

Intro to nonsmooth inference

Eric B. Laber

Department of Statistics, North Carolina State University

April 2019

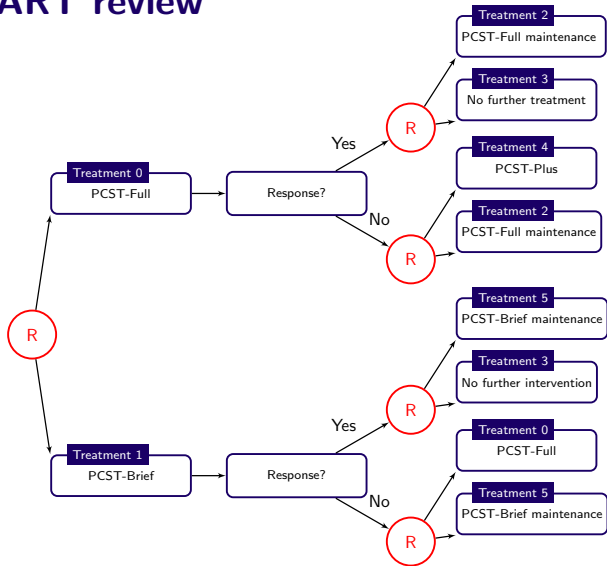
SAMSI



Last time

- ▶ SMARTs gold standard for est and eval of txt regimes
 - ▶ Highly configurable but choices driven by science
 - ▶ Looked at examples with varying scientific/clinical goals which lead to different timing, txt options, response criteria etc.
- ▶ Often powered by simple comparisons
 - ▶ First-stage response rates
 - ▶ Fixed regimes (most- vs. least-intensive)
 - ▶ First stage txts (problematic)
 - ▶ If test statistic is regular and asymptotically normal under null can use same basic template for power

Quick SMART review



Refresher

- ▶ Suppose that researchers are interested in comparing the embedded regimes:
 - (e_1) assign PCST-Full initially, assign PCST-Full maintenance to responders, and assign PCST-Plus to non-responders;
 - (e_2) assign PCST-Brief initially, assign no further intervention to responders, and assign PCST-Brief maintenance to responders.
- ▶ Recall our general template:
 - ▶ Test statistic: $\hat{V}_n(e_1) - \hat{V}_n(e_2)$, where \hat{V}_n is IPWE
 - ▶ Use $\sqrt{n}T_n/\hat{\sigma}_{e_1, e_2, n}^2$ asy normal and reject when this is large in magnitude

Goals for today

- ▶ Introduction to inference for txt regimes
 - ▶ Nonregular inference (and why we should care)
 - ▶ Basic strategies with a toy problem
 - ▶ Examples in one-stage problems

Warm up part I: quiz!

- ▶ Discuss with your stat buddy:
 - ▶ What are some common scenarios where series approx or the bootstrap cannot ensure correct op characteristics?
 - ▶ What is a local alternative?
 - ▶ How do we know if an asymptotic approx is adequate?
- ▶ True or false
 - ▶ If n is large asymptotic approximations can be trusted.
 - ▶ The top review of CLT on yelp complains about the burritos being too expensive.
 - ▶ The BBC produced an Hitler-themed sitcom titled 'Heil Honey, I'm home' in the 1950s.

On reality and fantasy

Your cat didn't say that. You know how I know? It's a cat. It doesn't talk. If you died, it would eat you. Starting with your face.

– Matt Zabka, recently single

Asymptotic approximations

- ▶ Basic idea: study behavior of statistical procedure in terms of dominating features while ignoring lower order ones
 - ▶ Often, but not always, consider diverging sample size
 - ▶ 'Dominating features' intentionally ambiguous
 - ▶ Generate new insights and general statistical procedures as large classes of problems share same dominating features
- ▶ Asymptotics mustn't be applied mindlessly
 - ▶ Disgusting trend in statistics: propose method, push through irrelevant asymptotics, handpick simulation experiments
 - ▶ Require careful thought about what op characteristics are needed scientifically and how to ensure these hold with the kind of data that are likely to be observed
 - ▶ No panacea \Rightarrow handcrafted construction and evaluation

Inferential questions in precision medicine

- ▶ Identify key tailoring variables
- ▶ Evaluate performance of true optimal regime
- ▶ Evaluate performance of estimated optimal regime
- ▶ Compare performance of two+ (possibly data-driven) regimes
- ▶ ...

Toy problem: max of means

- ▶ Simple problem that retains many of the salient features of inference for txt regimes
 - ▶ Non-smooth function of smooth functionals
 - ▶ Well-studied in the literature
- ▶ Basic notation
 - ▶ For $Z_1, \dots, Z_n \sim_{i.i.d.} P$ comprising n copies of $Z \sim P$ write $Pf(Z) = \int f(z)dP(z)$ and $\mathbb{P}_n f(Z) = n^{-1} \sum_{i=1}^n f(Z_i)$
 - ▶ Use ' \rightsquigarrow ' to denote convergence in distribution
 - ▶ Check: assuming requisite moments exist:

$$\sqrt{n}(\mathbb{P}_n - P)Z \rightsquigarrow \underline{\hspace{2cm}????\hspace{2cm}}$$

Max of means

- ▶ Observe $X_1, \dots, X_n \sim_{i.i.d.} P$ in \mathbb{R}^p with $\mu_0 = PX$, define

$$\theta_0 = \bigvee_{j=1}^p \mu_{0,j} = \max(\mu_{0,1}, \dots, \mu_{0,p})$$

- ▶ While we consider this estimand primarily for illustration, it corresponds to problem of estimating the mean outcome under an optimal one-size-fits-all treatment recommendation where $\mu_{0,j}$ is mean outcome under treatment $j = 1, \dots, p$.

Max of means: estimation

- ▶ Define $\hat{\mu}_n = \mathbb{P}_n X$, the plug-in estimator of θ_0 is

$$\hat{\theta}_n = \bigvee_{j=1}^p \hat{\mu}_{n,j}$$

- ▶ Warm-up:
 - ▶ Three minutes trying to derive limiting distn of $\sqrt{n}(\hat{\theta}_n - \theta_0)$
 - ▶ Three minutes discussing soln with your stat buddy

Max of means: first result

- ▶ For $v \in \mathbb{R}^p$ define $\mathfrak{L}(v) = \arg \max_j v_j$

Lemma

Assume regularity conditions under which $\sqrt{n}(\mathbb{P}_n - P)X$ is asymptotically normal with mean zero and variance-covariance matrix Σ . Then

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

where $Z \sim \text{Normal}(0, \Sigma)$.

Max of means: proof of first result

Extra page if needed

Max of means: discussion of first result

- ▶ Limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ depends abruptly on μ_0
 - ▶ If $\mu_0 = (0, 0)^T$ and $\Sigma = I_2$, the limiting distn is the max of two ind std normals
 - ▶ If $\mu_0 = (0, \epsilon)^T$ for $\epsilon > 0$ and $\Sigma = I_2$, the limiting distn is std normal even if $\epsilon = 1 \times 10^{-27}$!!
 - ▶ How can we use such an asymptotic result in practice?!
- ▶ Limiting distn of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ depends only on submatrix of Σ cor. to elements of $\mathcal{U}(\theta_0)$. What about in finite samples?

Max of means: discussion of first result cont'd

- ▶ Suppose $X_1, \dots, X_n \sim_{i.i.d.} \text{Normal}(\theta_0, I_p)$ and μ_0 has a unique maximizer, i.e., $\mathcal{L}(\mu_0)$ a singleton, say $\{\mu_{0,1}\}$
 - ▶ $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \text{Normal}(0, 1)$
 - ▶ $P\left\{\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \leq t\right\} = \Phi(t) \prod_{j=2}^p \Phi\left\{t + \sqrt{n}(\theta_0 - \mu_{0,j})\right\}$

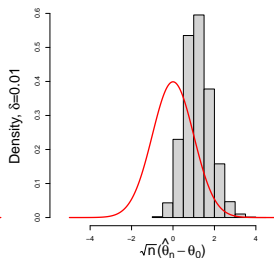
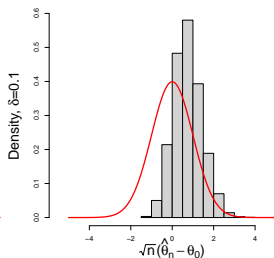
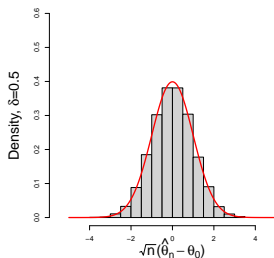
Quick break: derive this.

