Debiasing Evaluations That Are Biased by Evaluations

Jingyan Wang
Carnegie Mellon/Georgia Tech

"Debiasing Evaluations That Are Biased by Evaluations"
Jingyan Wang, Ivan Stelmakh, Yuting Wei, Nihar B. Shah
AAAI 2021

Motivation 1: teaching evaluation

- Students are asked to rate instructors' teaching effectiveness
- Correlation between ratings vs. teaching quality can be negative
  [Carrell & West, 2008; Braga et al., 2014; Boring et al., 2016]
- Highly biased by grading leniency:

> *"...the* ***effects of grades*** *on teacher–course evaluations are both* ***substantively and statistically important****..."*

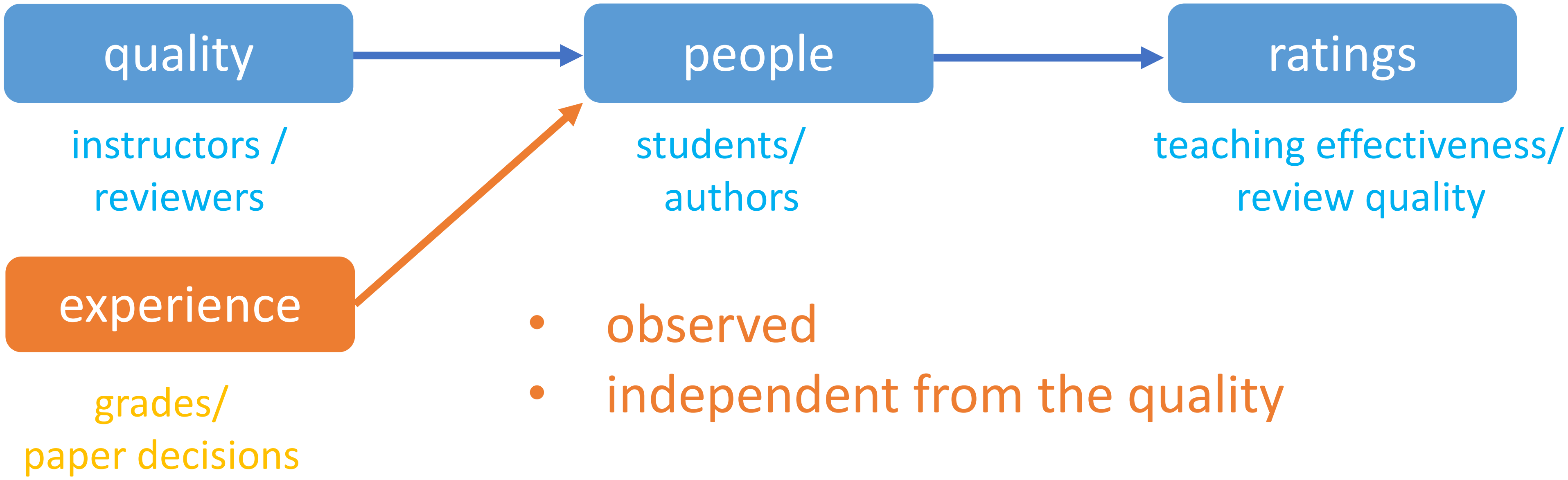
[Johnson, 2003]

Motivation 2: peer review

- Authors are asked to rate the reviews they receive
- Highly biased by positiveness of reviews: [Weber et al., 2002; Papagiannaki, 2007; Khosla, 2013]

> *"Satisfaction [of the author with the review] had a* ***strong, positive association with acceptance of the manuscript*** *for publication... Quality of the review of the manuscript was not associated with author satisfaction."*
>
> [Weber et al., 2002]

High-level problem

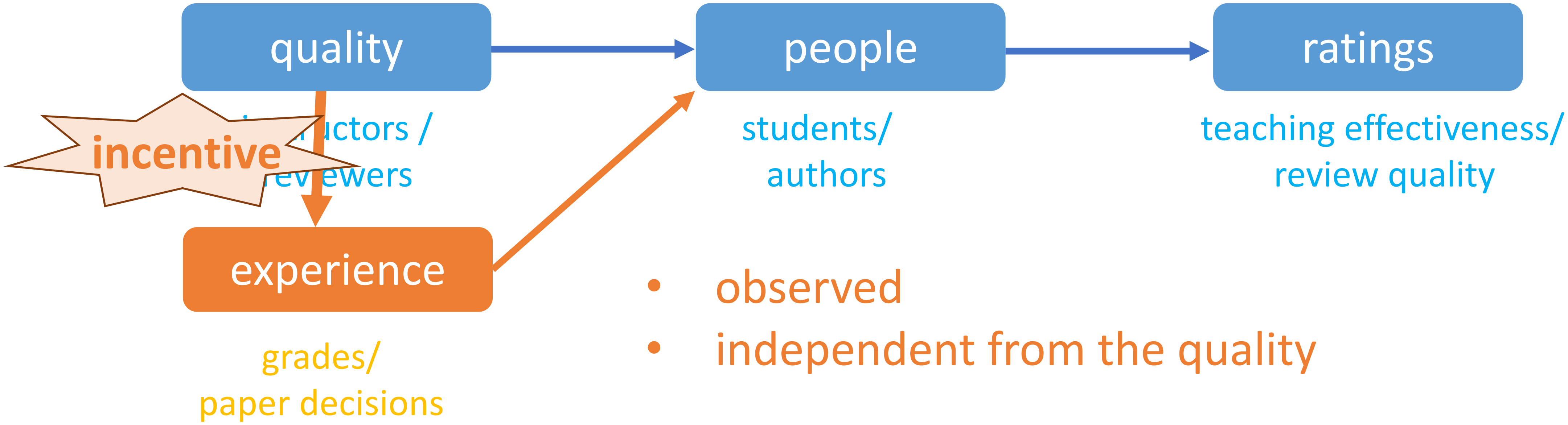


- observed
- independent from the quality

Unfair for rigorous and strict instructors

This work: correct experience-induced bias

Incentives



- observed
- independent from the quality

Introduce incentives for inflating grades, reducing content, “teaching to test” etc. [Carrell & West, 2008; Braga et al., 2014]

“... instructors can often ***double their odds*** *of receiving high evaluations from students simply by awarding A’s rather than B’s or C’s.”* [Johnson, 2003]

This work: Correcting experience-induced bias reduces such incentives.

