Distribution-free inference for regression: discrete, continuous, and in between

Rina Foygel Barber (joint work with Yonghoon Lee)

http://www.stat.uchicago.edu/~rina/
Thanks to support from NSF & ONR

Binary regression problem

Setting:

- Features $X \in \mathbb{R}^d$, response $Y \in \{0, 1\}$
- Unknown distribution $P = P_X \times \pi_P$

Marginal distrib. of X

$\pi_P(x) = \mathbb{P}\{Y = 1 \mid X = x\}$

Binary regression problem

Setting:

- Features $X \in \mathbb{R}^d$, response $Y \in \{0, 1\}$
- Unknown distribution $P = P_X \times \pi_P$

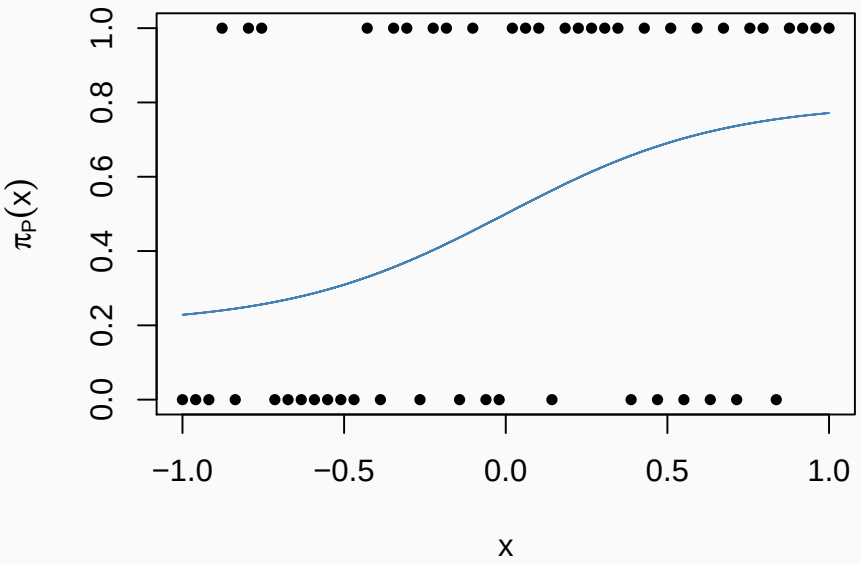
  P_X: Marginal distrib. of X

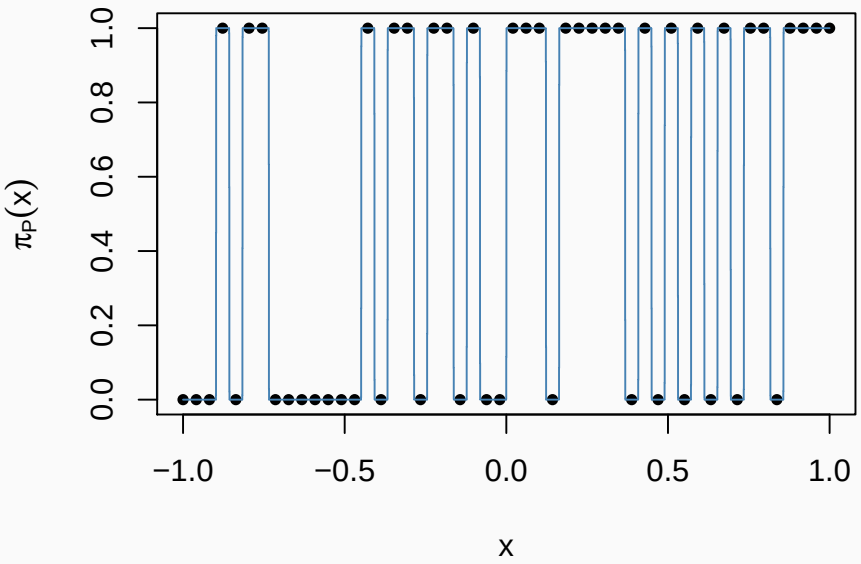
  $\pi_P(x) = \mathbb{P}\{Y = 1 \mid X = x\}$

- Training data $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P$
- Can we construct a confidence interval for $\pi_P(x)$, with no assumptions on P?

Background: smoothness & adaptivity

The challenge: data drawn from a smooth distribution
is also consistent with a highly non-smooth distribution





Background: smoothness & adaptivity

Extensive literature in nonparametric inference & adaptivity:

Low 1997, Györfi et al 2006, Genovese & Wasserman 2008,
Cai et al 2014, Hall & Horowitz 2013, Carpentier 2015,
Szabó et al 2015, Picard & Tribouley 2000, Hoffmann & Nickl 2011,
Giné & Nickl 2010, Giné & Nickl 2016, Bull & Nickl 2013,
Wahba 1983, Li 1989, Cai & Low 2006, ...

Background: smoothness & adaptivity

Extensive literature in nonparametric inference & adaptivity:

A partial summary...

If $\pi_P(x)$ is β-Hölder smooth with β known...

