Maximum Likelihood Estimation for Learning Populations of Parameters

Ramya Korlakai Vinayak

Postdoctoral Researcher

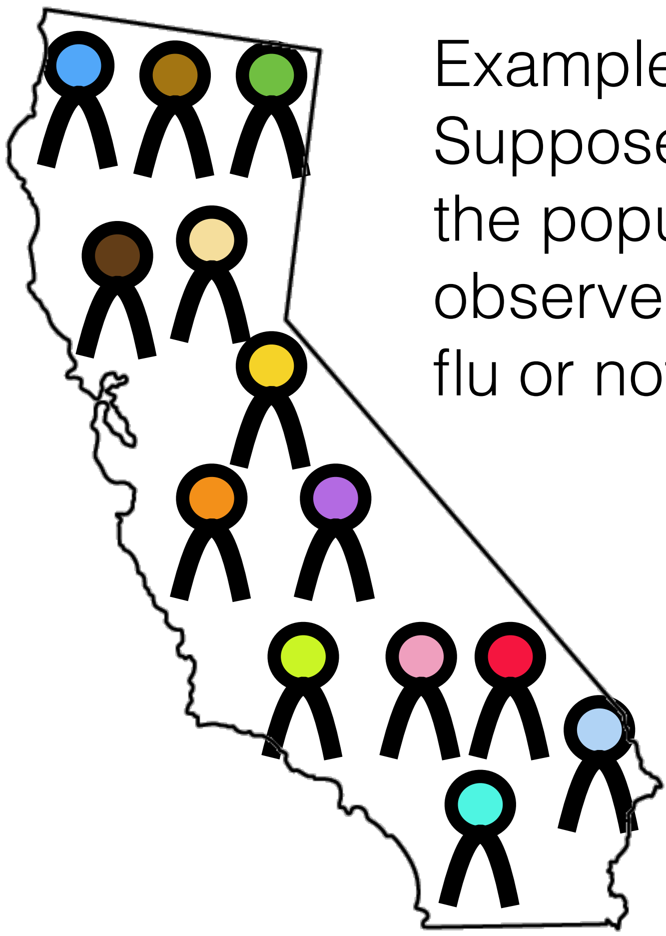
Paul G. Allen School of CSE

joint work with
Weihao Kong, Gregory Valiant, Sham Kakade

ramya@cs.washington.edu

Poster #189

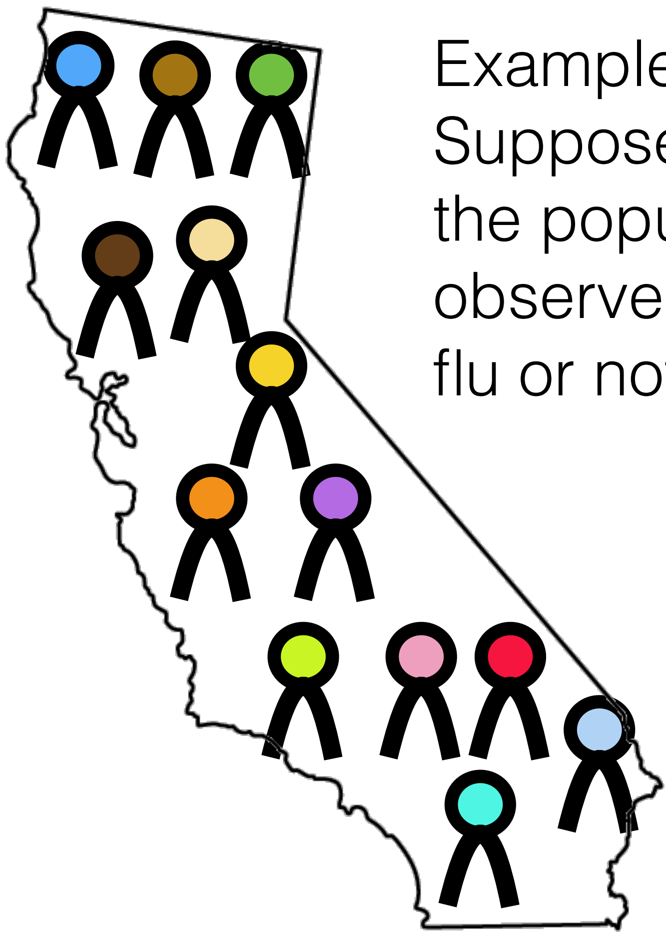
Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

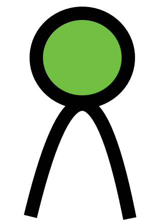
Poster #189

Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

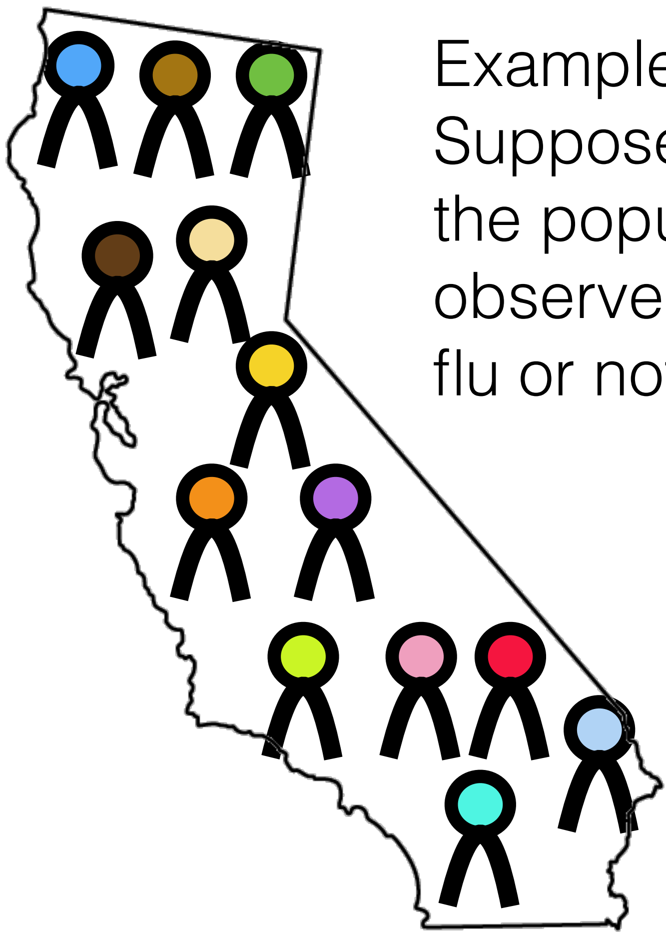
Probability of catching flu (unknown)



