Topic 4: Making decisions

Eric B. Laber

Department of Statistical Science, Duke University

Statistics 561

On choices

Dad always thought laughter was the best medicine, which I guess is why several of us died of tuberculosis.
—Robert Koch

On decisions

When you're ten years old, and a car drives by and splashes a puddle of water all over you, it's hard to decide if you should go to school like that or try to go home and change and probably be late. So while he was trying to decide, I drove by and splashed him again..
—Bertrand Russell

Warm-up (5 minutes)

- Explain to your group
  - What is a randomized clinical trial? Why do we randomize?
  - What is confounding?
  - What is a one-armed bandit?
- True or false
  - Regression + randomization = causality
  - Bandit problems were invented by computer scientists
  - Laber has had food poisoning from Pizza Hut twice

Decision problems

- Nearly all statistical analyses drive decisions
  - Estimate treatment effect $\Rightarrow$ treatment recommendations
  - Identify gene associated with disease $\Rightarrow$ follow-up study
  - Model click-through-rate as function of customer + ad attributes $\Rightarrow$ ad selection for website
  - Forecast product demand $\Rightarrow$ manufacturing decisions
  - Model wins-above-replacement $\Rightarrow$ contract decisions
  - ...

Roadmap

- One-stage decision problems
- K-stage decision problems
- Contextual bandits
- Markov Decision Problems
- Freedom!

One-stage setup

- Observe $\{(\boldsymbol{X}_i, A_i, Y_i)\}_{i=1}^n$ iid from P
  - $\boldsymbol{X} \in \mathcal{X} \subseteq \mathbb{R}^p$ covariates (decision context)
  - $A \in \mathcal{A} = \{-1, 1\}$ action (treatment, intervention, decision, etc.)
  - $Y \in \mathbb{R}$ utility (outcome, output, reward, etc.) higher is better
- Goal: select actions to maximize expected utility

Policies

- $\psi : \mathcal{X} \to 2^{\mathcal{A}}$ is set of allowable actions, i.e., $\psi(\boldsymbol{x}) \subseteq \mathcal{A} \setminus \emptyset$
- Policy $\pi : \mathcal{X} \to \mathcal{A}$ such that $\pi(\boldsymbol{x}) \in \psi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$
  - Under π decision maker will select action $\pi(\boldsymbol{x})$ in context $\boldsymbol{x}$
  - Define $V(\pi) \triangleq \mathbb{E}^{\pi} Y$ to be expected utility if actions are selected according the policy π
  - Optimal policy satisfies $V(\pi^{\mathrm{opt}}) \geq V(\pi)$ for all π

Formalizing the optimal policy

- Potential outcome $Y^*(a)$ under action a, i.e., the outcome under action a (which may be contrary to what was observed)
  - Imagine each individual has two potential outcomes $Y^*(1)$ and $Y^*(-1)$, one associated with each action
  - The potential outcome under policy π is

$$Y^*(\pi) = Y^*(1)1_{\pi(\boldsymbol{X})=1} + Y^*(-1)1_{\pi(\boldsymbol{X})=-1}$$

    formally, the value of a policy is $V(\pi) = \mathbb{E}Y^*(\pi)$

laber draws a table illustrating potential outcomes

Identifying the optimal policy

- Optimal policy defined in terms of potential outcomes, need to link to data-generating model
- Standard causal assumptions
  - No unmeasured confounders: $\{Y^*(1), Y^*(-1)\} \perp A \big| \boldsymbol{X}$
  - Consistency: $Y = Y^*(A)$, i.e., outcome is potential outcome under action taken
  - Positivity: there exists $\epsilon > 0$ such that $P(A = a|\boldsymbol{X} = \boldsymbol{x}) \geq \epsilon$ for (almost) all $\boldsymbol{x} \in \mathcal{X}$

  No interference: the action assigned to one unit do affect outcomes of others

No unmeasured confounders

- Actions may be selected according to $\boldsymbol{X}$ and the perceived impact on the outcome Y, e.g., clinical decisions
  - No unmeasured confounders says we captured all factors affecting action selection and the outcome
  - One minute: construct example in which this assumption is violated

Spillover effects

- One minute: generate three examples where this is violated
  -
  -
  -

A sad fact about your life

- Requisite causal assumption are not testable using observed data
  - There is no test, procedure, etc. that can be applied to the observed data (no matter how much there is) to evaluated needed causal conditions
  - Randomization ensures no unmeasured confounders by construction
  - Must use external information: i.e., knowledge of underlying science, richness of the features $\boldsymbol{X}$, etc.

Regression-based characterization of optimal regime

- Define $Q(\boldsymbol{x}, a) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}, A = a)$
- Under standard causal assumptions

$$V(\pi) = \mathbb{E}Q\left\{\boldsymbol{X}, \pi(\boldsymbol{X})\right\}$$

given this expression suggest an estimator of π^{opt}

Derivation of regression-based estimator

Q-learning: part I

- Bound: for any policy π it follows that

$$V(\pi) = \mathbb{E} Q\left\{\boldsymbol{X}, \pi(\boldsymbol{X})\right\} \leq \mathbb{E} \sup_{a \in \psi(\boldsymbol{X})} Q(\boldsymbol{X}, a)$$