- ▶ If the gaps $\theta_0 - \mu_{0,j}$ are small relative to \sqrt{n} , the finite sample behavior can be quite different from limit $\Phi(t)$ ¹

¹Note that the limiting distribution doesn't depend on these gaps at all!

Max of means: normal approximation in pictures

- Generate data from $\text{Normal}(\mu, I_6)$ with $\mu_1 = 2$ and $\mu_j = \mu_1 - \delta$ for $j = 2, \dots, 6$. Results shown for $n = 100$.



Choosing the right asymptotic framework

- ▶ Dangerous pattern of thinking:
 - ▶ *In practice, none of the txt effect differences are zero.*
 - ▶ *I'll build my asy approximations assuming a unique maximizer.*

Choosing the right asymptotic framework

- ▶ Dangerous pattern of thinking:
 - ▶ *In practice, none of the txt effect differences are zero.*
 - ▶ *I'll build my asy approximations assuming a unique maximizer.*
 - ▶ *There finitely many components so maximizer is well-separated.*
 - ▶ *Idea! Plug-in estimated mazimizer and use asy normal approx.*
- ▶ Preceding pattern happens frequently, e.g., oracle property in model selection, max eigenvalues in matrix , and txt regimes

Choosing the right asymptotic framework

- ▶ What goes wrong? After all, this thinking works well in many other settings, e.g., everything you learned in stat 101.

Choosing the right asymptotic framework

- ▶ What goes wrong? After all, this thinking works well in many other settings, e.g., everything you learned in stat 101.
- ▶ Finite sample behavior driven by small (not necessarily zero) differences in txt effectiveness
 - ▶ We saw this analytically in normal case
 - ▶ Intuition helped by thinking in extremes, e.g., what if all txts were equal? What if one were infinitely better than others?
 - ▶ Abrupt dependence of limiting distribution on $\mathfrak{L}(\mu_0)$ is a redflag. It is tempting to construct procedures that will recover this limiting distn even if some txt differences are exactly zero. This is asymptotics for asymptotics sake. Don't do it.

Asymptotic working assumptions

- ▶ A useful asy approximation should be robust to the setting where some (all) txt differences are zero
 - ▶ Necessary but not sufficient
 - ▶ Heuristic: in small samples, one cannot distinguish between small (but nonzero) txt differences so use an asy framework which allows for exact equality.

This heuristic has been misinterpreted and misused in lit.

- ▶ Some procedures we'll look at are designed for such robustness

Local asymptotics: horseshoes and hand grenades

- ▶ Allowing null txt differences problematic
 - ▶ Asymptotically, differences either zero or infinite²
 - ▶ Txt differences are (probably) not exactly zero
- ▶ Challenge: allow small differences to persist as n diverges
 - ▶ Local or moving parameter asy framework does this
 - ▶ Idea: allow gen model to change with n so that gaps $\theta_0 - \max_{j \notin \mathcal{U}(\mu_0)} \mu_{0,j}$ shrink to zero as n increases³

²In the stat sense that we have power one to discriminate between them.

³This idea should be familiar from hypothesis testing.

Triangular arrays

- ▶ For each n , $X_{1,n}, \dots, X_{n,n} \sim_{i.i.d.} P_n$

<u>Observations</u>					<u>Distribution</u>
$X_{1,1}$					P_1
$X_{1,2}$	$X_{2,2}$				P_2
$X_{1,3}$	$X_{2,3}$	$X_{3,3}$			P_3
$X_{1,4}$	$X_{2,4}$	$X_{3,4}$	$X_{4,4}$		P_4
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots

- ▶ Define $\mu_{0,n} = P_n X$ and $\theta_{0,n} = \bigvee_{j=1}^p \mu_{0,j}$
 - ▶ Assume $\mu_{0,n} = \mu_0 + s/\sqrt{n}$ where $s \in \mathbb{R}^p$ called local parameter
 - ▶ Assume $\sqrt{n}(\mathbb{P}_n - P_n)X \rightsquigarrow \text{Normal}(0, \Sigma)^4$

⁴This is true under very mild conditions on the sequence of distributions $\{P_n\}_{n \geq 1}$. However, given our limited time we will not discuss such conditions. See van der Vaart and Wellner (1996) for details.

Quick quiz

- ▶ Suppose that $X_{1,n}, \dots, X_{n,n} \sim_{i.i.d.} \text{Normal}(\mu_0 + s/\sqrt{n}, \Sigma)$
what is the distribution of $\sqrt{n}(\mathbb{P}_n - P_n)X$?

Local alternatives anticipate unstable performance

Lemma

Let $s \in \mathbb{R}^P$ be fixed. Assume that for each n we observe $\{X_{i,n}\}_{i=1}^n$ drawn i.i.d. from P_n which satisfies: (i) $P_n X = \mu u_0 + s/\sqrt{n}$, and (ii) $\sqrt{n}(\mathbb{P}_n - P_n)X \rightsquigarrow \text{Normal}(0, \Sigma)$. Then, under P_n ,

$$\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) \rightsquigarrow \bigvee_{j \in \mathfrak{L}(\mu_0)} (Z_j + s_j) - \bigvee_{j \in \mathfrak{L}(\mu_0)} s_j,$$

where $Z \sim \text{Normal}(0, \Sigma)$.

Local alternatives anticipate unstable performance

Lemma

Let $s \in \mathbb{R}^P$ be fixed. Assume that for each n we observe $\{X_{i,n}\}_{i=1}^n$ drawn i.i.d. from P_n which satisfies: (i) $P_n X = \mu_{0,n} + s/\sqrt{n}$, and (ii) $\sqrt{n}(\mathbb{P}_n - P_n)X \rightsquigarrow \text{Normal}(0, \Sigma)$. Then, under P_n ,

$$\sqrt{n}(\hat{\theta}_n - \theta_{0,n}) \rightsquigarrow \bigvee_{j \in \mathfrak{L}(\mu_0)} (Z_j + s_j) - \bigvee_{j \in \mathfrak{L}(\mu_0)} s_j,$$

where $Z \sim \text{Normal}(0, \Sigma)$.

- ▶ Discussion/observations on local limiting distn
 - ▶ Dependence of limiting distn on $s \Rightarrow$ nonregular
 - ▶ Set $\mathfrak{L}(\mu_0)$ represents set of near-maximizers though $s_j = 0$ corresponds to exact equality (so haven't ruled this out)

Proof of local limiting distribution

Extra page if needed

Intermission: more regularity after this



Comments on nonregularity

- ▶ Sensitivity of estimator to local alternatives cannot be rectified through the choice of a more clever estimator
 - ▶ Inherent property of the *estimand*⁵
 - ▶ This has not stopped some from trying...
- ▶ Remainder of today's notes: cataloging of confidence intervals

⁵See van der Vaart (1991), Hirano and Porter (2012), and L. et al. (2011, 2014, 2019)

Projection region

- ▶ Idea: exploit the following two facts
 - ▶ $\hat{\mu}_n$ is nicely behaved (reg. asy normal)
 - ▶ If μ_0 were known this would be trivial⁶
 - ▶ Given $\alpha \in (0, 1)$ denote acceptable error level and $\zeta_{n,1-\alpha}$ a confidence region for μ_0 , e.g.,

$$\zeta_{n,1-\alpha} = \left\{ \mu \in \mathbb{R}^p : n(\hat{\mu}_n - \mu)^T \hat{\Sigma}_n (\hat{\mu}_n - \mu) \leq \chi_{p,1-\alpha}^2 \right\},$$

$$\text{where } \hat{\Sigma}_n = \mathbb{P}_n(X - \hat{\mu}_n)(X - \hat{\mu}_n)^T$$

- ▶ Projection CI:

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

⁶In this problem, θ_0 is a function of μ_0 and is thus completely known when μ_0 is known. In more complicated problems, knowing the value of a nuisance parameter will make the inference problem of interest regular.

