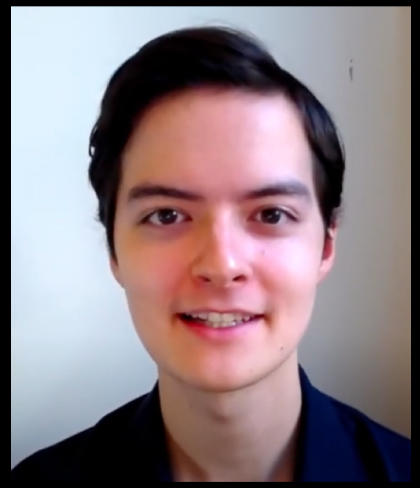
Estimation Beyond the IID Setting (Two Vignettes)

Gregory Valiant
Stanford

Part I
Worst Case Analysis for Randomly Collected Data

Justin Chen



Paul Valiant

Justin Chen, Gregory Valiant, Paul Valiant, *Worst Case Analysis for Randomly Collected Data, NeurIPS'20 (https://arxiv.org/abs/1911.03605)*

Warm-Up: Estimating Bias of a Coin (Worst-Case Optimal)

Guess the bias of my coin:

What guess minimizes the expected (squared) error for worst-case **p**?

1 flip: **0 → 1/4** **1 → 3/4**

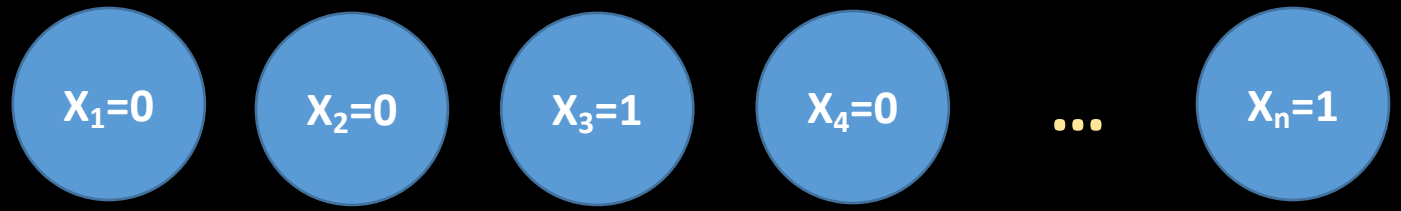
2 flips: **0,0 → 0.2071** **0,1 or 1,0 →1/2** **1,1, → 1- 0.2071**

k flips: $\text{mean(sample)}\frac{\text{sqrt(k)}}{\text{sqrt(k)+1}} + \frac{1}{2}\frac{1}{\text{sqrt(k)+1}}$ **[Hodges/Lehman, 1950]**

Estimating Bias of a Coin (Worst-Case Optimal)

Problem: Flip *k i.i.d.* coins with unknown bias *p*,
Goal: estimate *p* ***optimally*** (from perspective of worst case p)

Alternate formulation:



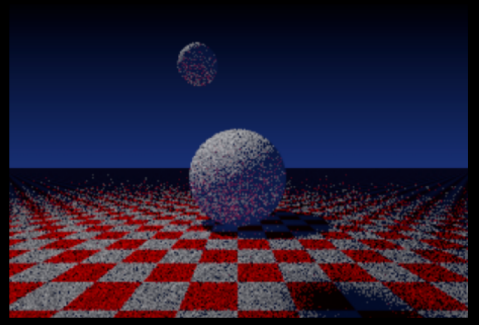
Observe a uniformly random set of *k:* | $X_{15} = 1, X_{19} = 1, X_{27} = 0, .. X_{145} = 1,$
Goal: estimate mean(X_1,…,X_n) so as to minimize worst-case expected error:

$$\underset{f:\ \text{obs}\to\text{pred}}{\text{Min}} \ \underset{x_1,\ldots,x_n}{\text{Max}} \ \underset{\text{obs}}{E} \ [\ (\text{prediction} - \text{mean}(x_1,\ldots,x_n)\ )^2\ ]$$

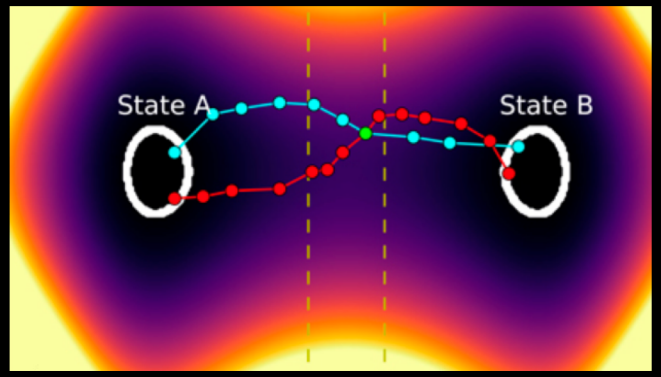
As n-> Infinity, turns into estimating bias of *k* iid flips of a coin.

What if observations are not a uniformly random set of size *k*?

Accurate estimation despite complex data collection processes

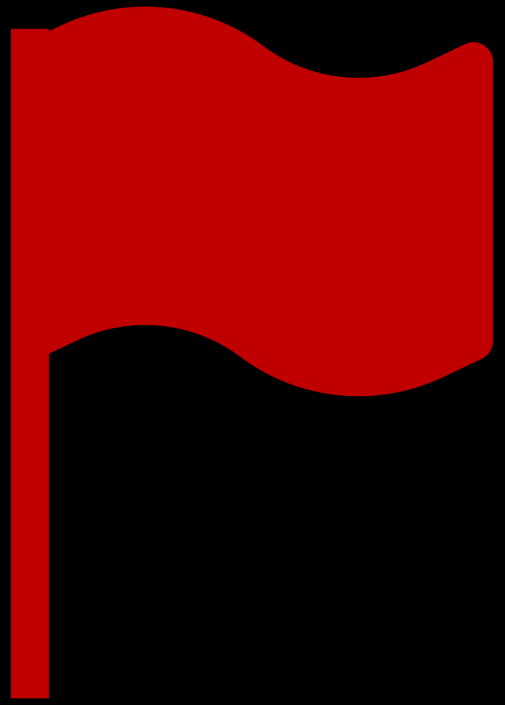


Non-uniform sampling
(e.g. importance sampling)



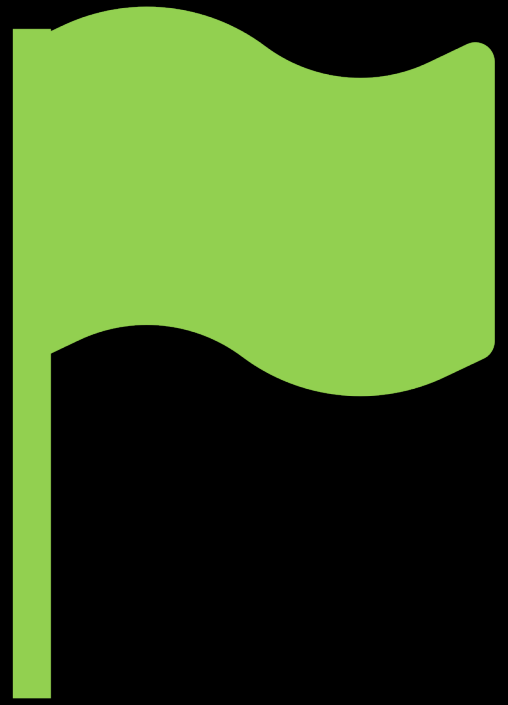
Dependent samples
(e.g. given by Markov process)

