Optimal Learning of Patterns from Discrete Samples

Yanjun Han (Stanford EE)

Joint work with:

Jiantao Jiao	Stanford EE
Tsachy Weissman	Stanford EE

July 7th, 2017

Outline

Pattern Learning Problem

Given n i.i.d samples drawn from a discrete distribution $P = (p_1, \cdots, p_S)$ with an *unknown* support size S, we would like to learn the patterns of P, including:

- the distribution P itself
- some functional of P, e.g., the entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$ and the support size $S(P) = \sum_{i=1}^{S} \mathbb{1}(p_i \neq 0)$

Pattern Learning Problem

Given n i.i.d samples drawn from a discrete distribution $P = (p_1, \cdots, p_S)$ with an *unknown* support size S, we would like to learn the patterns of P, including:

- the distribution P itself
- some functional of P, e.g., the entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$ and the support size $S(P) = \sum_{i=1}^{S} \mathbb{1}(p_i \neq 0)$

Remark

Things get interesting when S is large.

Our Problem

Target

Learn the “spectrum/histogram” of P, i.e., learn the distribution vector $P = (p_1, \cdots, p_S)$ up to permutation.

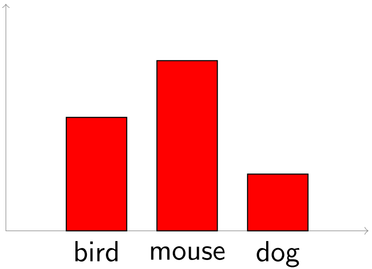
Our Problem

Target

Learn the "spectrum/histogram" of P, i.e., learn the distribution vector $P = (p_1, \cdots, p_S)$ up to permutation.

Example

Suppose our observation for animals on an island is {mouse, mouse, bird, dog, mouse, bird}, we would like to obtain:



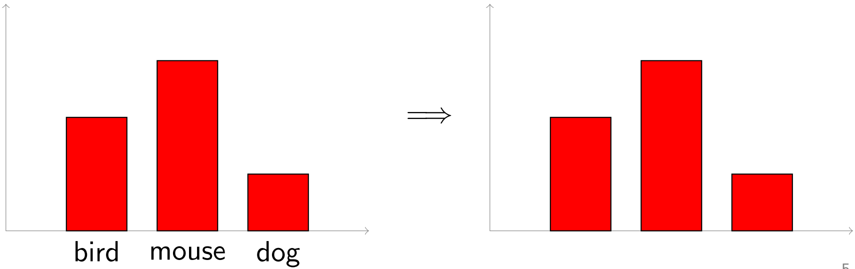
Our Problem

Target

Learn the "spectrum/histogram" of P, i.e., learn the distribution vector $P = (p_1, \cdots, p_S)$ up to permutation.

Example

Suppose our observation for animals on an island is {mouse, mouse, bird, dog, mouse, bird}, we would like to obtain:



Motivation

The spectrum contains some essential information of the distribution:

- shape of the distribution: unimodal or not, light-tail or heavy-tail, etc
- symmetric functional of the distribution: can be plugged into general functionals of the form $F(P) = \sum_{i=1}^{S} f(p_i)$

Two-step Learning of Distribution

Suppose now we would like to estimate P without permutation. We may decompose this process into two steps:

- Step 1: learn the distribution P without labeling (our target!)
- Step 2: assign labels to the unlabeled distribution obtained in Step 1.

Two-step Learning of Distribution

Suppose now we would like to estimate P without permutation. We may decompose this process into two steps:

- Step 1: learn the distribution P without labeling (our target!)
- Step 2: assign labels to the unlabeled distribution obtained in Step 1.

Question

Which step is more difficult?

A Non-trivial Answer

Theorem (Valiant and Valiant'16)

Even for $S = +\infty$, there is some estimator $\hat{P}$ of P such that for any discrete distribution P, and any oracle $\hat{P}^*$ who observes the same samples and knows P up to permutation,

$$\mathbb{E}_P \|\hat{P} - P\|_1 \leq \mathbb{E}_P \|\hat{P}^* - P\|_1 + o_n(1).$$

A Non-trivial Answer

Theorem (Valiant and Valiant'16)

Even for $S = +\infty$, there is some estimator $\hat{P}$ of P such that for any discrete distribution P, and any oracle $\hat{P}^*$ who observes the same samples and knows P up to permutation,

$$\mathbb{E}_P\|\hat{P} - P\|_1 \le \mathbb{E}_P\|\hat{P}^* - P\|_1 + o_n(1).$$

It seems that labeling is a hard task even if we knew the distribution...

Combining the Two Steps

Let $\mathcal{M}_S$ be the class of all probability distributions supported on at most S elements.

Combining the Two Steps

Let $\mathcal{M}_S$ be the class of all probability distributions supported on at most S elements.

Theorem (Optimal Learning of Labeled Distribution, H.–Jiao–Weissman'15, Kamath et al.'15)

The minimax ℓ_1 risk of distribution learning is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P\|_1 \asymp \sqrt{\frac{S}{n}}$$

and the upper bound is attained by the natural estimator.

Combining the Two Steps

Let $\mathcal{M}_S$ be the class of all probability distributions supported on at most S elements.

Theorem (Optimal Learning of Labeled Distribution, H.–Jiao–Weissman'15, Kamath et al.'15)

The minimax ℓ_1 risk of distribution learning is

$$\inf_{\hat{P}} \sup_{P\in\mathcal{M}_S} \mathbb{E}_P\|\hat{P}-P\|_1 \asymp \sqrt{\frac{S}{n}}$$

and the upper bound is attained by the natural estimator.

Corollary

Labeled distribution learning is possible if and only if $n \gg S$.

Proof of Upper Bound

By definition, we have $n\hat{p}_i \sim \mathrm{B}(n, p_i)$.

Proof of Upper Bound

By definition, we have $n\hat{p}_i \sim \mathrm{B}(n, p_i)$. Hence,

$$\begin{aligned}\mathbb{E}|\hat{p}_i - p_i| &\le \sqrt{\mathbb{E}(\hat{p}_i - p_i)^2} \\ &= \sqrt{\frac{p_i(1-p_i)}{n}} \\ &\le \sqrt{\frac{p_i}{n}}.\end{aligned}$$

Proof of Upper Bound

By definition, we have $n\hat{p}_i \sim \mathrm{B}(n, p_i)$. Hence,

$$
\begin{aligned}
\mathbb{E}|\hat{p}_i - p_i| &\le \sqrt{\mathbb{E}(\hat{p}_i - p_i)^2} \\
&= \sqrt{\frac{p_i(1-p_i)}{n}} \\
&\le \sqrt{\frac{p_i}{n}}.
\end{aligned}
$$

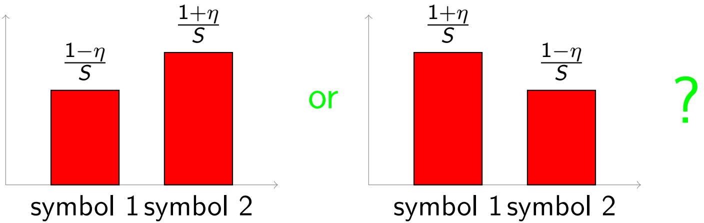
Summing up:

$$
\mathbb{E}_P\|\hat{P} - P\|_1 \le \sum_{i=1}^{S} \sqrt{\frac{p_i}{n}} \le \sqrt{\frac{S}{n}}.
$$

Proof of Lower Bound

Simple Fact

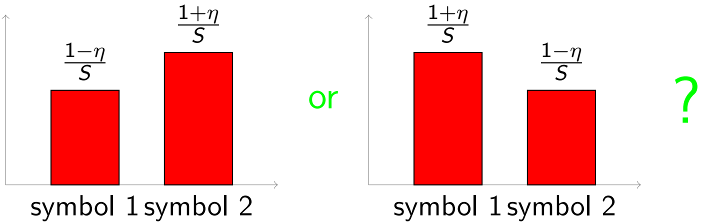
When $\eta \asymp \sqrt{\frac{S}{n}}$, the distributions $\mathrm{B}(n, \frac{1-\eta}{S})$ and $\mathrm{B}(n, \frac{1+\eta}{S})$ are indistinguishable using n samples.