i p_i

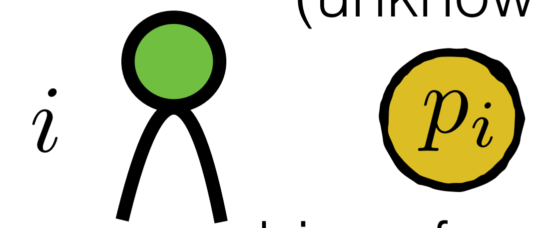
Poster #189

Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

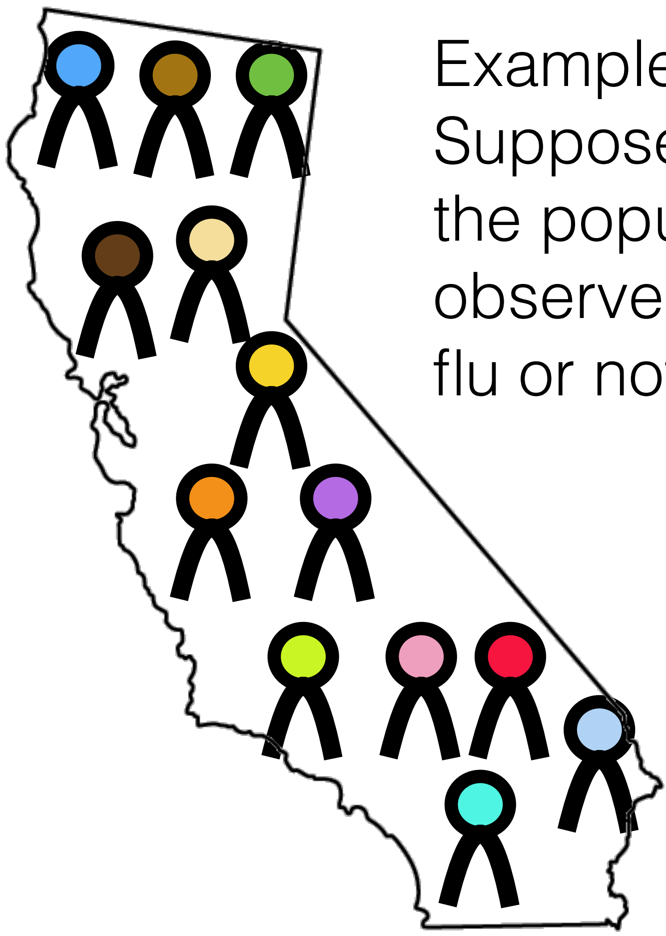
Probability of catching flu (unknown)



bias of coin i

Poster #189

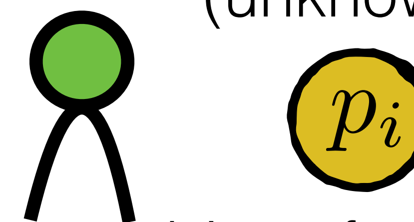
Motivation: Large yet Sparse Data



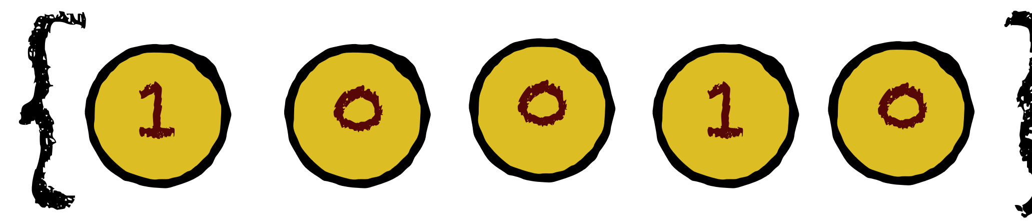
Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

Probability of catching flu (unknown)

i



bias of coin i

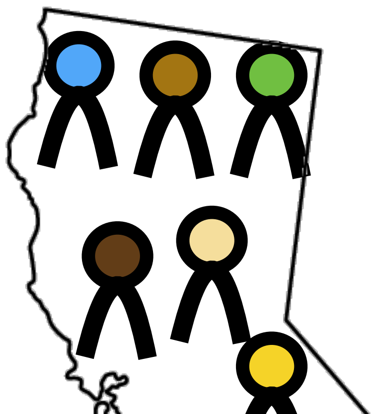


$$x_i = 2$$

$$\widehat{p}_i = \frac{x_i}{t} = 0.4 \pm 0.45$$

Poster #189

Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

Probability of catching flu (unknown)

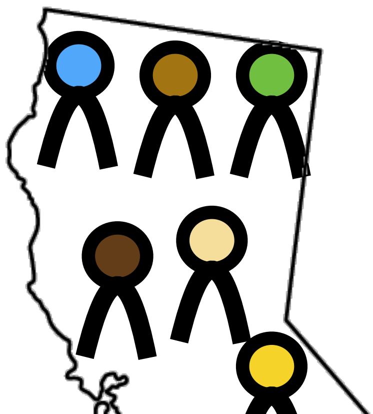
i

bias of coin i

Goal: Can we learn the distribution of the biases over the population?

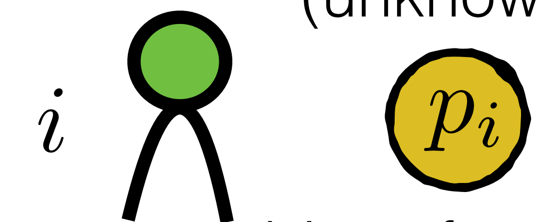
Poster #189

Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

Probability of catching flu (unknown)



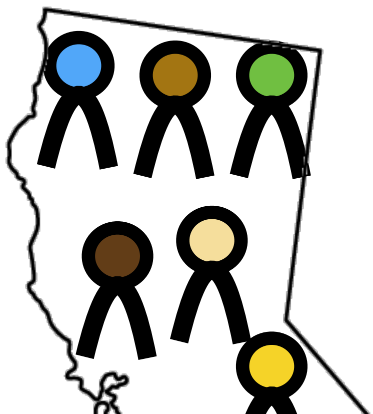
bias of coin i

Goal: Can we learn the distribution of the biases over the population?

- Application domains: Epidemiology, Social Sciences, Psychology, Medicine, Biology
- Population size is large, often hundreds of thousands or millions
- Number of observations per individual is limited (*sparse*) prohibiting accurate estimation of parameters of interest

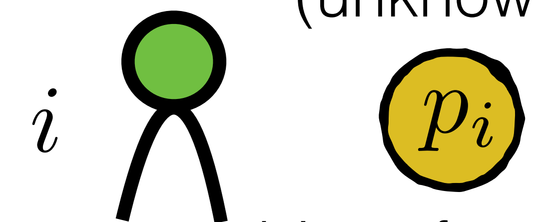
Poster #189

Motivation: Large yet Sparse Data



Example: Flu data
Suppose for a large random subset of the population in California, we observe whether a person caught the flu or not for last 5 years

Probability of catching flu (unknown)



bias of coin i

Goal: Can we learn the distribution of the biases over the population?

Why?

