

Joint Value Estimation and Bidding (JVEB) in First-Price Auctions (FPAs)

Yuxiao Wen

Computer Science Department, New York University

Joint work with Yanjun Han (Math and Data Science) and Zhengyuan Zhou (Stern Business School)

POMS-HK, Shenzhen

Janurary 2026

Talk Outline

- ▶ Introduction
- ▶ An “online + causal” formulation for JVEB
- ▶ Main results
- ▶ Conclusion and future directions

Talk Outline

- ▶ Introduction
- ▶ An “online + causal” formulation for JVEB
- ▶ Main results
- ▶ Conclusion and future directions

Digital advertising

Sponsored ads:

The screenshot shows a mobile search results page for "cat wet food" on Amazon. At the top, there's a search bar with the query and a camera icon. Below the search bar, the word "Canned" is highlighted. A "Subscribed" button is visible. The main content area displays two sponsored product listings:

Instinct Original Grain Free Recipe Wet Canned Cat Food by Nature's Variety
Natural, Variety Pack
4.0 ★★★★☆ (5.4k)
\$22.99 (80.64/ounce) You pay \$18.39 with coupon
FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon
Or fastest delivery Tomorrow, Dec 6

Tiki Cat After Dark, Variety Pack, High-Protein and 100% Non-GMO Ingredients, Wet Cat Food for...
Cat
4.4 ★★★★★ (3.2k)
\$24.99 (80.74/ounce) List: \$37.48 You pay \$20.89 with coupon
FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon
Or fastest delivery Tomorrow, Dec 6

At the bottom of the screen, there are navigation icons for home, search, cart, and account.

Digital advertising

Sponsored ads:

A screenshot of a mobile application interface. At the top, there's a search bar with the text "cat wet food". Below it, a "Canned" section shows two sponsored product cards. The first card for "Instinct Original Grain Free Recipe Wet Canned Cat Food by Nature's Variety Pack" has a 4.0 rating, \$22.99 price, and a green button saying "You pay \$18.39 with coupon". The second card for "Tiki Cat After Dark, Variety Pack" has a 4.4 rating, \$24.99 price, and a similar coupon offer. Both cards include delivery information and an "Add to cart" button.

A screenshot of a Google search results page for the query "singapore hotel". The search bar at the top contains "singapore hotel". Below the search bar, there are several search filters: AI Mode, All, Images, Maps, News, Videos, Shopping, More, and Tools. The main content area is titled "Sponsored results" and lists three hotel booking services: Booking.com, Book Near City Centre, and Expedia. Each listing includes a thumbnail image, the service name, a brief description, and a "Compare" button.

Digital advertising

Sponsored ads:

11:06

cat wet food

Canned

Subscribed

Instinct Original Grain Free Recipe Wet Canned Cat Food by Nature's Variety Natural, Natural Pack 4.0 ★★★★☆ (5.4k) \$22.99 (150.64 ounce) You pay \$18.39 with coupon FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon Or fastest delivery Tomorrow, Dec 6

Add to cart

Tiki Cat After Dark, Variety Pack, High-Protein and 100% Non-GMO Ingredients, Wet Cat Food for... Cat 4.4 ★★★★★ (3.2k) \$24.99 (160.74 ounce) List: \$37.49 You pay \$20.99 with coupon FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon Or fastest delivery Tomorrow, Dec 6

Add to cart

See all >

Home Account Cart Help

Google

singapore hotel

All Mode All Images Maps News Videos Shopping More Tools

Sponsored results

Booking.com Book Near City Centre Booking.com - Hotels Expedia Hotels

11:11

YouTube

Texas hold 'em Shinichi Kudō N

can hollow day as hard as this guy.
>> Look who I found on the roof.
>> A

0:24

Triple Treat Box

Get 2 medium pizzas, breadsticks, and a dessert starting at \$19.99.

Sponsored • Pizza Hut

Order Now

Digital advertising

Sponsored ads:

11:06

cat wet food

Canned

Subscribed

Instinct Original Grain Free Recipe Wet Canned Cat Food by Nature's Variety Pack
4.0 ★★★★☆ (5.4k)
\$22.99 (\$0.64/ounce)
You pay **\$18.39** with coupon
FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon
Or fastest delivery Tomorrow, Dec 6

Add to cart

Tiki Cat After Dark, Variety Pack, High-Protein and 100% Non-GMO Ingredients, Wet Cat Food for...
Cat
4.4 ★★★★☆ (3.2k)
\$24.99 (\$0.74/ounce) List: \$37.49
You pay **\$20.89** with coupon
FREE delivery Sun, Dec 7 on \$35 of pet supplies shipped by Amazon
Or fastest delivery Tomorrow, Dec 6

Add to cart

See all >

Home • Account • Cart

Google

singapore hotel

Sponsored results

Booking.com Book your Hotel in Singapore online. No reservation costs. Great rates. Best Price-Guarantee. Read Real Guest Reviews.

Hotels in Singapore

Booking.com - Hotels Book your Hotel in Singapore online. No reservation costs. Great rates

Expedia Book Now — Compare Hotel Rooms on Expedia. View Deals and Reserve Now. Get the Most Out of Your Trip...

11:11

YouTube

Texas hold 'em | Shinichi Kudō

can hollow day as hard as this guy.
>> Look who I found on the roof.
>> A

0:24

Triple Treat Box

Get 2 medium pizzas, breadsticks, and a dessert starting at \$19.99.

Sponsored • Pizza Hut

Order Now

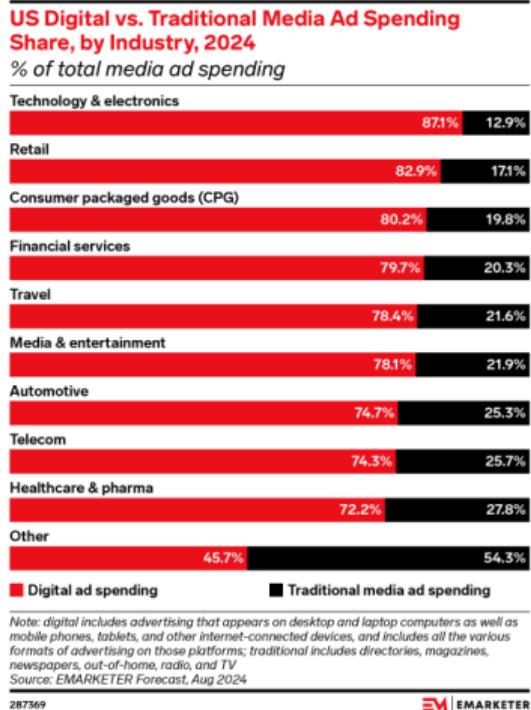
11:12

Visit advertiser

Durable Adjustable Tactical...
derostes.com

Shop now

Digital advertising v.s. traditional advertising



287369

EM | EMARKETER

Figure: Digital advertising over industries.