Sampling process based on underlying network (e.g. “Snowball sampling”/ “respondent-driven-sampling”)



Distributional assumptions about data **values**

Gaussian, i.i.d., exchangeable
Robust statistics

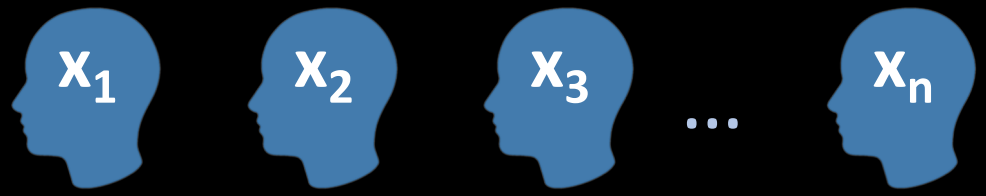


Leverage the data **collection** process

Without assumptions about the distribution of the data

Worst-Case Framework

n entities, each with a hidden value x_i (bounded real number)



Known distribution **P** over subsets of {1,2,…,n}

- Subset ***S*** $\subset$ {1,2,.., *n* } drawn from ***P***; observe ***S***, *values* x_S *indexed by* ***S***
- Goal: Estimate *mean*(x_1,..,x_n)

Simple Example

Political survey of 1000 people

P: Responds to survey w.p. 10% | Responds to survey w.p. 50%

Standard estimator: Weight each answer 5x | Weight each answer 1x

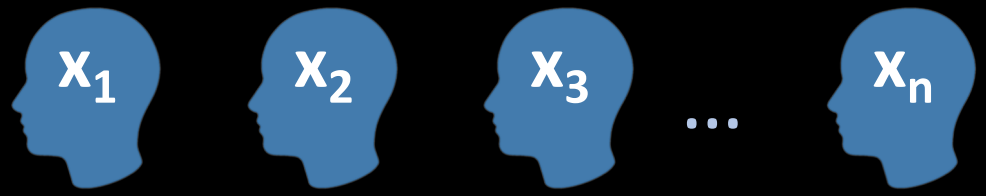
Return overall mean

Performance measure:
Worst-Case Expected Error

$$\min_{f} \max_{x_1,\ldots,x_n} \mathop{E}_{S \sim \mathbf{P}} \left[\left(f(P,S,x_S) - \text{mean}(x_1,..,x_n) \right)^2 \right]$$

Worst-Case Framework

n entities, each with a hidden value x_i (bounded real number)



Known distribution **P** over subsets of {1,2,…,n}

- Subset ***S*** $\subset$ {1,2,.., *n* } drawn from ***P***; observe ***S***, *values* x_S *indexed by* ***S***
- Goal: Estimate *mean*(x_1,..,x_n)

Two Main Tasks:

1) Given estimator f, compute worst-case expected error

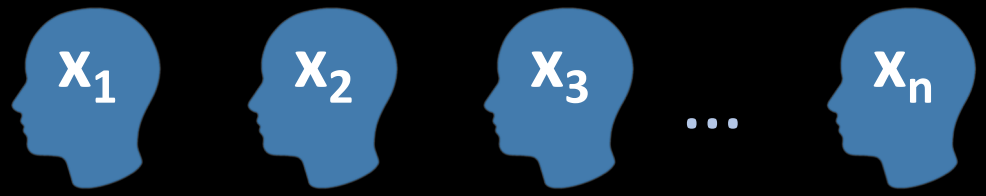
2) Find the best estimator f

Performance measure:
Worst-Case Expected Error

$$\min_{f} \max_{x_1,\ldots,x_n} \mathop{E}_{S \sim \mathbf{P}} [\, (\, f(P,S,x_S) - \text{mean}(x_1,..,x_n) \,)^2 \,]$$

Worst-Case Framework

n entities, each with a hidden value x_i (bounded real number)



Known distribution **P** over subsets of {1,2,…,n}

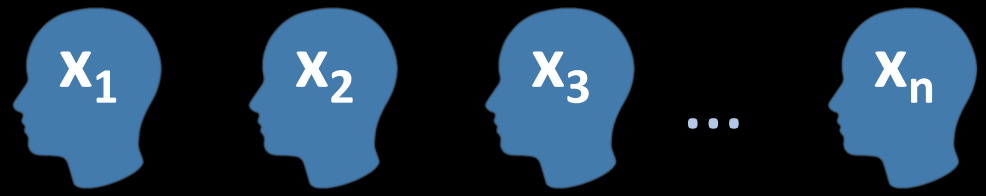
- Subset ***S*** ⊂{1,2,.., *n* } drawn from ***P***; observe ***S***, *values* x_S *indexed by* ***S***
- Goal: Estimate *mean*(x_1,..,x_n)

Performance measure:
Worst-Case Expected Error

$$\min_{f} \max_{x_1,\ldots,x_n} \mathop{E}_{(S,T)\sim P\,P} [\,(\,f(P,S,T,x_S) - \text{mean}(x_T)\,)^2\,]$$

Worst-Case Framework++

n entities, each with a hidden value x_i (bounded real number)



Known distribution **P** over **pairs of** subsets of…

- Subsets ***(S,T)*** ⊂{1,2,.., *n* } drawn from ***P***; observe ***S,T***, *values* x_S *indexed by* ***S***
- Goal: Estimate *mean*(x_T)

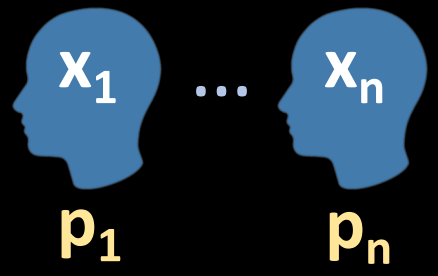
Performance measure:
Worst-Case Expected Error

***S*="Sample"**
***T*="Target"**

Illustrative Examples

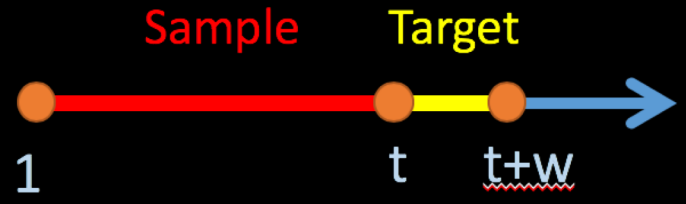
- Random sample of size *k* [ie estimating bias of a coin]
  P: uniformly random set of *k* individuals/coins appears in the sample
  Optimal Estimator for infinite *n*: [Hodges/Lehmann,1950]

- "Importance Sampling"
  P: each individual appears in the sample independently w.p. p_i



- "Snowball Sampling" (i.e. "Chain Referral Sampling")
  P: sample generated by a viral process on a social network