- Note that the policy

$$\pi^{\text{opt}}(\boldsymbol{x}) = \arg \max_{a \in \psi(\boldsymbol{x})} Q(\boldsymbol{x}, a)$$

attains this bound and is thus optimal

Q-learning: part I cont'd

- Idea: estimate $Q(\boldsymbol{x}, a)$ by regressing Y on $\boldsymbol{X}, A$ to obtain $\widehat{Q}_n(\boldsymbol{x}, a)$ and subsequently $\widehat{\pi}_n(\boldsymbol{x}) = \arg\max_{a \in \psi(\boldsymbol{x})} \widehat{Q}_n(\boldsymbol{x}, a)$
- Ex. suppose $\psi(\boldsymbol{x}) = \{-1, 1\}$ and posit linear model $Q(\boldsymbol{x}, a; \beta) = \boldsymbol{x}_0^\intercal \beta_0 + a \boldsymbol{x}_1^\intercal \beta_1$ where $\beta = (\beta_0^\intercal, \beta_1^\intercal)^\intercal$ and $\boldsymbol{x}_0, \boldsymbol{x}_1$ features of $\boldsymbol{x}$
  - $\widehat{\beta}_n \triangleq \arg\min_{\beta} \mathbb{P}_n \{Y - Q(\boldsymbol{X}, A; \beta)\}^2$
  - $\widehat{\pi}_n(\boldsymbol{x}) = \arg\max_a Q(\boldsymbol{x}, a; \widehat{\beta}_n) = \operatorname{sign}(\boldsymbol{x}_1^\intercal \widehat{\beta}_{1,n})$

Q-learning: flexible models

- No need to restrict to linear models
- Construct estimator $\widehat{Q}_n(\boldsymbol{x}, a)$ of $Q(\boldsymbol{x}, a) = \mathbb{E}(Y|\boldsymbol{X} = x, A = a)$ using random forest and then take $\widehat{\pi}_n(\boldsymbol{x}) = \arg \max_{a \in \psi(\boldsymbol{x})} \widehat{Q}_n(\boldsymbol{x}, a)$
- Quantum theory of decision rules: if a regression method exists, someone, somewhere has published a paper applying it in Q-learning
  - Boosting
  - Neural nets
  - Nearest neighbors
  - Gaussian processes
  - NP-Bayes
  - ...

Advantage learning v1

- Note that we can always write

$$Q(\boldsymbol{x}, a) = \mu(\boldsymbol{x}) + a\Delta(\boldsymbol{x})$$

where

$$\mu(\boldsymbol{x}) = \frac{Q(\boldsymbol{x}, 1) + Q(\boldsymbol{x}, -1)}{2}, \text{ and } \quad \Delta(\boldsymbol{x}) = \frac{Q(\boldsymbol{x}, 1) - Q(\boldsymbol{x}, -1)}{2}$$

only need to estimate Δ to identify optimal policy

Advantage learning v1 cont'd

- Write $Q(\boldsymbol{x}, a) = \widetilde{\mu}(\boldsymbol{x}) + \{(1+a)/2 - q(\boldsymbol{x})\}\,\widetilde{\Delta}(\boldsymbol{x})$, where $\widetilde{\mu}(\boldsymbol{x}) = \mu(\boldsymbol{x})(1/2 - q(\boldsymbol{x}))\Delta(\boldsymbol{x})$, $q(\boldsymbol{x}) = P(A = 1|\boldsymbol{X} = \boldsymbol{x})$, and $\widetilde{\Delta}(\boldsymbol{x}) = 2\Delta(\boldsymbol{x})$

- Advantage learning solves

$$\widehat{\mu}_n, \widehat{\Delta}_n = \arg\min_{\mu,\Delta} \mathbb{P}_n \left[Y - \mu(\boldsymbol{X}) + \{(A+1)/2 - q(\boldsymbol{X})\}\,\Delta(\boldsymbol{X})\right]^2$$

so that the estimated optimal policy is given by $\widehat{\pi}_n(\boldsymbol{x}) = \operatorname{sign}\left\{\widehat{\Delta}_n(\boldsymbol{x})\right\}$ if $\psi(\boldsymbol{x}) = \{-1, 1\}$

Why A-learning v1 works

Why A-learning v1 works

Advantage learning v2

- Let $A \in \mathcal{A} \subset \mathbb{R}$ be more general action space and for simplicity assume that $\psi(\boldsymbol{x}) = \mathcal{A}$ for all $\boldsymbol{x}$
- Define $\Gamma(\boldsymbol{x}, a) = Q(\boldsymbol{x}, a) - \max_a Q(\boldsymbol{x}, a)$ so that $\Gamma(\boldsymbol{x}, a) \leq 0$ and $\Gamma(\boldsymbol{x}, a) = 0$ if $a = \pi^{\text{opt}}(\boldsymbol{x})$ then

$$Q(\boldsymbol{x}, a) = \omega(\boldsymbol{x}) + \Gamma(\boldsymbol{x}, a),$$

  where $\omega(\boldsymbol{x}) = \max_a Q(\boldsymbol{x}, a)$
- $\Gamma(\boldsymbol{x}, a)$ is the advantage of selecting action a in context $\boldsymbol{x}$ and $\pi^{\text{opt}}(\boldsymbol{x}) = \arg\max_a \Gamma(\boldsymbol{x}, a)$

Advantage learning v2

- Goal: estimate Γ without estimating ω
- Claim: Γ satisfies

$$\Gamma = \arg\min_{\gamma} \mathbb{E}\left\{ Y - \gamma(\boldsymbol{X}, A) + \int \gamma(\boldsymbol{X}, a) p(a|\boldsymbol{X}) d\eta(a) \right\}^2,$$

  where η is a dominating measure

- Note there's no ω in the above expression!