Digital advertising v.s. traditional advertising

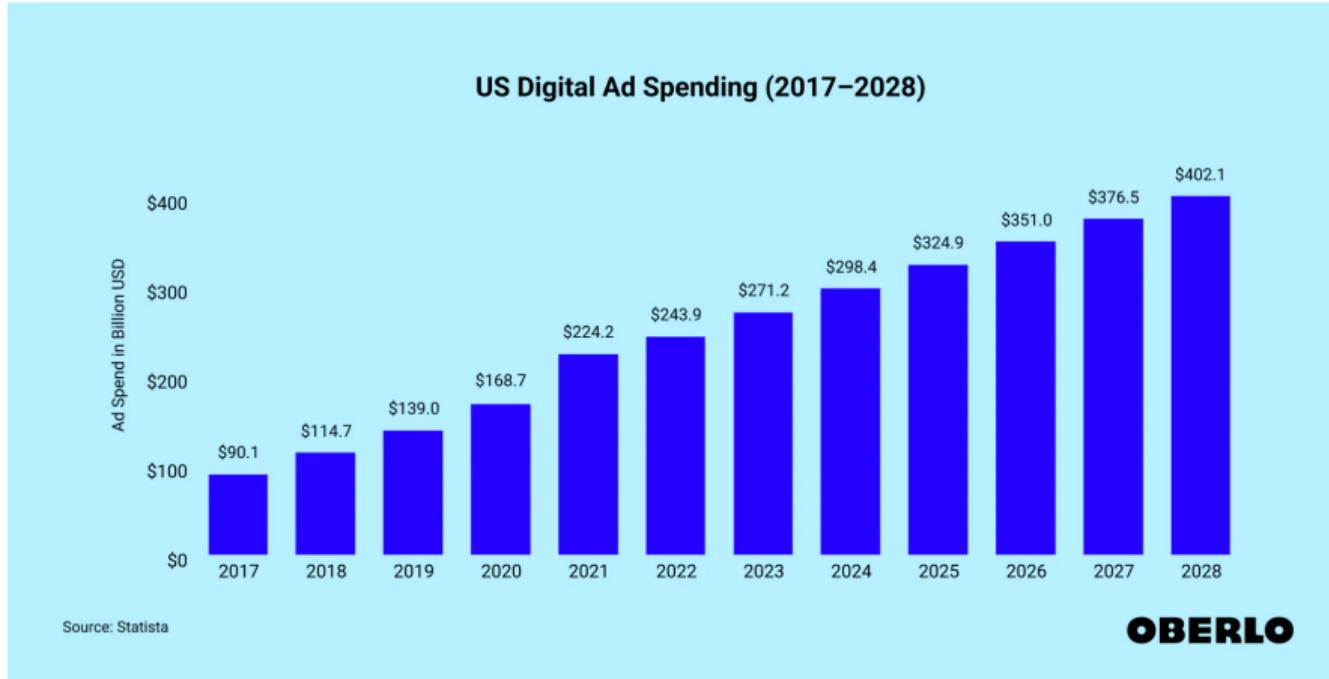


Figure: Media ad spending trend and forecast (updated 2024).

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

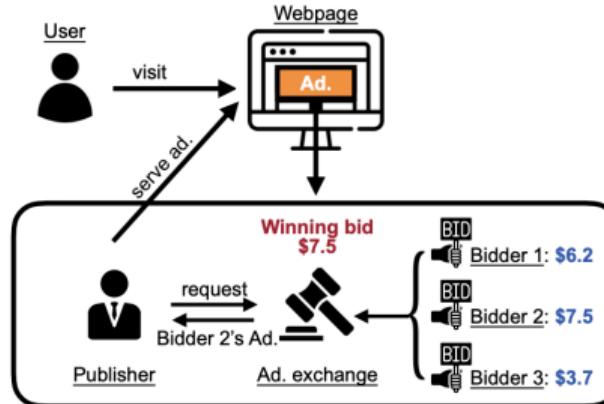


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

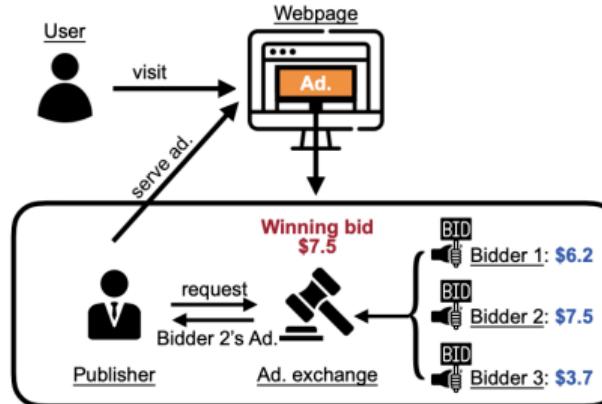


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Automated bidding algorithms.

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

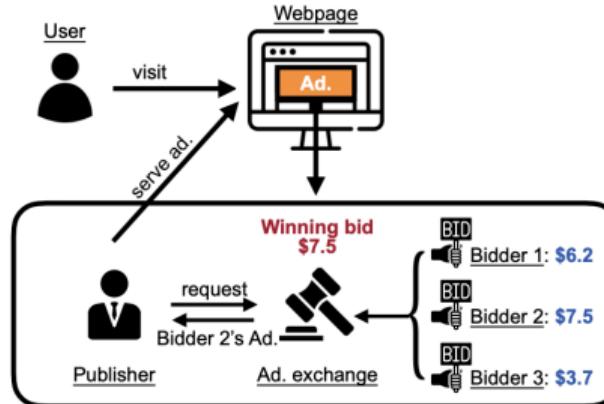


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Automated bidding algorithms.

E.g. User value = click-through-rate / conversion rate.
 $Bid \in [0, \text{value}]$.

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

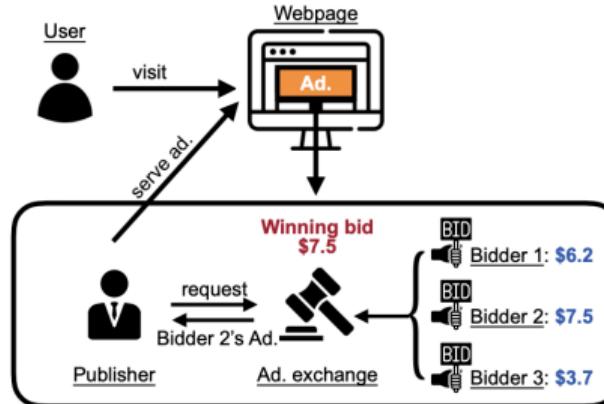


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Automated bidding algorithms.

E.g. User value = click-through-rate / conversion rate.
 $\text{Bid} \in [0, \text{value}]$.

Question: What if user was already familiar/exposed through other channels?

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

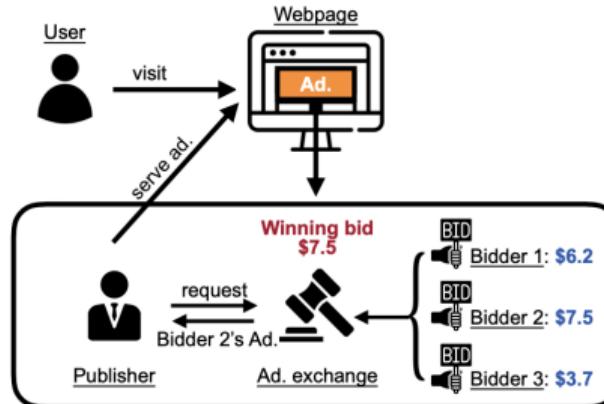


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Automated bidding algorithms.

E.g. User value = click-through-rate / conversion rate.
 $\text{Bid} \in [0, \text{value}]$.

Question: What if user was already familiar/exposed through other channels? \implies **wasted bids**.

Digital advertising with first-price auctions (FPAs)

Publisher \longleftrightarrow Ad Exchange (Auctions) \longleftrightarrow Bidders.