- "Selective Prediction" [e.g. Drucker'13, Qiao/Valiant'1[illegible]
  (Forecasting)
  P: samples corresponds to past data with prediction over future data



Main Results

$$\min_{f} \left[\max_{x_1,\ldots,x_n} \; \mathop{E}_{S \sim \mathbf{P}} \left[\left(f(P,S,x_S) - \text{mean}(x_1,..,x_n) \right)^2 \right] \right]$$

Thm 1: Given estimator *f*, in poly-time, with poly # samples from P, we can $\pi/2$-approximate the error of *f*.[†]

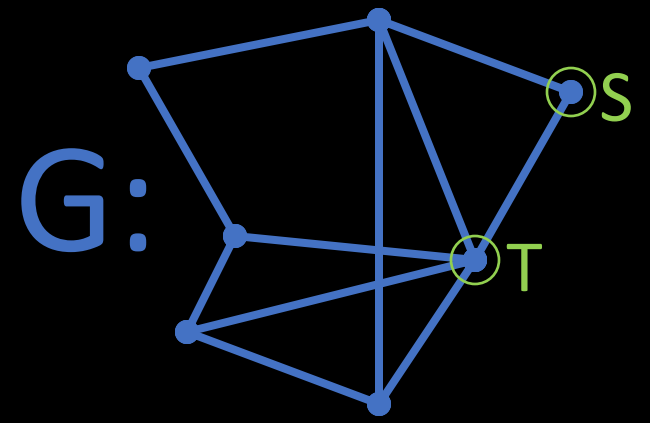
Thm 2: In poly-time, with poly # samples from P, we can find a $\pi/2$-optimal[†] estimator *f*.

[†]*We restrict f to the general class of "semilinear" estimators where* $f(P, S, x_S) = \langle a_{(P,S)}, x_S \rangle$

Techniques

Worst-Case Expected Error of *f:* $\max_{x_1,\ldots,x_n} \; \mathop{E}_{(S,T)\sim \mathbf{P}} [\,(\, f(P,S,T,x_S) - \text{mean}(x_T)\,)^2\,]$

e.g. Consider **P**: Given *n* node graph G,
Choose an edge (i,j) uniformly at random,
Let *S* = {i} *and* T = {j}. I.e. given x_i predict x_j.
Let *f* be the trivial algorithm that predicts $x_j = x_i$.



Worst-Case Expected Error of *f* = MAXCUT/|Edges|

NP-hard, but hope: const factor approx. via Goemans-Williamson SDP.

Key steps:

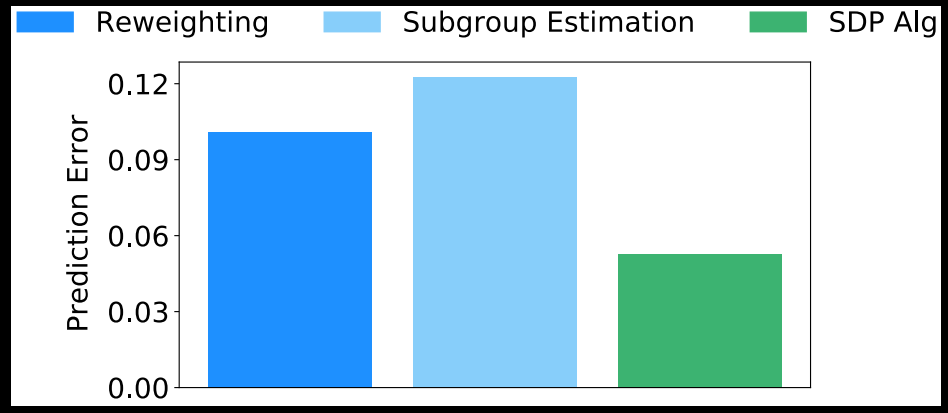
Exponential-sized SDP → subsampling techniques

Thm 2 "min max" → convex duality; analysis of PSD Grothendieck problem

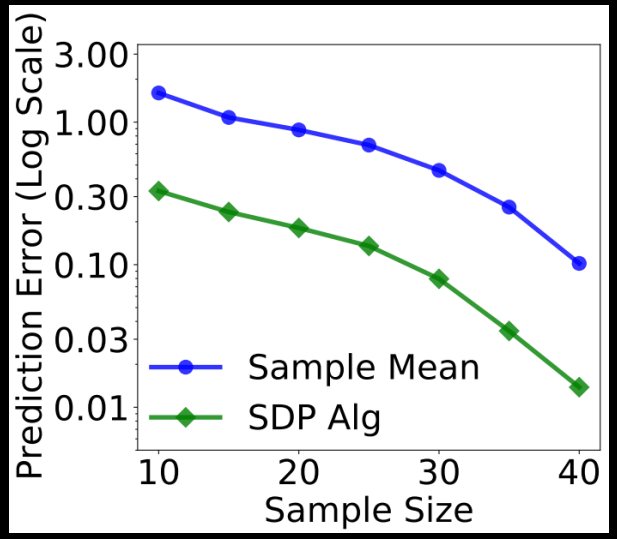
Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines

