from the rewards when $a_t = 1$

When $\mu^{(0)} \neq \mu^{(1)}$, we expect $\rho \not\approx 0$ since the rewards when $a_t = 0$ are systematically shifted relative to the rewards when $a_t = 1$

Suggests $\Phi(D) = 1\{\,|\rho| > c\}$, but how to find $c$ such that $\mathbb{P}(\,|\rho| > c) = 5\,\%$?

# Randomization testing

Want: $c$ such that $\mathbb{P}(|\rho| > c) = 5\,\%$

Suppose data is fully i.i.d., e.g., $a_t \sim \text{Bernoulli}(0.5)$ for all $t$

If we resample $a$ as $\tilde{a}_t \sim \text{Bernoulli}(0.5)$ and define $\tilde{a} := (\tilde{a}_0, \ldots, \tilde{a}_{T-1})$ and $\tilde{r} := r$,

our resampled data $\tilde{D} := (\tilde{r}, \tilde{a})$ has exactly the same distribution as $D := (r, a)$

Thus also $\tilde{\rho} := \dfrac{(\tilde{r} - \bar{\tilde{r}})^\top (\tilde{a} - \bar{\tilde{a}})}{\|\tilde{r} - \bar{\tilde{r}}\| \, \|\tilde{a} - \bar{\tilde{a}}\|}$ has the same distribution as $\rho$

Idea: independently sample $\tilde{\rho}_1, \ldots, \tilde{\rho}_{100{,}000}$ and use the $950{,}000^{\text{th}}$ largest $|\tilde{\rho}_i|$ as $c$

# Randomization tests with bandit data

Now what about when data is not i.i.d.?

Want: resampled data $\tilde{D} := (\tilde{r}, \tilde{a})$ to have same distribution as $D := (r, a)$

When $H_0$ is true, the $r_t$ are i.i.d. regardless of $a_t$, so can still set $\tilde{r} := r$, and just need to sample $\tilde{a}$ from the conditional distribution of $a \mid r$

But the $a_t$ in a bandit depends on previous $r_t$, so they are not i.i.d.

Idea: simply run bandit to choose $\tilde{a}_t$ as if the rewards you're getting are the $r_t$

Then this $\tilde{a}$ is exactly a sample from $a \mid r$, and we have the property we want:
$\tilde{D} := (\tilde{r}, \tilde{a})$ has same distribution as $D := (r, a)$

# Formal algorithm for resampling

A bandit algorithm $\mathscr{A}$ determines the arm sampling distribution given $H_t$:

$$\mathbb{P}_{\mathscr{A}}(\cdot \mid a_0, r_0, \ldots, a_{t-1}, r_{t-1})$$

To sample $\tilde{D} = (\tilde{r}, \tilde{a})$:

For $t = 0, \ldots, T - 1$:

Sample $\tilde{a}_t \sim \mathbb{P}_{\mathscr{A}}(\cdot \mid \tilde{a}_0, \tilde{r}_0, \ldots, \tilde{a}_{t-1}, \tilde{r}_{t-1})$

Set $\tilde{r}_t = r_t$

# Even better…

Since the $r_t$ are i.i.d. anyway, can randomize their order:

Sample permutation $p$ uniformly from permutations of $\{0,\ldots,T-1\}$

For $t = 0,\ldots,T-1$:

Sample $\tilde{a}_t \sim \mathbb{P}_{\mathscr{A}}(\cdot \mid \tilde{a}_0, \tilde{r}_0, \ldots, \tilde{a}_{t-1}, \tilde{r}_{t-1})$

Set $\tilde{r}_t = r_{p(t)}$

Can also add back context $x_t$, treat it like $r_t$ since it's i.i.d.:

Sample permutation $p$ uniformly from permutations of $\{0,\ldots,T-1\}$

For $t = 0,\ldots,T-1$:

Set $\tilde{x}_t = x_{p(t)}$

Sample $\tilde{a}_t \sim \mathbb{P}_{\mathscr{A}}(\cdot \mid \tilde{x}_0, \tilde{a}_0, \tilde{r}_0, \ldots, \tilde{x}_{t-1}, \tilde{a}_{t-1}, \tilde{r}_{t-1}, \tilde{x}_t)$

Set $\tilde{r}_t = r_{p(t)}$

# Test statistic

Nothing about previous argument used any properties of $\rho$, just that it was a function of the data and the same function could be computed on resampled data

Thus, test statistic can be **anything** we think would take different values when $H_0$ is true than when it is false

Without context we were looking for differences between $r_t$ when $a_t = 0$ versus when $a_t = 1$, generally by comparing estimates of means $\hat{\mu}_T^{(0)}$ and $\hat{\mu}_T^{(1)}$

With context, want to compare estimates of conditional means given context: The functions $\hat{\mu}_T^{(0)}(x)$ and $\hat{\mu}_T^{(1)}(x)$

$\hat{\mu}_T^{(0)}(x)$ and $\hat{\mu}_T^{(1)}(x)$ could be fitted via supervised learning in totally black-box way (e.g., neural networks)

# Summary

Resample the data by shuffling the $(x_t, r_t)$ pairs and sampling $a_t$ per your algorithm

Compute a test statistic $\rho$ on the original data and many resampled data sets

Reject $H_0$ if $\rho$ above the 95th percentile of resampled $\tilde{\rho}$

Works for **any** contextual bandit algorithm

Works for **any** test statistic

Makes **no assumptions** about the conditional reward distributions $\nu^{(k)}(x)$

# Extensions

Under some assumptions, same idea gives a confidence interval for difference between two arms: uncertainty quantification

With more than two arms, can test more specific hypotheses like $\nu^{(1)}(x) = \nu^{(3)}(x)$

Can also give prediction interval, i.e., interval that contains next (unseen) reward with high probability

For these prior two extensions, need sophisticated importance sampling!

Can extend beyond (contextual) bandits to MDPs, but it gets hard…

# Today

- ✓ Feedback from last lecture
- ✓ Recap
- ✓ Motivation: analyzing data from RL
- ✓ Hypothesis testing
- ✓ Randomization testing

# Summary:

- Uncertainty quantification is a critical aspect of RL, and independently useful
- Randomization testing can answer questions non-conservatively

- Thanks for a great semester, and good luck on your final projects!

Attendance:

bit.ly/3RcTC9T

Feedback:

bit.ly/3RHtlxy