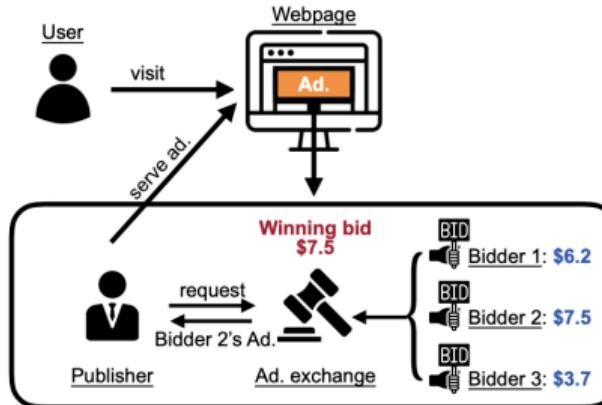


Figure: Example ad exchange diagram*. FPAs used by Google, Amazon, etc...

*Figure from Han and Weissman and Zhou (2025)

Automated bidding algorithms.

E.g. User value = click-through-rate / conversion rate.
 $Bid \in [0, \text{value}]$.

Question: What if user was already familiar/exposed through other channels? \implies **wasted bids**.

Waismann and Nair and Narrion (2025): model user value as **treatment effect**.

Digital advertising with first-price auctions (FPAs)

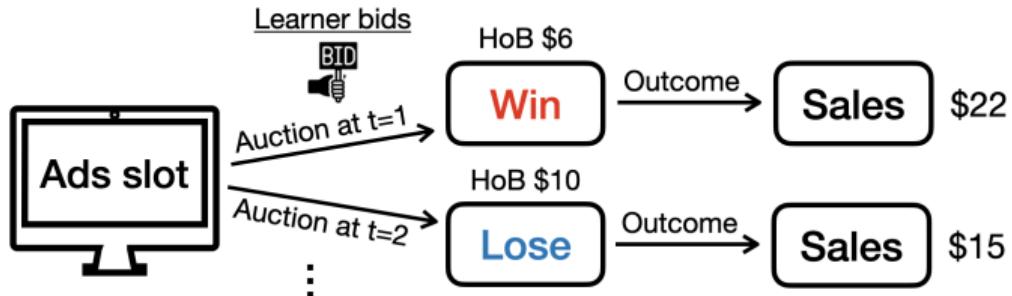


Figure: Value = outcome difference

Durable Adjustable Tactical...
derostes.com

Shop now

Digital advertising with first-price auctions (FPAs)

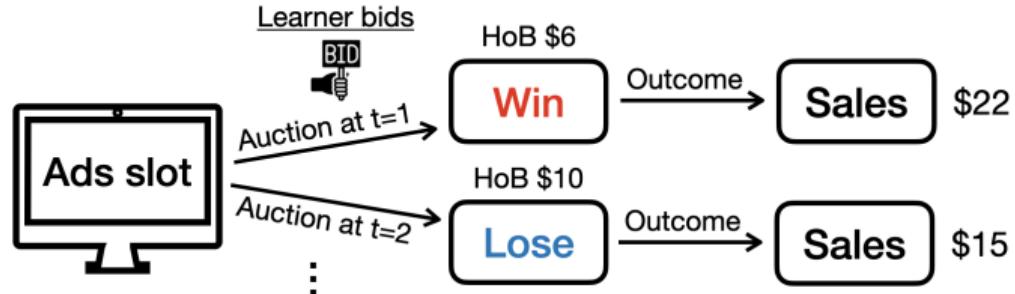


Figure: Value = outcome difference

- ▶ Value = *marginal* gain through this ad exposure.
- ▶ Bidder needs to learn both the **value** and the **highest other bid (HOB)** to bid optimally.

Digital advertising with first-price auctions (FPAs)

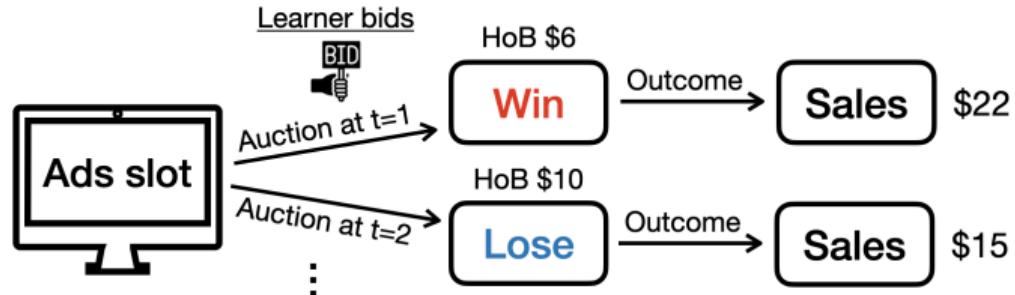


Figure: Value = outcome difference

- ▶ Value = *marginal* gain through this ad exposure.
- ▶ Bidder needs to learn both the **value** and the **highest other bid (HOB)** to bid optimally.
- ▶ Value is **never** observed!

Digital advertising with first-price auctions (FPAs)

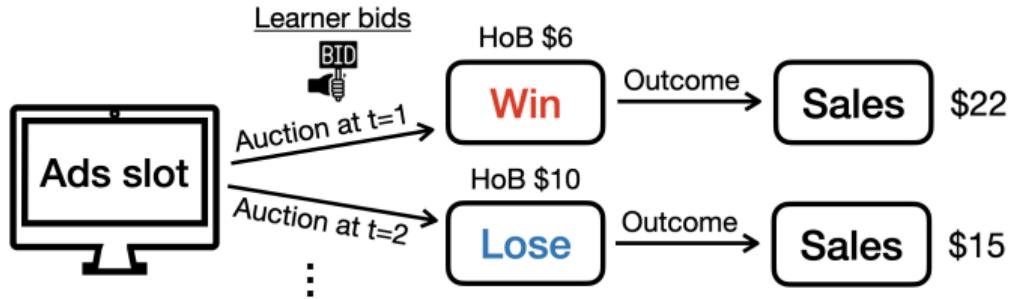


Figure: Value = outcome difference

- ▶ Know value precisely: Balseiro et al. (2019), Han et al. (2020), Badanidiyuru and Fend and Guruganesh (2023), Han and Weissman and Zhou (2025)

- ▶ Value is **never** observed!

Digital advertising with first-price auctions (FPAs)

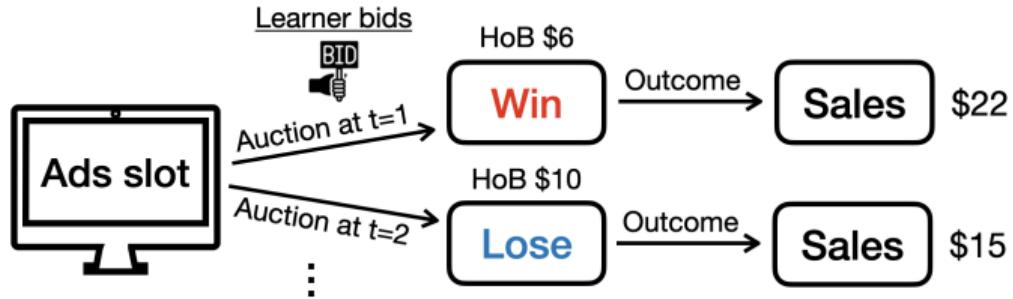


Figure: Value = outcome difference

- ▶ Know value precisely: Balseiro et al. (2019), Han et al. (2020), Badanidiyuru and Fend and Guruganesh (2023), Han and Weissman and Zhou (2025)
- ▶ Observe value if win: Feng and Podimata and Syrgkanis (2018), Cesa-Bianchi et al. (2024)
- ▶ Value is **never** observed!