Problem formulation

- n courses to evaluate: unknown true quality x_i^* for $i \in [n]$
- d students per course
- Student $j \in [d]$ in course $i \in [n]$ gives ratings:

$$y_{ij} = x_i^* + \text{bias} + \text{noise}$$

- **Noise:** iid zero-mean normal
- **Bias:** marginally distributed as normal

The observed experience gives structural information about the bias

- Higher grades → better ratings

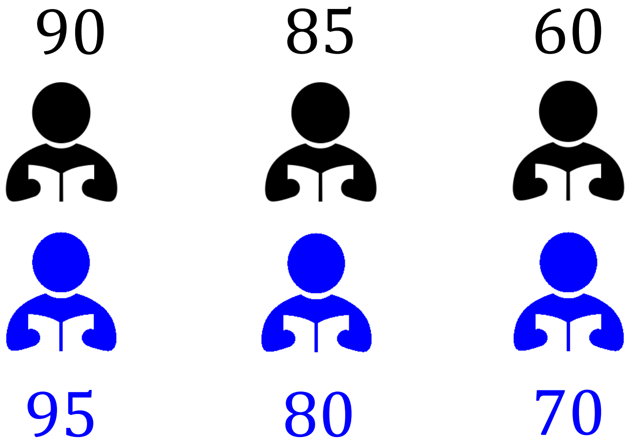
Problem formulation

Example 1: total ordering of grades

$n = 2, \;\; d = 3$

Course 1 (x_1^*)

Course 2 (x_2^*)



Bias: $b_{95} \geq b_{90} \geq b_{85} \geq b_{80} \geq b_{70} \geq b_{60}$

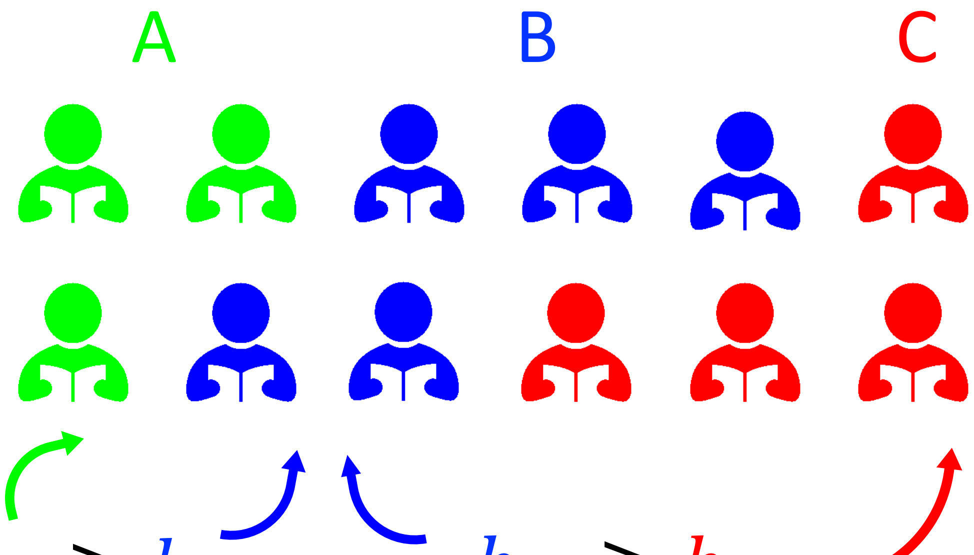
Problem formulation

Example 2: partial ordering of grades

$n = 2, \; d = 6$

Course 1 (x_1^*)

Course 2 (x_2^*)



Bias: $b_A \geq b_B \quad b_B \geq b_C$

Ratings: $Y = x1^T + B + \text{noise}$

Goal: estimate x^* (given Y and ordering)

Proposed estimator

$$\hat{x}^{(\lambda)} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad \min_{B \text{ obeys ordering}} \quad \|Y - x1^T - B\|_F^2 + \lambda \|B\|_F^2$$