Useful for downstream analysis:

Testing and estimating properties of the distribution

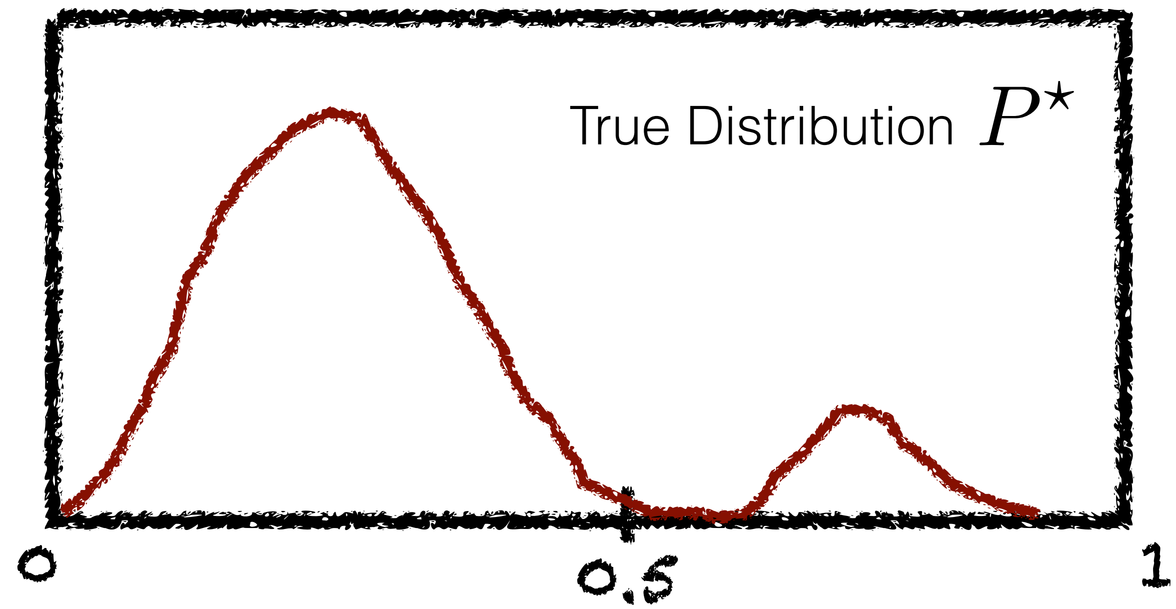
Model: Non-parametric Mixture of Binomials

Lord 1965, 1969

- N independent coins
  Each coin has its own bias drawn from $P^\star$

$i = 1, 2, ..., N$ $\quad p_i \sim P^\star$ (unknown)

(unknown)



Poster #189

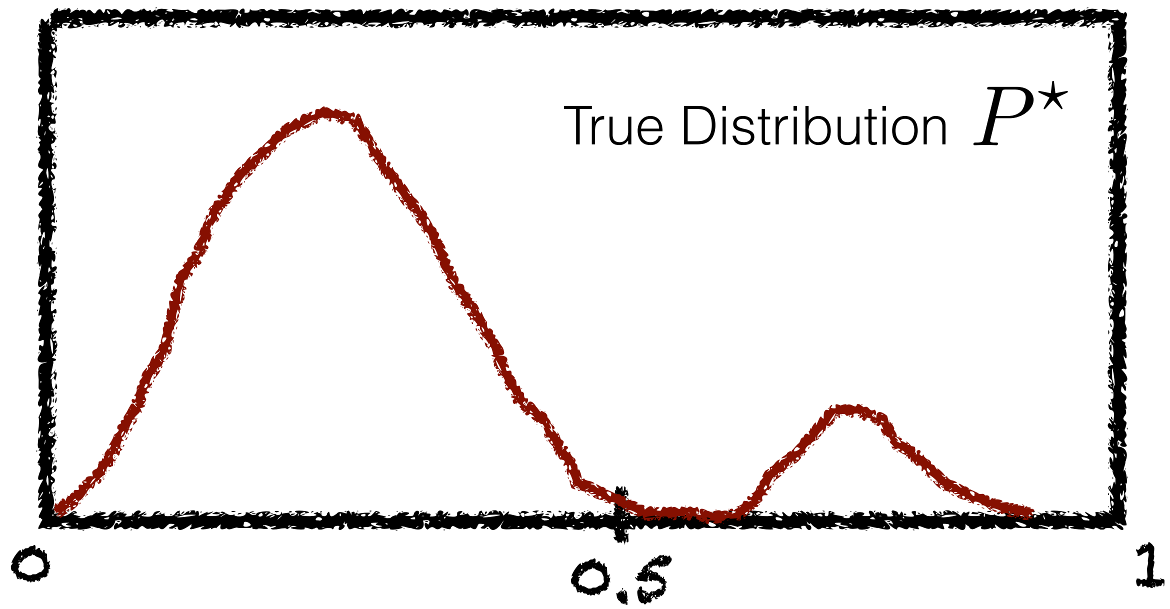
Model: Non-parametric Mixture of Binomials

Lord 1965, 1969

- N independent coins
  Each coin has its own bias drawn from $P^\star$

  $i = 1, 2, ..., N$ $\quad p_i \sim P^\star$ (unknown)

  (unknown)



- We get to observe t tosses for every coin

  Observations: $X_i \sim \mathrm{Bin}(t, p_i) \in \{0, 1, ..., t\}$

  $t = 5$ tosses $\{1 \quad 0 \quad 0 \quad 1 \quad 0\}$ $x_i = 2$

Poster #189

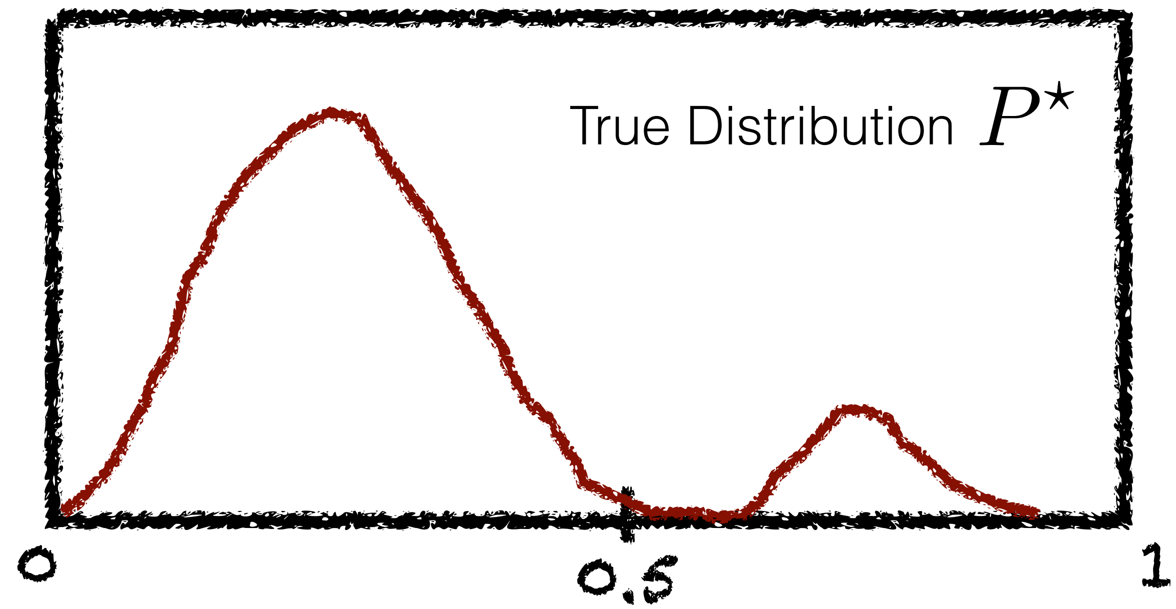
Model: Non-parametric Mixture of Binomials