Digital advertising with first-price auctions (FPAs)

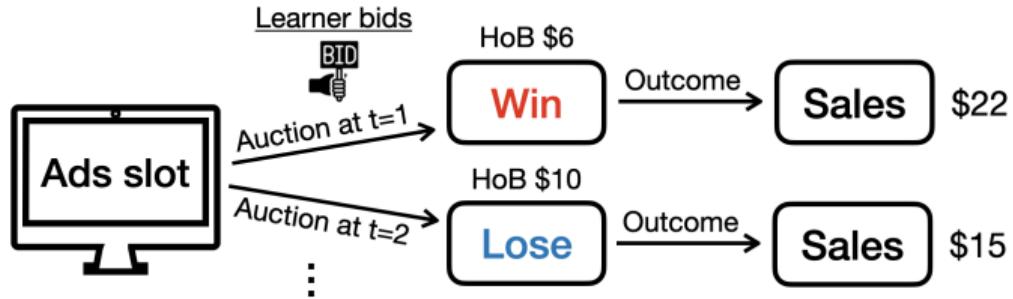


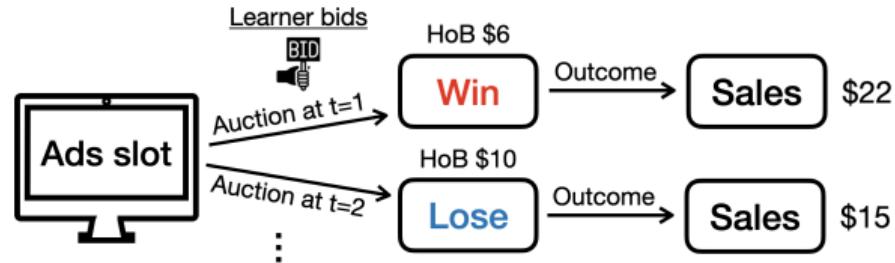
Figure: Value = outcome difference

- ▶ Know value precisely: Balseiro et al. (2019), Han et al. (2020), Badanidiyuru and Fend and Guruganesh (2023), Han and Weissman and Zhou (2025)
- ▶ Observe value if win: Feng and Podimata and Syrgkanis (2018), Cesa-Bianchi et al. (2024)
- ▶ Value is **never** observed! \implies causal approach.

Talk Outline

- ▶ Introduction
- ▶ An “online + causal” formulation for JVEB
- ▶ Main results
- ▶ Conclusion and future directions

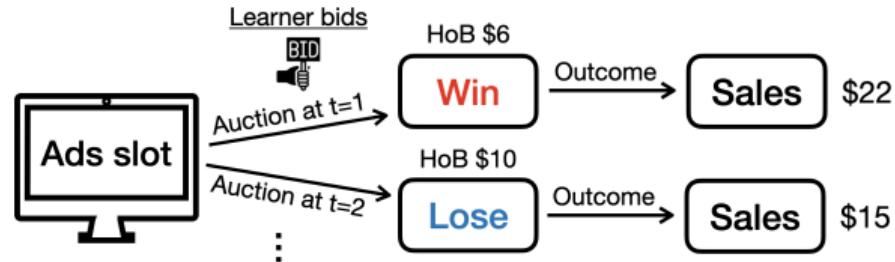
Bidding protocol



Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$,
- ▶ submit bid b_t ,
- ▶ receive outcome $v_{t,1}$ if $b_t \geq m_t$ (the HOB), or $v_{t,0}$ otherwise,
- ▶ expected reward $r_t(b_t) = \mathbb{E}[\mathbb{1}[b_t \geq m_t](v_{t,1} - b_t) + \mathbb{1}[b_t < m_t]v_{t,0}]$.

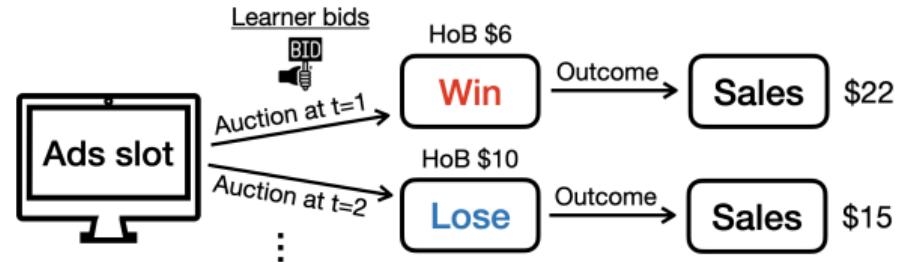
Bidding protocol



Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$, e.g. purchase history, user image, seasonal factors...
- ▶ submit bid b_t ,
- ▶ receive outcome $v_{t,1}$ if $b_t \geq m_t$ (the HOB), or $v_{t,0}$ otherwise,
- ▶ expected reward $r_t(b_t) = \mathbb{E}[\mathbb{1}[b_t \geq m_t](v_{t,1} - b_t) + \mathbb{1}[b_t < m_t]v_{t,0}]$.

Bidding protocol



Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$, e.g. purchase history, user image, seasonal factors...
- ▶ submit bid b_t ,
- ▶ receive outcome $v_{t,1}$ if $b_t \geq m_t$ (the HOB), or $v_{t,0}$ otherwise,
- ▶ expected reward $r_t(b_t) = \mathbb{E}[\mathbb{1}[b_t \geq m_t](v_{t,1} - b_t) + \mathbb{1}[b_t < m_t]v_{t,0}]$.

Regret minimization:

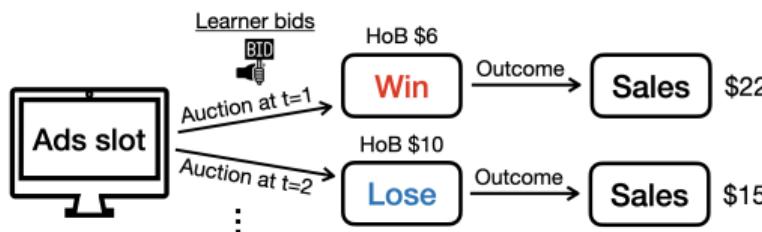
$$R(\pi) = \mathbb{E} \left[\sum_{t=1}^T \max_{b_t^*} r_t(b_t^*) - r_t(b_t) \right].$$

A causal framework

Neyman-Rubin potential outcome model:

Bidding in FPAs:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$



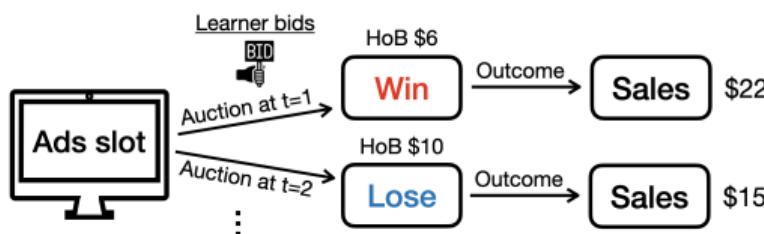
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$