Why A-learning v2 works

Why A-learning v2 works

Classification-based representation: quiz

- Warm-up: discuss with your stats group (3.25 minutes)
  - What is sampling bias? When does it occur?
  - What is the Horvitz-Thompson estimator?
  - What is cost-sensitive classification?
- True or false
  - Double sampling is when you use the same spoon twice in the same trough of bean dip (I'm looking at you Chad)
  - Survey sampling is mostly relegated to the census and marketing
  - Laber discovered 'Cart Narcs' on YouTube at 2AM and binged every episode could find (let's keep the tough questions to a minimum)

Classification-based representation overview

- Define the propensity score $P(A = 1|\boldsymbol{X} = \boldsymbol{x})$
- Under our standard causal conditions

$$V(\pi) = P\left\{\frac{Y 1_{A=\pi(\boldsymbol{X})}}{P(A|\boldsymbol{X})}\right\}$$

  this is the classic Horvitz-Thompson (HT) estimator from survey sampling!

- HT more commonly known as inverse probability weighted (IPW) representation; What's the intuition behind this estimator?

HT representation details

IPWE

- Inverse probability weighted estimator (IPWE) of $V(\pi)$

$$\widehat{V}_n^{\mathrm{IPWE}}(\pi) = \mathbb{P}_n \left\{ \frac{Y 1_{A=\pi(\boldsymbol{X})}}{\widehat{P}_n(A|\boldsymbol{X})} \right\},$$

  where $\widehat{P}_n(a|\boldsymbol{x})$ is the estimated propensity score, e.g., estimated using logistic regression, nnet, etc.

- Estimated optimal decision rule

$$\widehat{\pi}_n = \arg\max_{\pi \in \Pi} \mathbb{P}_n \left\{ \frac{Y 1_{A=\pi(\boldsymbol{X})}}{P(A|\boldsymbol{X})} \right\}$$

  feasible for some classes Π if n isn't too large but generally not tractable

Linking IPWE with cost-sensitive classification

- A tale of brilliance and the despair of being too late

$$\begin{aligned}
\arg\max_{\pi} \widehat{V}_n^{\mathrm{IPWE}}(\pi) &= \arg\max_{\pi} \mathbb{P}_n \left\{ \frac{Y 1_{A\pi(\boldsymbol{X})\geq 0}}{\widehat{P}_n(A|\boldsymbol{X})} \right\} \\
&= \arg\min_{\pi} \mathbb{P}_n \left\{ \frac{Y 1_{A\pi(\boldsymbol{X})< 0}}{\widehat{P}_n(A|\boldsymbol{X})} \right\} \\
&= \arg\min_{\pi} \mathbb{P}_n \widehat{W}_n 1_{YA\pi(\boldsymbol{X})<0},
\end{aligned}$$

where $\widehat{W}_n = |Y|/\widehat{P}_n(A\boldsymbol{X})$

- Egad! This looks like a weighted classification problem!

Derivation of weighted classification

Convex surrogates and optimal decisions

- Consider decision rules $\pi(\boldsymbol{x}) = \operatorname{sign}\{f(\boldsymbol{x})\}$ where $f : \mathbb{R}^p \to \mathbb{R}$ is a (generally smooth) fn, e.g., $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\beta$
- Let $\phi : \mathbb{R} \to \mathbb{R}_+$ be one of our convex surrogates from classification, e.g., hinge loss, logistic loss, exp loss, etc.
- Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^p$ into $\mathbb{R}$, the outcome weighted estimator (OWL) is given by $\widehat{\pi}_n(\boldsymbol{x}) = \operatorname{sign}\left\{\widehat{f}_n(\boldsymbol{x})\right\}$ where

$$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \mathbb{P}_n \widehat{W}_n \phi\{YAf(\boldsymbol{X})\}$$

OWL example

- OWL of linear decision rule is given by

$$\widehat{\beta}_n = \arg\min_{\beta} \mathbb{P}_n \widehat{W}_n \phi\left(YA\boldsymbol{X}^\intercal\beta < 0\right) + \lambda||\beta||^2$$

  so that $\widehat{\pi}_n^{\mathrm{OWL}}(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{x}^\intercal\widehat{\beta}_n)$

- Same theory applies as in classification!

Fact: IPWE is terrible

- IPWE is highly unstable b/c small propensities inflate variance and only a fraction of data is used, e.g., in a randomized trial only 1/2 the data appear (on average) in the weighted sum[1]

- Better: AIPWE which is given by

$$\widehat{V}_n^{\mathrm{AIPWE}}(\pi) = \mathbb{P}_n \left[\frac{Y 1_{A=\pi(\boldsymbol{X})}}{\widehat{P}_n(A|\boldsymbol{X})} - \frac{1_{A=\pi(\boldsymbol{X})} - \widehat{P}_N(A|\boldsymbol{X})}{\widehat{P}_n(A|\boldsymbol{X})} \widehat{Q}_n \left\{ \boldsymbol{X}, \pi(\boldsymbol{X}) \right\} \right]$$

[1] All the data are used in the estimation of the propensity score.

Why AIPWE work? Double robustness.

Why AIPWE work? Double robustness. cont'd

Roadmap

- One-stage decision problems
- **K-stage decision problems**
- Contextual bandits
- Markov Decision Problems
- Freedom!

Sequential decision problems

- Multi-stage treatment strategies
- Planning a transition to carbon-neutrality
- Navigation
- Optimizing vaccine distribution over the next three months
- ...