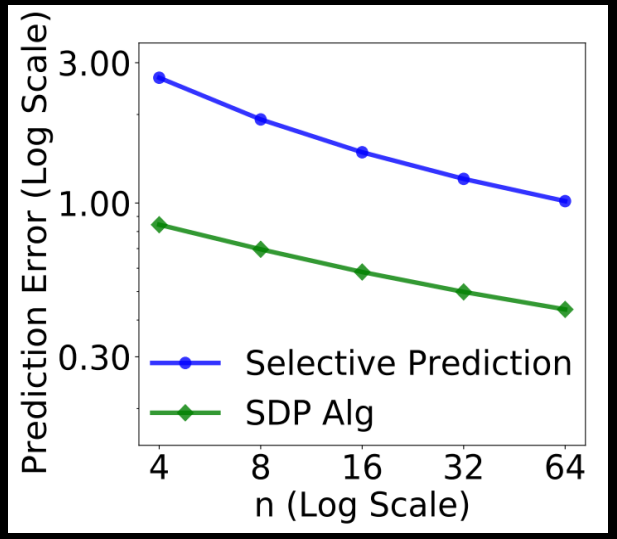
Importance Sampling



Snowball Sampling

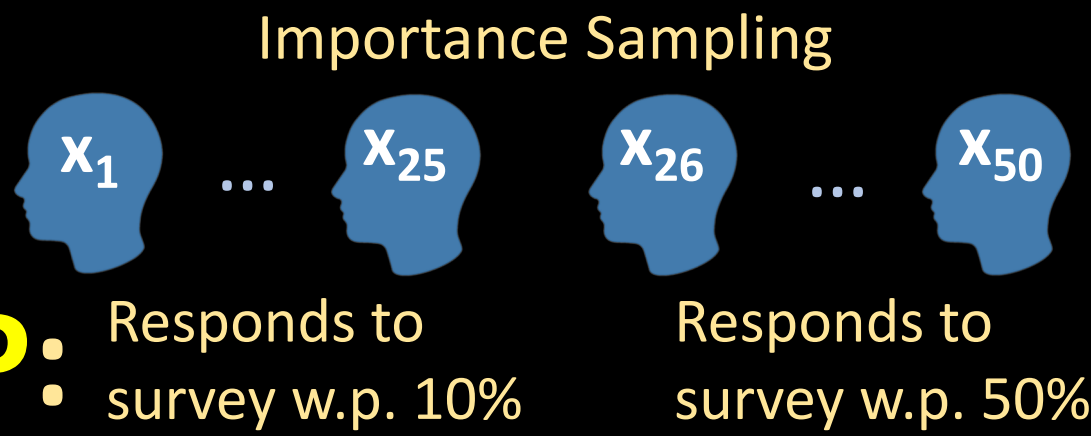


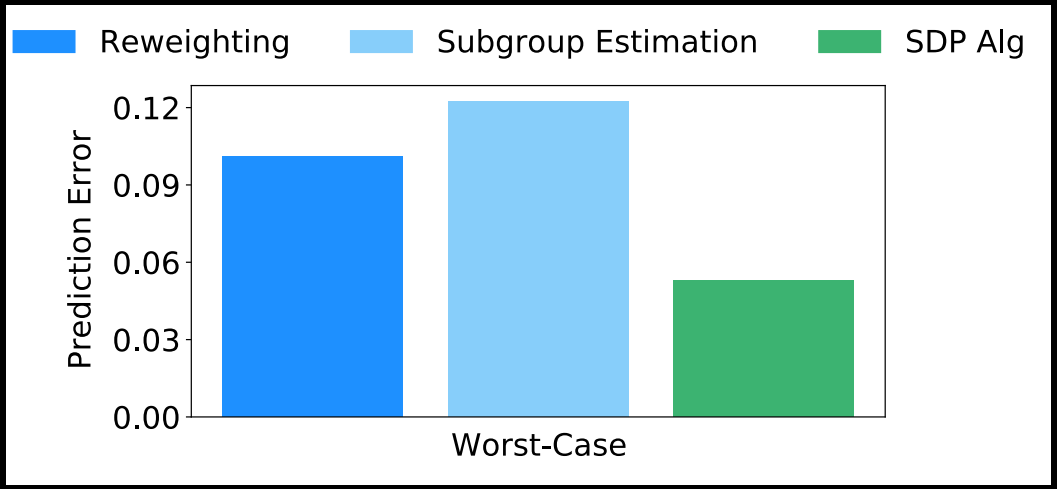
Forecasting



Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines



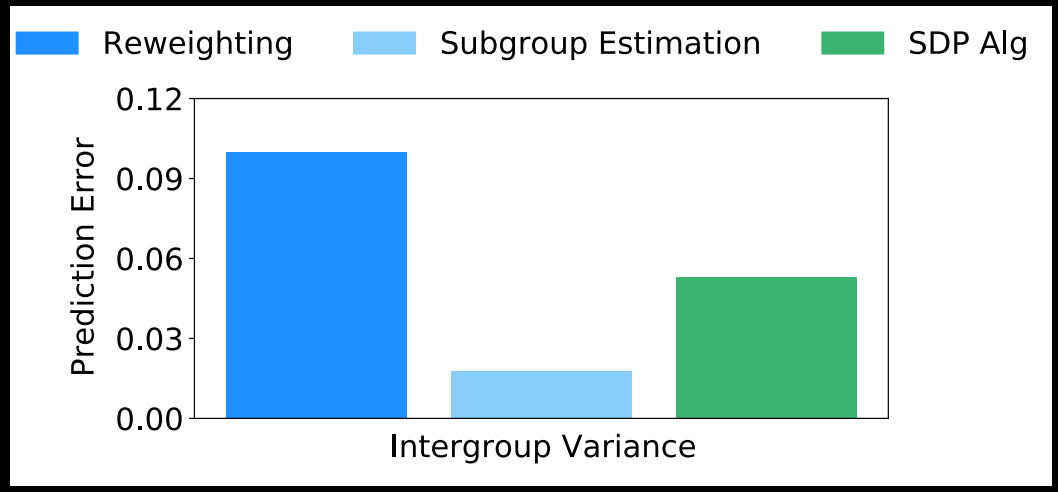


Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines

Importance Sampling

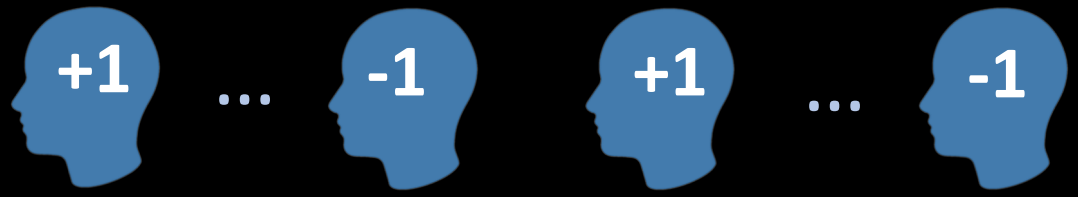
P: Responds to survey w.p. 10% | Responds to survey w.p. 50%



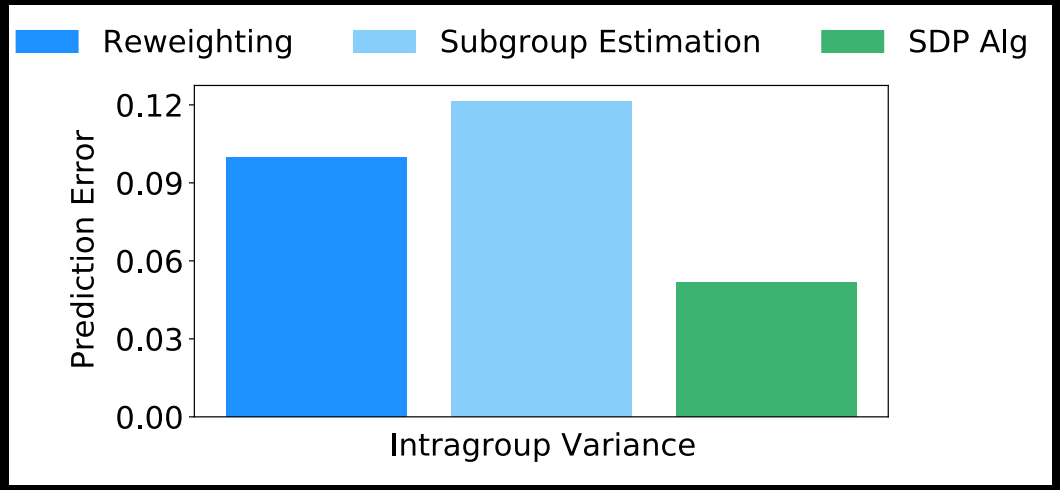
Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines