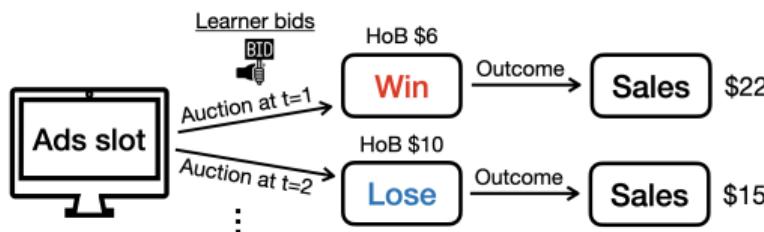
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$



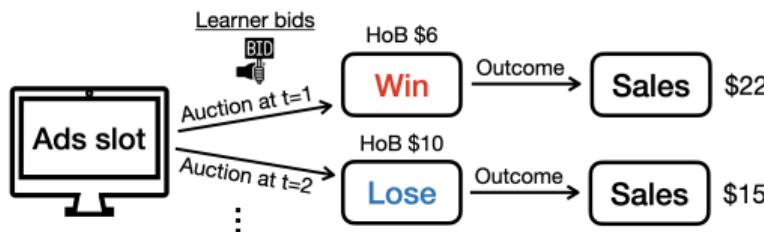
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$



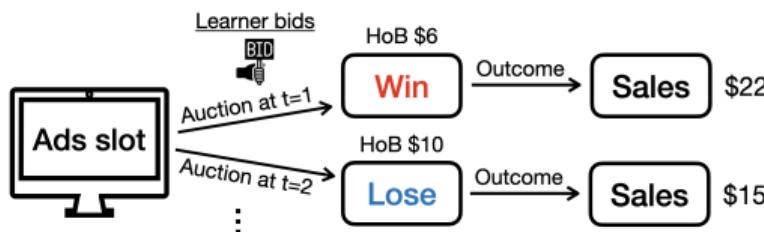
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$



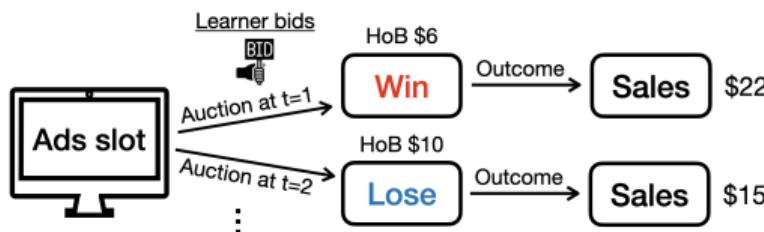
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$



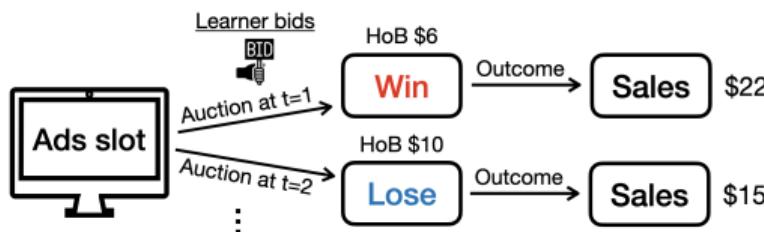
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$



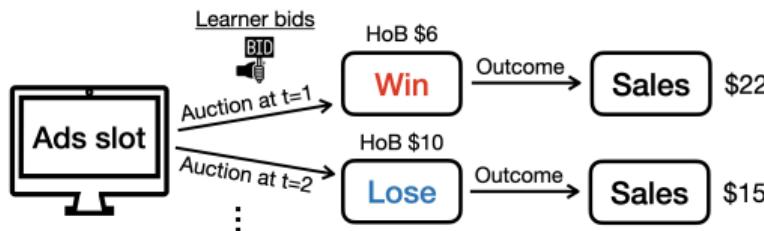
A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$

Bidding in FPAs:

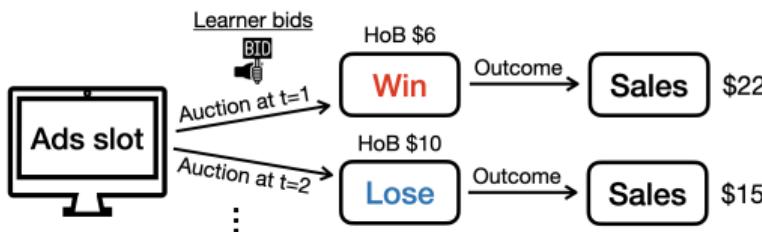
- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$
- * Goal: regret minimization



A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$
- * Overlap condition $e_i \in [c, 1 - c] \forall i$



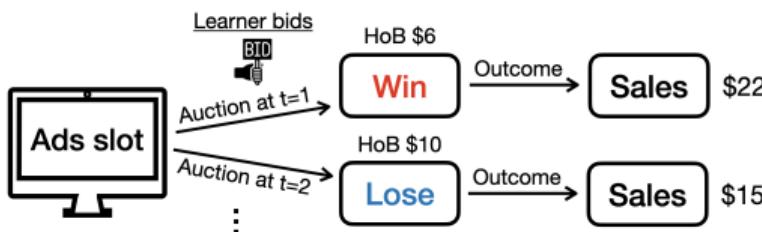
Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$
- * Goal: regret minimization

A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$
- * Overlap condition $e_i \in [c, 1 - c] \forall i$



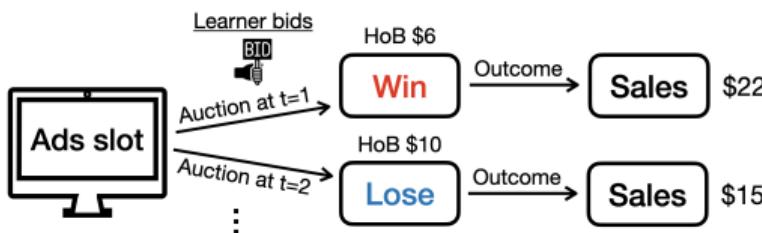
Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$
- * Goal: regret minimization
- * Overlap condition ?

A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$
- * Overlap condition $e_i \in [c, 1 - c] \forall i$



Bidding in FPAs:

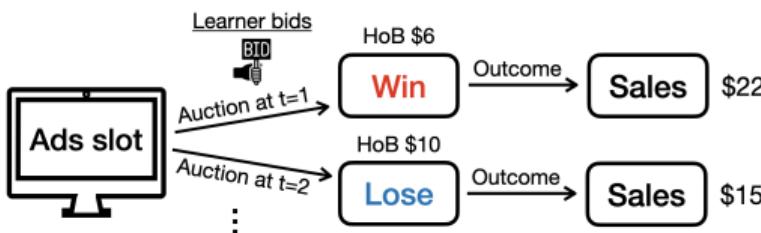
- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$
- * Goal: regret minimization
- * Overlap condition ?

*If $\mathbb{P}(b_t \geq m_t | x_t) \in [c, 1 - c]$, must win (and lose) at least cT rounds (randomized experiments).