- Conf. int. width $\asymp n^{-\frac{\beta}{2\beta+d}}$ (e.g., via k-NN with $k \asymp n^{\frac{2\beta}{2\beta+d}}$)

Background: smoothness & adaptivity

Extensive literature in nonparametric inference & adaptivity:

A partial summary...

If $\pi_P(x)$ is β-Hölder smooth with β known...

- Conf. int. width $\asymp n^{-\frac{\beta}{2\beta+d}}$ (e.g., via k-NN with $k \asymp n^{\frac{2\beta}{2\beta+d}}$)

If $\pi_P(x)$ is β-Hölder smooth with $\beta \in [a, b]$ and $b \leq 2a$...

- Relax definition of coverage—cover $\pi(x)$ for "most" x
- $\rightsquigarrow$ Conf. int. width $\asymp n^{-\frac{\beta}{2\beta+d}}$ (adapts to β)

Background: smoothness & adaptivity

Extensive literature in nonparametric inference & adaptivity:

A partial summary...

If $\pi_P(x)$ is β-Hölder smooth with β known...

- Conf. int. width $\asymp n^{-\frac{\beta}{2\beta+d}}$ (e.g., via k-NN with $k \asymp n^{\frac{2\beta}{2\beta+d}}$)

If $\pi_P(x)$ is β-Hölder smooth with $\beta \in [a, b]$ and $b \leq 2a$...

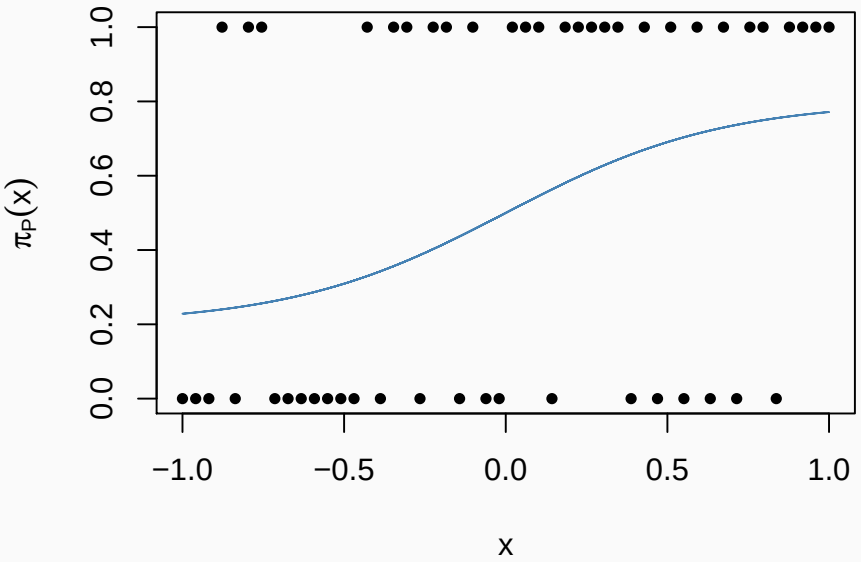
- Relax definition of coverage—cover $\pi(x)$ for "most" x
- $\rightsquigarrow$ Conf. int. width $\asymp n^{-\frac{\beta}{2\beta+d}}$ (adapts to β)

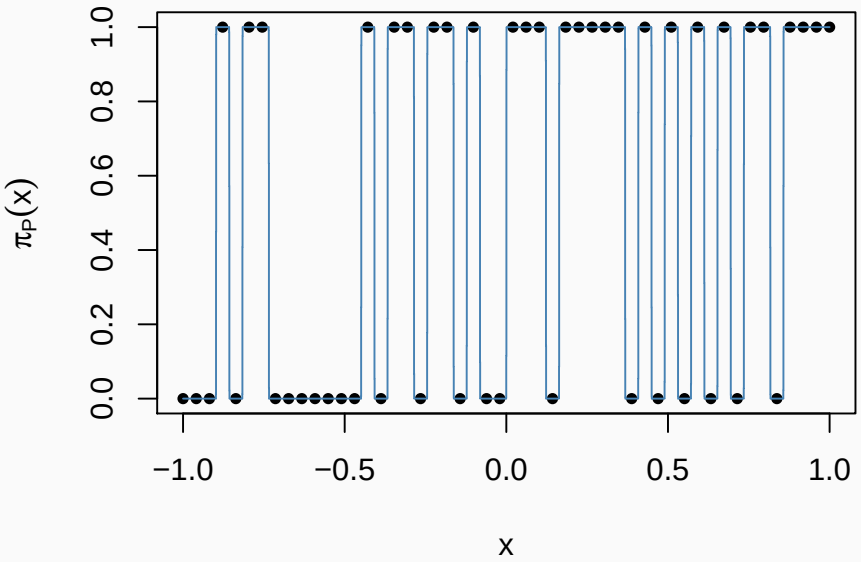
If $\pi_P(x)$ is β-Hölder smooth with $\beta \in [a, b]$ and $b > 2a$...