Prove the following with your stat buddy

► $P(\theta_0 \in \Gamma_{n,1-\alpha}) \geq 1 - \alpha + o_P(1)$

Comments on projection regions

- ▶ Useful when parameter of interest is a non-smooth functional of a smooth (regular) parameter
- ▶ Robust and widely applicable but conservative
 - ▶ Projection interval valid under local alternatives (why?)
 - ▶ Can reduce conservatism using pre-test (L. et al., 2014)
 - ▶ Berger and Boos (1991) and Robins (2004) for seminal papers
- ▶ Consider as a first option in new non-reg problem

Bound-based confidence intervals

- ▶ Idea: sandwich non-smooth functional between smooth upper and lower bounds than bootstrap bounds to form conf region
- ▶ Let $\{\tau_n\}_{n \geq 1}$ be seq of pos constants such that $\tau_n \rightarrow \infty$ and $\tau_n = o(\sqrt{n})$ as $n \rightarrow \infty$, define

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

where $\hat{\sigma}_{j,k,n}$ is est of asy variance of $\hat{\mu}_{n,k} - \hat{\mu}_{n,j}$.

Note* May help to think of $\hat{\mathfrak{U}}_n$ to be the indices of txts that we cannot distinguish from being optimal.

Bound-based confidence intervals cont'd

- Given $\hat{\mathfrak{U}}_n(\mu_0)$ define

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

then, it follows that

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

is an upper bound on $\sqrt{n}(\hat{\theta}_n - \theta_0)$. (Why?) A lower bound, \hat{L}_n is constructed by replacing sup with an inf.

Dad, where do bounds come from?

- ▶ \hat{U}_n obtained by taking sup over all local, i.e., order $1/\sqrt{n}$, perturbations of generative model
 - ▶ By construction, insensitive to local perturbations \Rightarrow regular
 - ▶ $\hat{\mathfrak{U}}_n(\mu_0)$ conservative est of $\mathfrak{U}(\mu_0)$, lets wave our hands:

Bootstrapping the bounds

- ▶ Both \hat{U}_n and \hat{L}_n are regular and their distns consistently estimated via nonpar bootstrap
 - ▶ Let $\hat{u}_{n,1-\alpha/2}^{(b)}$ be $(1 - \alpha/2) \times 100$ perc of bootstrap distn of \hat{U}_n and $\hat{\ell}_{n,\alpha/2}^{(b)}$ the $(\alpha/2) \times 100$ perc of bootstrap distn of \hat{L}_n
 - ▶ Bound based confidence interval

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

Bound-based intervals discussion

- ▶ General approach, applies to implicitly defined estimators as as those with closed form expressions like we considered here
- ▶ Less conservative than projection interval but still conservative, such conservatism is unavoidable
 - ▶ Bounds are tightest in some sense
 - ▶ Bounding quantiles directly rather than estimand may reduce conservatism though possibly at price of addl complexity
- ▶ See Fan et al. (2017) for other improvements/refinements

Bootstrap methods

- ▶ Bootstrap is not consistent without modification
 - ▶ Due to instability (nonregularity)
 - ▶ Nondifferentiability of max operator causes this instability (see Shao 1994 for a nice review)
- ▶ Bootstrap is appealing for complex problems
 - ▶ Doesn't require explicitly computing asy approximations⁷.
 - ▶ Higher order convergence properties

⁷There are exceptions to this, including parametric bootstrap and those based on quadratic expansions

How about some witchcraft?

- ▶ m -out-of- n bootstrap can be used to create valid confidence intervals for non-smooth functionals
- ▶ Idea: resample datasets of size $m_n = o(n)$ so sample-level parameters converge 'faster' than bootstrap analogs (i.e., witchcraft)



m-out-of-*n* bootstrap

- ▶ Accepting some components of witchcraft on faith
 - ▶ $\sqrt{m_n} \left(\hat{\mu}_{m_n}^{(b)} - \hat{\mu}_n \right) \rightsquigarrow \text{Normal}(0, \Sigma)$ conditional on the data
 - ▶ See Arcones and Gine (1989) for details
- ▶ An even toyier example than our toy example: W_1, \dots, W_n i.i.d. w/ (μ, σ^2) derive limit distns of $\sqrt{n}(|\overline{W}_n| - |\mu|)$ and $\sqrt{m_n}(|\overline{W}_n^{(b)}| - |\overline{W}_n|)$

m-out-of*n* bootstrap with max of means

- ▶ Derive limiting distribution of $\sqrt{m_n} \left(\hat{\theta}_{m_n}^{(b)} - \hat{\theta}_n \right)$:

Extra page if needed

Intermission

I wouldn't want to wind up hooked to a bunch of wires and tubes, unless somehow the wires and tubes were keeping me alive. —Don Alden Adams

Finally! Back to treatment regimes (briefly)

- ▶ Consider a one-stage problem with observed data $\{(X_i, A_i, Y_i)\}_{i=1}^n$ where $X \in \mathbb{R}^p$, $A \in \{-1, 1\}$, and $Y \in \mathbb{R}$
 - ▶ Assume requisite causal conditions hold
 - ▶ Assume linear rules $\pi(x) = \text{sign}(x^\top \beta)$, where $\beta \in \mathbb{R}^p$, and x might contain polynomial terms etc.

Warm-up! Derive limiting distn of parameters in linear Q-learning!!!!⁸

- ▶ Posit linear model $Q(x, a; \beta) = x_0^\top T + ax_1^\top T \beta$, indexed by $\beta = (\beta_0^\top T, \beta_1^\top T)^\top T$ and x_0, x_1 known features.
 - ▶ $\hat{\beta}_n = \arg \min_{\beta} \mathbb{P}_n \{Y - Q(X, A; \beta)\}^2$ and
 $\beta^* = \arg \min_{\beta} P \{Y - Q(X, A; \beta)\}$
 - ▶ Derive limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta^*)$
 - ▶ Construct confidence interval for $Q(x, a)$ assuming $Q(x, a) = Q(x, a; \beta^*)$

⁸He exclaimed.

Extra page if needed

Extra page if needed

Parameters in (1-stage) Q-learning are easy!

- ▶ Similar arguments show that coefficients indexing g -computation and outcome weighted learning are normal
- ▶ Preview: consider the (regression-based) estimator of the value of $\hat{\pi}_n(x) = \text{sign}(x_1^\top \hat{\beta}_{1,n})$, which you'll recall is

$$\begin{aligned}\hat{V}_n(\hat{\beta}_{1,n}) &= \mathbb{P}_n \max_a Q(X, a; \hat{\beta}_n) \\ &= \mathbb{P}_n X_0^\top \hat{\beta}_{0,n} + \mathbb{P}_n |X_1^\top \hat{\beta}_{1,n}|\end{aligned}$$

What is limit of $\sqrt{n} \left\{ \hat{V}_n(\hat{\beta}_{1,n}) - V(\hat{\beta}_{1,n}) \right\}$ and can we use it to derive CI for $V(\hat{\beta}_n)$? What about a CI for $V(\beta^*)$?

Parameters in (1-stage) OWL are easy!

- ▶ To illustrate, assume $P(A = 1|X) = P(A = -1|X) = 1/2$ wp1
- ▶ Recall OWL based on cvx relaxation of IPWE

$$\begin{aligned}\widehat{V}_n(\beta) &= \mathbb{P}_n \left[\frac{Y \mathbb{1} \{A = \text{sign}(X^\top \beta)\}}{P(A|X)} \right] \\ &= 2\mathbb{P}_n Y \mathbb{1} \{A \text{sign}(X^\top \beta) > 0\}\end{aligned}$$

- ▶ Let $\ell : \mathbb{R} \rightarrow \mathbb{R}$ be cvx, OWL estimator is

$$\widehat{\beta}_n = \arg \min_{\beta \in \mathbb{R}^p} \mathbb{P}_n |Y| \ell(W^\top \beta),$$

where $W = \text{sign}(Y)AX$

Extra page if needed

Some facts about OWL (and more generally convex M-estimators)

- ▶ $\mathbb{E}_n |y| \ell(w^\top \beta)$ is composition of linear and cvx function and thus cvx in β for each $(y, w) \Rightarrow$ greatly simplifies inference!
- ▶ Regularity conditions
 - ▶ $\beta^* = \arg \min_{\beta} P |Y| \ell(W^\top \beta)$ exists and unique
 - ▶ Map $\beta \mapsto P |Y| \ell(W^\top \beta)$ differentiable in nbrhd of β^* ⁹
 - ▶ Under these conditions, $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is regular¹⁰ and asymptotically normal \Rightarrow many results from Q-learning port

⁹More formally, require

$|y| \ell \left\{ w^\top (\beta^* + \delta) \right\} - |y| \ell(w^\top \beta^*) = S(y, w, ; \beta^*)^\top \delta + R(y, w, \delta; \beta^*)$ where $PS(Y, W; \beta^*) = 0$, $\Sigma_O = PS(Y, W; \beta^*) S(Y, W; \beta^*)^\top$ finite, and $PR(Y, W, \delta; \beta^*) = (1/2) \delta^\top \Omega_O \delta + o(\|\delta\|^2)$ (Haberman, 1989; Nemiro, 1992; Hjort and Pollard, 2011).