Lord 1965, 1969

- N independent coins
  Each coin has its own bias drawn from $P^\star$

$i = 1, 2, ..., N$ $\quad p_i \sim P^\star$ (unknown)

(unknown)



- We get to observe t tosses for every coin

Observations: $X_i \sim \mathrm{Bin}(t, p_i) \in \{0, 1, ..., t\}$

$t = 5$ tosses $\{1 \quad 0 \quad 0 \quad 1 \quad 0\}$ $x_i = 2$

- Given $\{X_i\}_{i=1}^N$, return $\hat{P}$ estimate of $P^\star$

Poster #189

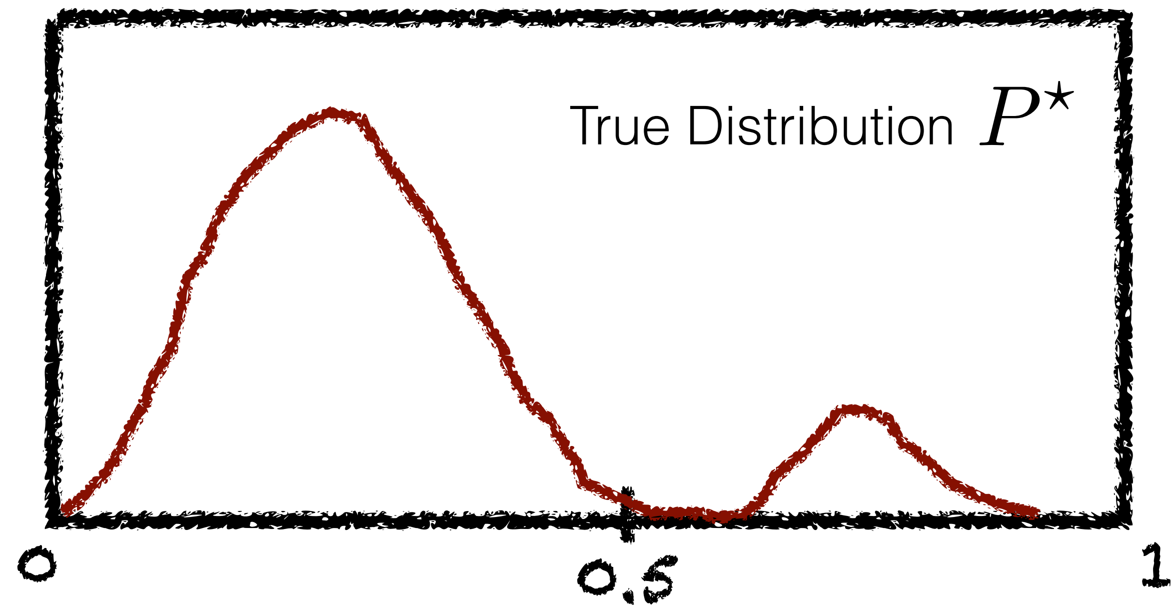
Model: Non-parametric Mixture of Binomials

Lord 1965, 1969

- N independent coins
  Each coin has its own bias drawn from $P^\star$

$i = 1, 2, ..., N$ $\quad p_i \sim P^\star$ (unknown)

(unknown)



- We get to observe t tosses for every coin

Observations: $X_i \sim \mathrm{Bin}(t, p_i) \in \{0, 1, ..., t\}$

$t = 5$ tosses $\{1\ 0\ 0\ 1\ 0\}$ $x_i = 2$

- Given $\{X_i\}_{i=1}^N$, return $\hat{P}$ estimate of $P^\star$

- Wasserstein-1 distance
  (Earth Mover's Distance) $\quad W_1\left(P^\star, \hat{P}\right)$

Poster #189

Learning with Sparse Observations is Non-trivial

- Empirical plug-in estimator is bad $\hat{P}_{\text{plug-in}} = \text{histogram}\left\{\frac{X_1}{t}, \ldots \frac{X_i}{t}, \ldots, \frac{X_N}{t}\right\}$

  When $t \ll N$ incurs error of $\Theta\left(\frac{1}{\sqrt{t}}\right)$

$N =$ Number of coins $t =$ Number of tosses per coin

Poster #189

Learning with Sparse Observations is Non-trivial

- Empirical plug-in estimator is bad $\hat{P}_{\text{plug-in}} = \text{histogram}\left\{\frac{X_1}{t}, \ldots \frac{X_i}{t}, \ldots, \frac{X_N}{t}\right\}$

  When $t \ll N$ incurs error of $\Theta\left(\frac{1}{\sqrt{t}}\right)$

  $N =$ Number of coins $t =$ Number of tosses per coin

- Many recent works on estimating symmetric properties of a discrete distribution with sparse observations

  Paninski 2003, Valiant and Valiant 2011, Jiao et. al. 2015, Orlitsky et. al. 2016, Acharya et. al. 2017 ….

  The setting in this work is different

Poster #189

Learning with Sparse Observations is Non-trivial

- Empirical plug-in estimator is bad $\hat{P}_{\text{plug-in}} = \text{histogram}\left\{\frac{X_1}{t}, ... \frac{X_i}{t}, ..., \frac{X_N}{t}\right\}$

  When $t \ll N$ incurs error of $\Theta\left(\frac{1}{\sqrt{t}}\right)$

  $N =$ Number of coins $t =$ Number of tosses per coin

- Many recent works on estimating symmetric properties of a discrete distribution with sparse observations

  Paninski 2003, Valiant and Valiant 2011, Jiao et. al. 2015, Orlitsky et. al. 2016, Acharya et. al. 2017 ….