Difference between raw ratings y vs. experience-corrected ratings $x + b$

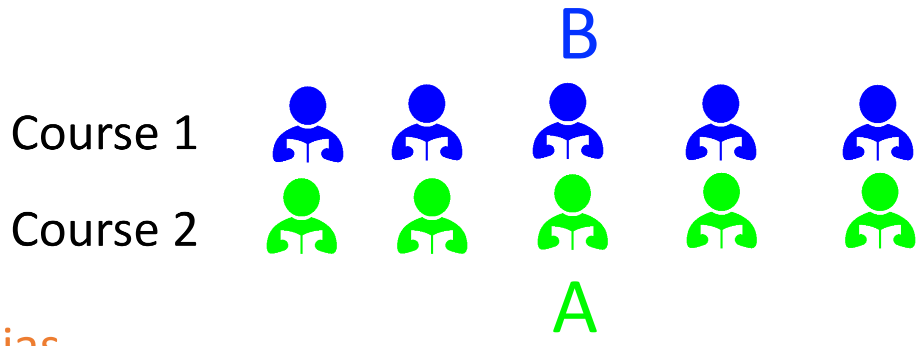
Regularization on magnitude of b

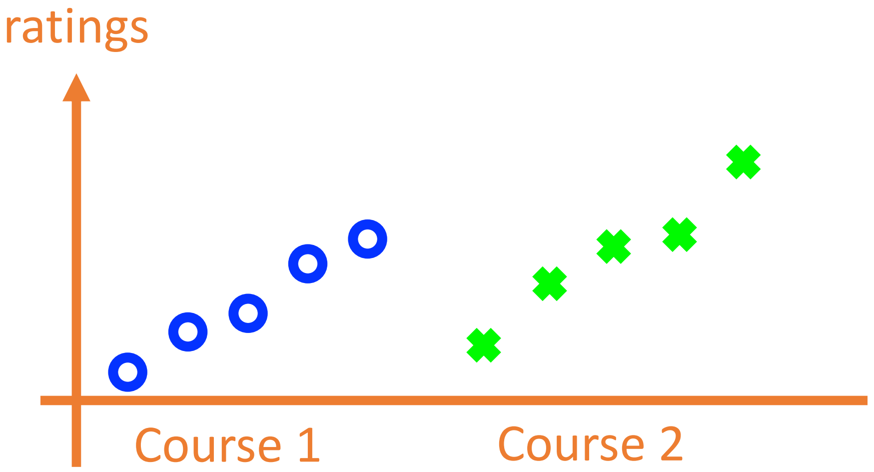
- Analyze two extremal cases: $\lambda = 0$ and $\lambda = \infty$
- Choose λ based on the data

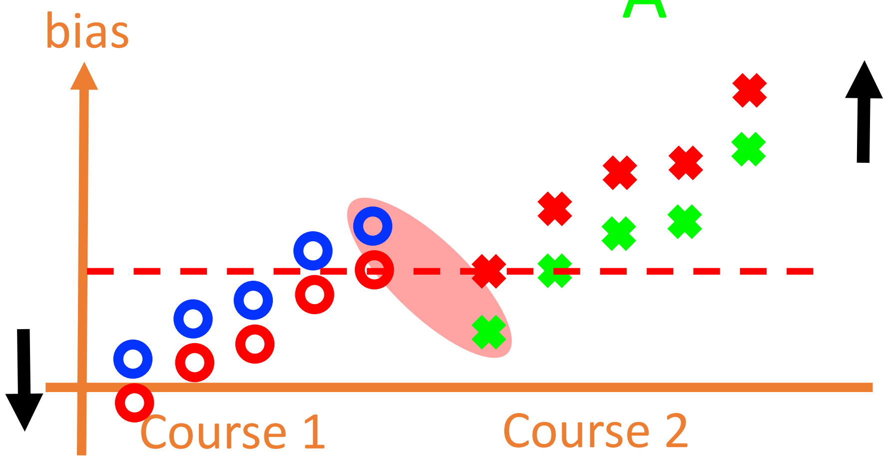
Extremal case 1: $\lambda = 0$

$$\hat{x}^{(\lambda)} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \ \min_{\substack{B \text{ obeys} \\ \text{ordering}}} \ \|Y - x1^T - B\|_F^2 + \lambda\|B\|_F^2$$

- No regularization means we "explain" the ratings as much as possible by B
- Closed-form solution







Extremal case 1: $\lambda = 0$

$$\hat{x}^{(\lambda)} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \min_{B \text{ obeys ordering}} \|Y - x1^T - B\|_F^2 + \lambda \|B\|_F^2$$

- No regularization means we "explain" the ratings as much as possible by B
- Closed-form solution
- Works well when there is no/little noise

Theorem 1 (informal). Our estimator (with $\lambda = 0$) is consistent when there is no noise.

- Sample mean is not consistent

Extremal case 2: $\lambda \to \infty$

$$\hat{x}^{(\lambda)} \in \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \quad \min_{B \text{ obeys ordering}} \quad \|Y - x1^T - B\|_F^2 + \lambda \|B\|_F^2$$

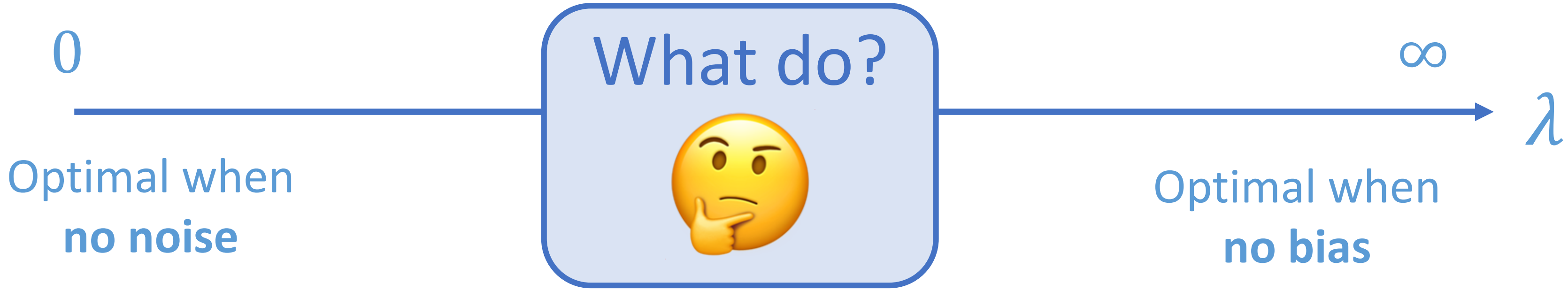
- $\mathrm{B} \approx 0$
- $\hat{x}^{(\infty)} \approx \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \|Y - x1^T\|_F^2 =$ taking sample mean
- Formally, define $\hat{x}^{(\infty)} = \lim_{\lambda \to \infty} \hat{x}^{(\lambda)}$

Theorem 2. $\hat{x}^{(\infty)}$ is equivalent to taking the sample mean.

- Our class of estimators includes one of the most commonly-used methods
- Minimax optimal when there is no bias. [Wainwright 2019]

Choosing λ

- $\lambda = 0$ and $\lambda = \infty$ work well respectively when there is no noise and no bias.