¹⁰I am being a bit loose with language in this course by referring to both estimands and rescaled estimators as 'regular' or 'non-regular.'

Value function(s)

- ▶ Three ways to measure performance
 - ▶ Conditional value: $V(\hat{\pi}_n) = PY^*(\hat{\pi}_n) = \mathbb{E} \{ Y^*(\hat{\pi}_n) | \hat{\pi}_n \}$, measures the performance of an estimated decision rule as if it were to be deployed in popn (note* this is a random variable)
 - ▶ Unconditional value: $\mathcal{V}_n = \mathbb{E}V(\hat{\pi}_n)$, measures the *average* performance of the algorithm used to construct $\hat{\pi}_n$ with sample of size n
 - ▶ Population-level value: $V(\pi^*)$, where $\pi^*(x) = \text{sign}(x^\top \beta^*)$, measures the potential of applying precision medicine strategy in given domain if algorithm for constructing $\hat{\pi}_n$ will be used
- ▶ Discuss these measures with your stat buddy. Is there a meaningful distinction as the sample size grows large?

It's a wacky world out there

- ▶ The three value measures need not coincide asymptotically
- ▶ Let $\hat{\pi}_n(x) = \text{sign}(x^\top \hat{\beta}_n)$ and suppose $\sqrt{n}\hat{\beta}_n \rightsquigarrow \text{Normal}(0, \Sigma)$ so that $\beta^* \equiv 0$ and $\pi^*(x) \equiv -1$. With stat buddy, compute:
 - ▶ $\hat{V}_n(\hat{\beta}_n) \rightsquigarrow \underline{\hspace{1cm}????\hspace{1cm}}$
 - ▶ $\mathcal{V}_n = \mathbb{E}V(\hat{\beta}_n) \rightarrow \underline{\hspace{1cm}????\hspace{1cm}}$
 - ▶ $V(\beta^*) = \underline{\hspace{1cm}????\hspace{1cm}}$

Calculon!

Extra page if needed

Have some confidence you useless pile!

- ▶ We'll construct confidence sets for $V(\hat{\beta}_n)$ and $V(\beta^*)$ as these are most commonly of interest in application
- ▶ Starting with conditional value fn assume that the data-generating model is a triangular array P_n such that:

$$(A0) \quad \hat{\pi}_n(x) = \text{sign}(x^\top \hat{\beta}_n)$$

$$(A1) \quad \exists \beta_n^* \text{ s.t. } \beta_n^* = \beta^* + s/\sqrt{n} \text{ for some } s \in \mathbb{R}^p \text{ and} \\ \sqrt{n}(\hat{\beta}_n - \beta_n^*) = \sqrt{n}(\mathbb{P}_n - P_n)u(X, A, Y) + o_{P_n}(1), \text{ where } u \\ \text{does not depend on } s, \sup_n P_n \|u(X, A, Y)\|^2 < \infty, \text{ and} \\ \text{Cov}\{u(X, A, Y)\} \text{ is p.d.}$$

$$(A2) \quad \text{If } \mathcal{F} \text{ is uniformly bounded Donsker class and } \sqrt{n}(pn - P) \rightsquigarrow \mathbb{T} \\ \text{in } \ell^\infty(\mathcal{F}) \text{ under } P \text{ then } \sqrt{n}(\mathbb{P}_n - P_n) \rightsquigarrow \mathbb{T} \text{ in } \ell^\infty(\mathcal{F}) \text{ under } P_n.$$

$$(A3) \quad \sup_n P_n \|Y\|^2 < \infty.$$

- ▶ Detailed discussion is beyond the scope of this class. Laber will wave his hands a bit. Our goal is understand key

Building block: joint distribution before nonsmooth operator

- Define class of functions

$$\mathcal{G} = \left\{ g(X, A, Y; \delta) = Y \mathbb{1} \left\{ AX^{\top T} \delta > 0 \right\} \mathbb{1} \left\{ X^{\top T} \beta^* = 0 \right\} : \delta \in \mathbb{R}^p \right\}$$

view $\sqrt{n}(\mathbb{P}_n - P_n)$ as random element of $\ell^\infty(\mathbb{R}^p)$.

Lemma

Assume (A0)-(A3). Then

$$\sqrt{n} \begin{bmatrix} \mathbb{P}_n - P_n \\ \hat{\beta}_n - \beta^* \\ (\mathbb{P}_n - P_n) Y \mathbb{1} \left\{ AX^{\top T} \beta^* > 0 \right\} \end{bmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{T} \\ \mathbb{Z} \\ \mathbb{W} \end{pmatrix}$$

in $\ell^\infty(\mathbb{R}^p) \times \mathbb{R}^p \times \mathbb{R}$ under P_n .

Limiting distn of $V(\hat{\beta}_n)$

Corollary

Assume (A0)-(A3). Then,

$$\sqrt{n} \left\{ \hat{V}_n(\hat{\beta}_n) - V(\hat{\beta}_n) \right\} \rightsquigarrow \mathbb{T}(\mathbb{Z} + s) + \mathbb{W}.$$

► Notes

- Presence of s shows this is nonregular
- \mathbb{T} is a Brownian bridge indexed by \mathbb{R}^p
- \mathbb{W} and \mathbb{Z} are normal

Hand-waving!

Extra page if needed

Bound-based confidence interval

- ▶ Limiting distribution: $\mathbb{T}(\mathbb{Z} + s) + \mathbb{W}$
 - ▶ Local parameter only appears in first term
 - ▶ (Asy) bound should only affect this term
- ▶ Schematic for constructing a bound
 - ▶ Partition input space into those that are 'near' the decision boundary $x^T \beta^* = 0$ vs. those that are 'far' from boundary
 - ▶ Take sup/inf over local perturbations of points in 'near' group

Upper bound

- Let $\widehat{\Sigma}_n$ be estimator of asy var of $\widehat{\beta}_n$ an upper bound on $\sqrt{n} \left\{ \widehat{V}_n(\widehat{\beta}_n) - V(\widehat{\beta}_n) \right\}$ is

$$U_n = \sup_{\omega \in \mathbb{R}^p} \sqrt{n}(\mathbb{P}_n - P_n)Y \mathbb{1} \left\{ AX^{\top T} \omega > 0 \right\} \mathbb{1} \left\{ \frac{n(X^{\top T} \widehat{\beta}_n)^2}{X^{\top T} \widehat{\Sigma}_n X} \leq \tau_n \right\} \\ + \sqrt{n}(\mathbb{P}_n - P_n)Y \mathbb{1} \left\{ AX^{\top T} \widehat{\beta}_n > 0 \right\} \mathbb{1} \left\{ \frac{n(X^{\top T} \widehat{\beta}_n)^2}{X^{\top T} \widehat{\Sigma}_n X} > \tau_n \right\},$$

where τ_n is seq of tuning parameters s.t. $\tau_n \rightarrow \infty$ and $\tau_n = o(n)$ as $n \rightarrow \infty$. Lower bound constructed by replacing sup with inf.

Limiting distribution of bounds

Theorem

Assume (A0)-(A3). Then

$$(L_n, U_n) \rightsquigarrow \left\{ \inf_{\omega \in \mathbb{R}^p} \mathbb{T}(\omega) + \mathbb{W}, \sup_{\omega \in \mathbb{R}^p} \mathbb{T}(\omega) + \mathbb{W} \right\}$$

under P_n .