A causal framework

Neyman-Rubin potential outcome model:

- ▶ Units $i = 1, 2, \dots, N$
- ▶ Covariates X_i
- ▶ Treatments $\tau_i \in \{0, 1\}$
- ▶ Propensity $e_i = \mathbb{P}(\tau_i = 1 | X_i)$
- ▶ Potential outcomes $\{Y_i(0), Y_i(1)\}$
- ▶ Outcomes $Y_i(\tau_i)$
- * Goal: estimate $\mathbb{E}[Y(1) - Y(0)|X]$
- * Overlap condition $e_i \in [c, 1 - c] \forall i$



Bidding in FPAs:

- ▶ Time $t = 1, 2, \dots, T$
- ▶ Context $x_t \in \mathbb{R}^d$
- ▶ Ad-presence $\mathbb{1}[b_t \geq m_t]$
- ▶ $\mathbb{P}(\mathbb{1}[b_t \geq m_t] = 1 | x_t) = \mathbb{P}(b_t \geq m_t | x_t)$
- ▶ Potential outcomes $\{v_{t,1}, v_{t,0}\}$
- ▶ Outcomes $v_{t,\mathbb{1}[b_t \geq m_t]}$
- * Goal: regret minimization
- * Overlap condition ?

*If $\mathbb{P}(b_t \geq m_t | x_t) \in [c, 1 - c]$, must win (and lose) at least cT rounds (randomized experiments).

Our result: optimal regret with no/few randomized experiments!

Model formulation

HOB modeling:

Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$,
 - ▶ submit bid b_t ,
 - ▶ win if $b_t \geq m_t$;
 - ▶ receive $v_{t,1}$ or $v_{t,0}$.
-

Model formulation

HOB modeling:

- (Stochasticity) m_t drawn from Lipschitz CDF G_t .

Over horizon $t = 1, 2, \dots, T$:

- observe user context $x_t \in \mathbb{R}^d$,
- submit bid b_t ,
- win if $b_t \geq m_t$;
- receive $v_{t,1}$ or $v_{t,0}$.

Model formulation

HOB modeling:

- ▶ (Stochasticity) m_t drawn from Lipschitz CDF G_t .

Value modeling:

Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$,
- ▶ submit bid b_t ,
- ▶ win if $b_t \geq m_t$;
- ▶ receive $v_{t,1}$ or $v_{t,0}$.

Model formulation

HOB modeling:

- ▶ (Stochasticity) m_t drawn from Lipschitz CDF G_t .

Value modeling:

- ▶ User-ad interaction $\mathbb{E}[v_{t,1} - v_{t,0}] = \theta_*^\top x_t.$

Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$,
- ▶ submit bid b_t ,
- ▶ win if $b_t \geq m_t$;
- ▶ receive $v_{t,1}$ or $v_{t,0}$.

Model formulation

HOB modeling:

- ▶ (Stochasticity) m_t drawn from Lipschitz CDF G_t .

Value modeling:

- ▶ User-ad interaction $\mathbb{E}[v_{t,1} - v_{t,0}] = \theta_*^\top x_t$. ^a

Over horizon $t = 1, 2, \dots, T$:

- ▶ observe user context $x_t \in \mathbb{R}^d$,
- ▶ submit bid b_t ,
- ▶ win if $b_t \geq m_t$;
- ▶ receive $v_{t,1}$ or $v_{t,0}$.

^aThis also aligns with recent causal literature assuming that the treatment effect has simpler structures than the baselines.

Key challenges: joint value estimation and bidding (JVEB)

Inter-dependent components:

Key challenges: joint value estimation and bidding (JVEB)

Inter-dependent components:

- (1) **Value estimation**: treatment effect estimation for θ_* ;

Key challenges: joint value estimation and bidding (JVEB)

Inter-dependent components:

- (1) **Value estimation**: treatment effect estimation for θ_* ;
- (2) **Bidding**: reward maximization;

Key challenges: joint value estimation and bidding (JVEB)

Inter-dependent components:

- (1) **Value estimation**: treatment effect estimation for θ_* ;
- (2) **Bidding**: reward maximization;
- (3) **HOB estimation**: learning G_t (propensity score).

Talk Outline

- ▶ Introduction
- ▶ An “online + causal” formulation for JVEB
- ▶ Main results
- ▶ Conclusion and future directions

Fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

Fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

- ▶ Estimating G_t decoupled from JVEB (value estimation + bidding).

Fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

- ▶ Estimating G_t decoupled from JVEB (value estimation + bidding).
- ▶ Allow us to focus on JVEB.

Fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

- ▶ Estimating G_t decoupled from JVEB (value estimation + bidding).
- ▶ Allow us to focus on JVEB.

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\forall b, \quad \left| \hat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

- ▶ Estimating G_t decoupled from JVEB (value estimation + bidding).
- ▶ Allow us to focus on JVEB.

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\forall b, \quad \left| \hat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

- ▶ Faster convergence in smaller variance regime.

Main results: fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Theorem 1. After implementing this oracle, there is an experiment-free policy π achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T}\right)$$

with $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

Main results: fully observed HOB

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Theorem 1. After implementing this oracle, there is an experiment-free policy π achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T}\right)$$

with $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

- ▶ d : dimension of $\theta_* \in \mathbb{R}^d$;
- ▶ Δ : error parameter from oracle.

Examples of oracle

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Define $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

Examples of oracle

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Define $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

(I.i.d. HOB) $G_t \equiv G$. By Bernstein's concentration, $\delta_t = \tilde{O}\left(\frac{1}{\sqrt{t}}\right)$, $\Delta = \tilde{O}(1)$.

► Regret $R(\pi) = \tilde{O}(\sqrt{dT})$.

Examples of oracle

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Define $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

(I.i.d. HOB) $G_t \equiv G$. By Bernstein's concentration, $\delta_t = \tilde{O}\left(\frac{1}{\sqrt{t}}\right)$, $\Delta = \tilde{O}(1)$.

► Regret $R(\pi) = \tilde{O}(\sqrt{dT})$.

(Linear HOB) $m_t = \varphi_*^\top x_t + \varepsilon_t$ with "light-tail" noise. We show $\Delta = \tilde{O}(d)$.

► Regret $R(\pi) = \tilde{O}(d\sqrt{T})$.

Examples of oracle

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Define $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

(I.i.d. HOB) $G_t \equiv G$. By Bernstein's concentration, $\delta_t = \tilde{O}\left(\frac{1}{\sqrt{t}}\right)$, $\Delta = \tilde{O}(1)$.

► Regret $R(\pi) = \tilde{O}(\sqrt{dT})$.

(Linear HOB) $m_t = \varphi_*^\top x_t + \varepsilon_t$ with "light-tail" noise. We show $\Delta = \tilde{O}(d)$.

► Regret $R(\pi) = \tilde{O}(d\sqrt{T})$.

Theorem 2. Even if G_t is known ($\Delta = 1$), it holds that $\inf_\pi \sup_\nu R(\pi; \nu) = \Omega(\sqrt{dT})$ with sup over all problem instances ν .

Examples of oracle

Full-info feedback: bidder observes m_t regardless of winning/losing (e.g. Google).

(HOB Estimation Oracle) For every time t , we have w.h.p.

$$\left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t \sqrt{G_t(b)(1 - G_t(b))} + \delta_t^2.$$

Define $\Delta = 1 + \sum_{t=1}^T \delta_t^2$.

(I.i.d. HOB) $G_t \equiv G$. By Bernstein's concentration, $\delta_t = \tilde{O}\left(\frac{1}{\sqrt{t}}\right)$, $\Delta = \tilde{O}(1)$.

► Regret $R(\pi) = \tilde{O}(\sqrt{dT})$.

(Linear HOB) $m_t = \varphi_*^\top x_t + \varepsilon_t$ with "light-tail" noise. We show $\Delta = \tilde{O}(d)$.

► Regret $R(\pi) = \tilde{O}(d\sqrt{T})$.