- Adaptivity is impossible...
- ...unless assume stronger conditions to exclude difficult cases

Background: smoothness & adaptivity

The challenge: data drawn from a smooth distribution
is also consistent with a highly non-smooth distribution,
and it's impossible to estimate the smoothness of P





Distribution-free coverage

Training data $\{(X_i, Y_i)\}_{i=1,\ldots,n}$ $\rightsquigarrow$ algorithm $\widehat{C}_n$ $\rightsquigarrow$ Conf. intervals $\{\widehat{C}_n(x)\}_{x \in \mathbb{R}^d}$

Definition

An algorithm $\widehat{C}_n$ is a $(1-\alpha)$ distribution-free confidence interval if for every distribution P,

$$\mathbb{P}\left\{\pi_P(X_{n+1}) \in \widehat{C}_n(X_{n+1})\right\} \geq 1-\alpha$$

with respect to $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1}) \overset{\text{iid}}{\sim} P$.

a.k.a. "weakly valid probability estimators" (Vovk et al 2005)

(Related to distribution-free prediction — Papadopoulos et al 2002, Vovk et al 2005, Lei & Wasserman 2014, Lei et al 2018, Sadinle et al 2019, Barber et al 2019, ...)

Distribution-free coverage

Main question

Does there exist a distribution-free confidence interval $\widehat{C}_n$, such that length$(\widehat{C}_n(x)) \to 0$ for "nice" distributions P?

Distribution-free coverage

Main question

Does there exist a distribution-free confidence interval $\widehat{C}_n$, such that $\text{length}(\widehat{C}_n(x)) \to 0$ for "nice" distributions P?

For a distribution P & coverage level $1-\alpha$...

$$\text{lengthDF}_{n,\alpha}(P) = \inf_{\substack{\text{Distrib.-free} \\ \widehat{C}_n \text{ at level } 1-\alpha}} \mathbb{E}_{(X_i,Y_i)\overset{\text{iid}}{\sim}P}\left[\text{length}(\widehat{C}_n(X_{n+1}))\right]$$

$\widehat{C}_n$ must have coverage for every distribution P

What is $\widehat{C}_n$'s length for a particular distribution P?

Overview of new results

Our results study this problem in two settings...

- Part 1 — if P_X is nonatomic ("never see the same X twice")

  Is distribution-free inference possible for binary regression?
  `arXiv:2004.09477`

- Part 2 — if P_X is discrete or mixed discrete+nonatomic

  Distribution-free inference for regression: discrete, continuous, and in between
  `arXiv:2105.14075` (joint with Yonghoon Lee)

Part 1: the nonatomic case

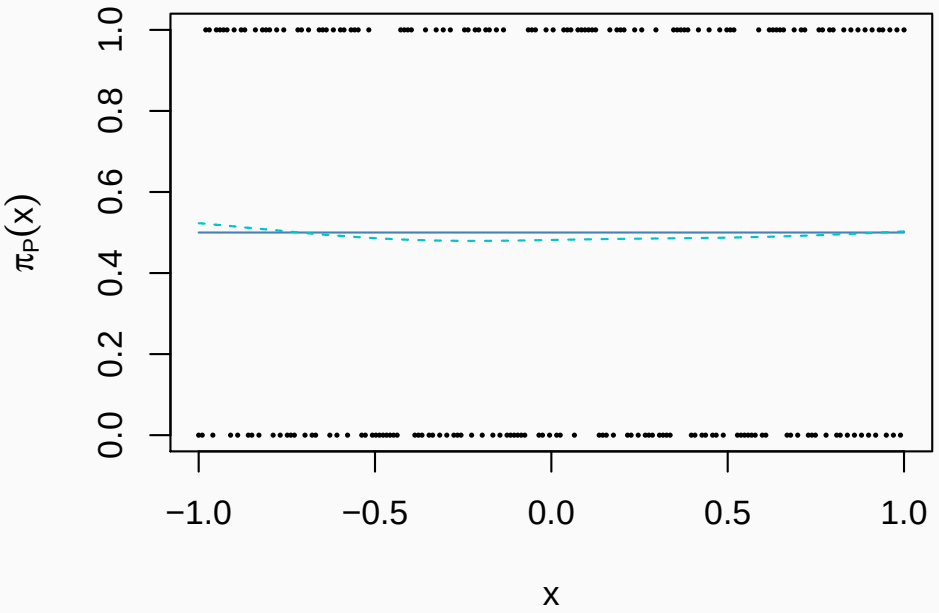
A trivial upper bound...

$$\text{lengthDF}_{n,\alpha}(P) \leq 1 - \alpha$$

Proof: let $\widehat{C}_n(x) = \begin{cases} [0,1], & \text{with probability } 1-\alpha, \\ \varnothing, & \text{with probability } \alpha. \end{cases}$

Part 1: the nonatomic case

Lower bound warmup: what if $X \perp\!\!\!\perp Y$ with $Y \sim \text{Bernoulli}(0.5)$?



Part 1: the nonatomic case

Lower bound warmup: what if $X \perp\!\!\!\perp Y$ with $Y \sim \text{Bernoulli}(0.5)$?