  The setting in this work is different

- Tian et. al 2017 proposed a moment matching based estimator which achieves optimal error of $\mathcal{O}\left(\frac{1}{t}\right)$ when $t < c\log N$

  Weakness of moment matching estimator is that it fails to obtain optimal error when $t > c \log N$ due to higher variance in larger moments

Poster #189

Learning with Sparse Observations is Non-trivial

- Empirical plug-in estimator is bad $\hat{P}_{\text{plug-in}} = \text{histogram}\left\{\frac{X_1}{t}, \ldots \frac{X_i}{t}, \ldots, \frac{X_N}{t}\right\}$

  When $t \ll N$ incurs error of $\Theta\left(\frac{1}{\sqrt{t}}\right)$

  $N =$ Number of coins $t =$ Number of tosses per coin

What about Maximum Likelihood Estimator?

Weakness of moment matching estimator is that it fails to obtain optimal error when $t > c \log N$ due to higher variance in larger moments

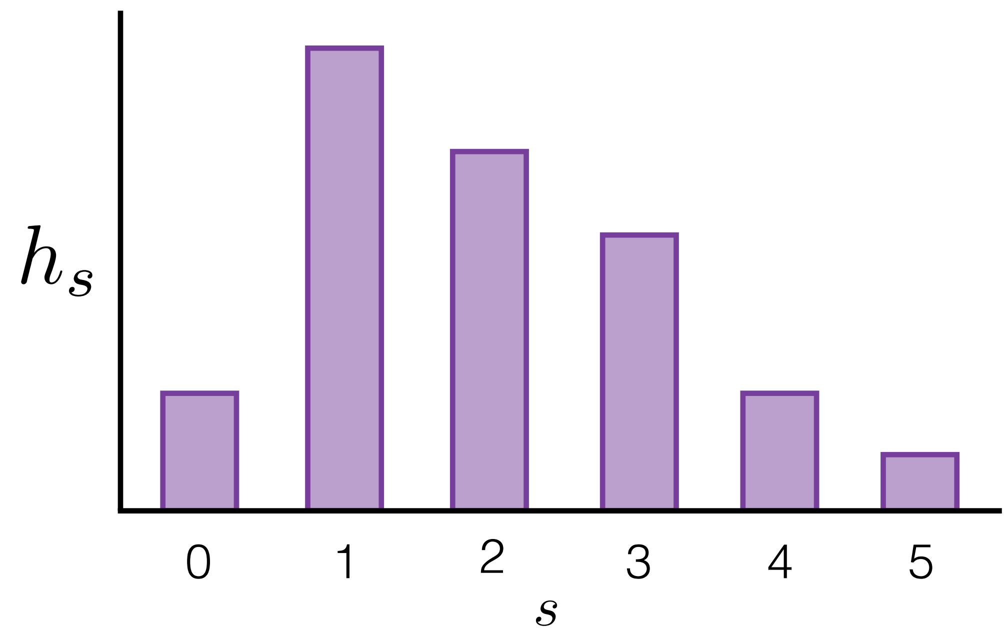
Poster #189

Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

$$h_s = \frac{\#\text{ coins that show s heads}}{N} \qquad s = 0, 1, ..., t$$

$\mathbf{h} = [h_0, h_1, ..h_s, .., h_t]$ fingerprint vector



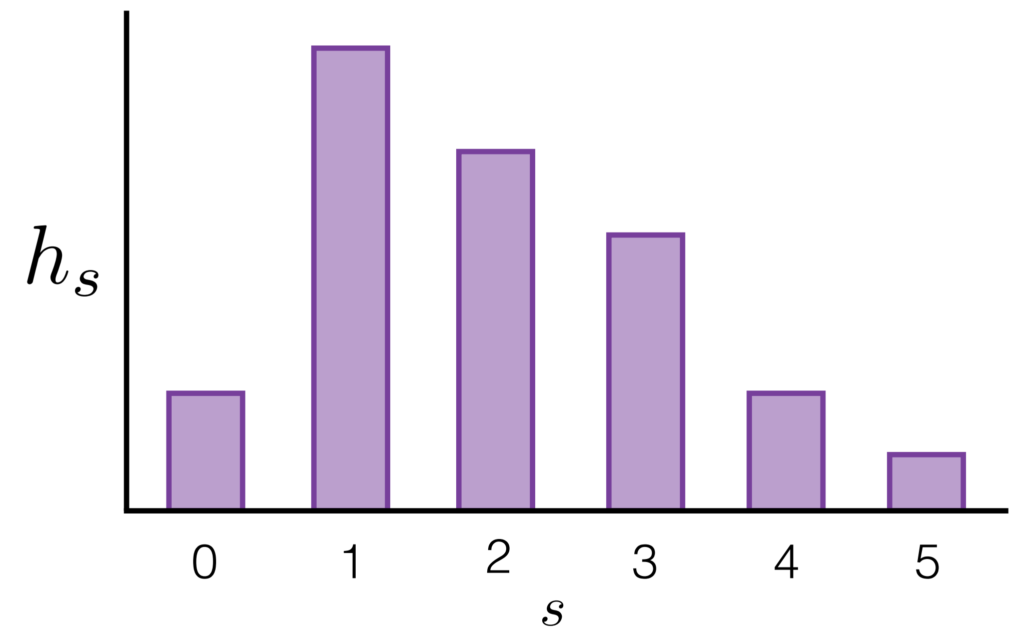
Poster #189

Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

$$h_s = \frac{\text{\# coins that show s heads}}{N} \qquad s = 0, 1, ..., t$$

$\mathbf{h} = [h_0, h_1, ..h_s, .., h_t]$ fingerprint vector



$$\hat{P}_{\text{mle}} \in \underset{Q \in \text{dist}[0,1]}{\arg\min} \ \mathrm{KL}\left(\text{Observed } \mathbf{h} \ , \ \text{Expected } \mathbf{h} \text{ under the distribution } Q\right)$$