- ▶ Recall limit distn of $\sqrt{n} \left\{ \widehat{V}_n(\widehat{\beta}_n) - V(\widehat{\beta}_n) \right\}$ is $\mathbb{T}(\mathbb{Z} + s) + \mathbb{W}$
 - ▶ Bounds equiv to sup/inf over local perturbations
 - ▶ If all subject have large txt effects, bound are tight
 - ▶ Bootstrap bounds to construct confidence bound, theoretical results for the bootstrap bounds given in book

Note on tuning

- ▶ Seq $\{\tau_n\}_{n \geq 1}$ can affect finite sample performance
 - ▶ Idea: tune using double bootstrap, i.e., bootstrap the bootstrap samples to estimate coverage and adapt τ_n
 - ▶ Double bootstrap considered computationally expensive but not much of a burden in most problems with modern computing infrastructure
 - ▶ Tuning can be done without affecting theoretical results

Algy the friendly tuning algorithm

Input: $\{(X_i, A_i, Y_i)\}_{i=1}^n, M, \alpha \in (0, 1), \{\tau_n^{(1)}, \dots, \tau_n^{(L)}\}$

```
1  $\mathcal{V} = \hat{\mathcal{V}}_n(\hat{d}_n)$ 
2 for  $j = 1, \dots, L$  do
3    $c^{(j)} = 0$ 
4   for  $b = 1, \dots, M$  do
5     Draw a sample of size  $n$ , say  $S_n^{(b)}$ , from  $\{(X_i, A_i, Y_i)\}_{i=1}^n$ 
      with replacement
6     Compute bound-based confidence set,  $\zeta_{m_n}^{(b)}$ , using sample  $S_n^{(b)}$ 
      and critical value  $\tau_n^{(j)}$ 
7     if  $\mathcal{V} \in \zeta_{m_n}^{(b)}$  then
8        $c^{(j)} = c^{(j)} + 1$ 
9     end
10  end
11 end
12 Set  $j^* = \arg \min_{j: c^{(j)} \geq M(1-\alpha)} c^{(j)}$ 
Output: Return  $\tau_n^{(j^*)}$ 
```

Intermission

If any man says he hates war more than I do, he better have a knife, that's all I have to say. –Ghandi

m-out-of-*n* bootstrap

- ▶ Bound-based intervals complex (conceptually and technically)
 - ▶ Subsampling easier to implement and understand¹¹
 - ▶ Does not require specialized code etc.
- ▶ Let $m_n = o(n)$ be resample size s.t. $m_n \rightarrow \infty$ and $m_n = o(n)$, let $\mathbb{P}_{m_n}^{(b)}$ be bootstrap empirical distn
 - ▶ Approximate $\sqrt{n} \left\{ \hat{V}_n(\hat{\beta}_n) - V(\hat{\beta}_n) \right\}$ with its bootstrap analog $\sqrt{m_n} \left\{ \hat{V}_{m_n}^{(b)}(\hat{\beta}_{m_n}^{(b)}) - \hat{V}_n(\hat{\beta}_n) \right\}$ (Laber might draw picture)
 - ▶ Let $\hat{\ell}_{m_n}$ and \hat{u}_{m_n} be the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of $\sqrt{m_n} \left\{ \hat{V}_{m_n}^{(b)}(\hat{\beta}_{m_n}^{(b)}) - \hat{V}_n(\hat{\beta}_n) \right\}$, ci given by

$$\left[\hat{V}_n(\hat{\beta}_n) - \hat{u}_{m_n}/\sqrt{m_n}, \hat{V}_n(\hat{\beta}_n) + \hat{\ell}_{m_n}/\sqrt{m_n} \right]$$

¹¹Though the theory underpinning subsampling can be non-trivial so 'understand' here is meant more mechanically.

Emmy the subsampling bootstrap algo

Input: $m_n, \{X_i, A_i, Y_i\}_{i=1}^n, M, \alpha \in (0, 1)$

- 1 **for** $b = 1, \dots, M$ **do**
- 2 Draw a sample of size m_n , say $S_{m_n}^{(b)}$, from $\{X_i, A_i, Y_i\}_{i=1}^n$ with replacement
- 3 Compute $\hat{\beta}_{m_n}^{(b)}$ on $S_{m_n}^{(b)}$
- 4 $\Delta_{m_n}^{(b)} = \sqrt{m_n} \left[\sum_{i \in S_{m_n}^{(b)}} Y_i \mathbf{I} \left\{ A_i X_i^T \hat{\beta}_{m_n}^{(b)} > 0 \right\} - \sum_{k=1}^n Y_k \mathbf{I} \left\{ A_k X_k^T \hat{\beta}_{m_n}^{(b)} > 0 \right\} \right]$
- 5 **end**
- 6 Relabel so that $\Delta_{m_n}^{(1)} \leq \Delta_{m_n}^{(2)} \leq \dots \leq \Delta_{m_n}^{(B)}$
- 7 $\hat{\ell}_{m_n} = \Delta_{m_n}^{(\lfloor B\alpha/2 \rfloor)}$
- 8 $\hat{u}_{m_n} = \Delta_{m_n}^{(\lfloor B(1-\alpha/2) \rfloor)}$

Output: $\left[\hat{V}_n(\hat{d}_n) - \hat{u}_{m_n}/\sqrt{m_n}, \hat{V}_n(\hat{d}_n) - \hat{\ell}_{m_n}/\sqrt{m_n} \right]$

m -out-of- n cont'd

- ▶ Provides valid confidence intervals under (A0)-(A3)
 - ▶ Proof omitted (tedious)
 - ▶ Can tune m_n using double bootstrap
 - ▶ Reliance on asymptotic tomfoolery makes me hesitant to use this in practice¹²

¹²I did not always think this way, see Chakraborty, L., and Zhao (2014ab). Also, I should not be so dismissive of these methods. Some of the work in this area has been quite deep and produced general uniformly convergent methods. See work by Romano and colleagues.

Confidence interval for opt regime within a class

- ▶ Let π^* denote the optimal txt regime within a given class, our goal is to construct a CI for $V(\pi^*)$
 - ▶ Were π^* known, one could use $\sqrt{n} \left\{ \hat{V}_n(\pi^*) - V(\pi^*) \right\}$, with your stats buddy, compute this limiting distn
 - ▶ Suppose that we could construct a valid confidence region for π^* , suggest a method for CI for $V(\pi^*)$

Projection interval

- ▶ For any fixed π , let $\zeta_{n,1-\nu}(\pi)$ be a $(1 - \nu) \times 100\%$ confidence set for $V(\pi)$, e.g., using asymptotic approx on previous slide
- ▶ Let $\mathcal{D}_{n,1-\eta}$ denote a $(1 - \eta) \times 100\%$ confidence set for π^* , then a $(1 - \eta - \nu) \times 100\%$ confidence region for $V(\pi^*)$ is

$$\bigcup_{\pi \in \mathcal{D}_{n,1-\eta}} \zeta_{n,1-\nu}(\pi)$$

Why?

Ex. projection interval for linear regime

- Consider regimes of the form $\pi(x; \beta) = \text{sign}(x^\top \beta)$, then

$$\begin{aligned}\sqrt{n} \left\{ \widehat{V}_n(\beta) - V(\beta) \right\} &= \sqrt{n} (\mathbb{P}_n - P) \frac{Y 1_{AX^\top \beta > 0}}{P(A|X)} \\ &\rightsquigarrow \text{Normal} \{0, \sigma^2(\beta)\},\end{aligned}$$

take

$$\zeta_{n,1-\nu}(\beta) = \left[\widehat{V}_n(\beta) - \frac{z_{1-\nu/2} \widehat{\sigma}_n(\beta)}{\sqrt{n}}, \widehat{V}_n(\beta) + \frac{z_{1-\nu/2} \widehat{\sigma}_n(\beta)}{\sqrt{n}} \right]$$

- If $\sqrt{n}(\widehat{\beta}_n - \beta^*) \rightsquigarrow \text{Normal}(0, \Sigma)$, take

$$\mathcal{D}_{n,1-\eta} = \left\{ \beta : n(\widehat{\beta}_n - \beta)^\top \widehat{\Sigma}_n^{-1} (\widehat{\beta}_n - \beta) \leq \chi_{p,1-\eta}^2 \right\}$$

- Projection interval:
$$\bigcup_{\beta \in \mathcal{D}_{n,1-\eta}} \zeta_{n,1-\nu}(\beta)$$

Quiz break!

- ▶ What does the 'Q' in Q-learning stand for?
- ▶ In txt regimes, which of the following is not yet a thing: A-learning, B-learning, C-Learning, D-learning, E-learning?
- ▶ Write down the two-stage Q-learning algorithm assuming binary treatments and linear models at each stage
- ▶ True or false:
 - ▶ I would rather have a zombie ice dragon than two live fire dragons.
 - ▶ The story of the Easter bunny is based on the little known story of Jesus swapping the internal organs of chickens and rabbits to prevent a widespread famine
 - ▶ Q-learning has been used to obtain state-of-the-art performance in game-playing domains like chess, backgammon, and atari

Inference for two-stage linear Q-learning

- ▶ Learning objectives
 - ▶ Identify source of nonregularity
 - ▶ Understand implications on coverage and asy bias
 - ▶ Intuition behind bounds
- ▶ Hopefully this will be trivial for you now!