Theorem 2. Even if G_t is known ($\Delta = 1$), it holds that $\inf_\pi \sup_\nu R(\pi; \nu) = \Omega(\sqrt{dT})$ with sup over all problem instances ν . \implies matches upper bound up to Δ .

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(\mathbf{b}) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(\mathbf{b})^2 + \xi.$$

Theorem 3. After implementing this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(\mathbf{b}_t)^2$.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(\mathbf{b}) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(\mathbf{b})^2 + \xi.$$

- Bias and **bid-dependent** confidence from limited feedback.

Theorem 3. After implementing this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(\mathbf{b}_t)^2$.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(\mathbf{b}) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(\mathbf{b})^2 + \xi.$$

- Bias and **bid-dependent** confidence from limited feedback.

Theorem 3. After implementing this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(\mathbf{b}_t)^2$.

- E.g. $\delta_t(b_t) \asymp \frac{1}{\sqrt{1 + \#\{b_\tau \in I : \tau < t\}}}$ for some $I \subseteq [0, 1]$. Then $\Delta = \tilde{O}(1)$.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(\mathbf{b}) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(\mathbf{b})^2 + \xi.$$

- Bias and **bid-dependent** confidence from limited feedback.

Theorem 3. After implementing this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(\mathbf{b}_t)^2$.

- E.g. $\delta_t(b_t) \asymp \frac{1}{\sqrt{1 + \#\{b_\tau \in I : \tau < t\}}}$ for some $I \subseteq [0, 1]$. Then $\Delta = \tilde{O}(1)$.
- Use $\tilde{O}(\Delta)$ experiments.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(b) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(b)^2 + \xi.$$

- Bias and **bid-dependent** confidence from limited feedback.

Theorem 3. Given this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(b_t)^2$.

- **(I.i.d. HOB)** We find Oracle II such that $R(\pi) = \tilde{O}(d^{\frac{1}{3}} T^{\frac{2}{3}})$ with $\Delta = \tilde{O}(T^{\frac{1}{3}})$.

Main results: binary HOB

Binary feedback: bidder observes win/loss $\mathbb{1}[b_t \geq m_t]$.

(HOB Oracle II) For every time t , we have w.h.p.

$$\forall b, \quad \left| \widehat{G}_t(b) - G_t(b) \right| \leq \delta_t(b) \sqrt{G_t(b)(1 - G_t(b))} + \delta_t(b)^2 + \xi.$$

- Bias and **bid-dependent** confidence from limited feedback.

Theorem 3. Given this oracle, there is a policy π with **few** experiments achieving

$$R(\pi) = \tilde{O}\left(\sqrt{\Delta d T} + \xi T\right)$$

with “self-normalized” $\Delta = 1 + \sup_{b_1, \dots, b_T} \sum_{t=1}^T \delta_t(b_t)^2$.

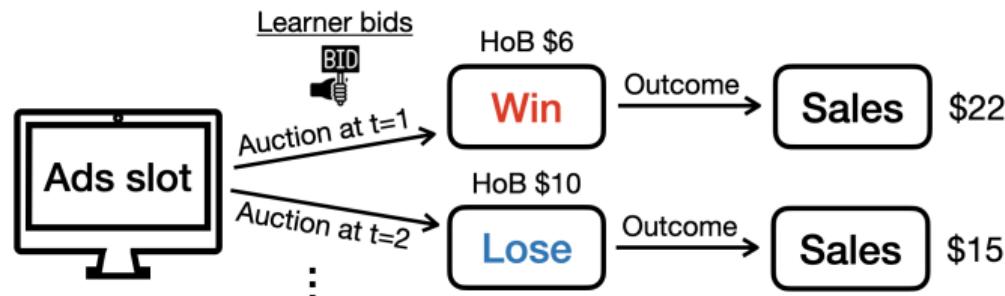
- **(I.i.d. HOB)** We find Oracle II such that $R(\pi) = \tilde{O}(d^{\frac{1}{3}} T^{\frac{2}{3}})$ with $\Delta = \tilde{O}(T^{\frac{1}{3}})$.
- Lower bound $\Omega(T^{\frac{2}{3}})$ even without contexts by Balseiro et al. (2019).

Talk Outline

- ▶ Introduction
- ▶ An “online + causal” formulation for JVEB
- ▶ Main results
- ▶ Summary and future directions

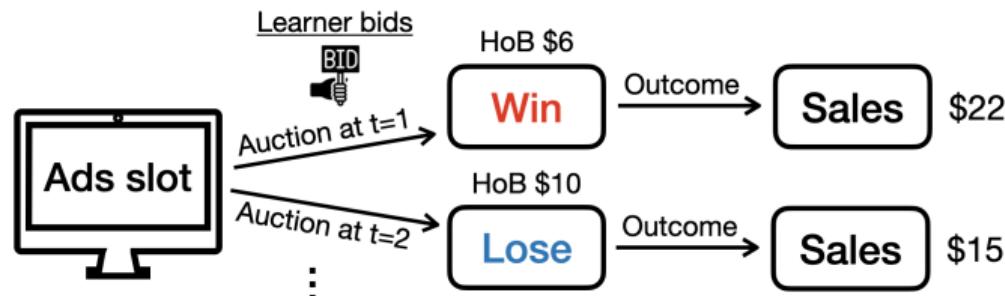
Summary

- User value = *marginal* gain from ad exposure



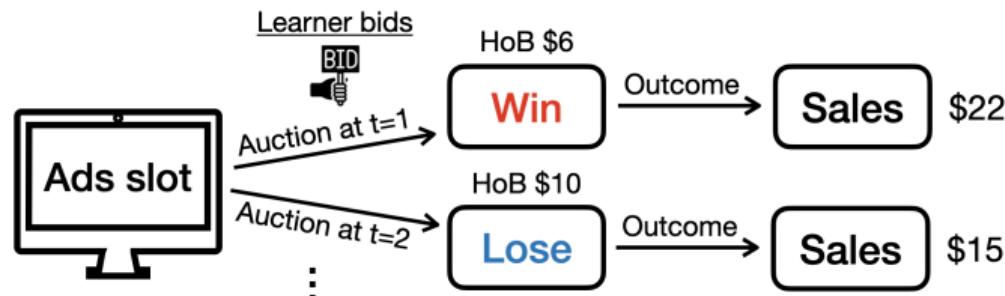
Summary

- User value = *marginal* gain from ad exposure
- Existing works cannot handle the two potential outcomes $v_{t,0}, v_{t,1} \implies$ wasteful bids.



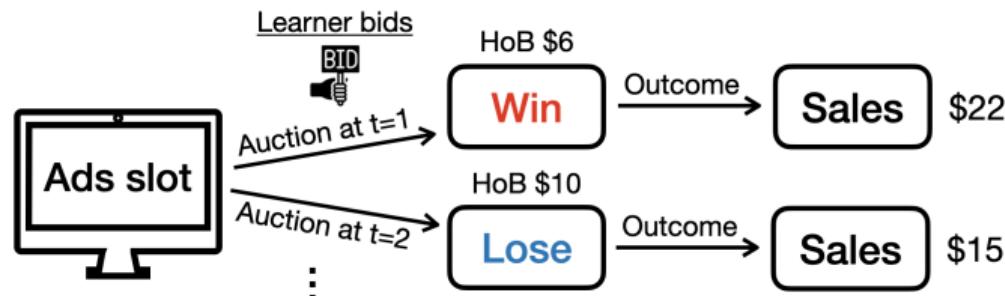
Summary

- ▶ User value = *marginal* gain from ad exposure
- ▶ Existing works cannot handle the two potential outcomes $v_{t,0}, v_{t,1} \implies$ wasteful bids.
- ▶ Bidders need an “online + causal” framework.