Setup

- Observe $\{\boldsymbol{X}_{1,i}, A_{1,i}, \boldsymbol{X}_{2,i}, A_{2,i}, \ldots, \boldsymbol{X}_{T,i}, A_{T,i}, Y_i\}_{i=1}^n$ comprising n i.i.d. trajectories drawn from unknown distn P
  - $\boldsymbol{X}_t \in \mathcal{X} \subseteq \mathbb{R}^p$ measurements at time t
  - $A_t \in \mathcal{A} = \{-1, 1\}$ action/decision/txt at time t
  - $Y \in \mathbb{R}$ outcome/utility coded so that higher is better
- Define $\boldsymbol{H}_1 = \boldsymbol{X}_1$ and $\boldsymbol{H}_t = (\boldsymbol{H}_{t-1}, A_{t-1}, \boldsymbol{X}_t)$ to be history, i.e., info available to decision maker before decision at time t

Multi-stage policy

- Policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_T)$ so that $\pi_t : \operatorname{supp} \boldsymbol{H}_t \to \mathcal{A}$
  - Given info $\boldsymbol{H}_t = \boldsymbol{h}_t$, select decision $\pi_t(\boldsymbol{h}_t)$
  - Indices $t = 1, \ldots, T$ need not correspond to fixed calendar times
- Optimal regime $\boldsymbol{\pi}^{\text{opt}}$ maximizes mean utility to select decisions at each decision point; formalize using potential outcomes

Ex. Treatment regime: mHealth for PTSD in cancer patients (PI S. Smith)

First stage decision rule

If distress ≥ 3 **then:** Cancer Distress Coach (CDC)

Else if PTSD symptom score ≥ 20 **then:** CDC

Else: usual care

Second stage decision rule

If responder **then:** continue first stage treatment

Else if using CDC and PSTD change ≥ 3 **then:** add mCoaching

Else if using CDC and distress ≥ 4 **then:** add FaceTime CBT

Else FaceTime CBT only

Characterizing the optimal policy

- Let $\bar{\boldsymbol{a}}_t = (a_1, \ldots, a_t)$ denote decision sequence
  - $\boldsymbol{H}_t^*(\bar{\boldsymbol{a}}_{t-1})$ potential history under $\bar{\boldsymbol{a}}_{t-1}$
  - $Y^*(\bar{\boldsymbol{a}}_t)$ potential outcome under $\bar{\boldsymbol{a}}_T$
  - Potential outcome under policy π is thus

$$Y^*(\pi) = \sum_{\bar{\boldsymbol{a}}_T} Y^*(\bar{\boldsymbol{a}}_T) \prod_{v=1}^{T} 1_{\pi_v\{\boldsymbol{H}_v^*(\bar{\boldsymbol{a}}_{v-1})\}=a_v}$$

  - Define $V(\pi) \triangleq \mathbb{E}Y^*(\pi)$ opt policy satisfies $V(\pi^{\text{opt}}) \geq V(\pi)$ for all other policies π

Standard causal assumptions (generalized)

- No unmeasured confounders (sequential ignorability)

$$\{Y^*(\bar{\boldsymbol{a}}_T), \boldsymbol{H}_t^*(\bar{\boldsymbol{a}}_{t-1}), t = 1, \ldots, T\} \perp A_t \big| \boldsymbol{H}_t$$

  for all $t = 1, \ldots, T$

- Consistency, $Y = Y^*(\overline{\boldsymbol{A}}_T)$ and $\boldsymbol{H}_t = \boldsymbol{H}_t^*(\overline{\boldsymbol{A}}_{t-1})$ for all $t = 1, \ldots, T$

- Positivity, $P(A_t = a|\boldsymbol{H}_t) > 0$ with probability one for all $a \in \mathcal{A}$

Regression-based characterization of optimal regime

- Define Q-functions

$$Q_T(\boldsymbol{h}_T, a_T) \triangleq \mathbb{E}(Y|\boldsymbol{H}_T = \boldsymbol{h}_T, A_T = a_T)$$

$$Q_t(\boldsymbol{h}_t, a_t) \triangleq \mathbb{E}\left\{\max_{a_{t+1}} Q_{t+1}(\boldsymbol{H}_{t+1}, a_{t+1})\Big|\boldsymbol{H}_t = \boldsymbol{h}_t, A_t = a_t\right\}$$

$$\text{for} \quad t = T-1, \ldots, 1$$

- Dynamic programming (Bellman 1957)

$$\pi_t^{\text{opt}}(\boldsymbol{h}_t) = \arg\max_{a_t} Q_t(\boldsymbol{h}_t, a_t)$$

for $t = 1, \ldots, T$

Q-learning (again)

- Obvious idea: estimate Q-functions using regression
- E.g., posit linear models $Q_t(\boldsymbol{h}_t, a_t; \beta_t) = \boldsymbol{h}_{t,0}^{\mathsf{T}} \beta_{t,0} + a_t \boldsymbol{h}_{t,1}^{\mathsf{T}} \beta_{t,1}$ where $\boldsymbol{h}_{t,0}$ and $\boldsymbol{h}_{t,1}$ are features of $\boldsymbol{h}_t$ and $\beta_t = (\beta_{t,0}^{\mathsf{T}}, \beta_{t,1}^{\mathsf{T}})^{\mathsf{T}}$ are unknown coeffs
  - Compute $\widehat{\beta}_{T,n} = \arg\min_{\beta} \mathbb{P}_n \{Y - Q(\boldsymbol{H}_T, A_T; \beta_T)\}^2$
  - Recursively for $t = T-1, \ldots, 1$ compute

$$\widehat{\beta}_{t,n} = \arg\min_{\beta_t} \mathbb{P}_n \left\{ \max_{a_{t+1}} Q(\boldsymbol{H}_{t+1}, a_{t+1}; \widehat{\beta}_{t+1,n}) - Q_t(\boldsymbol{H}_t, A_t; \beta_t) \right\}^2$$

  - $\widehat{\pi}_t(\boldsymbol{h}_t) = \arg\max_{a_t} Q_t(\boldsymbol{h}_t, a_t; \widehat{\beta}_{t,n})$

Q-learning cont'd

- No need to stick to linear models can use ML/nonpar models
- Highly extensible, e.g., if regression works, Q-learning works (mostly)
  - Censoring
  - High-dimensional inputs
  - Image, text, other complex data structures
  - Functional inputs
  - ...

