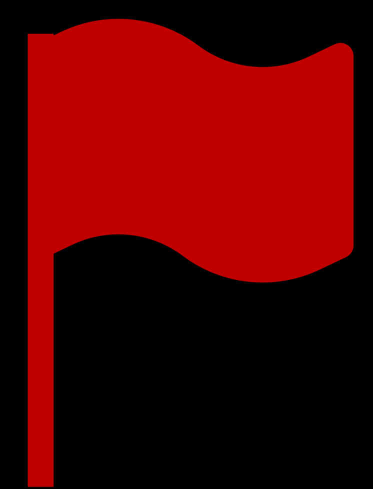
Worst Case Analysis for Randomly Collected Data

Justin Chen — MIT
Gregory Valiant — Stanford
Paul Valiant — IAS, Purdue

Traditional Statistical Estimation



Distributional assumptions on data **values**

e.g.: Gaussian, i.i.d., exchangeable, Robust statistics

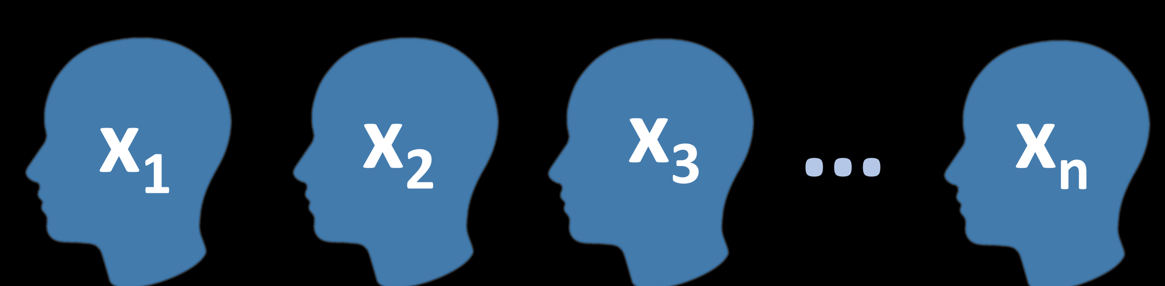
Alternate View

Leverage the data **collection** process –

without assumptions about the distribution of the data values

Our Framework

n entities, each with a hidden value $\mathbf{x}_i$ (bounded real number)



