

Last Time:

- Reward learning
- Policy learning

Lecture 8

HRI, FALL '25

Andrea Bojarsky

This Time:

- Behavior Cloning
- Experimental Design

Behavior Cloning (BC)

The "simplest" algorithm to run on your dataset $\mathcal{D} := \{\langle s_i \rangle\}$

Train a model (e.g. NN) to predict expert actions given observed states, treating it as if it was supervised learning:

$$\min_{\theta} \sum_{(s^i, a^i) \in \mathcal{D}} l(a, \pi_\theta(s))$$

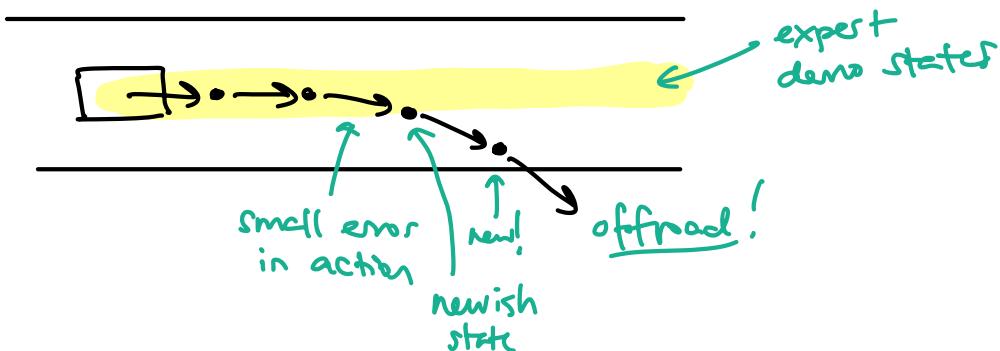
⇒ issues w/ BC:

In standard SL, we assume that the data pts. are independent & identically distributed (IID) ...

However, IL is in a sequential domain where current actions influence future states ⇒ assumption is wrong!

⇒ this means BC is prone to error accumulation & can't recover as it drifts away from demonstration distribution

ex. ALVINN (CMU '88)



Solutions to BC Error Accumulation:

↳ Dagger (Dataset Aggregation, 2011), DART (2017), GAIL (2016) + more!

User Studies & Experiments

What is a user study or experiment?

↳ user study is a broad term for any research conducted to understand users

↳ user experiment is a specific type of user study that tests hypotheses via a controlled design, treating users as participants in a structured investigation

ex. interviewing users about their experiences w/a voice assistant

vs.

measuring whether AI response time affects perceived trust in the AI system.

! Terms are often used interchangeably in the field, but it's important to know the difference.

Why do user studies / experiments?

- validate that a system works as expected
- compare 2+ systems or algorithms
- explore a phenomenon to develop a research Q
- collect training data for a model / algorithm

How to conduct user studies/ experiments:

1. Define the research question and hypotheses
2. Design a study to address Qs
3. Execute the study
4. Analyze data from the study
5. Draw conclusions from the analysis

EXPERIMENTAL DESIGN

What is an experiment vs. what is a Good experiment?

1. What is an experiment?

"The Arrangement of Field Experiments"

⇒ ORIGINS : - Agriculture: comparing different fertilizers to determine which ⇒ better crops (Fisher, 1926)

- Medical Research: testing a drug's effectiveness

⇒ OUR FOCUS : designing experiments to compare robot behaviors / algorithms / policies / models ; their effects on human collaboration / perception

⇒ COMPONENTS :

treatments
(conditions)

- drug vs. placebo
- random vs. optimal
- IRL policy vs. BC policy

responses
(measures)

- symptom progression
- comfort
- success rate

experimental units
(subjects)

- patients
- human users
- MDP problems

assignment methods
(subject allocation)

- random (drug, placebo)
- every user sees both
- every MDP "sees" both

let's operationalize these :

• Independent Variables (IV) → Conditions

↳ what you manipulate

↳ IV's have levels : - 2 levels : drug vs. no drug
- 3 levels : 100mg vs. 200mg vs. 300mg

- Dependent Variables (DV) → measures

↳ what you measure

↳ DV's can be objective (success rate, time) AND
subjective (surveys)

- Population → subjects

↳ how many (i.e. size of population)?

↳ who (e.g. age, gender, education, tech. experience)?