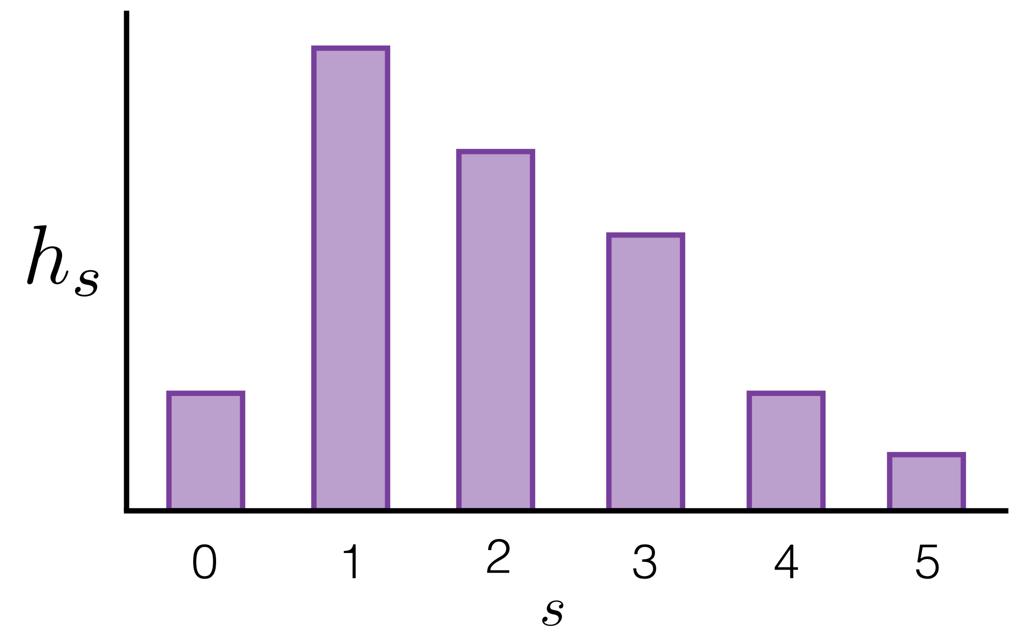
Poster #189

Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

$$h_s = \frac{\#\text{ coins that show s heads}}{N} \qquad s = 0, 1, ..., t$$

$\mathbf{h} = [h_0, h_1, ..h_s, .., h_t]$ fingerprint vector

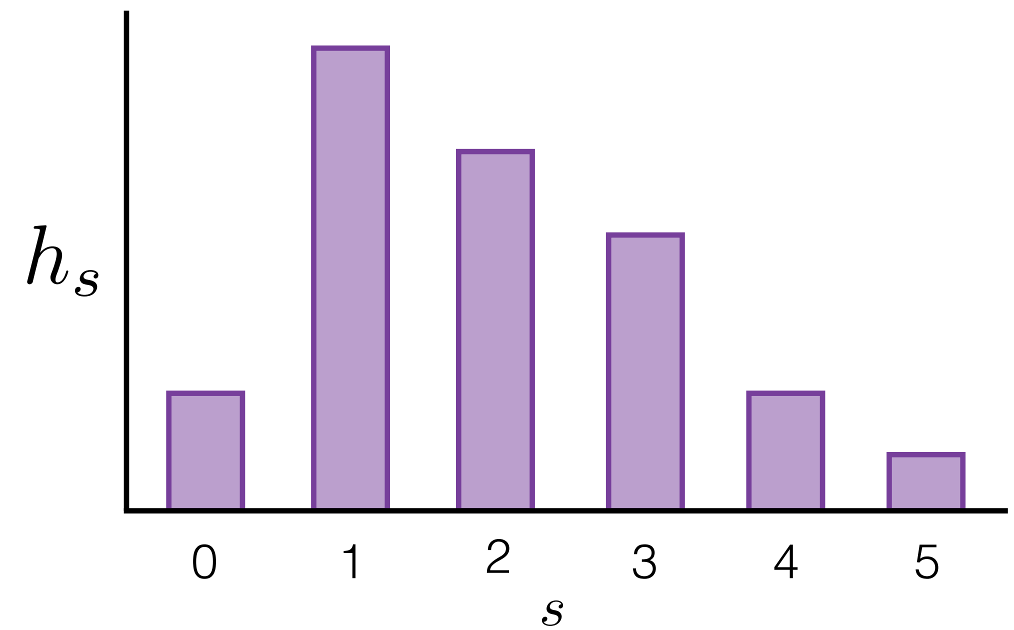


$$\hat{P}_{\text{mle}} \in \underset{Q \in \text{dist}[0,1]}{\arg\min} \; \text{KL}\left(\text{Observed } \mathbf{h} \;,\; \text{Expected } \mathbf{h} \text{ under the distribution } Q\right)$$

- NOT the empirical estimator
- Convex optimization: Efficient (polynomial time)

Poster #189

Maximum Likelihood Estimator



Sufficient statistic: Fingerprint

$$h_s = \frac{\#\text{ coins that show s heads}}{N} \qquad s = 0, 1, ..., t$$

$\mathbf{h} = [h_0, h_1, ..h_s, .., h_t]$ fingerprint vector

$$\hat{P}_{\text{mle}} \in \underset{Q \in \text{dist}[0,1]}{\arg\min} \ \text{KL}\left(\text{Observed } \mathbf{h}, \ \text{Expected } \mathbf{h} \text{ under the distribution } Q\right)$$

- NOT the empirical estimator
- Convex optimization: Efficient (polynomial time)
- Proposed in late 1960's by Frederic Lord in the context of psychological testing. Several works study the geometry and identifiability and uniqueness of the solution of the MLE

  Lord 1965,1969, Turnbull 1976, Laird 1978, Lindsay 1983, Wood 1999

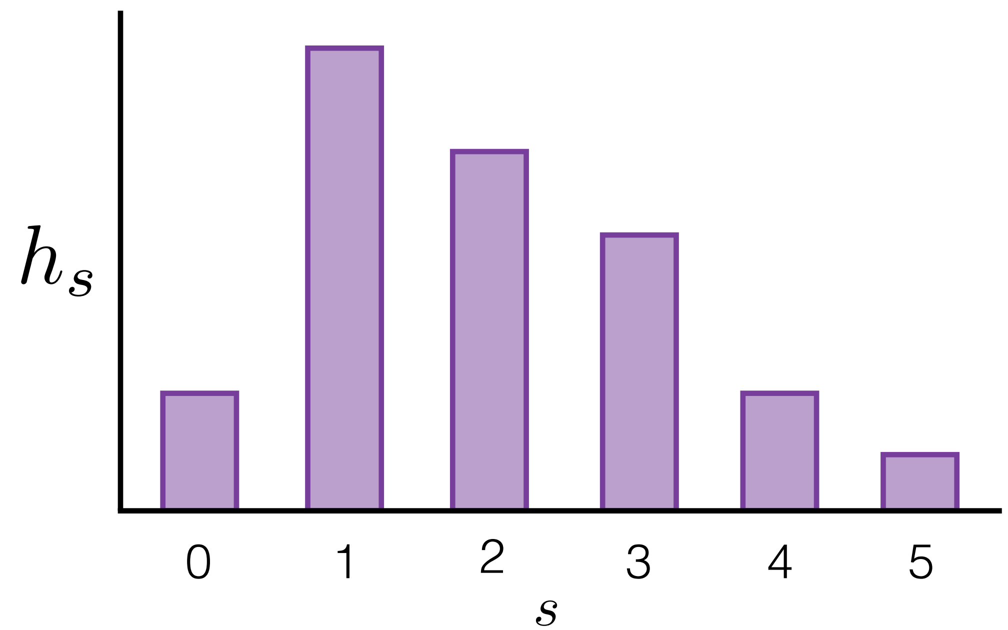
Poster #189

Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

$$h_s = \frac{\text{\# coins that show s heads}}{N} \qquad s = 0, 1, ..., t$$

$\mathbf{h} = [h_0, h_1, ..h_s, .., h_t]$ fingerprint vector



$$\hat{P}_{\text{mle}} \in \underset{Q \in \text{dist}[0,1]}{\text{arg min}} \ \text{KL}\left(\text{Observed } \mathbf{h} \ , \ \text{Expected } \mathbf{h} \text{ under the distribution } Q\right)$$

How well does the MLE recover the distribution?