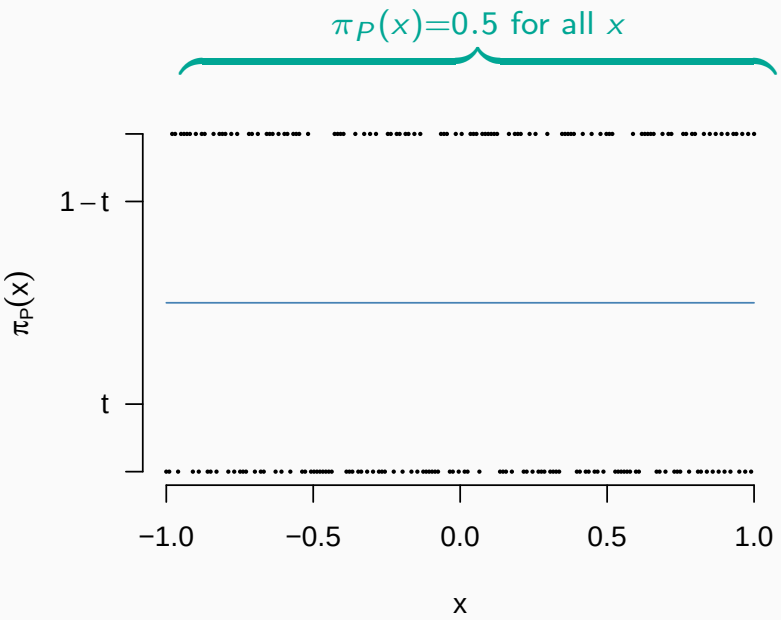
Part 1: the nonatomic case

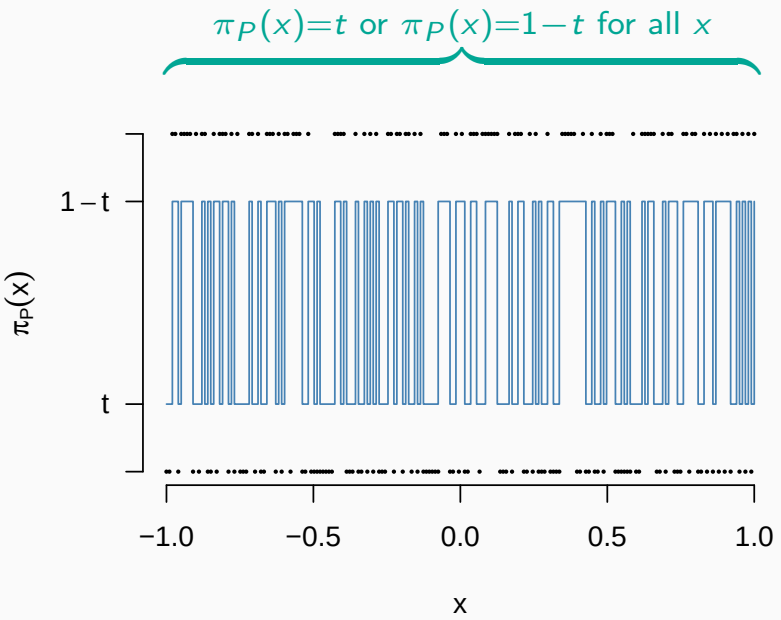
Lower bound warmup: what if $X \perp\!\!\!\perp Y$ with $Y \sim \text{Bernoulli}(0.5)$?

$$\rightsquigarrow \text{If } P_X \text{ is nonatomic, then } \text{lengthDF}_{n,\alpha}(P) \geq 1 - \alpha$$

Proof idea: show that

$$\frac{1}{2}\mathbb{P}\left\{t \in \widehat{C}_n(X_{n+1})\right\} + \frac{1}{2}\mathbb{P}\left\{1 - t \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha \quad \forall t \in [0, 1]$$





Main result: lower bound

Key lemma

Let $Z \in [0,1]$ satisfy $Z \perp\!\!\!\perp (X_1, Y_1), \ldots, (X_n, Y_n)$ and

$$\mathbb{E}\left[Z \mid X_{n+1}\right] = \pi_P(X_{n+1}).$$

Then if P_X is nonatomic,

$$\mathbb{P}\left\{Z \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

Main result: lower bound

Key lemma

Let $Z \in [0,1]$ satisfy $Z \perp\!\!\!\perp (X_1, Y_1), \ldots, (X_n, Y_n)$ and

$$\mathbb{E}[Z \mid X_{n+1}] = \pi_P(X_{n+1}).$$

Then if P_X is nonatomic,

$$\mathbb{P}\left\{Z \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

Proof idea: compare the true distribution P:

$$X \sim P_X$$
$$Y|X \sim \text{Bernoulli}(\pi_P(X))$$

with an equivalent distribution:

$$X \sim P_X$$
$$Z|X \sim (\text{any distrib. on } [0,1] \text{ with mean } \pi_P(X))$$
$$Y|X, Z \sim \text{Bernoulli}(Z)$$

Main result: lower bound

Theorem (simple case)

If $P = P_X \times \text{Bernoulli}(t)$ (i.e., $X \perp\!\!\!\perp Y$), and if P_X is nonatomic, then

$$\text{lengthDF}_{n,\alpha}(P) \geq \ell(t, \alpha).$$

Does not vanish as $n \to \infty$

Main result: lower bound

Theorem (simple case)