Challenge: don't know the amount of bias vs. noise

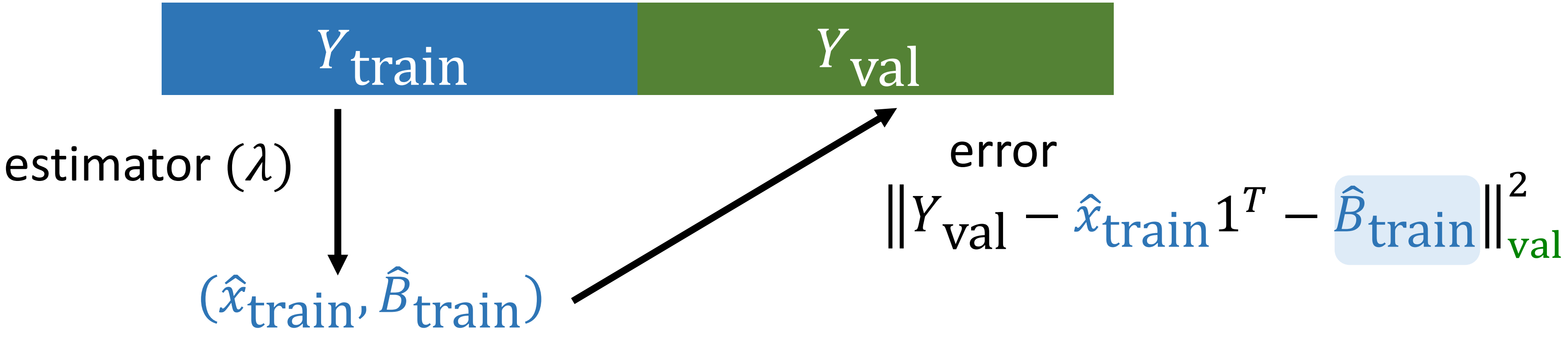
Idea: carefully design a cross-validation algorithm to choose λ

Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a “balanced” way

Algorithm (sketch)

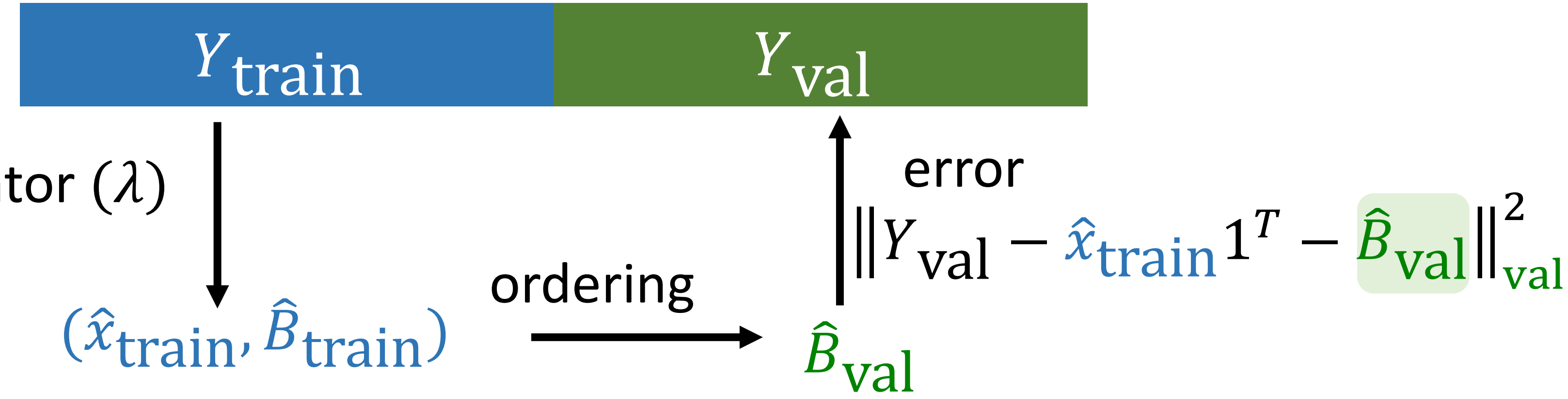
1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