- Convex optimization: Efficient (polynomial time)
- Proposed in late 1960's by Frederic Lord in the context of psychological testing. Several works study the geometry and identifiability and uniqueness of the solution of the MLE

  Lord 1965,1969, Turnbull 1976, Laird 1978, Lindsay 1983, Wood 1999

Poster #189

Main Results: MLE is Minimax Optimal in Sparse Regime

Non-asymptotic guarantees

Theorem 1

The MLE achieves following error bounds: $\text{w. p. } \geq 1-\delta$

- Small Sample Regime:

$$W_1\left(P^\star, \hat{P}_{\text{mle}}\right) = \mathcal{O}_\delta\left(\frac{1}{t}\right) \quad \text{when} \quad t < c\log N$$

- Medium Sample Regime:

$$W_1\left(P^\star, \hat{P}_{\text{mle}}\right) = \mathcal{O}_\delta\left(\frac{1}{\sqrt{t\log N}}\right) \quad \text{when} \quad c\log N \leq t \leq N^{2/9-\epsilon}$$

$N =$ Number of coins

$t =$ Number of tosses per coin

Sparse Regime

$t \ll N$

Poster #189

Main Results: MLE is Minimax Optimal in Sparse Regime

Non-asymptotic guarantees

Theorem 1

The MLE achieves following error bounds: $\text{w. p. } \geq 1-\delta$

- Small Sample Regime:

$$W_1\left(P^\star, \hat{P}_{\text{mle}}\right) = \mathcal{O}_\delta\left(\frac{1}{t}\right) \quad \text{when} \quad t < c\log N$$

- Medium Sample Regime:

$$W_1\left(P^\star, \hat{P}_{\text{mle}}\right) = \mathcal{O}_\delta\left(\frac{1}{\sqrt{t\ \log N}}\right) \quad \text{when} \quad c\log N \leq t \leq N^{2/9-\epsilon}$$

$N =$ Number of coins

$t =$ Number of tosses per coin

Sparse Regime
$t \ll N$

Theorem 2

- Matching Minimax Lower Bounds

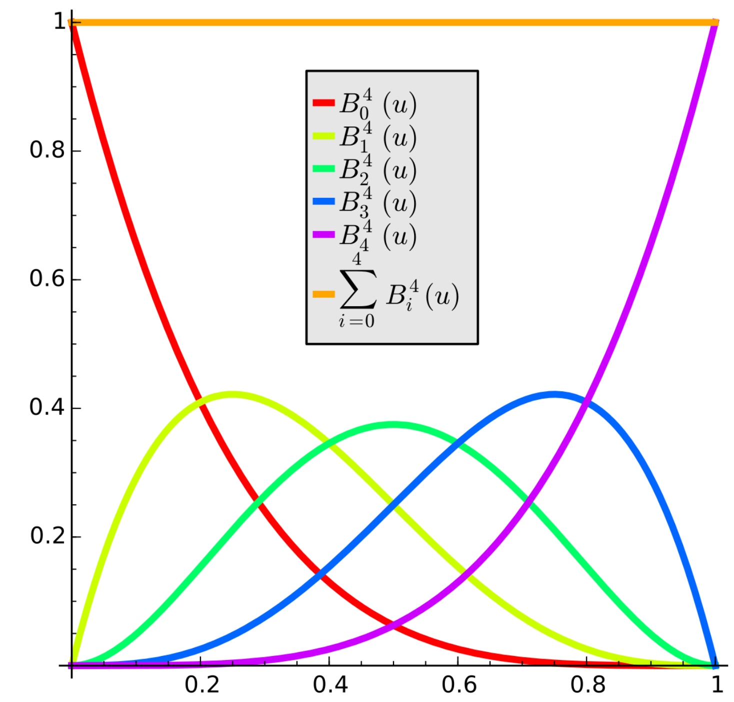
$$\inf_f \sup_P \; \mathrm{E}\left[W_1(P, f(\mathbf{X}))\right] > \Omega\left(\frac{1}{t}\right) \vee \Omega\left(\frac{1}{\sqrt{t\ \log N}}\right)$$

Poster #189

Novel Proof: Polynomial Approximations

New bounds on coefficients of Bernstein polynomials approximating Lipschitz-1 functions on [0, 1]