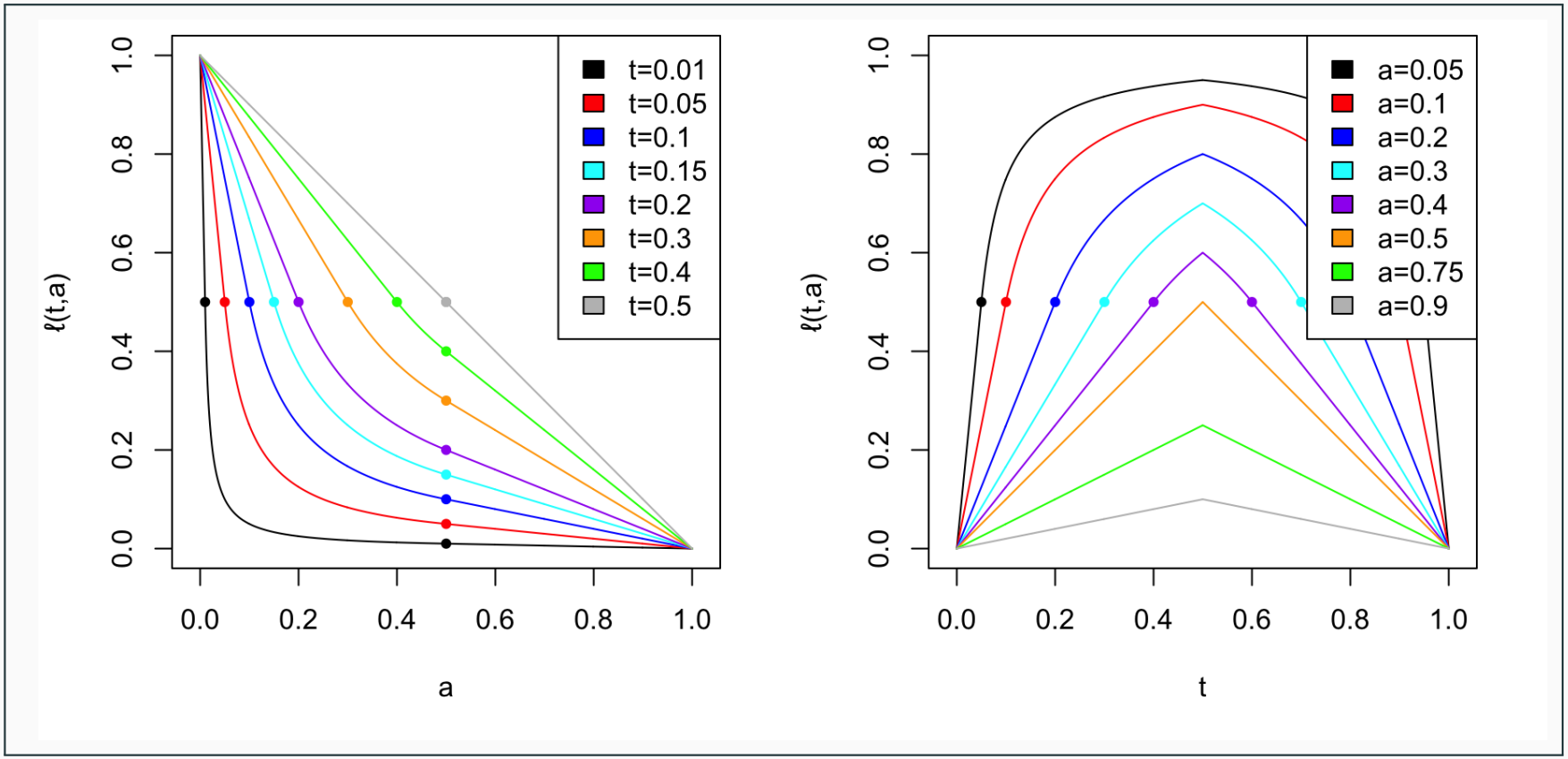
If $P = P_X \times \text{Bernoulli}(t)$ (i.e., $X \perp\!\!\!\perp Y$), and if P_X is nonatomic, then

$$\text{lengthDF}_{n,\alpha}(P) \geq \ell(t, \alpha).$$

Does not vanish as $n \to \infty$

Proof idea: apply the Key Lemma to a random variable Z that is a mixture of a point mass & a uniform

Main result: lower bound



$$\ell(t,a) = \begin{cases} 2(1-a)t, & t \le \frac{1}{2} \le a, \\ t/2a, & 0 < t \le a < \frac{1}{2}, \\ 1 - a/2t, & a < t \le \frac{1}{2}, \\ 0, & a = t = 0, \\ (\text{symmetric}), & t \ge \frac{1}{2} \end{cases}$$

Main result: lower bound

Theorem (general case)

For any P, if P_X is nonatomic, then $\text{lengthDF}_{n,\alpha}(P) \geq L_\alpha(P)$, where

$$L_\alpha(P) = \inf_{a:\mathbb{R}^d \to [0,1]} \left\{ \mathbb{E}_P \left[\ell(\pi_P(X), a(X)) \right] \ : \ \mathbb{E}_P \left[a(X) \right] \leq \alpha \right\}.$$

Does not vanish as $n \to \infty$

Main result: lower bound

Main question

Does there exist a distribution-free confidence interval $\widehat{C}_n$, such that length$(\widehat{C}_n(x)) \to 0$ for "nice" distributions P?