Reminder: setup and notation

- ▶ Observe $\{(X_{1,i}, A_{1,i}, X_{2,i}, A_{2,i}, Y_i)\}_{i=1}^n$, *i.i.d.* from P
 - ▶ $X_1 \in \mathbb{R}^{p_1}$: baseline subj. info.
 - ▶ $A_1 \in \{0, 1\}$: first treatment
 - ▶ $X_2 \in \mathbb{R}^{p_2}$: interim subj. info. during course of A_1
 - ▶ $A_2 \in \{0, 1\}$: second treatment
 - ▶ $Y \in \mathbb{R}$: outcome, higher is better
- ▶ Define history $H_1 = X_1$, $H_2 = (X_1, A_1, X_2)$
- ▶ DTR $\pi = (\pi_1, \pi_2)$ where

$$\pi_t : \text{supp } H_t \rightarrow \text{supp } A_t,$$

patient presenting with $H_t = h_t$ assigned treatment $\pi_t(h_t)$

Characterizing optimal DTR

- ▶ Optimal regime maximizes value $\mathbb{E} Y^*(\pi)$
- ▶ Define Q -functions

$$Q_2(h_2, a_2) = \mathbb{E}(Y | H_2 = h_2, A_2 = a_2)$$

$$Q_1(h_1, a_1) = \mathbb{E} \left\{ \max_{a_2} Q_2(H_2, a_2) | H_1 = h_1, A_1 = a_1 \right\}$$

- ▶ Dynamic programming (Bellman, 1957)
 $\pi_t^{\text{opt}}(h_t) = \arg \max_{a_t} Q_t(h_t, a_t)$

Q-learning

- ▶ Regression-based dynamic programming algorithm

(Q0) Postulate working models for Q -functions

$$Q_t(h_t, a_t; \beta_t) = h_{t,0}^\top \beta_{t,0} + a_t h_{t,1}^\top \beta_{t,1}, \quad h_{t,0}, h_{t,1} \text{ features of } h_t$$

(Q1) Compute $\hat{\beta}_2 = \arg \min_{\beta_2} \mathbb{P}_n \{Y - Q_2(H_2, A_2; \beta_2)\}^2$

(Q2) Compute

$$\hat{\beta}_1 = \arg \min_{\beta_1} \mathbb{P}_n \left\{ \max_{a_2} Q_2(H_2, A_2; \hat{\beta}_2) - Q_1(H_1, A_1; \beta_1) \right\}^2$$

(Q3) $\hat{\pi}_t(h_t) = \arg \max_{a_t} Q_t(h_t, a_t; \hat{\beta}_t)$

- ▶ Population parameters β_t^* obtained by replacing \mathbb{P}_n with P

- ▶ Inference for β_2^* standard, just OLS

- ▶ Focus on confidence intervals for $c^\top \beta_1^*$ for fixed $c \in \mathbb{R}^{dim, \beta_1^*}$

Q-learning

- ▶ Regression-based dynamic programming algorithm

(Q0) Postulate working models for Q -functions

$$Q_t(h_t, a_t; \beta_t) = h_{t,0}^\top \beta_{t,0} + a_t h_{t,1}^\top \beta_{t,1}, \quad h_{t,0}, h_{t,1} \text{ features of } h_t$$

(Q1) Compute $\hat{\beta}_2 = \arg \min_{\beta_2} \mathbb{P}_n \{Y - Q_2(H_2, A_2; \beta_2)\}^2$

(Q2) Compute

$$\hat{\beta}_1 = \arg \min_{\beta_1} \mathbb{P}_n \left\{ \max_{a_2} Q_2(H_2, A_2; \hat{\beta}_2) - Q_1(H_1, A_1; \beta_1) \right\}^2$$

(Q3) $\hat{\pi}_t(h_t) = \arg \max_{a_t} Q_t(h_t, a_t; \hat{\beta}_t)$

- ▶ Population parameters β_t^* obtained by replacing \mathbb{P}_n with P

- ▶ Inference for β_2^* standard, just OLS

- ▶ Focus on confidence intervals for $c^\top \beta_1^*$ for fixed $c \in \mathbb{R}^{dim, \beta_1^*}$

Inference for $c^T \beta_1^*$

- ▶ Non-smooth max operator makes $\hat{\beta}_1$ non-regular
 - ▶ Dstn of $c^T \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ sensitive to small perturbations of P
 - ▶ Limiting dstn does not have mean zero (asymptotic bias)
 - ▶ Occurs with small second stage txt effects, $H_{2,1}^T \beta_{2,1}^* \approx 0$
- ▶ Confidence intervals based on series approximations or bootstrap can perform poorly; proposed remedies include:
 - ▶ Apply shrinkage to reduce asymptotic bias
 - ▶ Form conservative estimates to tail probabilities of $c^T \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$

Characterizing asymptotic bias

Definition

For constant $c \in \mathbb{R}^{\dim \beta_1^*}$ and \sqrt{n} -consistent estimator $\tilde{\beta}_1$ of β_1^* with $\sqrt{n}(\tilde{\beta}_1 - \beta_1^*) \rightsquigarrow \mathbb{M}$, define the c -directional asymptotic bias

$$\text{Bias}(\tilde{\beta}_1, c) \triangleq \mathbb{E} c^\top \mathbb{M}.$$

Characterizing asymptotic bias cont'd

Theorem (Asymptotic bias Q-learning)

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ be fixed. Under moment conditions:

$$\text{Bias}(\hat{\beta}_1, c) = \frac{c^\top \Sigma_{1,\infty}^{-1} P \left(B_1 \sqrt{H_{2,1}^\top \Sigma_{21,21} H_{2,1}} 1_{H_{2,1}^\top \beta_{2,1}^* = 0} \right)}{\sqrt{2\pi}},$$

where $B_1 = (H_{1,0}^\top, A_1 H_{1,1}^\top)^\top$, $\Sigma_{1,\infty} = P B_1 B_1^\top$, and $\Sigma_{21,21}$ is the asy. cov. of $\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$.

Characterizing asymptotic bias cont'd

Theorem (Asymptotic bias Q-learning)

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ be fixed. Under moment conditions:

$$\text{Bias}(\hat{\beta}_1, c) = \frac{c^\top \Sigma_{1,\infty}^{-1} P \left(B_1 \sqrt{H_{2,1}^\top \Sigma_{21,21} H_{2,1}} 1_{H_{2,1}^\top \beta_{2,1}^* = 0} \right)}{\sqrt{2\pi}},$$


where $B_1 = (H_{1,0}^\top, A_1 H_{1,1}^\top)^\top$, $\Sigma_{1,\infty} = P B_1 B_1^\top$, and $\Sigma_{21,21}$ is the asy. cov. of $\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$.

► Asymptotic bias for Q-learning

- Ave. of $c^\top \Sigma_{1,\infty} B_1$ with wts $\propto \sqrt{\text{Var} \left(H_{2,1}^\top \hat{\beta}_{2,1} 1_{H_{2,1}^\top \beta_{2,1}^* = 0} \mid H_{2,1} \right)}$
- May be reduced by shrinking $h_{2,1}^\top \hat{\beta}_{2,1}$ when $h_{2,1}^\top \beta_{2,1}^* = 0$


Reducing asymptotic bias to improve inference

- ▶ Shrinkage is a popular method for reducing asymptotic bias with goal of improving interval coverage
 - ▶ Chakraborty et al. (2009) apply soft-thresholding
 - ▶ Moodie et al. (2010) apply hard-thresholding
 - ▶ Goldberg et al. (2013) and Song et al.(2015) use lasso-type penalization
- ▶ Shrinkage methods target

$$\max_{a_2} Q_2 \left(h_2, a_2; \hat{\beta}_2 \right) = h_{2,0}^T \hat{\beta}_{2,0} + \max_{a_2 \in \{0,1\}} a_2 h_{2,1}^T \hat{\beta}_{2,1}$$