Goal: Estimate *mean*$(x_1,..,x_n)$

Modeling data collection via distribution *P* over possible samples

Subset $S \subset \{1,2,.., n\}$ drawn from ***P***

Observe ***S***, *values* x_S *indexed by* ***S***, *return* $f(P,S,x_S)$

Performance measure: *Worst-Case Expected Error*

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

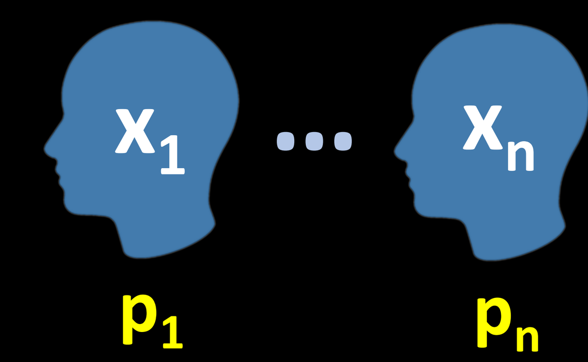
Worst-case analysis over data values

Expectation over sampling process described by P

Illustrative Examples

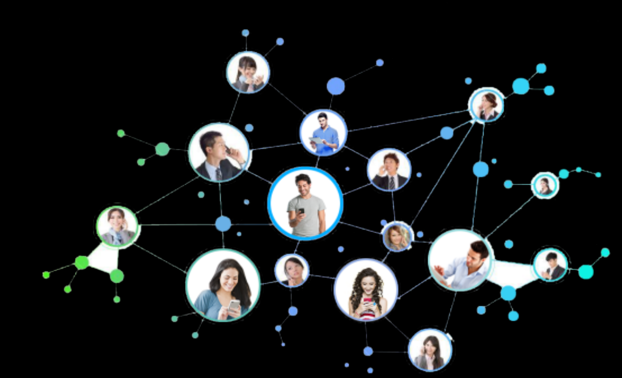
Importance Sampling

P: each individual appears in the sample independently w.p. p_i



Snowball Sampling

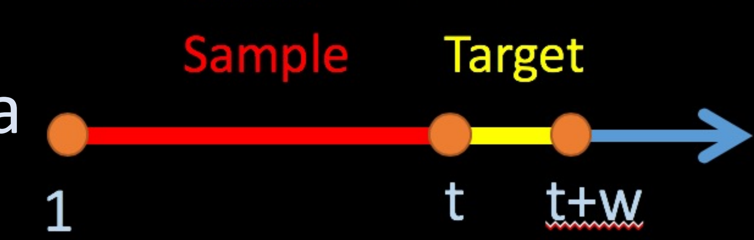
P: sample generated by a viral process on a social network



Selective Prediction (Forecasting)

P: samples corresponds to past data with prediction over future data

[Drucker'12,Qiao/V'19]



Main Results

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

Thm 1 (evaluation): Given estimator *f*, in poly-time, with poly # samples from P, we can π/2-approximate the error of *f*.†

Thm 2 (optimization): In poly-time, with poly # samples from P, we can find a π/2-optimal† estimator *f*.

†*We restrict f to the general class of "semilinear" estimators where the estimate is a linear combination of the sampled data (weights depending arbitrarily on P and S)*

$$f(P, S, x_S) = \langle a_{(P,S)}, x_S \rangle$$

Techniques

Exact evaluation and optimization of estimators in this regime are **NP-hard** (reduction to Max-Cut and semidefinite Grothendieck problem)

Given full description (exp size) of P, Goemans-Williamson SDP relaxation gives approximation

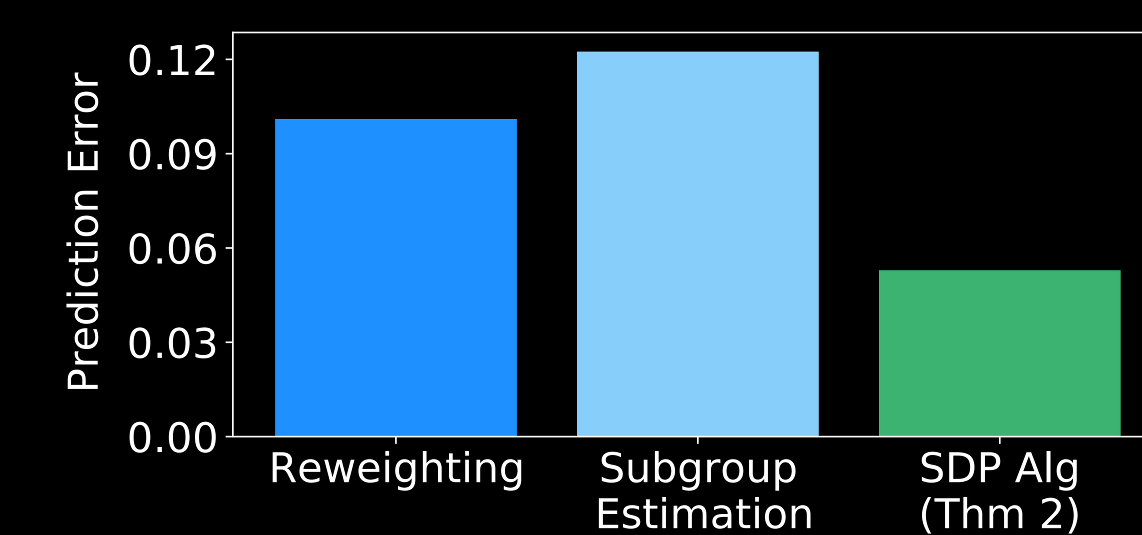
More work involving subsampling and convex duality give us efficient algorithms for Thms 1,2