Main result: lower bound

Main question

Does there exist a distribution-free confidence interval $\widehat{C}_n$, such that length$(\widehat{C}_n(x)) \to 0$ for "nice" distributions P?

Answer:

- If P_X is nonatomic, then no —
  $L_\alpha(P)$ is an explicit lower bound on the length of $\widehat{C}_n(X_{n+1})$ that depends only on the distrib. of $\pi_P(X)$ & not on n
  - Smoothness of $x \mapsto \pi_P(x)$ doesn't help! Worst case $\pi_P(X) \equiv 0.5$
- (Part 2 of this talk — what if P_X is not nonatomic?)

Main result: lower bound

An aside...

Key lemma

Let $Z \in [0,1]$ satisfy $Z \perp\!\!\!\perp (X_1, Y_1), \ldots, (X_n, Y_n)$ and

$$\mathbb{E}[Z \mid X_{n+1}] = \pi_P(X_{n+1}).$$

Then

$$\mathbb{P}\left\{Z \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

A corollary: by taking $Z = Y_{n+1}$,

$$\mathbb{P}\left\{Y_{n+1} \in \widehat{C}_n(X_{n+1})\right\} \geq 1 - \alpha$$

$\Rightarrow$ any d.f. confidence interval is a d.f. prediction interval

(See Vovk et al 2005, Gupta, Podkopaev, & Ramdas 2020 for related results)

A matching upper bound?

Intuition...

To construct a d.f. confidence interval with length $\approx L_\alpha(P)$...

- Estimate $\pi_P(X)$ using half of the data, & partition $\mathbb{R}^d$ into bins $\mathcal{X}_1 \cup \cdots \cup \mathcal{X}_M$, with $\pi_P(X) \approx$ constant in each bin
- Estimate $\mathbb{P}\{Y = 1 \mid X \in \mathcal{X}_m\}$ in each bin w/ remaining data, & use this to construct a C.I. (see paper for details)

A matching upper bound?

Distribution-free upper bound

The proposed binning-based algorithm $\widehat{C}_n$ satisfies

- Distribution-free validity, i.e., coverage $\geq 1 - \alpha$ w.r.t. P
- Near-optimal length if the partition is "good":

$$\mathbb{E}_P\left[\text{length}(\widehat{C}_n(X))\right] \leq L_\alpha(P) + \sqrt{2\alpha^{-1} \cdot \mathbb{E}_P\left[|\pi_P(X) - \pi_{m(X)}|\right]} + \mathcal{O}\left(\sqrt{\frac{M \log n}{\alpha n}}\right)$$

Part 2: beyond the nonatomic case

Suppose P_X is instead discrete.

At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$

Part 2: beyond the nonatomic case

Suppose P_X is instead discrete.

At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$ ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$

Part 2: beyond the nonatomic case

Suppose P_X is instead discrete.

At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$ ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$ ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$

Part 2: beyond the nonatomic case

Suppose P_X is instead discrete.

At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$ ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$ ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$ ↘ indistinguishable from nonatomic, so width is $\asymp 1$

Part 2: beyond the nonatomic case

Suppose P_X is instead discrete.

At sample size n, intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$ ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$ ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$ ↘ our initial guess: width $\asymp 1$ — incorrect!
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$ ↘ indistinguishable from nonatomic, so width is $\asymp 1$

Part 2: beyond the nonatomic case

Our work is inspired by the literature on testing properties of discrete distributions:

- *Instance Optimal Learning*, Valiant & Valiant
- *Optimal Algorithms for Testing Closeness of Discrete Distributions*, Chan, Diakonikolas, Valiant, & Valiant

- *Optimal Testing for Properties of Distributions*, Acharya, Daskalakis, & Kamath
- *Testing Closeness With Unequal Sized Samples*, Bhattacharya & Valiant
- *Estimating Renyi Entropy of Discrete Distributions*, Acharya, Orlitsky, Suresh, & Tyagi

(Thanks to John Lafferty for suggesting this connection!)

Part 2: beyond the nonatomic case

Our work is inspired by the literature on testing properties of discrete distributions:

- *Instance Optimal Learning*, Valiant & Valiant
- *Optimal Algorithms for Testing Closeness of Discrete Distributions*, Chan, Diakonikolas, Valiant, & Valiant ↘ p, q supported on M points; test $p = q$ vs $\mathrm{d}_{\mathrm{TV}}(p, q) > \epsilon$
- *Optimal Testing for Properties of Distributions*, Acharya, Daskalakis, & Kamath
- *Testing Closeness With Unequal Sized Samples*, Bhattacharya & Valiant
- *Estimating Renyi Entropy of Discrete Distributions*, Acharya, Orlitsky, Suresh, & Tyagi

(Thanks to John Lafferty for suggesting this connection!)