Importance Sampling



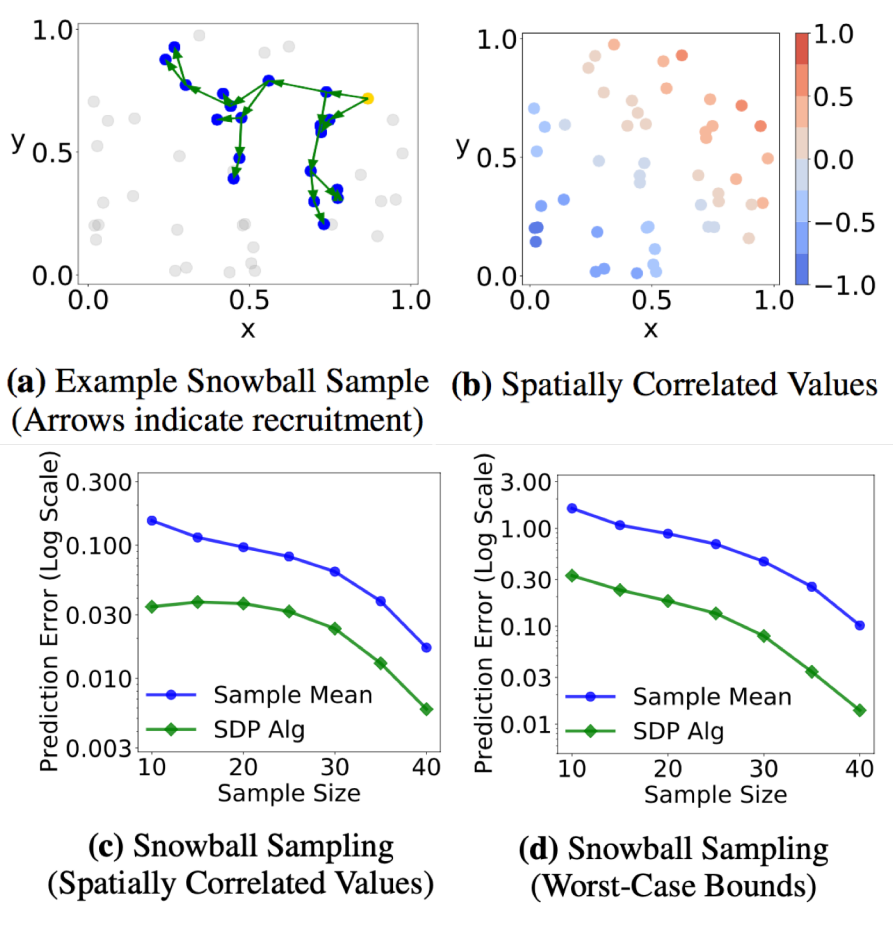
P: Responds to survey w.p. 10% | Responds to survey w.p. 50%



Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines

Snowball Sampling



(a) Example Snowball Sample (Arrows indicate recruitment)

(b) Spatially Correlated Values

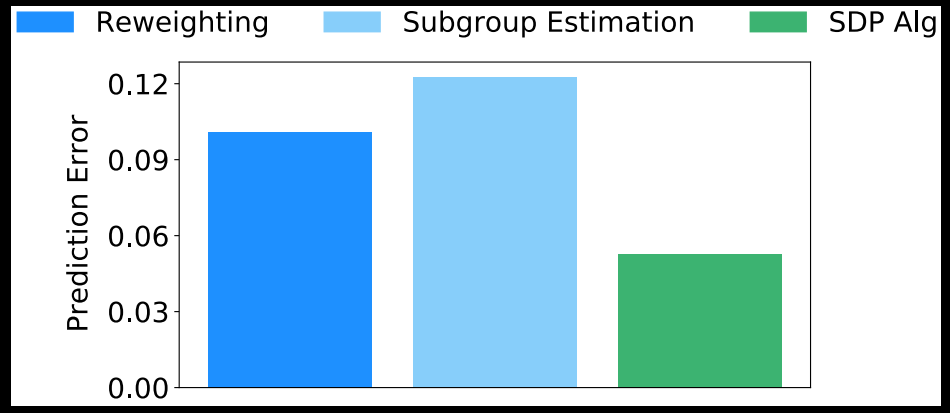
(c) Snowball Sampling (Spatially Correlated Values)

(d) Snowball Sampling (Worst-Case Bounds)

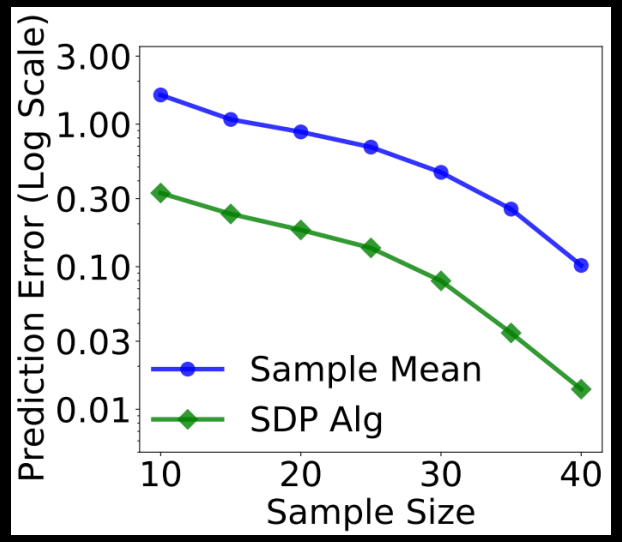
Experiments

- Comparison of worst-case mean squared error (MSE) in 3 very different settings
- Our algorithm (SDP Alg) gives 2-7x improvement over baselines
- Optimizing for worst-case expected error produces unintuitive estimators

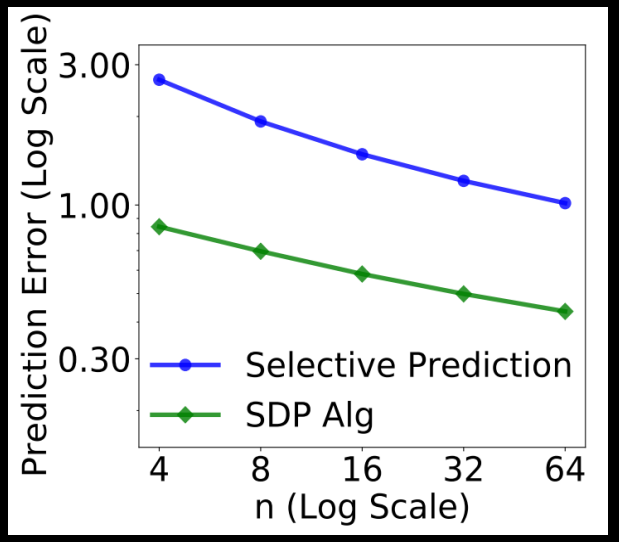
Importance Sampling



Snowball Sampling



Forecasting



Use Cases

1) We **know P**

- Evaluate and find estimators custom tailored to **P**

2) We **control P** (e.g. clinical trial or poll)

- Design sampling process to optimize estimability

3) We **don't know P**

- Evaluate stability of estimators with respect to plausible **P's**

Extensions and Open Questions

1) From mean-estimation to **regression?**

    1a) How to estimate other functionals?

2) Today: data values are bounded – optimize error in terms of $|x|_\infty$; analogous framework for L_2 norm has nice properties too.

3) New alg with performance depending on sample size not domain size n?