Proof of Lower Bound

Simple Fact

When $\eta \asymp \sqrt{\frac{S}{n}}$, the distributions $\mathrm{B}(n, \frac{1-\eta}{S})$ and $\mathrm{B}(n, \frac{1+\eta}{S})$ are indistinguishable using n samples.



Implication

Each symbol contributes error $\frac{\eta}{S}$, and thus $\eta \asymp \sqrt{\frac{S}{n}}$ error in total.

Loss Criterion for Our Problem

Let $P_< = (p_{(1)}, p_{(2)}, \cdots, p_{(S)})$ with $p_{(1)} < p_{(2)} < \cdots < p_{(S)}$ be the sorted version of P. We would like to minimize the sorted ℓ_1 loss:

Minimax Sorted ℓ_1 Risk

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1$$

Main Result

Theorem (Optimal Learning of Unlabeled Distribution, H.–Jiao–Weissman'17)

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \| \hat{P} - P_< \|_1 \asymp \sqrt{\frac{S}{n \ln n}} + \tilde{\Theta}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{S}{n}} \right)$$

where $\tilde{\Theta}(\cdot)$ *neglects* $o(\text{poly}(n))$ *factors, and our estimator (to be presented) attains the upper bound.*

Main Result

Theorem (Optimal Learning of Unlabeled Distribution, H.–Jiao–Weissman'17)

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{S}{n \ln n}} + \tilde{\Theta}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{S}{n}}\right)$$

where $\tilde{\Theta}(\cdot)$ neglects $o(\text{poly}(n))$ factors, and our estimator (to be presented) attains the upper bound.

Corollary

Unlabeled distribution learning is possible if and only if $n \gg \frac{S}{\ln S}$.

Main Result

Theorem (Optimal Learning of Unlabeled Distribution, H.–Jiao–Weissman'17)

The minimax sorted ℓ_1 risk of learning unlabeled distribution is

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \asymp \sqrt{\frac{S}{n \ln n}} + \tilde{\Theta}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{S}{n}}\right)$$

where $\tilde{\Theta}(\cdot)$ neglects $o(\mathrm{poly}(n))$ factors, and our estimator (to be presented) attains the upper bound.

Corollary

Unlabeled distribution learning is possible if and only if $n \gg \frac{S}{\ln S}$.

Alert

Uniform improvements over the natural estimator is possible only when $S \gg \tilde{\Theta}(n^{\frac{1}{3}})$.

First Let's Make Everything Simple...

Let's assume:

- support size S is known;
- each p_i is small; more specifically, $p_i \in [0, \frac{\ln n}{n}]$

First Let's Make Everything Simple...

Let's assume:

- support size S is known;
- each p_i is small; more specifically, $p_i \in [0, \frac{\ln n}{n}]$

A thought experiment

We have

$$\text{unlabeled distribution} \Longrightarrow \text{symmetric functional.}$$

How about the opposite direction?

Idea: Moment Matching

Suppose we could find some $Q = (q_1, \cdots, q_S)$ such that $q_1, \cdots, q_S \in [0, \frac{\ln n}{n}]$, and

$$
\begin{aligned}
q_1^0 + q_2^0 + \cdots + q_S^0 &= p_1^0 + p_2^0 + \cdots + p_S^0 \\
q_1^1 + q_2^1 + \cdots + q_S^1 &= p_1^1 + p_2^1 + \cdots + p_S^1 \\
q_1^2 + q_2^2 + \cdots + q_S^2 &= p_1^2 + p_2^2 + \cdots + p_S^2 \\
&\cdots\cdots \\
q_1^K + q_2^K + \cdots + q_S^K &= p_1^K + p_2^K + \cdots + p_S^K
\end{aligned}
$$

for some K.

Idea: Moment Matching

Suppose we could find some $Q = (q_1, \cdots, q_S)$ such that $q_1, \cdots, q_S \in [0, \frac{\ln n}{n}]$, and

$$q_1^0 + q_2^0 + \cdots + q_S^0 = p_1^0 + p_2^0 + \cdots + p_S^0$$
$$q_1^1 + q_2^1 + \cdots + q_S^1 = p_1^1 + p_2^1 + \cdots + p_S^1$$
$$q_1^2 + q_2^2 + \cdots + q_S^2 = p_1^2 + p_2^2 + \cdots + p_S^2$$
$$\cdots\cdots$$
$$q_1^K + q_2^K + \cdots + q_S^K = p_1^K + p_2^K + \cdots + p_S^K$$

for some K. How about using $Q_<$ as an estimate of $P_<$?

Idea: Moment Matching

Suppose we could find some $Q = (q_1, \cdots, q_S)$ such that $q_1, \cdots, q_S \in [0, \frac{\ln n}{n}]$, and

$$
\begin{aligned}
q_1^0 + q_2^0 + \cdots + q_S^0 &= p_1^0 + p_2^0 + \cdots + p_S^0 \\
q_1^1 + q_2^1 + \cdots + q_S^1 &= p_1^1 + p_2^1 + \cdots + p_S^1 \\
q_1^2 + q_2^2 + \cdots + q_S^2 &= p_1^2 + p_2^2 + \cdots + p_S^2 \\
&\cdots\cdots \\
q_1^K + q_2^K + \cdots + q_S^K &= p_1^K + p_2^K + \cdots + p_S^K
\end{aligned}
$$

for some K. How about using $Q_<$ as an estimate of $P_<$?

Goal

Show that

$$\text{moment matching} \Longrightarrow \text{distribution closeness.}$$

Wasserstein Distance

Definition (Wasserstein Distance)

Let (S, d) be a separable metric space, and P, Q be two probability measures on S. The Wasserstein Distance between P and Q is defined as

$$W(P, Q) \triangleq \inf_{\mathcal{L}(X)=P, \mathcal{L}(Y)=Q} \mathbb{E} d(X, Y)$$

where X, Y are random variables taking values in S.

Wasserstein Distance

Definition (Wasserstein Distance)

Let (S, d) be a separable metric space, and P, Q be two probability measures on S. The Wasserstein Distance between P and Q is defined as

$$W(P, Q) \triangleq \inf_{\mathcal{L}(X)=P, \mathcal{L}(Y)=Q} \mathbb{E} d(X, Y)$$

where X, Y are random variables taking values in S.

Theorem (Dual Representation, Kantorovich–Rubinstein'58)

Define the Lipschitz norm in (S, d) as $\|f\|_{Lip} \triangleq \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$, then

$$W(P, Q) = \sup_{f:\|f\|_{Lip} \leq 1} \mathbb{E}_P f - \mathbb{E}_Q f.$$

Rearrangement Inequality

Let μ_P be the uniform distribution on the multiset $\{p_1, \cdots, p_S\}$, and similarly for μ_Q.