Nonregularity: the talk you parents never gave you

- Inference for Q-functions notoriously difficult
  - Max operator $\Rightarrow$ non-smooth functional of distn
  - No asymptotically unbiased estimator exists
  - No regular estimator exists
  - Bootstrap, normal approximations, and credible regions can perform poorly

laber derives limiting distribution in Q-learning linear models

laber derives limiting distribution in Q-learning linear models cont'd

Simple treatment selection problem

- Let P be fixed but unknown distn on $\mathbb{R}^p$ ($p \geq 2$) and define:

$$\boldsymbol{\mu}_0 \triangleq P\boldsymbol{X} = \int \boldsymbol{x} dP(\boldsymbol{x})$$

$$\theta_0 \triangleq \bigvee_{j=1}^{p} \mu_{0,j} = \max\left(\mu_{0,1}, \mu_{0,2}, \ldots, \mu_{0,p}\right)$$

- Observe $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ drawn i.i.d. from P and define:

$$\widehat{\boldsymbol{\mu}}_n \triangleq \mathbb{P}_n\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$$

$$\widehat{\theta}_n \triangleq \bigvee_{j=1}^{p} \widehat{\mu}_{n,j} = \max\left(\widehat{\mu}_{n,1}, \widehat{\mu}_{n,2}, \ldots, \widehat{\mu}_{n,p}\right)$$

- Inference for θ_0 based on limiting distn of $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$

Fixed parameter asymptotics for $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$

- For any $\nu \in \mathbb{R}^p$ define $\mathfrak{U}(\nu) = \arg\max_j \nu_j$ which may contain more than one element if no unique maximizer

Fixed parameter asymptotics for $\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right)$

- For any $\nu \in \mathbb{R}^p$ define $\mathfrak{U}(\nu) = \arg\max_j \nu_j$ which may contain more than one element if no unique maximizer

Lemma

Assume regularity conditions ensuring that $\sqrt{n}(\mathbb{P}_n - P)\boldsymbol{X} \rightsquigarrow N(0, \boldsymbol{\Sigma})$. *Then*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \bigvee_{j \in \mathfrak{U}(\mu_0)} Z_j,$$

where $\boldsymbol{Z} \sim N(0, \boldsymbol{\Sigma})$.

Proof overview

Proof

Let $\mathcal{E}_n = 1\left\{ \max_{k \notin \mathfrak{U}(\mu_0)} \widehat{\mu}_{n,k} \geq \max_{k \in \mathfrak{U}(\mu_0)} \widehat{\mu}_{n,k} \right\}$ then $\mathcal{E}_n = o_P(1)$. Write

$$\begin{aligned}
\sqrt{n}(\widehat{\theta}_n - \theta_0) &= \bigvee_{j \in \mathfrak{U}(\mu_0)} \sqrt{n}\,(\widehat{\mu}_{n,j} - \theta_0) \\
&\quad + \left\{ \bigvee_{k \notin \mathfrak{U}(\mu_0)} \sqrt{n}\,(\widehat{\mu}_{n,k} - \theta_0) - \bigvee_{j \in \mathfrak{U}(\mu_0)} \sqrt{n}\,(\widehat{\mu}_{n,j} - \theta_0) \right\} \mathcal{E}_n \\
&= \bigvee_{j \in \mathfrak{U}(\mu_0)} \sqrt{n}\,(\widehat{\mu}_{n,j} - \mu_{0,j}) + r_n,
\end{aligned}$$

where we have used $\mu_{0,j} = \theta_0$ for all $j \in \mathfrak{U}(\mu_0)$. The result follows from the CLT, continuous mapping theorem, and showing that $r_n = o_P(1)$.

Looking closer at $\sqrt{n}(\widehat{\theta}_n - \theta_0)$

- Lemma states $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \bigvee_{j \in \mathfrak{U}(\mu_0)} Z_j$, where $Z_j \sim N(0, \boldsymbol{\Sigma})$
- Limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ depends abruptly on μ_0
  - Ex., if $p = 2$ and $\mu_0 = (0, 0)^\intercal$ and $\boldsymbol{\Sigma} = \boldsymbol{I}_2$ then the limit is the max of two independent normals but if $\mu_0 = (0, \epsilon)^\intercal$ for $\epsilon > 0$, say $\epsilon = 10^{-20}$, the limit is standard normal!
  - Asymptotic distn depends on maximizers of $\boldsymbol{\mu}_0$ and associated submatrix of $\boldsymbol{\Sigma}$ but not gaps in $\boldsymbol{\mu}_0$ this is problematic as the finite sample distn can depend critically on these gaps

Finite sample behavior of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$

- Suppoze $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N(\mu_0, \boldsymbol{I}_p)$ and μ_0 has a unique maximizer $\mu_1 > \mu_j$ for $j = 2, \ldots, p$, then

$$P\left\{\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \leq t\right\} = \Phi(t) \prod_{j=2}^{p} \Phi\left\{t + \sqrt{n}(\theta_0 - \mu_{0,j})\right\}$$

  which can be far from asymptotic limit, $\Phi(t)$, when gaps $\theta_0 - \mu_{0,j}$ are small relative to $\sqrt{n}$