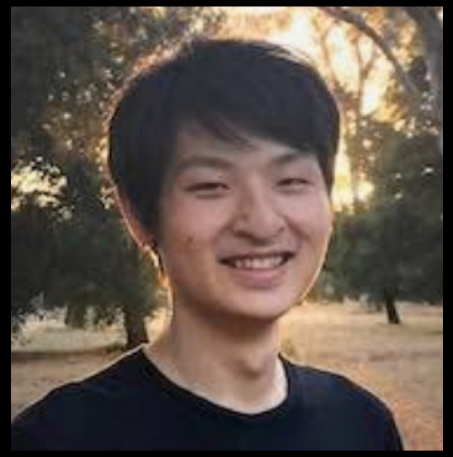
4) Instead of estimator depending on entire distribution D, is it enough to know low-order moments of D (correlation matrix? k-way marginals?)

5) Conjecture: The general class of algorithms we consider ("semilinear algs") is constant-factor optimal (1.004x gap is biggest we know)

6) New + practical blends of this model with more traditional models?

Part II
Stronger Calibration Lower Bounds via Sidestepping

Mingda Qiao



Mingda Qiao and Gregory Valiant. *Stronger Calibration Lower Bounds via Sidestepping,* STOC'21.

Setting: Sequential Binary Prediction

- Sequential game between "forecaster" and "nature/adversary"
- At each step t $= 1, 2, \ldots, T$:
  - Adversary picks $b(t) \in \{0,1\}$, forecaster predicts $p(t) \in [0,1]$
  - *b(t)* and *p(t)* can depend on (*b(1),p(1)*), (*b(2),p(2)*), …, (*b(t-1),p(t-1)*)

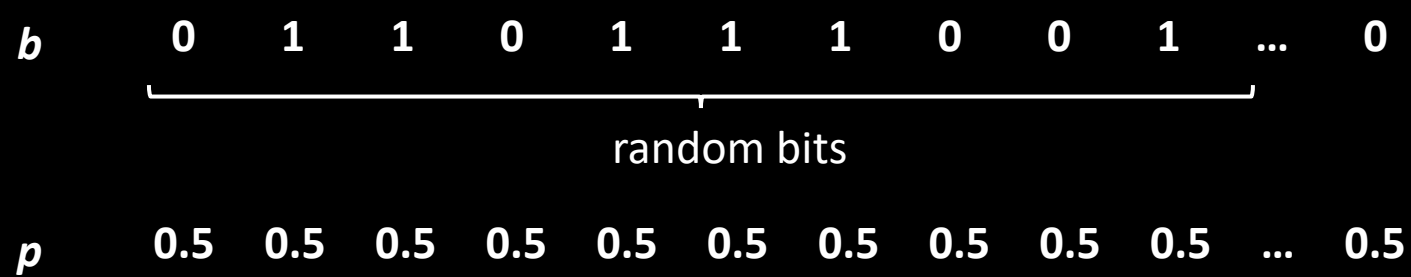
- Prediction loss: "how similar are *b(t)* and *p(t)*?" $\sum_{t=1}^{T} |b(t) - p(t)|$
- Calibration: "To what extent does $p(t)$ represent $\Pr[b(t) = 1]$?"

$$\text{calerr}(T) := \sum_{p \in [0,1]} |m_p(T) - p \cdot n_p(T)|$$

$n_p(T)$: #times the forecaster predicts p

$m_p(T)$: #time *b(t)*=1 when p is predicted

Pop Quiz: Calibration loss vs Prediction loss



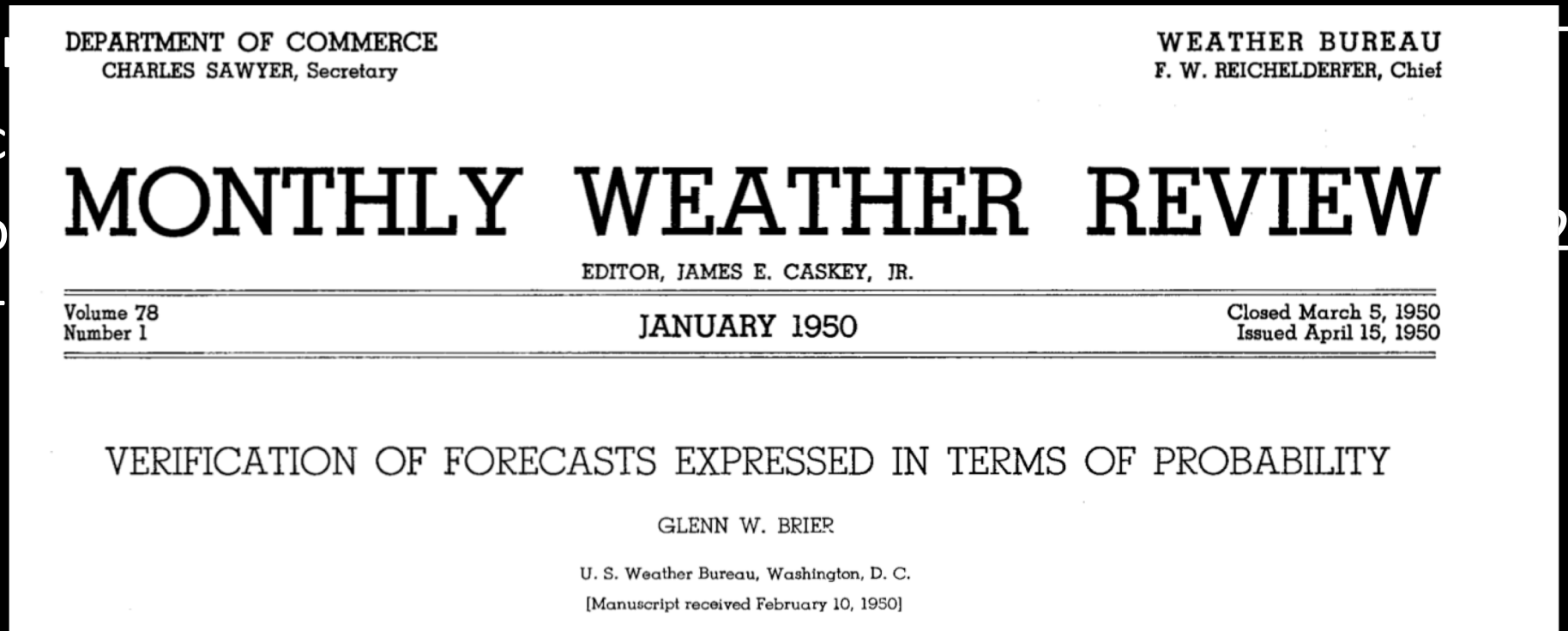
- Prediction loss: ~ T/2
- Calibration loss: ~ sqrt(T)

Calibration

- [Brier, 1950] suggested:

  "When a series of forecasts has been made using probability statements a study can also be made to determine whether the forecast probabilities are related to the relative frequency of the events' occurrence."