Summary

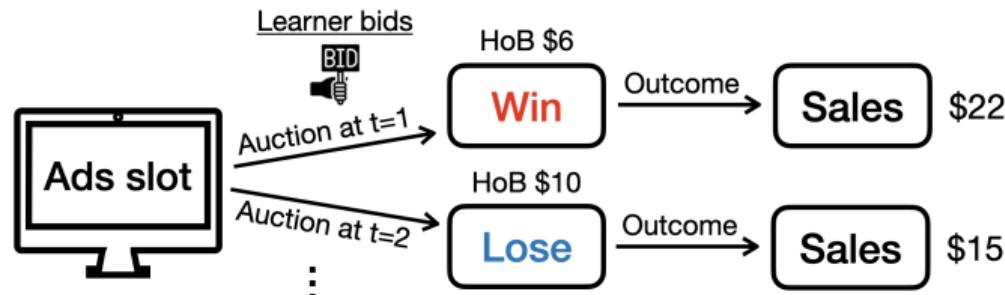
- ▶ User value = *marginal* gain from ad exposure
- ▶ Existing works cannot handle the two potential outcomes $v_{t,0}, v_{t,1} \implies$ wasteful bids.
- ▶ Bidders need an “online + causal” framework.
- ▶ Randomized experiments could be unnecessary (and suboptimal).



Summary

- User value = *marginal* gain from ad exposure
- Existing works cannot handle the two potential outcomes $v_{t,0}, v_{t,1} \implies$ wasteful bids.
- Bidders need an “online + causal” framework.
- Randomized experiments could be unnecessary (and suboptimal).

HOB feedback	Upper Bound	I.i.d. HOB	Linear HOB	Lower Bound
Full-info	$\sqrt{\Delta dT}$	\sqrt{dT}	$d\sqrt{T}$	\sqrt{dT}
Binary	$\sqrt{\Delta dT} + \xi T$	$d^{\frac{1}{3}} T^{\frac{2}{3}}$		



Future directions

- ▶ Is it optimal to consider oracle abstractions?
- ▶ Optimal regrets under second-price auctions (SPAs)?
- ▶ Bidding with constraints, e.g. budget, ROI, ...
- ▶ Bidding with delayed feedback?

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

- Unbounded variance: when $G_t(b_t) \rightarrow 0$ or 1 .

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

- ▶ Unbounded variance: when $G_t(b_t) \rightarrow 0$ or 1 .
- ▶ Unknown propensity.

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

- ▶ Unbounded variance: when $G_t(b_t) \rightarrow 0$ or 1 .
- ▶ Unknown propensity.

Bid experiments: Sample $b_t \sim \text{Unif}\{0, b_{\max}\}$ for learning θ_* , then commit.

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

- ▶ Unbounded variance: when $G_t(b_t) \rightarrow 0$ or 1 .
- ▶ Unknown propensity.

Bid experiments: Sample $b_t \sim \text{Unif}\{0, b_{\max}\}$ for learning θ_* , then commit.

- ▶ Regret = $\tilde{O}(T^{\frac{2}{3}})$ even under full-info HOB feedback (suboptimal).

θ_* estimation

Recall $\mathbb{E}[\nu_{t,1} - \nu_{t,0}] = \theta_*^\top x_t$; HOB CDF G_t .

Inverse-propensity-weighted (IPW) estimator: with propensity $G_t(b_t)$,

$$\widehat{\Delta\nu}_t(b_t) := \frac{\mathbb{1}[b_t \geq m_t]\nu_{t,1}}{G_t(b_t)} - \frac{\mathbb{1}[b_t < m_t]\nu_{t,0}}{1 - G_t(b_t)}, \quad \mathbb{E}[\widehat{\Delta\nu}_t(b_t)] = \theta_*^\top x_t.$$

Challenges:

- ▶ Unbounded variance: when $G_t(b_t) \rightarrow 0$ or 1 .
- ▶ Unknown propensity.

Bid experiments: Sample $b_t \sim \text{Unif}\{0, b_{\max}\}$ for learning θ_* , then commit.

- ▶ Regret = $\tilde{O}(T^{\frac{2}{3}})$ even under full-info HOB feedback (suboptimal).

Exploration-exploitation trade-off: exploratory $G_t(b_t)$ v.s. good b_t

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)}\rho_t$.
- $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)})\rho_t$.

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)}\rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)})\rho_t$.

(UCB algo) Choose $b_{t,0} = \arg \max_b \widehat{r}_{t,0}(b) + w_{t,0}(b)$, suffer regret $w_{t,0}(b_t)$.

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)}\rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)})\rho_t$.

(UCB algo) Choose $b_{t,0} = \arg \max_b \widehat{r}_{t,0}(b) + w_{t,0}(b)$, suffer regret $w_{t,0}(b_t)$.

Choose $b_{t,1} = \arg \max_b \widehat{r}_{t,1}(b) + w_{t,1}(b)$, suffer regret $w_{t,1}(b_t)$.

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)}\rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)})\rho_t$.

Bad estimation when variance large

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)} \rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)}) \rho_t$.

Bad estimation when variance large $\implies \textcolor{red}{G_t(b_t)^{-1}}$ or $(1 - G_t(b_t))^{-1}$ large

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)} \rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)}) \rho_t$.

Bad estimation when variance large $\implies \textcolor{red}{G_t(b_t)^{-1}}$ or $(1 - G_t(b_t))^{-1}$ large
 \implies accurate $\widehat{r}_{t,0}(b_t)$ or $\widehat{r}_{t,1}(b_t)$

Better of two Upper Confidence Bounds (UCBs)

Linear Regression: $|\widehat{\theta}_t^\top x_t - \theta_*^\top x_t| \lesssim \rho_t$.

Given x_t , maximize reward

$$r_t(b) = \underbrace{G_t(b)(\theta_*^\top x_t - b)}_{=: r_{t,0}(b)} + \mathbb{E}[v_{t,0}] = \underbrace{(1 - G_t(b))(b - \theta_*^\top x_t) - b}_{=: r_{t,1}(b)} + \mathbb{E}[v_{t,1}].$$

Plug-in estimators

$$\widehat{r}_{t,0}(b) := G_t(b)(\widehat{\theta}_t^\top x_t - b), \quad \widehat{r}_{t,1}(b) := (1 - G_t(b))(b - \widehat{\theta}_t^\top x_t) - b$$

Confidence widths:

- ▶ $|r_{t,0}(b) - \widehat{r}_{t,0}(b)| \lesssim w_{t,0}(b) := \textcolor{red}{G_t(b)} \rho_t$.
- ▶ $|r_{t,1}(b) - \widehat{r}_{t,1}(b)| \lesssim w_{t,1}(b) := (\textcolor{red}{1 - G_t(b)}) \rho_t$.

Bad estimation when variance large $\implies \textcolor{red}{G_t(b_t)^{-1}}$ or $(1 - G_t(b_t))^{-1}$ large
 \implies accurate $\widehat{r}_{t,0}(b_t)$ or $\widehat{r}_{t,1}(b_t)$ \implies "correct" UCB: $\min\{w_{t,0}(b_t), w_{t,1}(b_t)\}$.