Effective support size

P might be discrete, nonatomic, or a mixture —
how to unify these three cases?

$M_\gamma(P_X) =$ minimum # of points needed to capture $\geq 1 - \gamma$ probability

- P_X discrete, & supported on M points $\rightsquigarrow M_\gamma(P_X) \leq M$
- P_X nonatomic $\rightsquigarrow M_\gamma(P_X) = \infty$

Main result: lower bound

Theorem

For any P with $\pi_P(X) \in [t, 1-t]$ almost surely,

$$\text{lengthDF}_{n,\alpha}(P) \geq \tfrac{1}{3} t(1-t)(\gamma-\alpha)^2 \cdot \min\left\{ \frac{\left(M_\gamma(P_X)\right)^{1/4}}{n^{1/2}}, 1 \right\}.$$

Vanishing width iff $M_\gamma(P_X) \ll n^2$

Main result: lower bound

Theorem

For any P with $\pi_P(X) \in [t, 1-t]$ almost surely,

$$\text{lengthDF}_{n,\alpha}(P) \geq \tfrac{1}{3} t(1-t)(\gamma-\alpha)^2 \cdot \min\left\{ \frac{\left(M_\gamma(P_X)\right)^{1/4}}{n^{1/2}}, 1 \right\}.$$

Vanishing width iff $M_\gamma(P_X) \ll n^2$

A more general version in the paper...

- The response Y can be $\in [0,1]$ rather than binary
  $\rightarrow$ now assume $\text{Var}(Y|X) \geq c > 0$ almost surely
- Can relax to: $\text{Var}(Y|X) \geq c > 0$ with probability $> 1 - (\gamma - \alpha)$

Proof sketch for lower bound

The distribution P can be described as:

- Draw $X \sim P_X$
- Draw $Z \sim \text{Bernoulli}(0.5)$
- Draw the response Y:

$$\begin{cases} \text{If } Z = 1, \text{ draw } Y \text{ conditional on } \{Y \geq \text{Median}(Y|X)\} \\ \text{If } Z = 0, \text{ draw } Y \text{ conditional on } \{Y \leq \text{Median}(Y|X)\} \end{cases}$$

Proof sketch for lower bound

The distribution ~~P~~ P_ϵ can be described as:

- Draw $X \sim P_X$
- Draw $Z \sim \text{Bernoulli}(0.5)$ ~~X~~ $0.5 + \epsilon \cdot A(X)$, where $A(x) \overset{\text{iid}}{\sim} \{\pm 1\}$ for each x
- Draw the response Y:

$$\begin{cases} \text{If } Z = 1, \text{ draw } Y \text{ conditional on } \{Y \geq \text{Median}(Y|X)\} \\ \text{If } Z = 0, \text{ draw } Y \text{ conditional on } \{Y \leq \text{Median}(Y|X)\} \end{cases}$$

Proof sketch for lower bound

The distribution ~~P~~ P_ϵ can be described as:

- Draw $X \sim P_X$
- Draw $Z \sim \text{Bernoulli}(0.5)$ ~~0.5~~ $0.5 + \epsilon \cdot A(X)$, where $A(x) \overset{\text{iid}}{\sim} \{\pm 1\}$ for each x
- Draw the response Y:

$$\begin{cases} \text{If } Z = 1, \text{ draw } Y \text{ conditional on } \{Y \geq \text{Median}(Y|X)\} \\ \text{If } Z = 0, \text{ draw } Y \text{ conditional on } \{Y \leq \text{Median}(Y|X)\} \end{cases}$$

Key lemma

$$\text{d}_{\text{TV}}(n \text{ samples from } P, \; n \text{ samples from } P_\epsilon) \leq 2n\epsilon^2 \sqrt{\sum_x p(x)^2}$$

Proof sketch for lower bound

The distribution ~~P~~ P_ϵ can be described as:

- Draw $X \sim P_X$
- Draw $Z \sim \text{Bernoulli}(0.5)$ ~~0.5~~ $0.5 + \epsilon \cdot A(X)$, where $A(x) \overset{\text{iid}}{\sim} \{\pm 1\}$ for each x
- Draw the response Y:

$$\begin{cases} \text{If } Z = 1, \text{ draw } Y \text{ conditional on } \{Y \geq \text{Median}(Y|X)\} \\ \text{If } Z = 0, \text{ draw } Y \text{ conditional on } \{Y \leq \text{Median}(Y|X)\} \end{cases}$$

Key lemma

$$\text{d}_{\text{TV}}(n \text{ samples from } P,\ n \text{ samples from } P_\epsilon) \leq 2n\epsilon^2 \sqrt{\sum_x p(x)^2}$$

$= o(1)$ if $p(x) \lesssim 1/M$ and $\epsilon \ll \frac{M^{1/4}}{n^{1/2}}$

Proof sketch for lower bound

Distribution-free coverage $\Rightarrow$ $\widehat{C}(X_{n+1})$ must contain $\pi_P(X_{n+1})$

Key lemma $\Rightarrow$ since P & P_ϵ are indistinguishable,
$\widehat{C}(X_{n+1})$ must also contain $\pi_{P_\epsilon}(X_{n+1})$

Since $|\pi_P(x) - \pi_{P_\epsilon}(x)| \asymp \epsilon$, this proves width$(\widehat{C}(X_{n+1})) \gtrsim \epsilon$

A matching upper bound?