- Onli… …T11]
- Mac…
- Algo… …20, JLP+…



DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

Calibration

- Positive result by [Foster and Vohra, 1998]:
  - A randomized forecaster with $\mathbb{E}[\text{calerr}(T)] = O\left(T^{2/3}\right)$

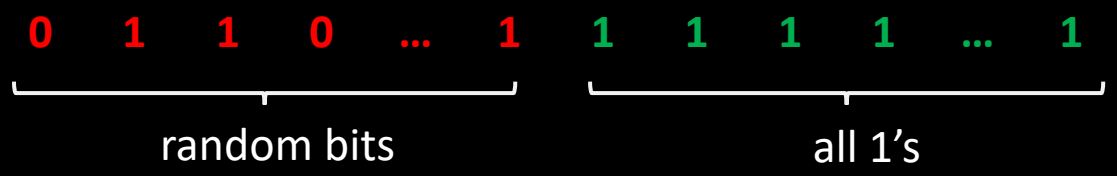
- Easy lower bound (folklore): $\mathbb{E}[\text{calerr}(T)] = \Omega\left(T^{1/2}\right)$
  - Suppose outcomes are coin flips and predictions are $1/2$
  - $\mathbb{E}[\text{calerr}(T)] = \mathbb{E}\left[\left|m_{1/2}(T) - (1/2) \cdot n_{1/2}(T)\right|\right] = \Omega\left(T^{1/2}\right)$

$\sim \text{Binomial}(T, 1/2)$ $= T/2$

- **[Qiao/Valiant, STOC'21]:** $\mathbb{E}[\text{calerr}(T)] = \Omega(T^{0.528})$
  - First super-$\sqrt{T}$ lower bound

Obstacle: “Cover-ups”

- Forecaster may act strategically to “cover up” previous mistakes



- Truthful forecaster: predict 1/2 on red, 1 on green
  - $\mathbb{E}[\text{calerr}(T)] \sim \sqrt{T}$
- Strategic forecaster: predict “0.6” on red, drive calerr to 0 using green
  - E.g., if frac 1’s on red < 0.6, predict “0.6” on green until calerr=0, then predict “1”
- Also known as “backcasting” [Foster and Hart, 2020]
  - i.e., make previous predictions look calibrated

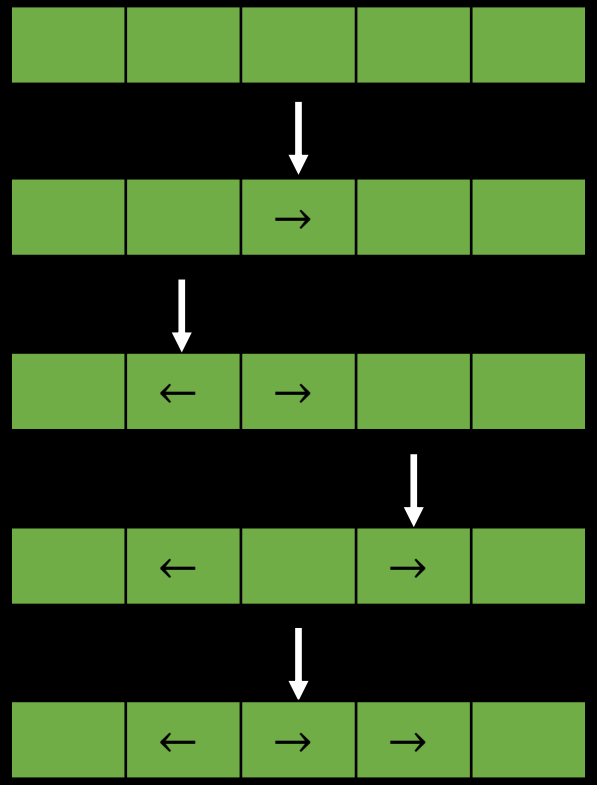
Detour: Sidestepping Game

- Game starts with k empty cells arranged in a row

- In each round:
  - Adversary picks empty cell $j \in \{1,2, \dots, k\}$
  - Forecaster places an arrow (L or R) into cell j
  - Any arrow "pointing to" cell j gets removed

- Zero-sum game: adversary maximizes #arrows
- Define $\mathrm{opt}(k, r)$ as outcome of a game with k cells and r rounds

Detour: Sidestepping Game



A picks cell 3, F places R-arrow

A picks cell 2, F places L-arrow

A picks cell 4, F places R-arrow, cell 3 becomes empty

A picks cell 3, F places R-arrow

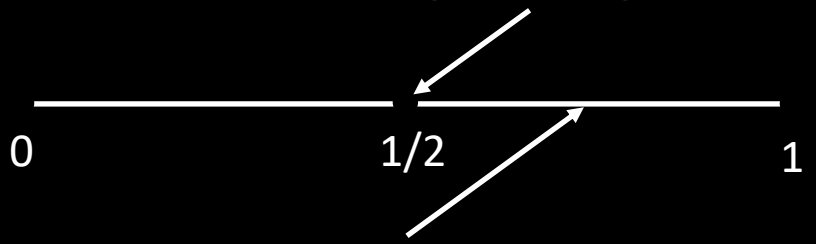
Detour: Sidestepping Game

$n_{1/2}$: #times 1/2 is predicted
$m_{1/2}$: #positives among them

Recall: calerr contains a $\left|m_{1/2} - n_{1/2} \cdot 1/2\right|$ term

Prediction Setting

Suppose $m_{1/2} - n_{1/2} \cdot 1/2 > 0$

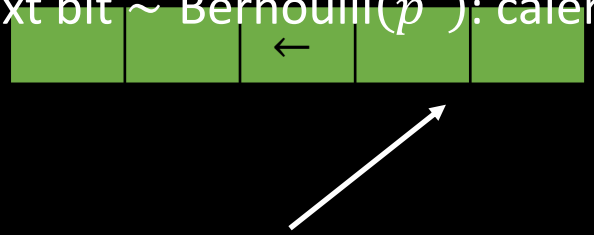


Adversary would choose $p^* > 1/2$

Assuming next prediction = 1/2:

- ~~Next bit = 0: calerr down by 1/2~~
- Next bit = 1: calerr up by 1/2
- Next bit $\sim$ Bernoulli(p^*): calerr up by $p^* - 1/2$

Sidestepping Game



Adversary would pick cell > 3

Main Lemma

Definition: Pair (α, β) is **admissible** if $\text{opt}(k, k^\alpha) = \Omega\left(k^\beta\right)$ for all k

- Recall: $\text{opt}(k, r)$ = outcome of a game with k cells and r rounds

Theorem: For every admissible (α, β), there is a lower bound

$$\mathbb{E}[\text{calerr}(T)] = \widetilde{\Omega}(T^c),$$

where exponent $c = \frac{2\beta+1}{\alpha+2\beta+2}$.

Note: $c > 1/2 \Leftrightarrow \beta > \alpha/2$ (can get to $c = 3/5$ if $(1,1)$ is admissible)

Finding Admissible Pairs

Fact 1: For integer t, $\text{opt}(2^t - 1, t) = t$.

Proof: "binary search"

This proves $\text{opt}(k, \log k) = \log k$.

Not enough: need $\text{opt}(k, k^\alpha) = \Omega(k^\beta)$ for $\alpha, \beta > 0$.