Rearrangement Inequality

Let μ_P be the uniform distribution on the multiset $\{p_1, \cdots, p_S\}$, and similarly for μ_Q.

Lemma (Rearrangement Inequality)

For $(S, d) = ([0,1], |\cdot|)$, *we have*

$$\begin{aligned} \|P_< - Q_<\|_1 &= S \cdot \inf_{\mathcal{L}(X)=\mu_P, \mathcal{L}(Y)=\mu_Q} \mathbb{E}|X - Y| \\ &= S \cdot W(\mu_P, \mu_Q). \end{aligned}$$

Rearrangement Inequality

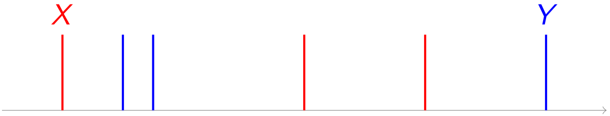
Let μ_P be the uniform distribution on the multiset $\{p_1, \cdots, p_S\}$, and similarly for μ_Q.

Lemma (Rearrangement Inequality)

For $(S, d) = ([0,1], |\cdot|)$, *we have*

$$\|P_< - Q_<\|_1 = S \cdot \inf_{\mathcal{L}(X)=\mu_P, \mathcal{L}(Y)=\mu_Q} \mathbb{E}|X - Y|$$
$$= S \cdot W(\mu_P, \mu_Q).$$

Example



Rearrangement Inequality

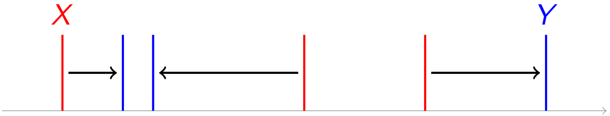
Let μ_P be the uniform distribution on the multiset $\{p_1, \cdots, p_S\}$, and similarly for μ_Q.

Lemma (Rearrangement Inequality)

For $(S, d) = ([0,1], |\cdot|)$, *we have*

$$\begin{aligned}\|P_< - Q_<\|_1 &= S \cdot \inf_{\mathcal{L}(X)=\mu_P, \mathcal{L}(Y)=\mu_Q} \mathbb{E}|X - Y| \\ &= S \cdot W(\mu_P, \mu_Q).\end{aligned}$$

Example



Using Moment Matching

$$\begin{aligned}
\|P_< - Q_<\|_1 &= S \cdot W(\mu_P, \mu_Q) \\
&= S \cdot \sup_{f:\|f\|_{\text{Lip}}\le 1} \mathbb{E}_{\mu_P} f - \mathbb{E}_{\mu_Q} f \qquad \text{[dual representation]} \\
&= \sup_{f:\|f\|_{\text{Lip}}\le 1} \sum_{i=1}^{S} f(p_i) - f(q_i) \qquad \text{[by definition of } \mu_P, \mu_Q\text{]} \\
&= \sup_{f:\|f\|_{\text{Lip}}\le 1} \inf_{\deg P \le K} \sum_{i=1}^{S} (f(p_i) - P(p_i)) - (f(q_i) - P(q_i)) \\
&\qquad \text{[moment matching up to order } K\text{]} \\
&\le \sup_{f:\|f\|_{\text{Lip}}\le 1} \inf_{\deg P \le K} \sum_{i=1}^{S} |f(p_i) - P(p_i)| + |f(q_i) - P(q_i)| \\
&\qquad \text{[triangle inequality]}
\end{aligned}$$

Polynomial Approximation of Lipschitz Function

Theorem (Jackson's Inequality, Devore'76)

Let f be any 1-Lipschitz function on $[a, b]$. There exists a degree-K polynomial P such that for any $x \in (a, b)$,

$$|f(x) - P(x)| \lesssim \frac{\sqrt{(b-a)(x-a)}}{K}.$$

Polynomial Approximation of Lipschitz Function

Theorem (Jackson's Inequality, Devore'76)

Let f be any 1-Lipschitz function on $[a, b]$. There exists a degree-K polynomial P such that for any $x \in (a, b)$,

$$|f(x) - P(x)| \lesssim \frac{\sqrt{(b-a)(x-a)}}{K}.$$

Choosing $[a, b] = [0, \frac{\ln n}{n}], K \asymp \ln n$, we have

$$\|P_< - Q_<\|_1 \lesssim \sum_{i=1}^{S} \sqrt{\frac{p_i}{n \ln n}} + \sqrt{\frac{q_i}{n \ln n}} \lesssim \sqrt{\frac{S}{n \ln n}}.$$

Implication

Implication

For unlabeled distribution learning, it suffices to match moments up to order $\ln n$.

Implication

Implication

For unlabeled distribution learning, it suffices to match moments up to order $\ln n$.

Questions

- What to do since we do not know the true moments $\sum_{i=1}^{S} p_i^k$?

Implication

Implication

For unlabeled distribution learning, it suffices to match moments up to order $\ln n$.

Questions

- What to do since we do not know the true moments $\sum_{i=1}^{S} p_i^k$?
- How to match moments and solve for Q efficiently? What if there is no solution?

Implication

Implication

For unlabeled distribution learning, it suffices to match moments up to order $\ln n$.

Questions

- What to do since we do not know the true moments $\sum_{i=1}^{S} p_i^k$?
- How to match moments and solve for Q efficiently? What if there is no solution?
- What if not all p_i lie in the interval $[0, \frac{\ln n}{n}]$?

Implication

Implication

For unlabeled distribution learning, it suffices to match moments up to order $\ln n$.

Questions

- What to do since we do not know the true moments $\sum_{i=1}^{S} p_i^k$?
- How to match moments and solve for Q efficiently? What if there is no solution?
- What if not all p_i lie in the interval $[0, \frac{\ln n}{n}]$?
- What if the support size S is unknown?

Q1: How to Know the True Moments $\sum_{i=1}^{S} p_i^k$?

Answer

Apply an unbiased estimator of the moments.

Q1: How to Know the True Moments $\sum_{i=1}^{S} p_i^k$?

Answer

Apply an unbiased estimator of the moments.

Fact

For $X \sim \mathrm{B}(n, p)$, we have

$$\mathbb{E}_p\left[\frac{X(X-1)\cdots(X-k+1)}{n(n-1)\cdots(n-k+1)}\right] = p^k, \qquad 1 \le k \le n.$$

Just use the support size S for $k = 0$.

Q1: How to Know the True Moments $\sum_{i=1}^{S} p_i^k$?

Answer

Apply an unbiased estimator of the moments.

Fact

For $X \sim \mathrm{B}(n,p)$, we have

$$\mathbb{E}_p\left[\frac{X(X-1)\cdots(X-k+1)}{n(n-1)\cdots(n-k+1)}\right] = p^k, \qquad 1 \le k \le n.$$

Just use the support size S for $k=0$.

Alert

If the plug-in idea $\sum_{i=1}^{S} \hat{p}_i^k$ were used, the moment matching process would return the empirical distribution!

Q1: How Much Do We Lose?

Instead of exact moment matching, now we have:

$$\mathbb{E}\left|\sum_{i=1}^{S} q_i^k - \sum_{i=1}^{S} p_i^k\right| \lesssim \tilde{\mathcal{O}}\left(\frac{1}{n^{k-\frac{1}{2}}}\right)$$

Tracing back to the proof, this incurs a negligible additional error $\tilde{\mathcal{O}}(n^{-\frac{1}{2}})$ to the original problem.