Reducing asymptotic bias to improve inference

- ▶ Shrinkage is a popular method for reducing asymptotic bias with goal of improving interval coverage
 - ▶ Chakraborty et al. (2009) apply soft-thresholding
 - ▶ Moodie et al. (2010) apply hard-thresholding
 - ▶ Goldberg et al. (2013) and Song et al.(2014) use lasso-type penalization
- ▶ Shrinkage methods target


$$\max_{a_2} Q_2 \left(h_2, a_2; \hat{\beta}_2 \right) = h_{2,0}^T \hat{\beta}_{2,0} + \left[h_{2,1}^T \hat{\beta}_{2,1} \right]_+$$

Soft-thresholding (Chakraborty et al., 2009)

- ▶ In Q-learning, replace $\max_{a_2} Q_2(H_2, A_2; \hat{\beta}_2)$ with

$$H_{2,0}^\top \hat{\beta}_{2,0} + \left[H_{2,1}^\top \hat{\beta}_{2,1} \right]_+ \left\{ 1 - \frac{\sigma H_{2,1}^\top \hat{\Sigma}_{21,21} H_{2,1}}{n \left(H_{2,1}^\top \hat{\beta}_{2,1} \right)^2} \right\}_+$$

- ▶ Amount of shrinkage governed by $\sigma > 0$
- ▶ Penalization schemes (Goldberg et al., 2013, Song et al., 2014) reduce to this estimator under certain designs
- ▶ No theoretical justification in Chakraborty et al. (2009) but improved coverage of bootstrap intervals in some settings

Soft-thresholding and asymptotic bias

Theorem

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ and let $\hat{\beta}_1^\sigma$ denote the soft-thresholding estimator. Under moment conditions:

1. $|\text{Bias}(\hat{\beta}_1^\sigma, c)| \leq |\text{Bias}(\hat{\beta}_1, c)|$ for any $\sigma > 0$.
2. If $\text{Bias}(\hat{\beta}_1, c) \neq 0$, then for $\sigma > 0$

$$\frac{\text{Bias}(\hat{\beta}_1^\sigma, c)}{\text{Bias}(\hat{\beta}_1, c)} = \exp\left(-\frac{\sigma}{2}\right) - \sigma \int_{\sqrt{\sigma}}^{\infty} \frac{1}{x} \exp\left(-\frac{x^2}{2}\right) dx$$

Soft-thresholding and asymptotic bias cont'd

- ▶ Is thresholding useful in reducing asymptotic bias?
 - ▶ Preceding theorem says yes, and more shrinkage is better
 - ▶ Chakraborty et al. suggest $\sigma = 3$, which corresponds to 13-fold decrease in asymptotic bias
 - ▶ However, the preceding theorem is based on pointwise, i.e., fixed parameter, asymptotics and may not faithfully reflect small sample performance

Local generative model

- ▶ Use local asymptotics approximate small sample behavior of soft-thresholding
- ▶ Assume:
 1. For any $s \in \mathbb{R}^{dim \beta_{2,1}^*}$ there exists sequence of distributions P_n so that

$$\int \left\{ \sqrt{n} \left(dP_n^{1/2} - dP^{1/2} \right) - \frac{1}{2} \nu_s dP^{1/2} \right\}^2 \rightarrow 0,$$

for some measurable function ν_s .

2. $\beta_{2,1,n}^* = \beta_{2,1}^* + s/\sqrt{n}$, where
$$\beta_{2,n}^* = \arg \min_{\beta_2} P_n \{ Y - Q_2(H_2, A_2; \beta_2) \}^2$$

Local asymptotics view of soft-thresholding

Theorem

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ be fixed. Under the local generative model and moment conditions:

1. $\sup_{s \in \mathbb{R}^{\dim \beta_{2,1}^*}} |\text{Bias}(\hat{\beta}_1, c)| \leq K < \infty.$
2. $\sup_{s \in \mathbb{R}^{\dim \beta_{2,1}^*}} |\text{Bias}(\hat{\beta}_1^\sigma, c)| \rightarrow \infty$ as $\sigma \rightarrow \infty.$

Local asymptotics view of soft-thresholding

Theorem

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ be fixed. Under the local generative model and moment conditions:

1. $\sup_{s \in \mathbb{R}^{\dim \beta_{2,1}^*}} |\text{Bias}(\hat{\beta}_1, c)| \leq K < \infty.$
2. $\sup_{s \in \mathbb{R}^{\dim \beta_{2,1}^*}} |\text{Bias}(\hat{\beta}_1^\sigma, c)| \rightarrow \infty$ as $\sigma \rightarrow \infty.$

- Thresholding can be infinitely worse than doing nothing if done too aggressively in small samples

Data-driven tuning

- ▶ Is it possible to construct a data-driven choice of σ that consistently leads to less asymptotic bias than no shrinkage?
- ▶ Consider data from a two-arm randomized trial $\{(A_i, Y_i)\}_{i=1}^n$, $A \in \{0, 1\}$, $Y \in \mathbb{R}$ coded so that higher is better¹³
 - ▶ Define $\mu_a^* \triangleq \mathbb{E}(Y|A=a)$ and $\hat{\mu}_a = \mathbb{P}_n Y 1_{A=a} / \mathbb{P}_n 1_{A=a}$
 - ▶ Mean outcome under optimal treatment assignment $\theta^* = \max(\mu_0^*, \mu_1^*)$, corresponding estimator

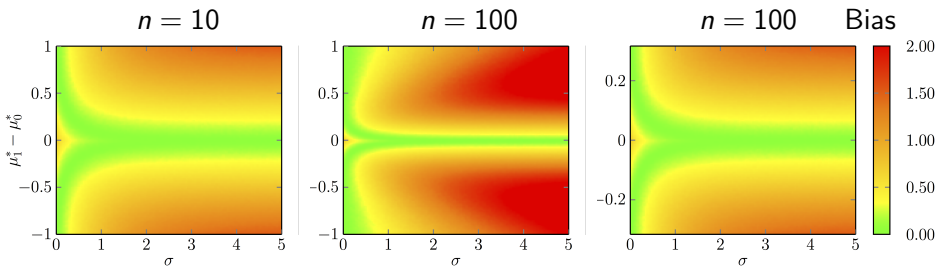
$$\hat{\theta} = \max(\hat{\mu}_0, \hat{\mu}_1) = \hat{\mu}_0 + [\hat{\mu}_1 - \hat{\mu}_0]_+$$

- ▶ Soft-thresholding estimator

$$\hat{\theta}^\sigma = \hat{\mu}_0 + [\hat{\mu}_1 - \hat{\mu}_0]_+ \left\{ 1 - \frac{4\sigma}{n(\hat{\mu}_1 - \hat{\mu}_0)^2} \right\}_+$$

¹³This is equivalent to two-stage Q-learning with no covariates and a single first stage treatment.

Data-driven tuning: toy example



- ▶ Optimal value of σ depends on $\mu_1^* - \mu_0^*$
 - ▶ Variability in $\hat{\mu}_1 - \hat{\mu}_0$ prevents identification of optimal value of σ , using plug-in estimator may lead to large bias
 - ▶ Data-driven σ that significantly improves asymptotic bias over no shrinkage is difficult

Asymptotic bias: discussion

- ▶ Asymptotic bias exists in Q -learning
- ▶ Local asymptotics show that aggressively shrinking to reduce asymptotic bias can be infinitely worse than no shrinkage
- ▶ Data-driven tuning seems to require choosing σ very small or risking large bias

Confidence intervals for $c^\top \beta_1^*$

- ▶ Possible to construct valid confidence intervals in presence of asymptotic bias
- ▶ Idea: construct regular bounds on $c^\top \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$
 - ▶ Bootstrap bounds to form confidence interval
 - ▶ Tightest among all regular bounds \Rightarrow automatic adaptivity
 - ▶ Local uniform convergence
 - ▶ Can also obtain conditional properties (Robins and Rotnitzky, 2014) and global uniform convergence (Wu, 2014)

Regular bounds on $c^{\top T} \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$