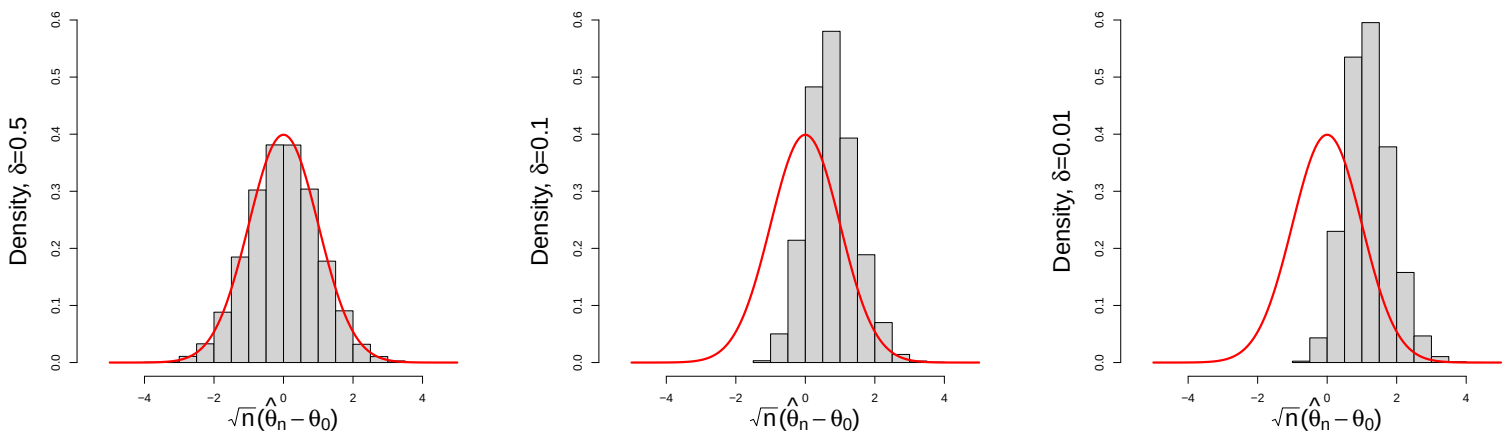
Finite sample behavior of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$

- Suppoze $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N(\mu_0, \boldsymbol{I}_p)$ and μ_0 has a unique maximizer $\mu_1 > \mu_j$ for $j = 2, \ldots, p$, then

$$P\left\{\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \leq t\right\} = \Phi(t) \prod_{j=2}^{p} \Phi\left\{t + \sqrt{n}(\theta_0 - \mu_{0,j})\right\}$$

  which can be far from asymptotic limit, $\Phi(t)$, when gaps $\theta_0 - \mu_{0,j}$ are small relative to $\sqrt{n}$

- Ex. set $p = 6$, $n = 100$, and $\mu_{0,j} = \mu_{0,1} - \delta$ for $j \geq 2$



Problems with fixed parameter asymptotics

- Asymptotic approximation to distn of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ poor as it fails to account for uncertainty in $\mathfrak{U}(\mu_0)$
  - No uncertainty about $\mathfrak{U}(\mu_0)$ as $n \to \infty$, i.e., power of $H_{0,j} : \mu_{0,j} = \theta_0$ goes to one
  - Mismatch of asymptotic framework and finite sample behavior
- Need asymptotic framework that retains uncertainty (instability) about $\mathfrak{U}(\mu_0)$ as n diverges

Moving parameter asymptotics

- Idea: consider sequence of generative models so that key features of sampling distribution retained asymptotically
  - Commonly used for comparing power of hypothesis tests
  - $1/\sqrt{n}$ perturbations of generative model[2]

- Consider $\boldsymbol{X}_{n,1}, \ldots, \boldsymbol{X}_{n,n}$ drawn i.i.d from P_n

$$\begin{aligned} \mu_{0,n} &\triangleq P_n \boldsymbol{X} = \mu_0 + \boldsymbol{s}/\sqrt{n} \\ \theta_{0,n} &\triangleq \bigvee_{j=1}^{p} \mu_{0,n,j} = \max\left(\mu_{0,n,1}, \ldots, \mu_{0,n,p}\right), \end{aligned}$$

  where $\boldsymbol{s} \in \mathbb{R}^p$ is a called local parameter

[2] It can also be useful to think of local alternatives as a kind of derivative, as we'll note later.

Moving parameter asymptotics cont'd

- Moving parameter asymptotic allow for gaps $\theta_{0,n} - \mu_{0,n,j}$ that shrink with sample size $\Rightarrow$ uncertainty about $\mathfrak{U}(\mu_0)$ persists

Lemma

Let $\boldsymbol{s} \in \mathbb{R}^p$ be fixed. Assume for each n that $\{\boldsymbol{X}_{n,i}\}_{i=1}^n$ are drawn i.i.d. from P_n, which satisfy (i) $P_n\boldsymbol{X} = \mu_0 + s/\sqrt{n}$, and (ii) $\sqrt{n}(\mathbb{P}_n - P_n)\boldsymbol{X} \rightsquigarrow N(0, \boldsymbol{\Sigma})$. Then, under P_n,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \bigvee_{j \in \mathfrak{U}(\mu_0)} (Z_j + s_j) - \bigvee_{j \in \mathfrak{U}(\mu_0)} s_j,$$

where $\boldsymbol{Z} \sim N(0, \boldsymbol{\Sigma})$.