Remark

The unbiased estimator is used to avoid bias accumulation (where the variance cancels out).

Q2: How to Implement Efficient Moment Matching?

Answer

Compute a continuous density μ_Q instead of a discrete vector Q.

Q2: How to Implement Efficient Moment Matching?

Answer

Compute a continuous density μ_Q instead of a discrete vector Q.

Algorithm

Solve the following feasibility problem: check whether the system

$$\left| S \cdot \int_0^{\frac{\ln n}{n}} x^k \mu_Q(dx) - \sum_{i=1}^{S} \frac{n\hat{p}_i(n\hat{p}_i - 1)\cdots(n\hat{p}_i - k + 1)}{n(n-1)\cdots(n-k+1)} \right| \lesssim \tilde{\mathcal{O}}\left(\frac{1}{n^{k-\frac{1}{2}}}\right)$$

for all $k = 1, \cdots, K$ contains a feasible probability measure μ_Q.
Choose any one if there are multiple solutions.

Q2: Analysis of the Algorithm

$$\left| S \cdot \int_0^{\frac{\ln n}{n}} x^k \mu_Q(dx) - \sum_{i=1}^{S} \frac{n\hat{p}_i(n\hat{p}_i - 1)\cdots(n\hat{p}_i - k + 1)}{n(n-1)\cdots(n-k+1)} \right| \lesssim \tilde{\mathcal{O}}\left(\frac{1}{n^{k-\frac{1}{2}}}\right)$$

Q2: Analysis of the Algorithm

$$\left| S \cdot \int_0^{\frac{\ln n}{n}} x^k \mu_Q(dx) - \sum_{i=1}^{S} \frac{n\hat{p}_i(n\hat{p}_i - 1)\cdots(n\hat{p}_i - k + 1)}{n(n-1)\cdots(n-k+1)} \right| \lesssim \tilde{\mathcal{O}}(\frac{1}{n^{k-\frac{1}{2}}})$$

Key Observation

There is a feasible solution with overwhelming probability since μ_P is!

Q2: Analysis of the Algorithm

$$\left| S \cdot \int_0^{\frac{\ln n}{n}} x^k \mu_Q(dx) - \sum_{i=1}^{S} \frac{n\hat{p}_i(n\hat{p}_i - 1)\cdots(n\hat{p}_i - k + 1)}{n(n-1)\cdots(n-k+1)} \right| \lesssim \tilde{\mathcal{O}}\left(\frac{1}{n^{k-\frac{1}{2}}}\right)$$

Key Observation

There is a feasible solution with overwhelming probability since μ_P is!

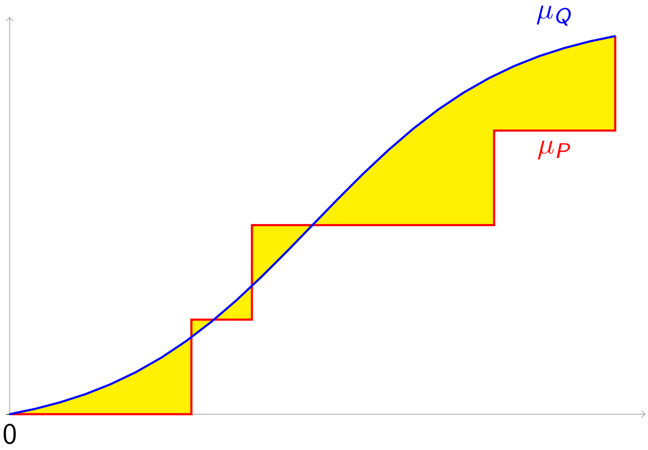
Implementation

- A linear program in μ_Q, but infinite dimensional
- Can transform into a finite-dimensional LP by quantizing μ_Q

Q2: From Continuous μ_Q to Discrete Q

Proof via figure:

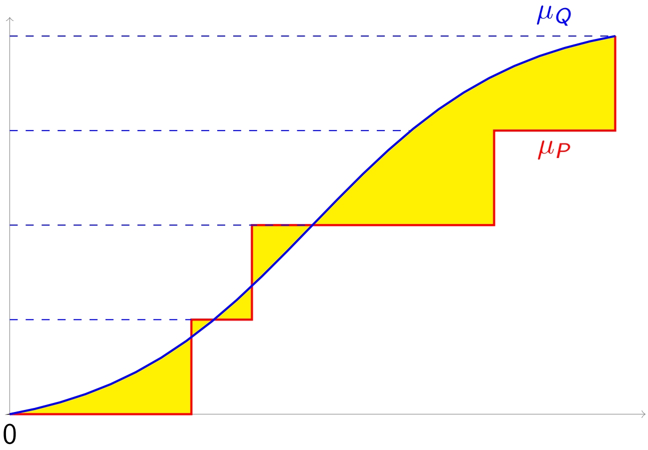
$$W(\mu_P, \mu_Q) = \text{yellow area} = \mathbb{E}W(\mu_P, \mu_Q')$$



Q2: From Continuous μ_Q to Discrete Q

Proof via figure:

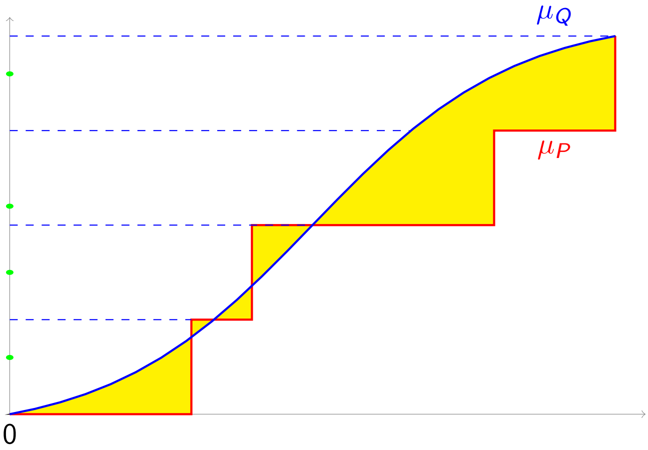
$$W(\mu_P, \mu_Q) = \text{yellow area} = \mathbb{E}W(\mu_P, \mu_Q')$$



Q2: From Continuous μ_Q to Discrete Q

Proof via figure:

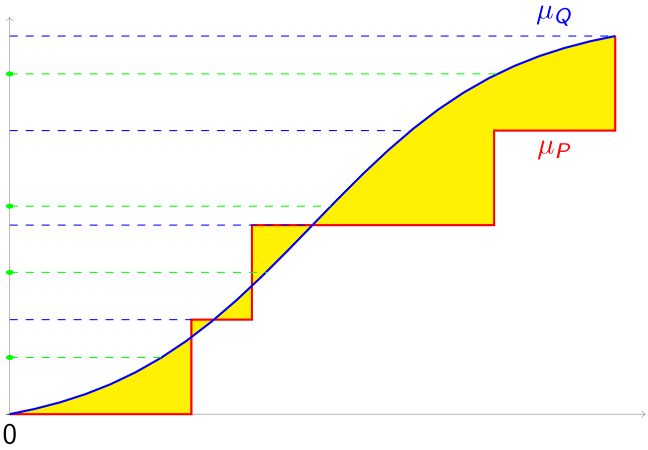
$$W(\mu_P, \mu_Q) = \text{yellow area} = \mathbb{E}W(\mu_P, \mu_Q')$$



Q2: From Continuous μ_Q to Discrete Q

Proof via figure:

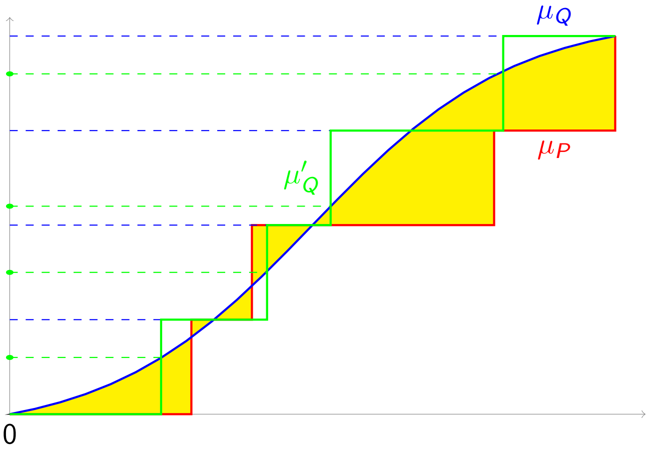
$$W(\mu_P, \mu_Q) = \text{yellow area} = \mathbb{E}W(\mu_P, \mu_Q')$$



Q2: From Continuous μ_Q to Discrete Q

Proof via figure:

$$W(\mu_P, \mu_Q) = \text{yellow area} = \mathbb{E}W(\mu_P, \mu_Q')$$



Q3: Not All Rare Symbols

Answer

Generalize polynomial approximation idea to other intervals.

Q3: Not All Rare Symbols

Answer

Generalize polynomial approximation idea to other intervals.

Fact

The same technique applies to the case where all p_i lie in $I_p \triangleq [p - \sqrt{\frac{p \ln n}{n}}, p + \sqrt{\frac{p \ln n}{n}}]$ for any p:

$$
\begin{aligned}
\|Q_< - P_<\|_1 &\lesssim S \cdot \sup_{f:\|f\|_{\mathrm{Lip}} \le 1} \inf_{\deg P \le K} \|f - P\|_{\infty, I_p} \\
&\lesssim S \cdot \frac{|I_p|}{K} \qquad \text{[Jackson's Inequality]} \\
&\lesssim S \cdot \sqrt{\frac{p}{n \ln n}} \qquad [K \asymp \ln n] \\
&\lesssim \sqrt{\frac{S}{n \ln n}} \qquad [pS \asymp 1]
\end{aligned}
$$

Q3: Partitioning and Moment Matching

Idea

Partitioning the whole interval $[0,1]$ into sub-intervals of the previous form, and match moments separately in each sub-interval.

Q3: Partitioning and Moment Matching

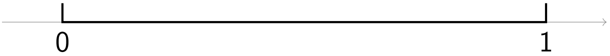
Idea

Partitioning the whole interval $[0, 1]$ into sub-intervals of the previous form, and match moments separately in each sub-interval.

Resulting Partition

Let $\eta_n = \frac{c \ln n}{n}$ with a suitable parameter c, the partition is

$$[0, \eta_n], [\eta_n, 4\eta_n], [4\eta_n, 9\eta_n], \cdots .$$



New Difficulty

Need to know which interval each probability mass p_i belongs to, which we actually do not know.

Q3: Partitioning and Moment Matching

Idea

Partitioning the whole interval $[0, 1]$ into sub-intervals of the previous form, and match moments separately in each sub-interval.

Resulting Partition

Let $\eta_n = \frac{c \ln n}{n}$ with a suitable parameter c, the partition is

$$[0, \eta_n], [\eta_n, 4\eta_n], [4\eta_n, 9\eta_n], \cdots .$$

New Difficulty

Need to know which interval each probability mass p_i belongs to, which we actually do not know.

Q3: Confidence Set

Definition (Confidence set)

Consider a statistical model $(P_\theta)_{\theta\in\Theta}$ and an estimator $\hat{\theta} \in \hat{\Theta}$ of θ, where $\Theta \subset \hat{\Theta}$. A confidence set of significant level $r \in [0,1]$, or an r-confidence set, is a collection of sets $\{U(x)\}_{x\in\hat{\Theta}}$, where $U(x) \subset \Theta$ for any $x \in \hat{\Theta}$, and

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\theta \notin U(\hat{\theta})) \le r.$$

Q3: Confidence Set

Definition (Confidence set)

Consider a statistical model $(P_\theta)_{\theta\in\Theta}$ and an estimator $\hat{\theta} \in \hat{\Theta}$ of θ, where $\Theta \subset \hat{\Theta}$. A confidence set of significant level $r \in [0,1]$, or an r-confidence set, is a collection of sets $\{U(x)\}_{x\in\hat{\Theta}}$, where $U(x) \subset \Theta$ for any $x \in \hat{\Theta}$, and

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta(\theta \notin U(\hat{\theta})) \le r.$$

- Confidence set always exists, but we seek for a small one
- Choice of significance: $r \asymp n^{-A}$

Q3: Confidence Set in Binomial Model

Example

0 ——— 1

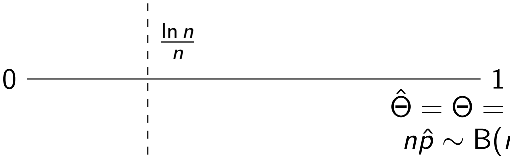
$$\hat{\Theta} = \Theta = [0, 1]$$
$$n\hat{p} \sim \mathrm{B}(n, p)$$

Remark

Each set in the partition is exactly a confidence set!

Q3: Confidence Set in Binomial Model

Example



$$\hat{\Theta} = \Theta = [0, 1]$$
$$n\hat{p} \sim \mathrm{B}(n, p)$$

Remark

Each set in the partition is exactly a confidence set!

Q3: Confidence Set in Binomial Model

Example

Remark

Each set in the partition is exactly a confidence set!

Q3: Confidence Set in Binomial Model

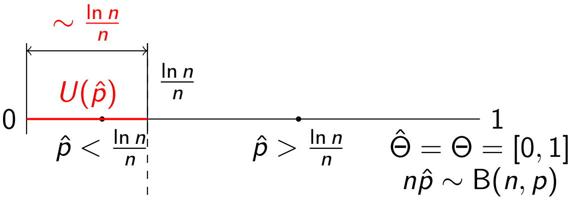
Example

Remark

Each set in the partition is exactly a confidence set!

Q3: Confidence Set in Binomial Model

Example

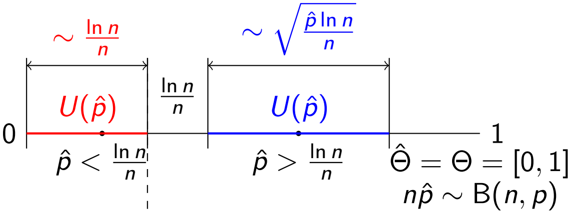


Remark

Each set in the partition is exactly a confidence set!

Q3: Confidence Set in Binomial Model

Example



Remark

Each set in the partition is exactly a confidence set!

Q3: Sample Splitting via Confidence Set

Sample Splitting Algorithm

Split the samples $X_1, \cdots, X_n \overset{i.i.d}{\sim} P$ into two halves:

- For each symbol i, use the empirical distribution of the first half to classify the partition set it belongs to;
- Match moments in each (slightly enlarged) partition set based on the classification in the first step.

Intuition

Since each partition set is also a confidence set, with overwhelming probability the true mass p_i lies in (an enlarged version of) the "told" region.