Experiments

2-7x improvements over baselines in 3 settings

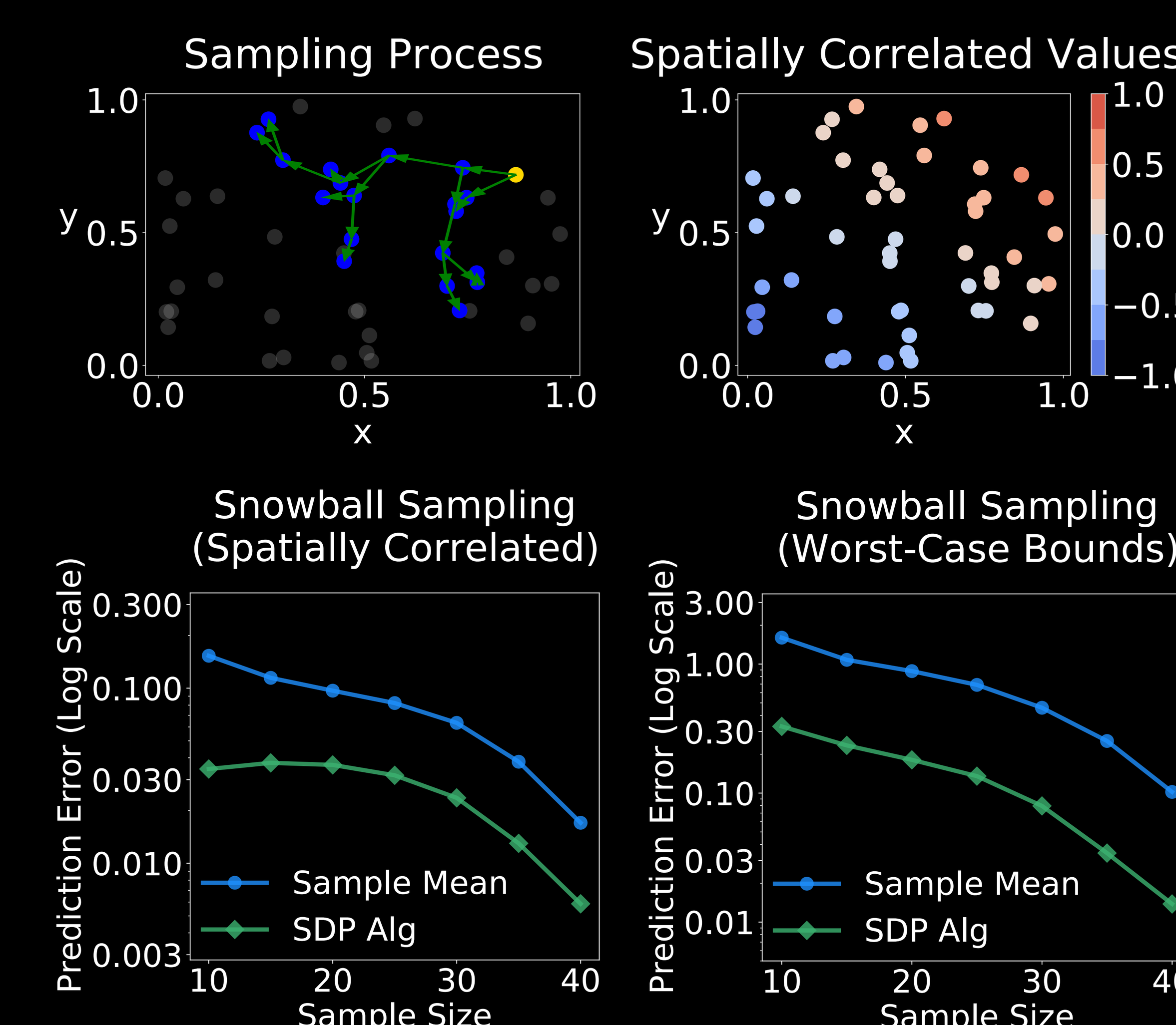
Importance Sampling

$p_1,\ldots,p_{25} = 0.1$
$p_{26},\ldots,p_{50} = 0.5$



Snowball Sampling

Points in unit square recruit nearby points



Many open questions within this framework and beyond - Ask Me!