⇒ B/c we work with people, we need to abide by research ethics; i.e. protect participants from physical / mental / emotional harm, violations of privacy, confidentiality, feeling forced to start or continue.

⇒ every study needs approval from Internal Review Board (IRB)

⇒ Before study starts, we obtain informed consent from all participants.

- tells user about task, risks, benefits, rights
- gets confirmation of voluntary participation

- Assignment

↳ between-subjects: one condition per user

→ useful if participants seeing multiple levels is a problem (i.e. bias)

→ good for large participant pools or short study sessions

↳ within-subjects: user experiences all conditions

→ accounts for interpersonal variability

- efficiently uses your participant pool
- susceptible to ordering effects (might want to randomize)

↳ mixed-subjects

- Hypothesis

↳ (weak) IV x affects DV

↳ (stronger) IV x positively affects DV

ex: B/c optimal motion is more predictable, we hypothesize:

H1: Optimal robot motions increase user comfort

↳ Good hypotheses:

- make specific predictions
(+ will be supported by data or not)
- are measurable ("better"; isn't measurable)
- address your research Q + extract a key insight

What is a good experiment?

↳ good experiments are controlled

ex. Left handedness (LH)

obs. study: - 2,000 people who recently died
(1980s, 1990s) - left handed died 9 yrs younger } \Rightarrow LH die younger

what's wrong? changing prevalence of reporting left handedness over time

! confound! left-handedness was condemned in older generations, so majority of people identified as right handed. But younger generations naturally show higher proportion of people who are left-handed

↳ good experiments avoids confounds

ex. artificial decrease of LH, gender, experience w/ robots, algorithm hyperparameters

↓ How to avoid confounds?

→ Between subjects: randomized group assignment

ex. if 50 participants and 2 conditions: rand. assign. 25 to condition 1 and 25 to condition 2.

→ Within subjects: counter-balance the condition

Order 1: Alg. A, Alg. B

N conditions $\Rightarrow N!$ orderings

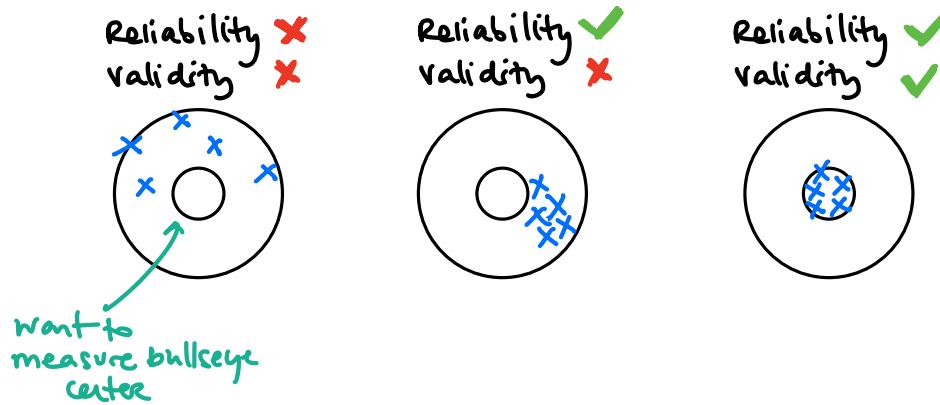
Order 2: Alg. B, Alg. A

→ pre-study practice: unrecorded familiarization stage

↳ good experiments are reliable

↳ good experiments have construct validity

ex. IQ test → intelligence?
rating of predictability → predictability?



↳ good experiments have external validity

↳ good experiments are factorial

ex. IV 1: Value Iter vs. Q-learning } 2x2 Design
IV 2: Sparse Rew. vs. Dense

	VI	Q
Sparse	C1	C2
Dense	C3	C4

common pitfall: 2 changes @ once

↳ w/o factorial design, we may only test Q + Dense vs. VI + sparse and then have a confound!

STATISTICAL ANALYSIS

Q "Has the DV changed as a result of manipulating IV?"

ex. 1 IV (alg), 2 levels (alg 0, alg 1), within subjects (2 MDP Problem)

MDP Problem	level	Total Reward	→	MDP	Diff Reward
				1	2
"participant 1"	0	10	→	1	2
	1	8			
"participant 2"	0	11	→	2	4
	1	7			

how participant behaved under 0
how they behaved under 1

↑ for each participant, calculate diff. btwn their pre & post intervention score

t-test: what's the probability 2 populations are different from each other?