Q3: Sample Splitting via Confidence Set

Sample Splitting Algorithm

Split the samples $X_1, \cdots, X_n \overset{i.i.d}{\sim} P$ into two halves:

- For each symbol i, use the empirical distribution of the first half to classify the partition set it belongs to;
- Match moments in each (slightly enlarged) partition set based on the classification in the first step.

Intuition

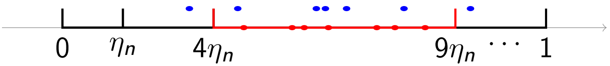
Since each partition set is also a confidence set, with overwhelming probability the true mass p_i lies in (an enlarged version of) the "told" region.

Q3: Sample Splitting via Confidence Set

Sample Splitting Algorithm

Split the samples $X_1, \cdots, X_n \overset{i.i.d}{\sim} P$ into two halves:

- For each symbol i, use the empirical distribution of the first half to classify the partition set it belongs to;
- Match moments in each (slightly enlarged) partition set based on the classification in the first step.



Intuition

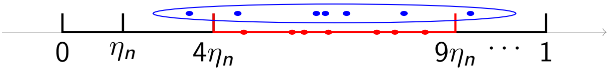
Since each partition set is also a confidence set, with overwhelming probability the true mass p_i lies in (an enlarged version of) the "told" region.

Q3: Sample Splitting via Confidence Set

Sample Splitting Algorithm

Split the samples $X_1, \cdots, X_n \overset{i.i.d}{\sim} P$ into two halves:

- For each symbol i, use the empirical distribution of the first half to classify the partition set it belongs to;
- Match moments in each (slightly enlarged) partition set based on the classification in the first step.



Intuition

Since each partition set is also a confidence set, with overwhelming probability the true mass p_i lies in (an enlarged version of) the "told" region.

Q3: Additional Loss

Observation

In each set of the partition, there is some loss due to the imperfect knowledge of the moments of μ_P.

Q3: Additional Loss

Observation

In each set of the partition, there is some loss due to the imperfect knowledge of the moments of μ_P.

Proposition

The loss incurred in the set A_j is given by

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{\sum_{p_i \in A_j} p_i}{n}}\right)$$

which gives the second term $\tilde{\Theta}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{S}{n}}\right)$ in the main theorem.

Q3: Additional Loss

Observation

In each set of the partition, there is some loss due to the imperfect knowledge of the moments of μ_P.

Proposition

The loss incurred in the set A_j is given by

$$\tilde{\mathcal{O}}\left(\sqrt{\frac{\sum_{p_i \in A_j} p_i}{n}}\right)$$

which gives the second term $\tilde{\Theta}\left(n^{-\frac{1}{3}} \wedge \sqrt{\frac{S}{n}}\right)$ in the main theorem.

Intuition

More improvements are possible if more symbols are grouped together.

Q4: Unknown Support Size S

Answer

Does not matter at all!

Q4: Unknown Support Size S

Answer

Does not matter at all!

Why?

- Support size has been made "known" by sample splitting
- Autofill zero in computing $\|\hat{P} - P_<\|_1$ if of different lengths

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;
- Partition $[0, 1]$ into $\cup_{r=0}^{R} A_r$ with $A_r = [r^2\eta_n, (r+1)^2\eta_n]$;

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;
- Partition $[0, 1]$ into $\cup_{r=0}^{R} A_r$ with $A_r = [r^2\eta_n, (r+1)^2\eta_n]$;
- Split samples and use the first half to classify the location of each symbol in the partition;

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;
- Partition $[0, 1]$ into $\cup_{r=0}^{R} A_r$ with $A_r = [r^2\eta_n, (r+1)^2\eta_n]$;
- Split samples and use the first half to classify the location of each symbol in the partition;
- Use the second half samples to compute the unbiased estimator of the k-moments in each partition set for $k = 1, 2, \cdots, K$;

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;
- Partition $[0,1]$ into $\cup_{r=0}^{R} A_r$ with $A_r = [r^2\eta_n, (r+1)^2\eta_n]$;
- Split samples and use the first half to classify the location of each symbol in the partition;
- Use the second half samples to compute the unbiased estimator of the k-moments in each partition set for $k = 1, 2, \cdots, K$;
- Match moments by solving the LP separately in each partition set;

Summary of the Estimator

- Choose suitable constant $c_1, c_2 > 0$, and let $\eta_n = \frac{c_1 \ln n}{n}$, $K = c_2 \ln n$;
- Partition $[0,1]$ into $\cup_{r=0}^{R} A_r$ with $A_r = [r^2\eta_n, (r+1)^2\eta_n]$;
- Split samples and use the first half to classify the location of each symbol in the partition;
- Use the second half samples to compute the unbiased estimator of the k-moments in each partition set for $k = 1, 2, \cdots, K$;
- Match moments by solving the LP separately in each partition set;
- Return the overall probability distribution.

When S Is Small

For small S, wish to prove:

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \gtrsim \sqrt{\frac{S}{n}}$$

When S Is Small

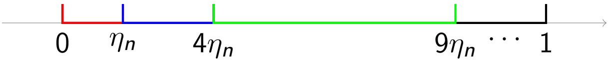
For small S, wish to prove:

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \gtrsim \sqrt{\frac{S}{n}}$$

Observation

Worst case occurs when each set in the partition contains at most one probability mass

- labeling step becomes easy
- essentially as hard as labeled distribution learning
- in this case, S cannot be too large



When S Is Small

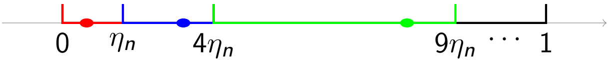
For small S, wish to prove:

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \gtrsim \sqrt{\frac{S}{n}}$$

Observation

Worst case occurs when each set in the partition contains at most one probability mass

- labeling step becomes easy
- essentially as hard as labeled distribution learning
- in this case, S cannot be too large



When S Is Large

For large S, wish to prove:

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \gtrsim \sqrt{\frac{S}{n \ln n}}$$

When S Is Large

For large S, wish to prove:

$$\inf_{\hat{P}} \sup_{P \in \mathcal{M}_S} \mathbb{E}_P \|\hat{P} - P_<\|_1 \gtrsim \sqrt{\frac{S}{n \ln n}}$$

Idea: Hypothesis Testing (Le Cam's Two Point Method)

Suffice to find $P, Q \in \mathcal{M}_S$ such that:

- $\|P_< - Q_<\|_1$ is large
- we cannot distinguish P, Q from observations $X_1, X_2, \cdots, X_n$

Fuzzy Hypothesis Testing

Generalized Le Cam's Method

Wish to find $\mu_P, \mu_Q \in \mathcal{P}(\mathcal{M}_S)$ such that:

- for $P \sim \mu_P, Q \sim \mu_Q$, $\|P_< - Q_<\|_1$ is probably large
- we cannot distinguish $P \sim \mu_P, Q \sim \mu_Q$ from observations $X_1, X_2, \cdots, X_n$

Fuzzy Hypothesis Testing

Generalized Le Cam's Method

Wish to find $\mu_P, \mu_Q \in \mathcal{P}(\mathcal{M}_S)$ such that:

- for $P \sim \mu_P, Q \sim \mu_Q$, $\|P_< - Q_<\|_1$ is probably large
- we cannot distinguish $P \sim \mu_P, Q \sim \mu_Q$ from observations $X_1, X_2, \cdots, X_n$

Try $\mu_P = \mu_1^{\otimes S}, \mu_Q = \mu_2^{\otimes S}$, where μ_1, μ_2 are both probability measures on $[0, 1]$:

Fuzzy Hypothesis Testing

Generalized Le Cam's Method

Wish to find $\mu_P, \mu_Q \in \mathcal{P}(\mathcal{M}_S)$ such that:

- for $P \sim \mu_P, Q \sim \mu_Q$, $\|P_< - Q_<\|_1$ is probably large
- we cannot distinguish $P \sim \mu_P, Q \sim \mu_Q$ from observations $X_1, X_2, \cdots, X_n$

Try $\mu_P = \mu_1^{\otimes S}, \mu_Q = \mu_2^{\otimes S}$, where μ_1, μ_2 are both probability measures on $[0,1]$:

Lemma (Wu–Yang'14, Jiao–H.–Weissman'17)

We cannot distinguish $P \sim \mu_1^{\otimes S}, Q \sim \mu_2^{\otimes S}$ from observations $X_1, X_2, \cdots, X_n$ if both μ_1, μ_2 are supported on $[p - \sqrt{\frac{p \ln n}{n}}, p + \sqrt{\frac{p \ln n}{n}}]$ for some $p \ge \frac{\ln n}{n}$, and

$$\mathbb{E}_{\mu_1} X^j = \mathbb{E}_{\mu_2} X^j, \qquad j = 0, 1, \cdots, K \asymp \ln n.$$

How Large Can the Difference Be?

Key Observation

By concentration of measure, for $P \sim \mu_1^{\otimes S}, Q \sim \mu_2^{\otimes S}$, $\|P_< - Q_<\|_1$ is close to the scaled Wasserstein distance $S \cdot W(\mu_1, \mu_2)$.

How Large Can the Difference Be?

Key Observation

By concentration of measure, for $P \sim \mu_1^{\otimes S}, Q \sim \mu_2^{\otimes S}$, $\|P_< - Q_<\|_1$ is close to the scaled Wasserstein distance $S \cdot W(\mu_1, \mu_2)$.

Duality

Wasserstein duality

$$W(\mu_1, \mu_2) = \sup_{f:\|f\|_{\text{Lip}} \le 1} \mathbb{E}_{\mu_1} f - \mathbb{E}_{\mu_2} f.$$

How Large Can the Difference Be?

Key Observation

By concentration of measure, for $P \sim \mu_1^{\otimes S}, Q \sim \mu_2^{\otimes S}$, $\|P_< - Q_<\|_1$ is close to the scaled Wasserstein distance $S \cdot W(\mu_1, \mu_2)$.

Duality

Wasserstein duality

$$W(\mu_1, \mu_2) = \sup_{f : \|f\|_{\text{Lip}} \le 1} \mathbb{E}_{\mu_1} f - \mathbb{E}_{\mu_2} f.$$

Implication: it suffices to find a suitable f with $\|f\|_{\text{Lip}} \le 1$.

Moment Matching and Another Duality

Moment Matching

For any f and two probability measures μ_1, μ_2 supported on $[a, b]$ with first K matching moments,

$$\mathbb{E}_{\mu_1} f - \mathbb{E}_{\mu_2} f = \inf_{\deg P \le K} \mathbb{E}_{\mu_1}(f - P) - \mathbb{E}_{\mu_2}(f - P)$$
$$\le 2 \cdot \inf_{\deg P \le K} \|f - P\|_{\infty,[a,b]}$$

Moment Matching and Another Duality

Moment Matching

For any f and two probability measures μ_1, μ_2 supported on $[a, b]$ with first K matching moments,

$$
\begin{aligned}
\mathbb{E}_{\mu_1} f - \mathbb{E}_{\mu_2} f &= \inf_{\deg P \le K} \mathbb{E}_{\mu_1}(f - P) - \mathbb{E}_{\mu_2}(f - P) \\
&\le 2 \cdot \inf_{\deg P \le K} \|f - P\|_{\infty,[a,b]}
\end{aligned}
$$

Lemma (Another Duality, Cai–Low'11)

There exist two probability measures μ_1^*, μ_2^* supported on $[a, b]$ with first K matching moments such that

$$
\mathbb{E}_{\mu_1^*} f - \mathbb{E}_{\mu_2^*} f = 2 \cdot \inf_{\deg P \le K} \|f - P\|_{\infty,[a,b]}.
$$

Another Viewpoint

Idea

Relate the unlabeled distribution learning problem to the functional estimation problem $\sum_{i=1}^{S} f(p_i)$ with $\|f\|_{\text{Lip}} \leq 1$.

Another Viewpoint

Idea

Relate the unlabeled distribution learning problem to the functional estimation problem $\sum_{i=1}^{S} f(p_i)$ with $\|f\|_{\text{Lip}} \leq 1$.

Observation

Functional estimation is *easier* than unlabeled distribution learning.

Another Viewpoint

Idea

Relate the unlabeled distribution learning problem to the functional estimation problem $\sum_{i=1}^{S} f(p_i)$ with $\|f\|_{\mathrm{Lip}} \leq 1$.

Observation

Functional estimation is *easier* than unlabeled distribution learning.

- By definition of Lipschitz property,

$$\left|\sum_{i=1}^{S} f(p_i) - f(q_i)\right| \leq \|P_< - Q_<\|_1.$$

Functional Estimation Problem

Given n i.i.d samples drawn from a discrete distribution $P = (p_1, \cdots, p_S)$ with an *unknown* support size S, we would like to estimate the functional of P of the form

$$F(P) = \sum_{i=1}^{S} f(p_i).$$

Functional Estimation Problem

Given n i.i.d samples drawn from a discrete distribution $P = (p_1, \cdots, p_S)$ with an *unknown* support size S, we would like to estimate the functional of P of the form

$$F(P) = \sum_{i=1}^{S} f(p_i).$$

- Shannon entropy $H(P) = \sum_{i=1}^{S} -p_i \ln p_i$
- power sum function $F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha$
- support size $S(P) = \sum_{i=1}^{S} \mathbb{1}(p_i \neq 0)$

Recent Breakthroughs

(Jiao–Venkat–H.–Weissman'14, Wu–Yang'14, Jiao–H.–Weissman'15, Wu–Yang'15)

	Minimax L_2 rate	L_2 rate of MLE
$H(P)=\sum_{i=1}^{S} -p_i \ln p_i$	$\frac{S^2}{(n\ln n)^2}+\frac{\ln^2 S}{n}$	$\frac{S^2}{n^2}+\frac{\ln^2 S}{n}$
$F_\alpha(P)=\sum_{i=1}^{S} p_i^\alpha, 0<\alpha<1/2$	$\frac{S^2}{(n\ln n)^{2\alpha}}$	$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P)=\sum_{i=1}^{S} p_i^\alpha, 1/2\le\alpha<1$	$\frac{S^2}{(n\ln n)^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}}+\frac{S^{2-2\alpha}}{n}$
$F_\alpha(P)=\sum_{i=1}^{S} p_i^\alpha, 1<\alpha<3/2$	$\frac{1}{(n\ln n)^{2(\alpha-1)}}$	$\frac{1}{n^{2(\alpha-1)}}$
$S(P)=\#\{p_i\neq 0\}, p_i\in\{0\}\cup[\frac{1}{S},1]$	$e^{-\Theta(\sqrt{\frac{n\ln n}{S}}\vee\frac{n}{S})}$	$e^{-\Theta(\sqrt{\frac{n}{S\ln S}}\vee\frac{n}{S})}$