Challenge: different bias on different individuals

Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



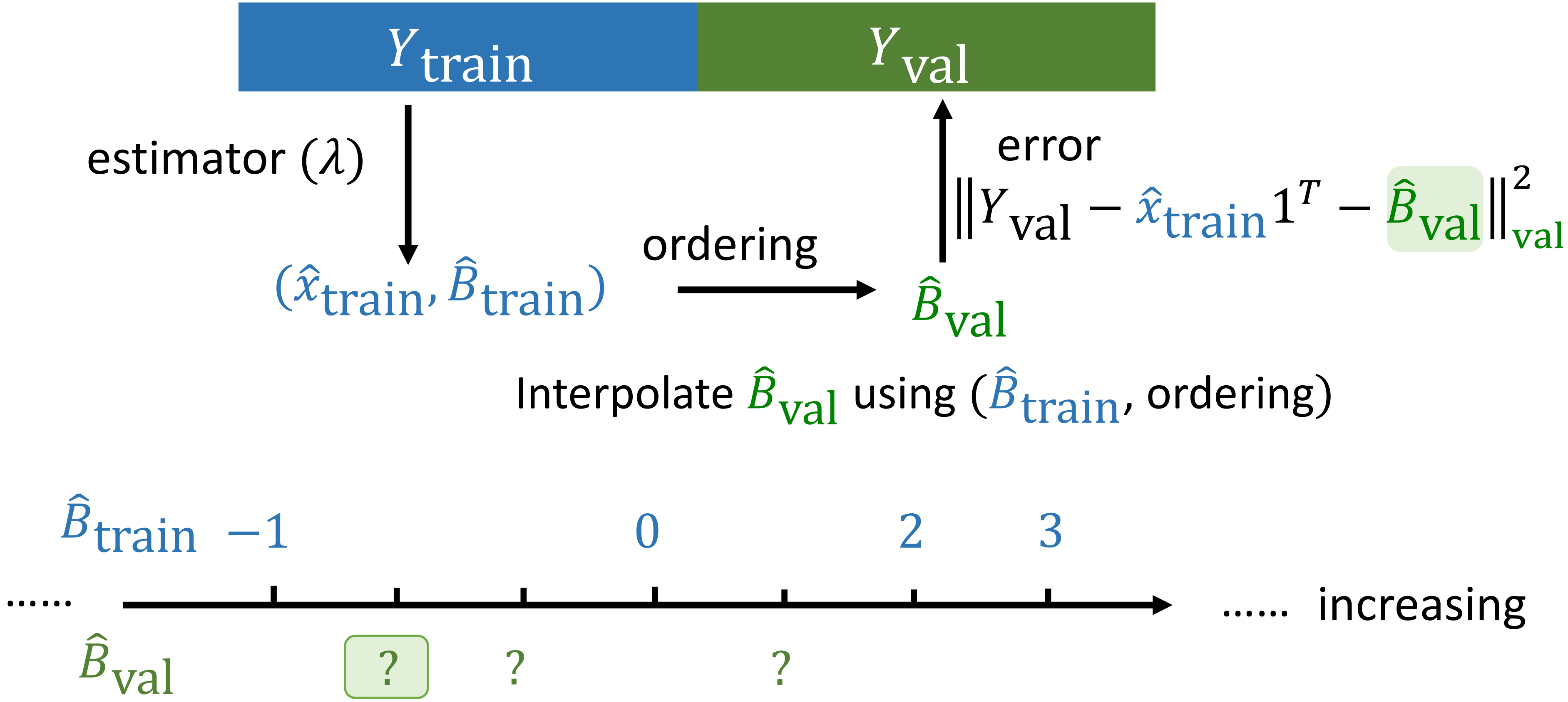
Interpolate $\hat{B}_{\text{val}}$ using $(\hat{B}_{\text{train}}$, ordering)

Challenge: different bias on different individuals 😔

Idea: interpolate train bias → val bias 🙂

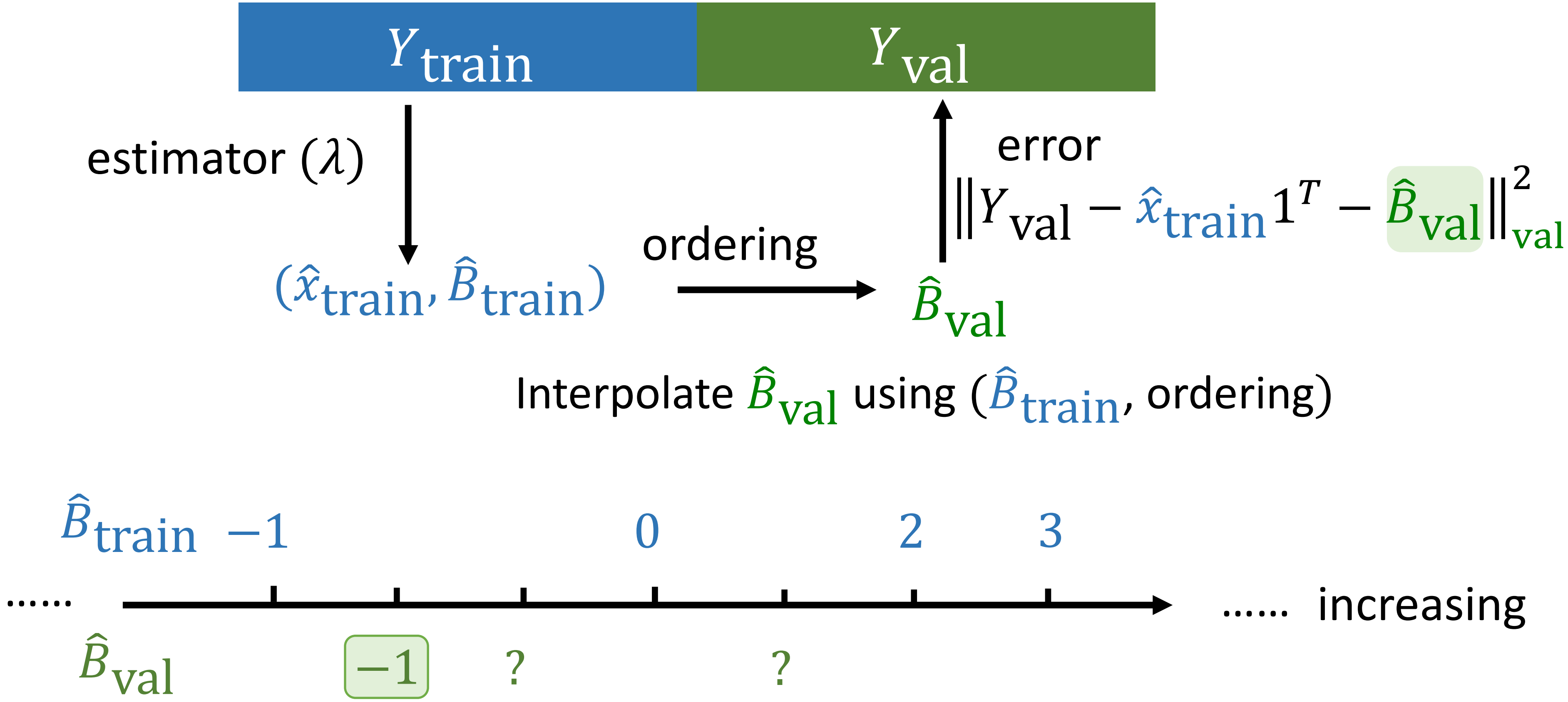
Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



