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
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.
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.

\[ p_i \]

Probability of catching flu (unknown)
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
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

\[
\hat{p}_i = \frac{x_i}{t} = 0.4 \pm 0.45
\]
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.

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

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
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.

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

- $N$ independent coins
  - Each coin has its own bias drawn from $P^*$
  - $i = 1, 2, ..., N$
  - $p_i \sim P^*$ (unknown)
  - (unknown)

Lord 1965, 1969
Model: Non-parametric Mixture of Binomials

- N independent coins
  Each coin has its own bias drawn from $P^*$
  
  $p_i \sim P^*$ (unknown)

- We get to observe t tosses for every coin

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

  $t = 5$ tosses
  
  $x_i = 2$
Model: Non-parametric Mixture of Binomials

- N independent coins
  Each coin has its own bias drawn from $P^*$
  
  $i = 1, 2, \ldots, N$ 
  
  $p_i \sim P^*$ (unknown)

- We get to observe $t$ tosses for every coin

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

  $t = 5$ tosses
  
  $x_i = 2$

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

Lord 1965, 1969
Model: Non-parametric Mixture of Binomials

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

\[ i = 1, 2, ..., N \quad p_i \sim P^* \quad (\text{unknown}) \]

- We get to observe t tosses for every coin

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

\[ t = 5 \text{ tosses} \quad \{1 \ 0 \ 0 \ 1 \ 0\} \quad x_i = 2 \]

- Given \( \{X_i\}_{i=1}^{N} \), return \( \hat{P} \) estimate of \( P^* \)

- Wasserstein-1 distance
  (Earth Mover’s Distance)

\[ W_1 \left( P^*, \hat{P} \right) \]
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 \( \Theta\left( \frac{1}{\sqrt{t}} \right) \)

\[ N = \text{Number of coins} \quad t = \text{Number of tosses per coin} \]
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 \( \Theta\left( \frac{1}{\sqrt{t}} \right) \)

\[ N = \text{Number of coins} \quad t = \text{Number of tosses per coin} \]

- 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
Learning with Sparse Observations is Non-trivial

- Empirical plug-in estimator is bad
  \[ \hat{P}_{\text{plug-in}} = \text{histogram} \left\{ \frac{X_1}{t}, \frac{X_i}{t}, \ldots, \frac{X_N}{t} \right\} \]
  When \( t \ll N \) incurs error of \( \Theta \left( \frac{1}{\sqrt{t}} \right) \)

- 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 \( O \left( \frac{1}{t} \right) \) 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
Learning with Sparse Observations is Non-trivial

• Empirical plug-in estimator is bad

\[ \hat{P}_{\text{plug-in}} = \text{histogram} \left\{ \frac{X_1}{t}, \frac{X_i}{t}, \ldots, \frac{X_N}{t} \right\} \]

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

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
Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

\[ h_s = \frac{\text{# coins that show } s \text{ heads}}{N} \quad s = 0, 1, \ldots, t \]

\[ h = [h_0, h_1, \ldots, h_s, \ldots, h_t] \quad \text{fingerprint vector} \]
Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

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

$$h = [h_0, h_1, \ldots, h_s, \ldots, h_t]$$ fingerprint vector

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

Sufficient statistic: Fingerprint

\[ h_s = \frac{\# \text{ coins that show } s \text{ heads}}{N} \quad s = 0, 1, ..., t \]

\[ h = [h_0, h_1, ... h_s, ..., h_t] \quad \text{fingerprint vector} \]

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

- NOT the empirical estimator

- Convex optimization: Efficient (polynomial time)
Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

\[ h_s = \frac{\# \text{ coins that show } s \text{ heads}}{N} \quad s = 0, 1, \ldots, t \]

\[ \mathbf{h} = [h_0, h_1, \ldots h_s, \ldots, h_t] \quad \text{fingerprint vector} \]

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

- 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


\[ h_s \]

\[ s = 0, 1, 2, 3, 4, 5 \]
Maximum Likelihood Estimator

Sufficient statistic: Fingerprint

\[ h_s = \frac{\# \text{ coins that show } s \text{ heads}}{N} \quad s = 0, 1, \ldots, t \]

\[ h = [h_0, h_1, \ldots, h_s, \ldots, h_t] \] fingerprint vector

\[ \hat{P}_{\text{mle}} \in \arg \min_{Q \in \text{dist}[0,1]} \text{KL}(\text{Observed } h, \text{ Expected } h \text{ under the distribution } Q) \]

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

Main Results: MLE is Minimax Optimal in Sparse Regime

Non-asymptotic guarantees

**Theorem 1**

The MLE achieves following error bounds: w. p. \( \geq 1 - \delta \)

- **Small Sample Regime:**
  \[
  W_1 \left( P^*, \hat{P}_{mle} \right) = \mathcal{O}_\delta \left( \frac{1}{t} \right) \text{ when } t < c \log N
  \]

- **Medium Sample Regime:**
  \[
  W_1 \left( P^*, \hat{P}_{mle} \right) = \mathcal{O}_\delta \left( \frac{1}{\sqrt{t \log N}} \right) \text{ when } c \log N \leq t \leq N^{2/9 - \epsilon}
  \]

\( N = \) Number of coins

\( t = \) Number of tosses per coin

Sparse Regime

\( t \ll N \)
Main Results: MLE is Minimax Optimal in Sparse Regime

Non-asymptotic guarantees

**Theorem 1**
The MLE achieves following error bounds: w. p. $\geq 1 - \delta$

- Small Sample Regime:
  \[ W_1 \left( P^*, \hat{P}_{\text{mle}} \right) = O_\delta \left( \frac{1}{t} \right) \text{ when } t < c \log N \]

- Medium Sample Regime:
  \[ W_1 \left( P^*, \hat{P}_{\text{mle}} \right) = O_\delta \left( \frac{1}{\sqrt{t \log N}} \right) \text{ when } c \log N \leq t \leq N^{2/9 - \epsilon} \]

**Theorem 2**
- Matching Minimax Lower Bounds
  \[ \inf_{f} \sup_{P} \mathbb{E}[W_1(P, f(X))] > \Omega \left( \frac{1}{t} \right) \lor \Omega \left( \frac{1}{\sqrt{t \log N}} \right) \]

$N =$ Number of coins
$t =$ Number of tosses per coin
Sparse Regime $t \ll N$

- Poster #189
Novel Proof: Polynomial Approximations

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

$$\hat{f}_t(x) = \sum_{j=0}^{t} b_j \binom{t}{j} x^j (1 - x)^{t-j}$$
Novel Proof: Polynomial Approximations

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

\[
\hat{f}_t(x) = \sum_{j=0}^{t} b_j \binom{t}{j} x^j (1 - x)^{t-j}
\]

Performance on Real Data

Bernstein polynomials
**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

ramya@cs.washington.edu