$$\widehat{f}_t(x) = \sum_{j=0}^{t} \boxed{b_j} \binom{t}{j} x^j (1-x)^{t-j}$$



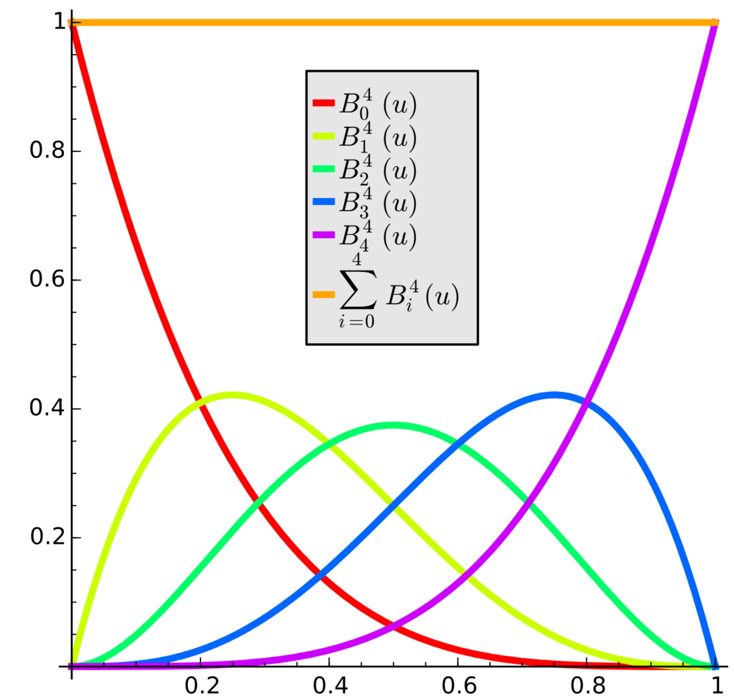
Bernstein polynomials

Poster #189

Novel Proof: Polynomial Approximations

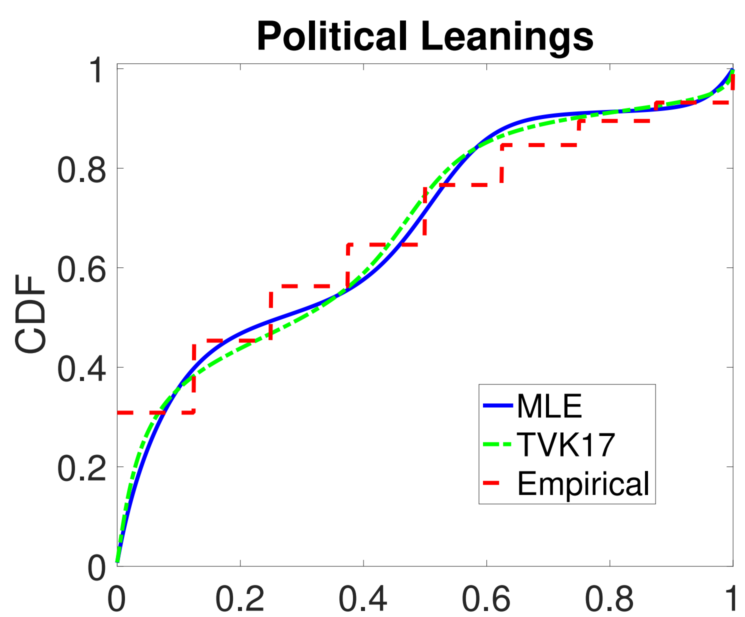
New bounds on coefficients of Bernstein polynomials approximating Lipschitz-1 functions on [0, 1]

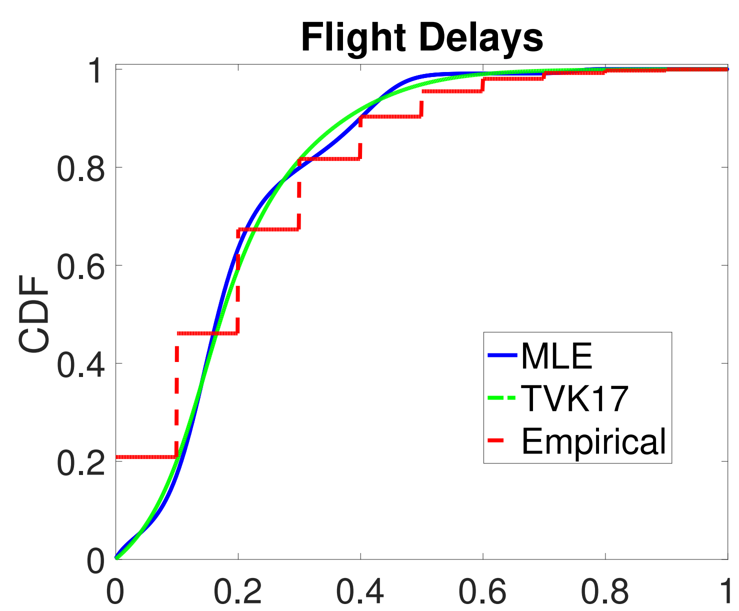
$$\widehat{f}_t(x) = \sum_{j=0}^{t} \boxed{b_j} \binom{t}{j} x^j (1-x)^{t-j}$$



Bernstein polynomials

Performance on Real Data



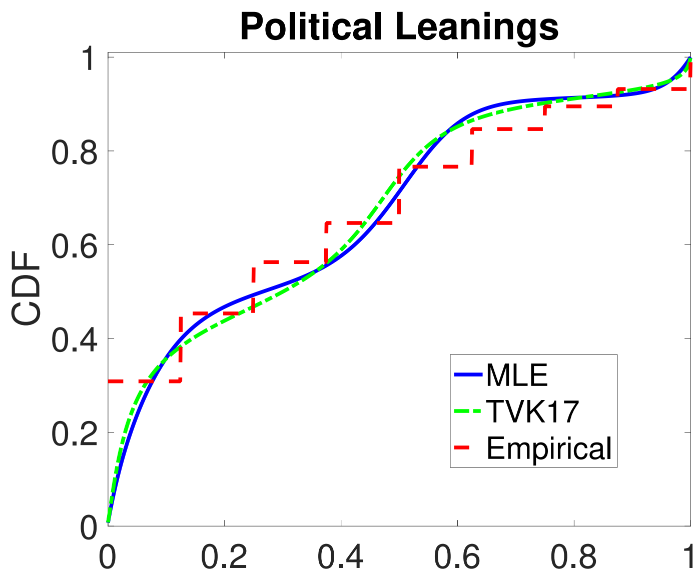


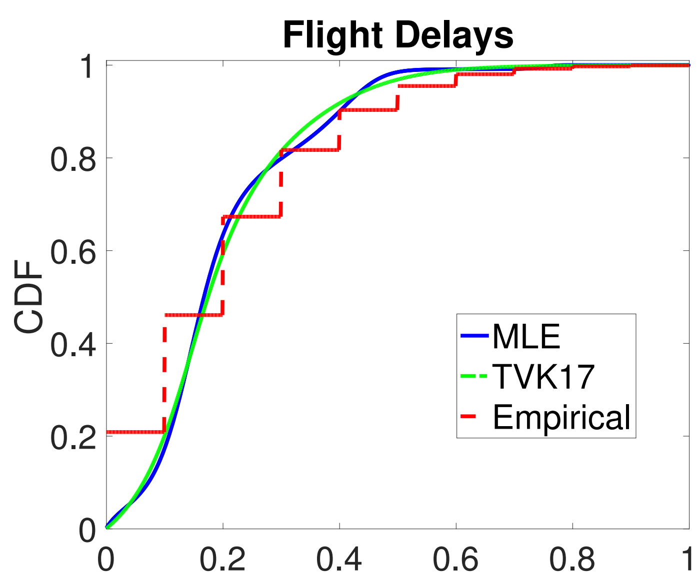
Poster #189

Summary

Learning distribution of parameters over a population with sparse observations per individual

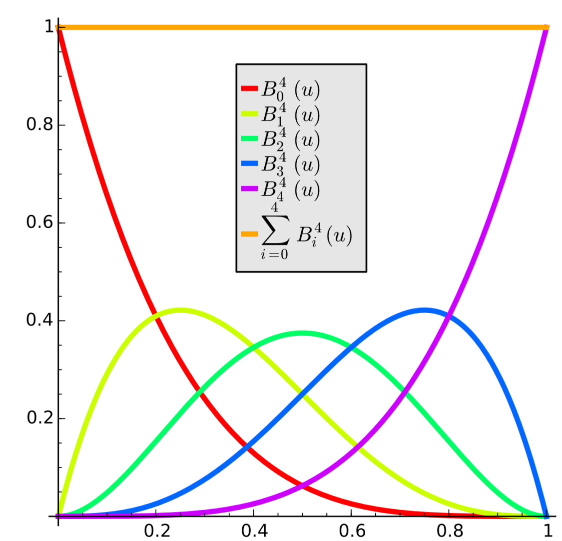
Performance on Real Data





MLE is Minimax Optimal even with sparse observations!

Novel proof: new bounds on coefficients of Bernstein polynomials approximating Lipschitz-1 functions



Poster #189

ramya@cs.washington.edu