► Define

$$\mathbb{V}_n(c, \gamma) = c^{\top T} \mathbb{S}_n + c^{\top T} \hat{\Sigma}_1^{-1} \mathbb{P}_n \mathbb{U}_n(\gamma),$$

where

$$\begin{aligned} \mathbb{S}_n &= \hat{\Sigma}_1^{-1} \sqrt{n} (\mathbb{P}_n - P) B_1 \left\{ H_{2,0}^{\top T} \beta_{2,0}^* + \left(H_{2,1}^{\top T} \beta_{2,1}^* \right)_+ - B_1^{\top T} \beta_1^* \right\} \\ &\quad + \hat{\Sigma}_1^{-1} \sqrt{n} \mathbb{P}_n H_{2,0}^{\top T} \left(\hat{\beta}_{2,0} - \beta_{2,0}^* \right), \end{aligned}$$

$$\mathbb{U}_n(\gamma) = B_1 \left[\left\{ H_{2,1}^{\top T} (\mathbb{Z}_n + \gamma) \right\}_+ - \left(H_{2,1}^{\top T} \gamma \right)_+ \right]$$

► \mathbb{S}_n is smooth and $\mathbb{U}_n(\gamma)$ is non-smooth

Regular bounds on $c^\top \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ cont'd

- ▶ It can be shown that $c^\top \sqrt{n}(\hat{\beta}_1 - \beta_1^*) = \mathbb{V}_n(c, \beta_{2,1}^*)$
- ▶ Use pretesting to construct upper bound

$$\begin{aligned}\mathcal{U}_n(c) = c^\top \mathbb{S}_n + c^\top \hat{\Sigma}_1^{-1} \mathbb{P}_n \mathbb{U}_n(\beta_{2,1}^*) 1_{T_n(H_{2,1}) > \lambda_n} \\ + \sup_{\gamma} c^\top \hat{\Sigma}_1^{-1} \mathbb{P}_n \mathbb{U}_n(\gamma) 1_{T_n(H_{2,1}) \leq \lambda_n},\end{aligned}$$

where $T_n(h_{2,1})$ test statistic for null $h_{2,1}^\top \beta_{2,1}^* = 0$ and λ_n is a critical value

- ▶ Lower bound, $\mathcal{L}_n(c)$ obtained by taking inf
- ▶ Bootstrap bounds to find confidence interval

Validity of the bounds

Theorem

Let $c \in \mathbb{R}^{\dim \beta_1^*}$ be fixed. Assume the local generative model, under moment conditions and conditions on the pretest:

1. $c^T \sqrt{n}(\hat{\beta}_1 - \beta_{1,n}^*) \rightsquigarrow c^T \mathbb{S}_\infty + c^T \Sigma_{1,\infty}^{-1} P B_1 H_{2,1}^T \mathbb{Z}_\infty 1_{H_{2,1}^T \beta_{2,1}^* > 0}$
 $+ c^T \Sigma_{1,\infty}^{-1} P B_1 \left[\left\{ H_{2,1}^T (\mathbb{Z}_\infty + s) \right\}_+ - \left(H_{2,1}^T s \right)_+ \right] 1_{H_{2,1}^T \beta_{2,1}^* = 0}$
2. $\mathcal{U}_n(c) \rightsquigarrow c^T \mathbb{S}_\infty + c^T \Sigma_{1,\infty}^{-1} P B_1 H_{2,1}^T \mathbb{Z}_\infty 1_{H_{2,1}^T \beta_{2,1}^* > 0}$
 $+ \sup_{\gamma} c^T \Sigma_{1,\infty}^{-1} P B_1 \left[\left\{ H_{2,1}^T (\mathbb{Z}_\infty + \gamma) \right\}_+ - \left(H_{2,1}^T \gamma \right)_+ \right] 1_{H_{2,1}^T \beta_{2,1}^* = 0}$

Validity of the bootstrap bounds

Theorem

Fix $\alpha \in (0, 1)$ and $c \in \mathbb{R}^{\dim \beta_1^*}$. Let $\hat{\ell}$ and \hat{u} denote the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of bootstrap distribution of the bounds and let P_M denote the distribution with respect to bootstrap weights. Under moment conditions and conditions on the pretest for any $\epsilon > 0$:

$$P \left\{ P_M \left(c^T \hat{\beta}_1 - \frac{\hat{u}}{\sqrt{n}} \leq c^T \beta_1^* \leq c^T \hat{\beta}_1 - \frac{\hat{\ell}}{\sqrt{n}} \right) < 1 - \alpha - \epsilon \right\} = o(1).$$

Uniform validity of the bootstrap bounds (Tianshuang Wu)

Theorem

Fix $\alpha \in (0, 1)$ and $c \in \mathbb{R}^{\dim \beta_1^*}$. Let $\hat{\ell}$ and \hat{u} denote the $(\alpha/2) \times 100$ and $(1 - \alpha/2) \times 100$ percentiles of bootstrap distribution of the bounds and let P_M denote the distribution with respect to bootstrap weights. Under moment conditions and conditions on the pretest for any $\epsilon > 0$:

$$\inf_{P \in \mathcal{P}} P \left\{ P_M \left(c^\top \hat{\beta}_1 - \frac{\hat{u}}{\sqrt{n}} \leq c^\top \beta_1^* \leq c^\top \hat{\beta}_1 - \frac{\hat{\ell}}{\sqrt{n}} \right) < 1 - \alpha - \epsilon \right\},$$

converges to zero for a large class of distributions \mathcal{P} .

Simulation experiments

- ▶ Class of generative models
 - ▶ $X_t \in \{-1, 1\}$, $A_t \in \{-1, 1\}$, $t \in \{1, 2\}$
 - ▶ $P(A_t = 1) = P(A_t = -1) = 0.5$, $t \in \{1, 2\}$
 - ▶ $X_1 \sim \text{Bernoulli}(0.5)$
 - ▶ $X_2 | X_1, A_1 \sim \text{Bernoulli} \{ \text{expit}(\delta_1 X_1 + \delta_2 A_1) \}$
 - ▶ $\epsilon \sim N(0, 1)$
 - ▶ $Y = \gamma_1 + \gamma_2 X_1 + \gamma_3 A_1 + \gamma_4 X_1 A_1 + \gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 + \epsilon$
- ▶ Vary parameters to obtain range of effects sizes, classify generative models as
 - ▶ Non-regular (NR)
 - ▶ Nearly non-regular (NNR)
 - ▶ Regular (R)

Simulation experiments cont'd

- ▶ Compare bounding confidence interval (ACI) with bootstrap (BOOT) and bootstrap thresholding (THRESH)
 - ▶ Compare in terms of width and coverage (target 95%)
 - ▶ Results based on 1000 Monte Carlo replications with datasets of size $n = 150$
 - ▶ Bootstrap computed with 1000 resamples
 - ▶ Tuning parameter λ_n chosen with double bootstrap

Simulation experiments: results

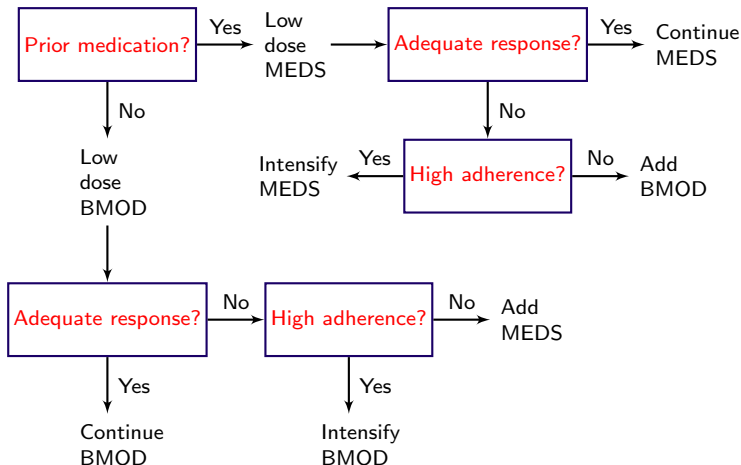
Coverage (target 95%)

Method	Ex. 1 NNR	Ex. 2 NR	Ex. 3 NNR	Ex. 4 R	Ex. 5 NR	Ex. 6 NNR
BOOT	0.935*	0.930*	0.933*	0.928*	0.925*	0.928*
THRESH	0.945	0.938	0.942	0.943	0.759*	0.762*
ACI	0.971	0.958	0.961	0.943	0.953	0.953

Average width

Method	Ex. 1 NNR	Ex. 2 NR	Ex. 3 NNR	Ex. 4 R	Ex. 5 NR	Ex. 6 NNR
BOOT	0.385*	0.430*	0.430*	0.436*	0.428*	0.428*
THRESH	0.339	0.426	0.427	0.436	0.426*	0.424*
ACI	0.441	0.470	0.470	0.469	0.473	0.473

Ex. DTR for ADHD without uncertainty



Ex. DTR for ADHD with uncertainty