↖ "null hypothesis" a statement of no effect

$H_0: \mu_1 = \mu_2$ (the population means are equal)

$H_1: \mu_1 \neq \mu_2$ (the population means are not equal)

Paired t-test:

$$t = \frac{\bar{X}_{\text{diff}}}{\bar{S}_{\text{diff}} / \sqrt{n}}$$

mean of sample differences

stdev. of the sample differences

sample size

ex. $n=2$, $\bar{X}_{\text{diff}} = \frac{2+4}{2} = 3$, $\bar{S}_{\text{diff}} = \sqrt{\frac{(2-3)^2 + (4-3)^2}{2}} = 1$

$$t = \frac{3}{1/\sqrt{2}} = 4.2426$$

$$df = n - 1 = 1$$

"degrees of freedom"

You use the combo of (t , df) to obtain a p-value (via a lookup table, or statistical software)

p-value: probability of obtaining a result at least as extreme as the results actually observed, assuming the null hypothesis is true.

→ Intuitively: determines if observed result is likely due to chance or a real effect

→ small p-value \Rightarrow observed results are unlikely to have occurred by chance; provides enough evidence to reject H_0
(e.g. ≤ 0.05)

→ large p-value \Rightarrow observed results are likely due to chance; can't reject H_0
(e.g. > 0.05)

! a small p-value does not prove alternative hypothesis is true, just suggests H_0 is unlikely

↳ in statistics, we aim to falsify the null hypothesis, not to prove the alternative

\bar{X} larger \Rightarrow more confident
 N larger \Rightarrow more confident
 S larger \Rightarrow less confident

→ if DV is binary (categorical rather than continuous):
(T/F) (yes/no/maybe) (reward & R)
use a chi-squared (χ^2) test instead of t-test

OTHER USEFUL SCENARIOS:

- 1 IV, 2 levels, between subjects

here, comparing 2 groups where the data is not paired (but in paired t-test it is)

independent t-test $\rightarrow t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\bar{S}_1^2}{N_1} + \frac{\bar{S}_2^2}{N_2}}}$

- 2 IVs, 2 levels each, between subjects



If we decided to run t-tests, for all pairwise comparisons of $k=4$ groups, we need:

$$\# \text{ comparisons} = \frac{k(k-1)}{2} = \frac{4(3)}{2} = 6$$

! Should we do this? 6 t-tests? No!

Increases Type I error rate (i.e. false positives)

ex. one t-test has a 5% chance ($\alpha = 0.05$) of falsely rejecting H_0 (i.e. detecting significance).

If you do 100 t-tests:

each test indep. of next
 $\alpha = \prod_i P(\text{correct}_i)$

$$P(\text{error}_1 \text{ or } \text{error}_2 \dots \text{ or } \text{error}_{100}) = 1 - P(\text{correct}_1, \text{AND} \dots, \text{correct}_{100})$$

" $P(\text{make } \geq 1 \text{ error})$ "
 $= 1 - 0.95^{100} = 0.9941!$

! 99.41% chance you falsely reject H_0

Some Solutions:

(A) Bonferroni Correction: adjust significance level by # of comparisons

if $\alpha = 0.05$ then now $\bar{\alpha} = \alpha/m$, $m = \# \text{ comparisons}$

ex. $m = 6$ from example above, $\bar{\alpha} = 0.008\bar{3}$

! very conservative (i.e. increased risk of Type II error)
 i.e. false negatives

(B) Tukey's HSD (Honesty Significant Differences):
 apply a single test to compare all groups @ once.

For Factorial Designs or IV with >2 levels

ANOVA (Analysis of variance)

- ⊗ always run an ANOVA first as a precursor to making multiple comparisons
 - ↳ it tells you if there is a difference, but not where the difference is

- main effect : the effect of each IV individually
- interaction effect: how IVs combine to influence DVs

If there is an interaction effect or the main effect is not clear, then run a "post-hoc" test (e.g. Tukey's HSD) to see which groups differ.

- Use a one-tailed ANOVA to test diff. in a specific direction (ex. method A leads to higher user efficiency than method B)
- Use a two-tailed ANOVA to test for any significant difference btwn. group means, whether its increase or decrease (ex. there is a diff. btwn method A and B)