Suppose P_X is supported on M points (this is generalized in the paper)
& we have an estimate π of the true π_P.

Define:

$$Z = \sum_{x \text{ observed } n_x \geq 2 \text{ times}} (n_x - 1) \cdot \left((\bar{y}_x - \pi(x))^2 - n_x^{-1} s_x^2 \right).$$

$\bar{y}_x$: sample mean & var. for the Y values at this x

A matching upper bound?

Suppose P_X is supported on M points (this is generalized in the paper)
& we have an estimate π of the true π_P.

Define:

$$Z = \sum_{x \text{ observed } n_x \geq 2 \text{ times}} (n_x - 1) \cdot \overbrace{\left((\bar{y}_x - \pi(x))^2 - n_x^{-1} s_x^2\right)}^{\mathbb{E} = (\pi_P(x) - \pi(x))^2}.$$

$\bar{y}_x$: sample mean & s_x^2: sample var. for the Y values at this x

(Inspired by similar statistics in the discrete testing literature,
e.g. *Optimal Algorithms for Testing Closeness of Discrete Distributions*,
Chan, Diakonikolas, Valiant, & Valiant)

A matching upper bound?

Construct the C.I.:

$$\widehat{C}(X_{n+1}) = \pi(X_{n+1}) \pm \mathcal{O}\left(\frac{M^{1/2}}{n} \cdot \left(Z^{1/2} + \left(\begin{array}{c}\#\ x\text{'s observed} \\ \geq 2 \text{ times}\end{array}\right)^{1/4}\right)\right)$$

A matching upper bound?

Construct the C.I.:

$$\widehat{C}(X_{n+1}) = \pi(X_{n+1}) \pm \mathcal{O}\left(\frac{M^{1/2}}{n} \cdot \left(Z^{1/2} + \left(\begin{array}{c}\#\ x\text{'s observed} \\ \geq 2 \text{ times}\end{array}\right)^{1/4}\right)\right)$$

$$\uparrow \qquad\qquad \uparrow$$

$$\mathbb{E} \asymp \frac{n}{M^{1/2}} \cdot \|\pi_P - \pi\| \qquad \mathbb{E} \asymp \frac{n^{1/2}}{M^{1/4}}$$

A matching upper bound?

Construct the C.I.:

$$\widehat{C}(X_{n+1}) = \pi(X_{n+1}) \pm \mathcal{O}\left(\frac{M^{1/2}}{n} \cdot \left(Z^{1/2} + \left(\begin{array}{c}\#\ x\text{'s observed} \\ \geq 2 \text{ times}\end{array}\right)^{1/4}\right)\right)$$

$$\uparrow \qquad\qquad \uparrow$$

$$\mathbb{E} \asymp \frac{n}{M^{1/2}} \cdot \|\pi_P - \pi\| \qquad \mathbb{E} \asymp \frac{n^{1/2}}{M^{1/4}}$$

$$\Rightarrow \text{C.I. width} \asymp \|\pi_P - \pi\| + \frac{M^{1/4}}{n^{1/2}}$$

A matching upper bound?

Suppose P_X is discrete, and we draw n data points.
Intuitively separates into distinct regimes...

- Trivial — finitely many possible X's
  $(P_X(x) \asymp 1)$ ↘ build a C.I. for each x, with width $\asymp n^{-1/2}$
- Easy — each possible X value is observed many times
  $(P_X(x) \gg n^{-1})$ ↘ build a C.I. for each x, with width $\asymp n_x^{-1/2}$
- Medium — some X's are repeated, but most are unique
  $(n^{-2} \ll P_X(x) \ll n^{-1})$ ↘ our initial guess: width $\asymp 1$ — incorrect!
- Hard — w.h.p. the data set has no repeated X's
  $(P_X(x) \ll n^{-2})$ ↘ indistinguishable from nonatomic, so width is $\asymp 1$

A matching upper bound?

Intuition for the "medium" regime...

- Suppose P_X is uniform over $M = n^a$ many points, for $a \in (1, 2)$
- $\asymp \frac{n^2}{M} = n^{2-a}$ many X values are observed multiple times
  $\rightsquigarrow$ we can estimate our error on this subset
- These X values are a random sample from P_X
  $\rightsquigarrow$ we can estimate our error on average over P

Summary & open questions

- If features X are nonatomic,
  distribution-free inference is "impossible" even as $n \to \infty$
- Discrete X behaves like nonatomic X if effective support size $\gg n^2$
- Distribution-free inference becomes meaningful
  once effective support size is $\ll n^2$

Lots of open questions...

- Are there interesting properties weaker than distrib.-free coverage, that are still meaningful without assumptions?
- Are there methods that achieve a weak property for all P, & a stronger property for "nice" P?

Thank you!