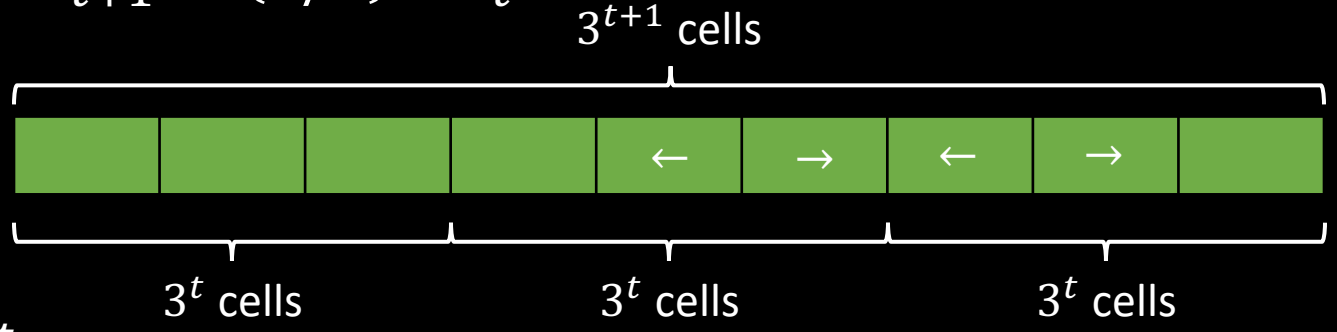
Finding Admissible Pairs

Fact 2: $\text{opt}(k, r) \geq c \Rightarrow \text{opt}(k^t, r^t) \geq \left(\frac{c+1}{2}\right)^t$ for integer t.

Proof (by example): $\text{opt}(3, 2) \geq 2 \Rightarrow \text{opt}(3^t, 2^t) \geq (3/2)^t$

Suffices to show $M_{t+1} \geq (3/2) \cdot M_t$

Define $M_t \coloneqq \text{opt}(3^t, 2^t)$



Step 1: Spend 2^t rounds in the middle block, $\geq M_t$ arrows survive.

Observation: $\geq M_t/2$ arrows in the same direction (say, left)

Step 2: Another 2^t rounds in the right block.

$\geq M_t/2 + M_t$ surviving arrows in the end, so $M_{t+1} \geq (3/2) \cdot M_t$.

Finding Admissible Pairs

Fact 1: For integer t, $\text{opt}(2^t - 1, t) = t$.

Fact 2: $\text{opt}(k, r) \geq c \Rightarrow \text{opt}(k^s, r^s) \geq \left(\frac{c+1}{2}\right)^s$ for integer s.

Combining two facts gives

$$\text{opt}\left((2^t - 1)^s, t^s\right) \geq \left(\frac{t+1}{2}\right)^s$$

Writing $k = (2^t - 1)^s$, we have $\text{opt}(k, k^\alpha) \geq k^\beta$ where

$$\alpha = \frac{\log(t)}{\log(2^t - 1)} \quad \text{and} \quad \beta = \frac{\log((t+1)/2)}{\log(2^t - 1)}$$

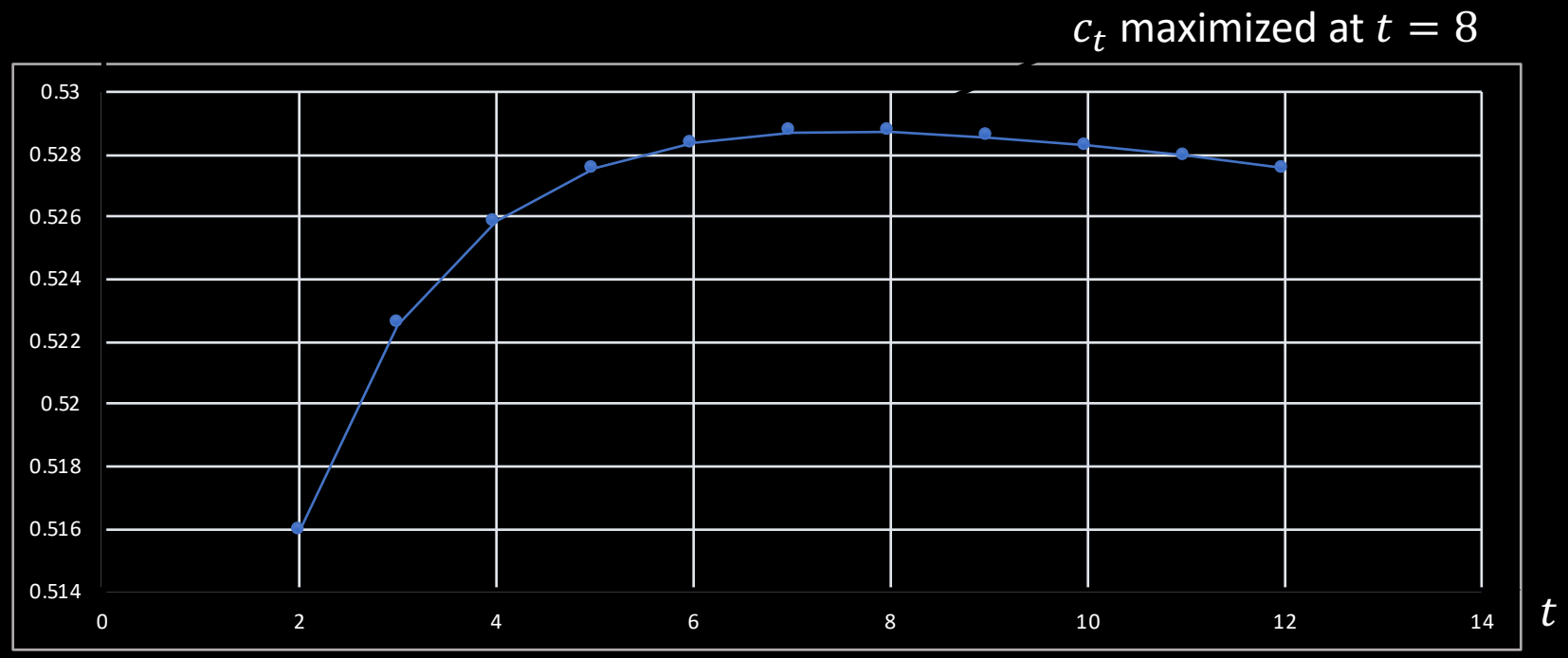
Finding Admissible Pairs

Lemma: For integer t, (α_t, β_t) is admissible where:

$$\alpha_t = \frac{\log(t)}{\log(2^t - 1)} \quad \text{and} \quad \beta_t = \frac{\log((t+1)/2)}{\log(2^t - 1)}$$

Optimizing t gives an exponent > 0.528:

$$c_t = \frac{2\beta_t + 1}{\alpha_t + 2\beta_t + 2}$$



Takeaways

- Calibration in binary prediction: [Still] one of the most basic open questions in sequential prediction. [$n^{2/3}$ vs $n^{0.53}$]
- Minimizing calibration error encourages “cover-ups”…
- Sidestepping: an interesting new game

Full paper:

Stronger Calibration Lower Bounds via Sidestepping, Mingda Qiao, Gregory Valiant, STOC 2021.

https://arxiv.org/abs/2012.03454