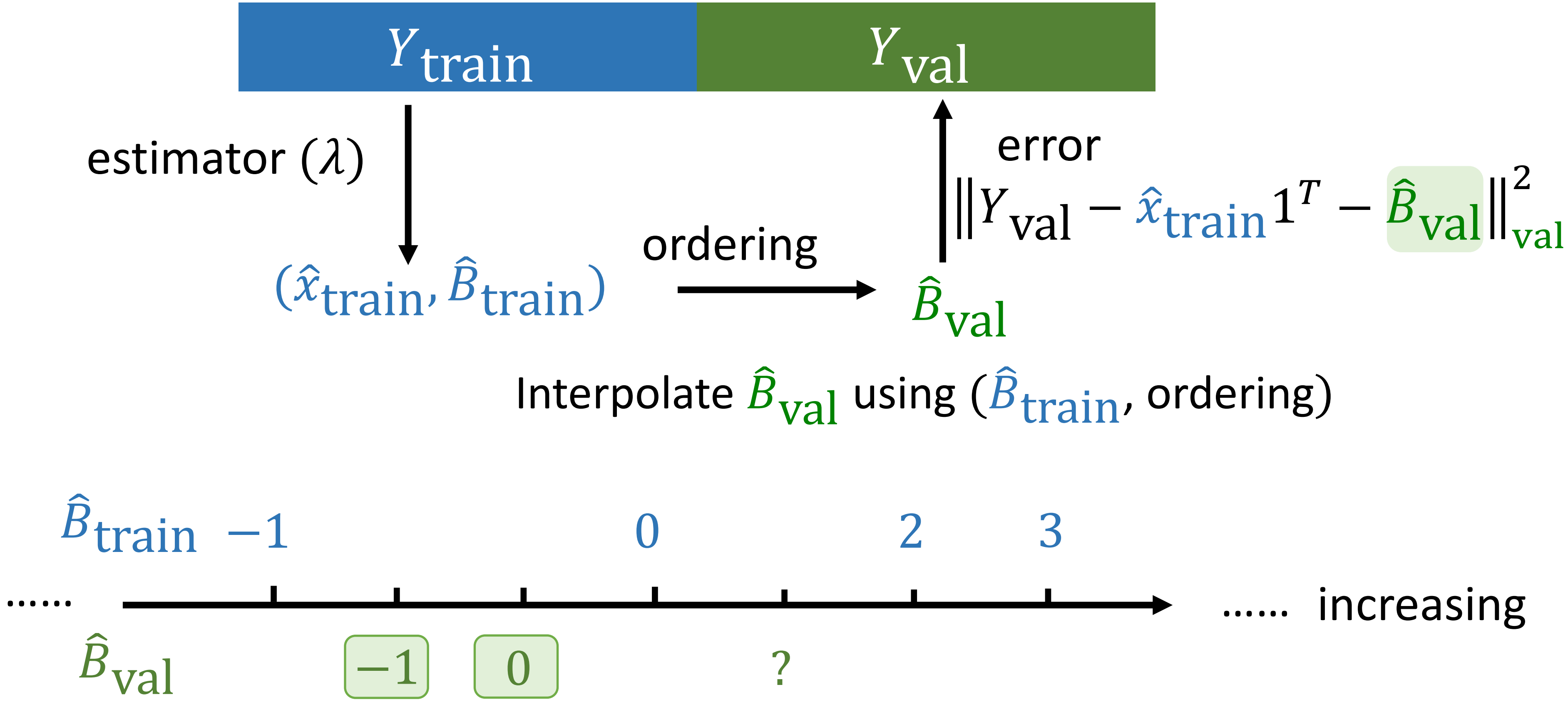
Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



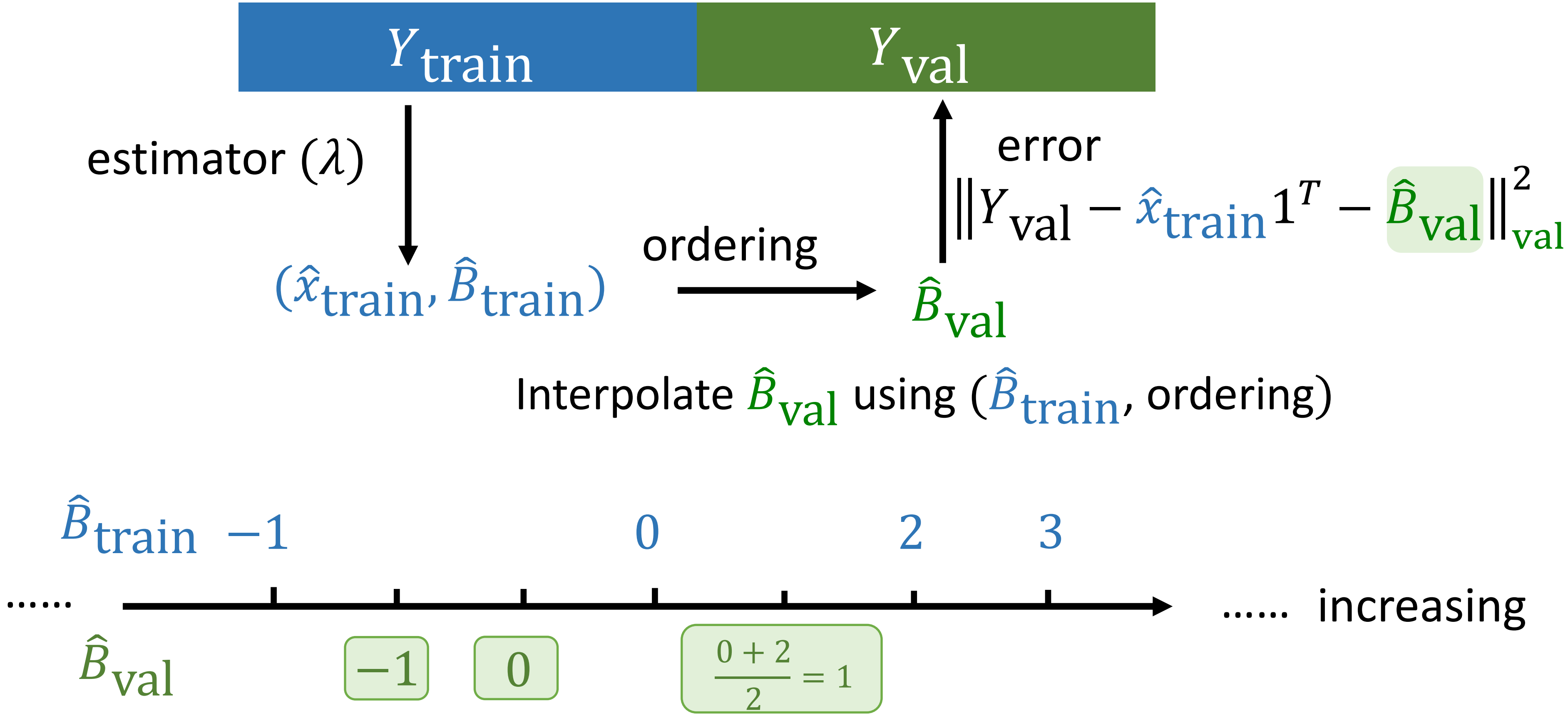
Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