Moving parameter limit of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$

- Under P_n: $\sqrt{n}(\widehat{\theta}_n - \theta_0) \rightsquigarrow \bigvee_{j \in \mathfrak{U}(\mu_0)} (Z_j + s_j) - \bigvee_{j \in \mathfrak{U}(\mu_0)} s_j$
- Limit depends on local parameter $\boldsymbol{s} \Rightarrow$ non-regular
  - Anticipates poor performance of asymptotic approximation
  - Bootstrap and series approximations invalid w/o modification
  - Non-regularity is ubiquitous in decision problems (think max operator in Bellman optimality eqns)
- Want inference procedures that are valid under moving parameter asymptotics

Valid inference via projection

- Building block idea: projection confidence set
  - $\boldsymbol{\mu}_0$ is regular
  - $\theta_0 = \theta_0(\boldsymbol{\mu}_0)$
- Projection confidence set for max of means
  - Construct $(1-\alpha)$ confidence region for $\boldsymbol{\mu}_0$

$$\zeta_{n,1-\alpha} = \{\boldsymbol{\mu} \in \mathbb{R}^p \,:\, n(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu})^\intercal \widehat{\boldsymbol{\Sigma}}_n^{-1} (\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) \leq \chi^2_{p,1-\alpha}\}$$

  then $P(\boldsymbol{\mu}_{0,n} \in \zeta_{n,1-\alpha}) \geq 1 - \alpha + o_{P_n}(1)$

  - Projection confidence set for θ_0

$$\Gamma_{n,1-\alpha} = \left\{ \theta \in \mathbb{R} \,:\, \theta = \bigvee_{j=1}^{p} \mu_j \text{ for some } \boldsymbol{\mu} \in \zeta_{n,1-\alpha} \right\}$$

Why the projection set works

- Error of projection set is

$$\begin{aligned} P\left(\theta_0 \notin \mathbf{\Gamma}_{n,1-\alpha}\right) &= P\left(\theta_0 \notin \mathbf{\Gamma}_{n,1-\alpha}, \boldsymbol{\mu}_0 \notin \boldsymbol{\zeta}_{n,1-\alpha}\right) \\ &\quad + P\left(\theta_0 \notin \mathbf{\Gamma}_{n,1-\alpha}, \boldsymbol{\mu}_0 \in \boldsymbol{\zeta}_{n,1-\alpha}\right) \\ &\leq P\left(\boldsymbol{\mu}_0 \notin \boldsymbol{\zeta}_{n,1-\alpha}\right) \\ &\leq \alpha + o_P(1), \end{aligned}$$

  holds under both fixed and moving parameter frameworks

- Pros and cons of projection sets
  - Robust, general, conceptually simple (one-line proof!)
  - Minimally adaptive $\Rightarrow$ extremely conservative in some settings

Bound-based inference

- Idea: sandwich non-smooth functional between two smooth bounds and use bootstrap or series approximations on bounds
- Bounds constructed by taking sup/inf over local perturbations
  - Valid under moving parameter frameworks by construction
  - Smallest possible among regular bounds[3]
  - Necessarily adaptive because if the functional is smooth (regular) it can serve as its own bound

[3] In some sense that can be made more precise.

Upper bound for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$

- Define conservative (overly inclusive) estimator of $\mathfrak{U}(\mu_0)$ as

$$\widehat{\mathfrak{U}}_n(\mu_0) = \left\{ j \,:\, \max_k \sqrt{n}(\widehat{\mu}_{n,k} - \mu_{n,j})/\widehat{\sigma}_{n,j,k} \leq \tau_n \right\},$$

  where $\tau_n \to \infty$, $\tau_n = o_p(\sqrt{n})$, and $\widehat{\sigma}_{n,j,k}$ an estimator of $n\mathrm{Var}(\widehat{\mu}_{n,j} - \widehat{\mu}_{n,k})$

- Local perturbations
  $\widehat{\mathcal{S}}_n(\mu_0) = \left\{ \boldsymbol{s} \in \mathbb{R}^p \,:\, s_j = \mu_{0,j} \text{ if } j \notin \widehat{\mathfrak{U}}_n(\mu_0) \right\}$

- Upper bound on $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is

$$\widehat{U}_n = \sup_{s \in \widehat{\mathcal{S}}_n(\mu_0)} \sqrt{n} \left[\bigvee_{j=1}^{p} \{(\widehat{\mu}_{n,j} - \mu_{0,j}) + s_j\} - \bigvee_{j=1}^{p} s_j \right]$$

  lower bound obtained by replace sup with an inf

Bound based confidence interval for θ_0

- Let $\widehat{u}^{(b)}_{n,1-\alpha/2}$ be the $(1-\alpha/2)\times 100$ percentile of $\widehat{U}_n$ and $\widehat{\ell}^{(b)}_{n,\alpha}$ the $(\alpha/2)\times 100$ percentile of the lower bound $\widehat{L}_n$ obtained via bootstrap, then

$$\left[\widehat{\theta}_n - \frac{\widehat{u}^{(b)}_{n,1-\alpha/2}}{\sqrt{n}}, \widehat{\theta}_n - \frac{\widehat{\ell}^{(b)}_{n,\alpha/2}}{\sqrt{n}}\right]$$

  is a $(1-\alpha)\times 100\%$ confidence interval for θ_0

Comments on bound-based interval

- Set $\widehat{\mathcal{S}}_n(\mu_0)$ only allows perturbations of means near the max
  - $\sim$ Pretest used in weakly identifiable problems in econometrics
  - Local parameter only appears in components that are near maximizers $\Rightarrow$ only want sup / inf to affect these
- Sup/inf adaptively 'smooths' $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ making it regular
- Alternative to bounding $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ and taking quantiles is to bound quantiles directly (may improve precision)

Discussion of max means problem

- Non-smooth max operator causes instability
  - Fixed parameter asymptotics provide poor approximation to the finite sample distribution $\Rightarrow$ std inference unreliable
  - Local (moving) parameter asymptotics faithfully capture instability as sample size grows large $\Rightarrow$ reliable for inference
  - Projection and bound-based sets two inference procedures that remain valid under local parameter asymptotics

Roadmap

- One-stage decision problems
- K-stage decision problems
- Contextual bandits
- **Markov Decision Problems**
- Freedom!

Setup and notation

- Observe $\left\{\boldsymbol{S}_i^1, A_i^1, \ldots, \boldsymbol{S}_i^T, A_i^T, \boldsymbol{S}^{T+1}\right\}_{i=1}^n$ drawn i.i.d. from P
  - T observation period, e.g., trial follow-up
  - $\boldsymbol{S}^t \in \mathcal{S} \subseteq \mathbb{R}^p$ state at time t
  - $A \in \mathcal{A} = \{1, \ldots, K\}$
- Utility $U^t = u(\boldsymbol{S}^t, A^t, \boldsymbol{S}^{t+1})$, higher is better
- Assume homogeneous MDP $\rightarrow$ focus on deterministic stationary policies $\pi : \mathcal{S} \rightarrow \mathcal{A}$

What is a homogeneous MDP?

Why is the MPD so powerful?

- Allows for estimation of optimal policy (to be applied indefinitely) from data collected over a finite horizon
- Joint distribution determined by
  - Initial state distn $f(\boldsymbol{s}^1)$
  - Transition kernel $f(\boldsymbol{s}'|\boldsymbol{s}, a)$, e.g., density of $\boldsymbol{S}_t$ given $\boldsymbol{S}_t$ and A_t
  - If we know these then we know everything!!

Defining an optimal policy

- Discounted reward $V(\boldsymbol{s};\pi) = \mathbb{E}^{\pi}\left(\sum_{v\geq 0}\gamma^{v}U^{t+v}\middle|\boldsymbol{S}^{t}=\boldsymbol{s}\right)$,

  optimal regime satisfies $V(\boldsymbol{s},\pi^{\text{opt}}) \geq V(\boldsymbol{s},\pi)$ for all $\boldsymbol{s}$ and π

- Model-based estimation (overview)
  - Posit model $f(\boldsymbol{s}^{t+1}|\boldsymbol{s}^{t},a;\boldsymbol{\theta})$ indexed by $\boldsymbol{\theta}\in\boldsymbol{\Theta}$
  - Construct likelihood $\mathcal{L}_n(\boldsymbol{\theta})$ compute $\widehat{\boldsymbol{\theta}}_n$ soln to $0=\nabla_{\boldsymbol{\theta}}\mathcal{L}_n(\boldsymbol{\theta})$
  - For any π, generate data under $\widehat{\boldsymbol{\theta}}_n$ construct estimator $V_{\widehat{\theta}_n}(\boldsymbol{s};\pi)$ with flexible approximation architecture (e.g., nnets)

Bellman equations

- Warm-up exercise, show:

$$V(\boldsymbol{S}^t;\pi) = \mathbb{E}^{\pi}\left\{U^t + \gamma V(\boldsymbol{S}^{t+1};\pi)\middle|\boldsymbol{S}^t\right\}$$

and thus for any function $\psi(\boldsymbol{S}^t)$ it follows that

$$0 = \mathbb{E}\left[\left(\frac{\delta_{A^t}(\boldsymbol{S}^t)}{\mu^t(A^t|\boldsymbol{H}^t)}\left\{U^t + \gamma V(\boldsymbol{S}^{t+1},\pi) - V(\boldsymbol{S}^t,\pi)\right\}\right)\psi(\boldsymbol{S}^t)\right],$$

where δ_w is a point-mass at w, $\boldsymbol{H}^t$ is the history at time t, and μ^t is the propensity

- Note: term inside parentheses (i.e., omitting ψ) is called importance-weighted TD-error

V-learning

- Model-free estimation (overview)
  - Posit model $V(\boldsymbol{s}, \pi; \boldsymbol{\lambda})$ indexed by $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$
  - Solve estimating eqns based on Bellman optimality eqns

$$0 = \mathbb{P}_n \sum_{t=1}^{T} \left[\frac{\delta_{A^t}(\boldsymbol{S}^t)}{\mu^t(A^t|\boldsymbol{H}^t)} \left\{ U^t + \gamma V(\boldsymbol{S}^{t+1}, \pi; \boldsymbol{\lambda}) - V(\boldsymbol{S}^t, \pi; \boldsymbol{\lambda}) \right\} \psi(\boldsymbol{S}^t, \pi; \boldsymbol{\lambda}) \right]$$

  - Optimal weights (per Godambe)

$$\psi(\boldsymbol{S}^t, \pi; \boldsymbol{\lambda}) = \mathbb{E}\left\{ \nabla_{\boldsymbol{\lambda}} g(\boldsymbol{\lambda}) \middle| \boldsymbol{S}^t \right\} / \mathbb{E}\left\{ g^2(\boldsymbol{\lambda}) \middle| \boldsymbol{S}^t \right\},$$

    where $g(\boldsymbol{\lambda})$ is importance-weighted TD-error
  - Est optimal weights, $\widehat{\psi}_{\widehat{\boldsymbol{\theta}}_n}(\boldsymbol{S}^t, \pi; \boldsymbol{\lambda})$ using fitted dynamics model

V-learning cont'd

- Let $\widehat{\boldsymbol{\lambda}}^{\pi}$ denote solution to est eqn at π
- Select reference distribution $\mathcal{R}$ over states[4]
- Compute estimated optimal policy in class Π as

$$\widehat{\pi} = \arg\max_{\pi \in \Pi} \int V(\boldsymbol{s}, \pi; \widehat{\boldsymbol{\lambda}}^{\pi}) d\mathcal{R}(\boldsymbol{s})$$

[4] How best to select this distribution is a bit of an open question. A typical choice is the estimated distribution of the initial state.

V-learning discussion

- Does not require correct dynamics model (i.e., model-free)
- Separate specification of class of policies, Π, which may be chosen to ensure intepretability, satisfaction of logistical/cost constraints, etc.
- (Blue)devil is in the details, optimization over space of policies can be non-trivial, typically stochastic search methods are used, e.g., SGD over parametric class

Thank you.

eric.laber@duke.edu

laber-labs.com