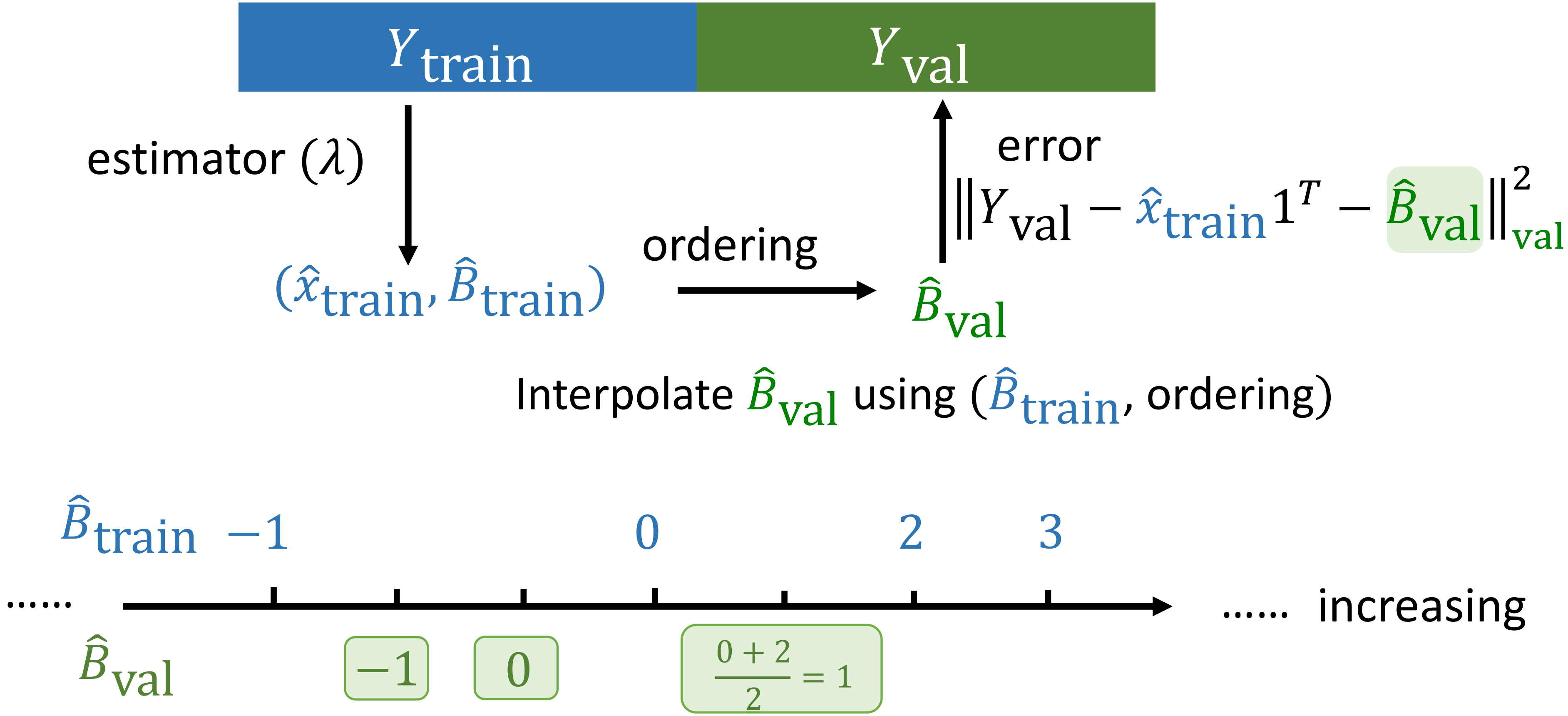
Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a "balanced" way
2. **Compute** validation error for each λ



Algorithm (sketch)

1. **Split** data to $(Y_{\text{train}}, Y_{\text{val}})$ in a “balanced” way
2. **Compute** validation error for each λ
3. **Choose** λ that minimizes the validation error



Theoretical guarantees

Theorem 3 (informal). In cases of common partial orderings,

- when there is **no noise**, we have

$$\hat{x}_{CV} \to \hat{x}^{(0)};$$

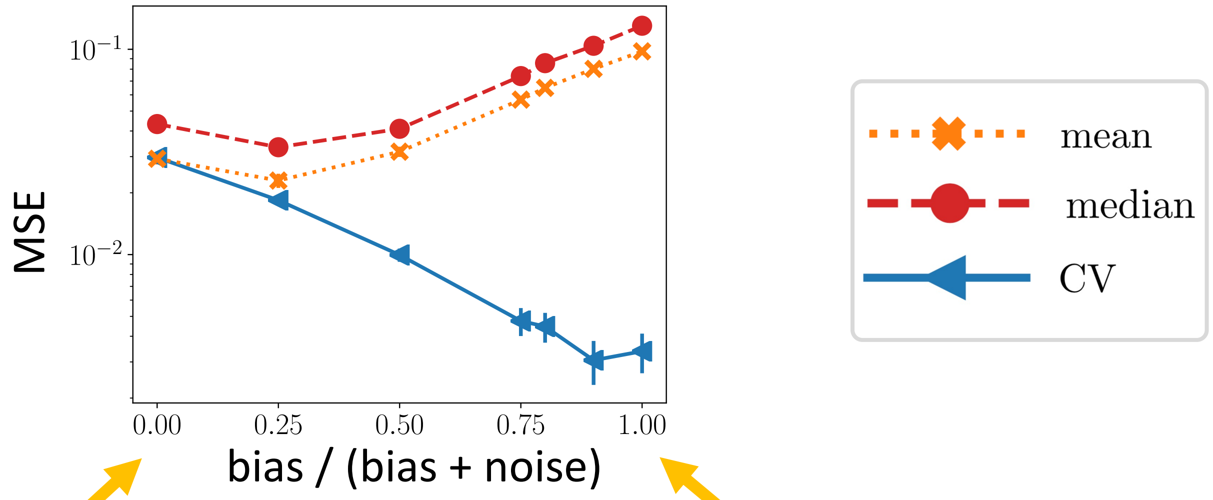
- when there is **no bias**, we have

$$\hat{x}_{CV} \to \hat{x}^{(\infty)}.$$

Our cross-validation successfully recovers the two extremal cases.

Experiment

- Indiana University Bloomington
- 10 sessions of a course
- Simulate bias and noise using real grading statistics



no bias:

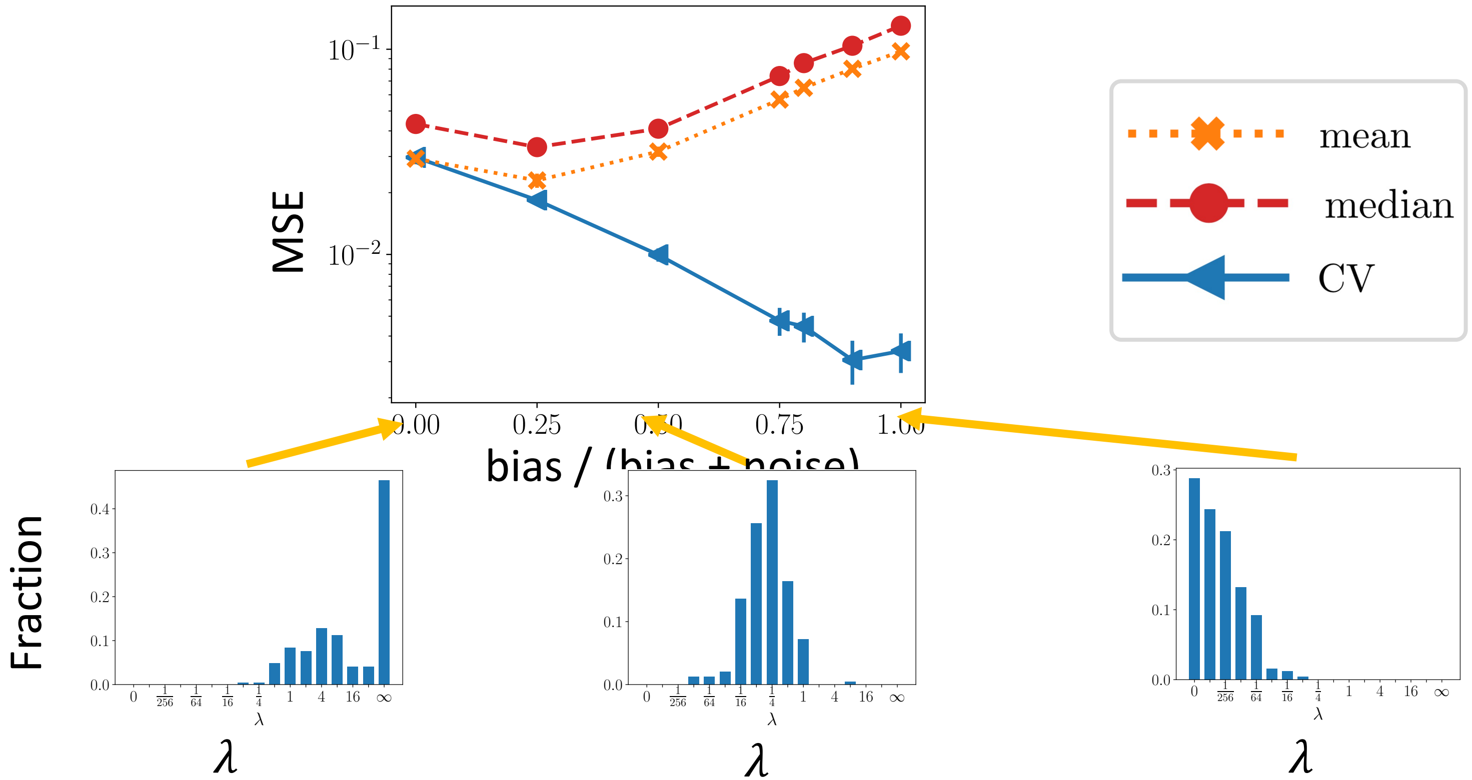
all estimators work well

lots of bias:

our estimator significantly better than {mean, median}

Experiment

- Indiana University Bloomington
- 10 sessions of a course
- Simulate bias and noise using real grading statistics



Take-aways

- Use an ordering constraint to model experience-induced bias, without making restrictive assumptions
- Design a novel CV algorithm to tease out bias vs noise

Future work

- Sharp statistical bounds on error rates / sample complexity + when there is both bias and noise
- Combining with a game-theoretic approach to design mechanisms

Thanks :)

jingyanw@cmu.edu