Recent Breakthroughs

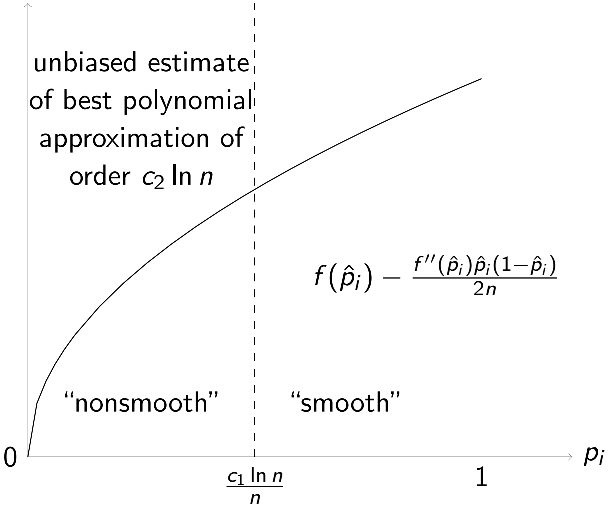
(Jiao–Venkat–H.–Weissman'14, Wu–Yang'14, Jiao–H.–Weissman'15, Wu–Yang'15)

	Minimax L_2 rate	L_2 rate of MLE
$H(P) = \sum_{i=1}^{S} -p_i \ln p_i$	$\frac{S^2}{(n \ln n)^2} + \frac{\ln^2 S}{n}$	$\frac{S^2}{n^2} + \frac{\ln^2 S}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 0 < \alpha < 1/2$	$\frac{S^2}{(n \ln n)^{2\alpha}}$	$\frac{S^2}{n^{2\alpha}}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 1/2 \le \alpha < 1$	$\frac{S^2}{(n \ln n)^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$	$\frac{S^2}{n^{2\alpha}} + \frac{S^{2-2\alpha}}{n}$
$F_\alpha(P) = \sum_{i=1}^{S} p_i^\alpha, 1 < \alpha < 3/2$	$\frac{1}{(n \ln n)^{2(\alpha-1)}}$	$\frac{1}{n^{2(\alpha-1)}}$
$S(P) = \#\{p_i \neq 0\}, p_i \in \{0\} \cup [\frac{1}{S}, 1]$	$e^{-\Theta(\sqrt{\frac{n \ln n}{S}} \vee \frac{n}{S})}$	$e^{-\Theta(\sqrt{\frac{n}{S \ln S}} \vee \frac{n}{S})}$

Similar results also hold for Rényi entropy estimation (Acharya–Orlitsky–Suresh–Tyagi'14), KL, Hellinger and χ^2-divergence estimation (H.–Jiao–Weissman'16), L_r norm estimation under Gaussian white noise model (H.–Jiao–Mukherjee–Weissman'16), L_1 distance estimation (Jiao–H.–Weissman'17)

Optimal estimator for $\sum_{i=1}^{S} f(p_i)$

$f(p_i) = -p_i \ln p_i$ or p_i^α



Past Insights

- Bias dominates in functional estimation
- Bias corresponds to polynomial approximation error
- Need to use the best polynomial approximation where the functional is non-smooth
- Plug-in approach is strictly sub-optimal

Main Results

Let $\hat{P}^* = (\hat{p}_1^*, \cdots, \hat{p}_S^*)$ be our optimal estimator for unlabeled distribution learning.

Main Results

Let $\hat{P}^* = (\hat{p}_1^*, \cdots, \hat{p}_S^*)$ be our optimal estimator for unlabeled distribution learning.

Theorem (H.–Jiao–Weissman'17)

For the Shannon entropy $H(P)$, the power sum function $F_\alpha(P)$ with $\alpha \in (0,1)$, and the support size function $S(P)$, the plug-in approach $F(\hat{P}^*)$ attains the optimal sample complexity (with $F = H, F_\alpha, S$ respectively).
Note that for the support size function $S(P)$, when forming $\hat{P}^*$ we should require that $\mu_Q((0, \frac{1}{S})) = 0$ in our LP.

Main Results

Let $\hat{P}^* = (\hat{p}_1^*, \cdots, \hat{p}_S^*)$ be our optimal estimator for unlabeled distribution learning.

Theorem (H.–Jiao–Weissman'17)

For the Shannon entropy $H(P)$, the power sum function $F_\alpha(P)$ with $\alpha \in (0,1)$, and the support size function $S(P)$, the plug-in approach $F(\hat{P}^*)$ attains the optimal sample complexity (with $F = H, F_\alpha, S$ respectively).
Note that for the support size function $S(P)$, when forming $\hat{P}^*$ we should require that $\mu_Q((0, \frac{1}{S})) = 0$ in our LP.

Plug-in becomes optimal!

Implicit Polynomial Approximation

Why New Plug-in Estimator Works

Suppose all $p_i \in [0, \frac{\ln n}{n}]$. We have $\mathbb{E}\sum_{i=1}^S (\hat{p}_i^*)^k \approx \sum_{i=1}^S p_i^k$ for $k = 0, 1, \cdots, K$ by construction, and thus

$$\mathbb{E}\sum_{i=1}^S f(\hat{p}_i^*) - f(p_i) \approx \inf_{\deg P \le K} \mathbb{E}\sum_{i=1}^S (f(\hat{p}_i^*) - P(\hat{p}_i^*)) - (f(p_i) - P(p_i))$$

yields to polynomial approximation

Implicit Polynomial Approximation

Why New Plug-in Estimator Works

Suppose all $p_i \in [0, \frac{\ln n}{n}]$. We have $\mathbb{E}\sum_{i=1}^S (\hat{p}_i^*)^k \approx \sum_{i=1}^S p_i^k$ for $k = 0, 1, \cdots, K$ by construction, and thus

$$\mathbb{E}\sum_{i=1}^S f(\hat{p}_i^*) - f(p_i) \approx \inf_{\deg P \le K} \mathbb{E}\sum_{i=1}^S (f(\hat{p}_i^*) - P(\hat{p}_i^*)) - (f(p_i) - P(p_i))$$

yields to polynomial approximation

Properties

- Implicit polynomial approximation: we did not construct any explicit polynomial in our estimator

Implicit Polynomial Approximation

Why New Plug-in Estimator Works

Suppose all $p_i \in [0, \frac{\ln n}{n}]$. We have $\mathbb{E}\sum_{i=1}^S (\hat{p}_i^*)^k \approx \sum_{i=1}^S p_i^k$ for $k = 0, 1, \cdots, K$ by construction, and thus

$$\mathbb{E}\sum_{i=1}^S f(\hat{p}_i^*) - f(p_i) \approx \inf_{\deg P \le K} \mathbb{E}\sum_{i=1}^S (f(\hat{p}_i^*) - P(\hat{p}_i^*)) - (f(p_i) - P(p_i))$$

yields to polynomial approximation

Properties

- Implicit polynomial approximation: we did not construct any explicit polynomial in our estimator
- Universal estimation: a single estimator works for multiple functionals

Implicit Polynomial Approximation

Why New Plug-in Estimator Works

Suppose all $p_i \in [0, \frac{\ln n}{n}]$. We have $\mathbb{E}\sum_{i=1}^{S}(\hat{p}_i^*)^k \approx \sum_{i=1}^{S} p_i^k$ for $k = 0, 1, \cdots, K$ by construction, and thus

$$\mathbb{E}\sum_{i=1}^{S} f(\hat{p}_i^*) - f(p_i) \approx \inf_{\deg P \le K} \mathbb{E}\sum_{i=1}^{S}(f(\hat{p}_i^*) - P(\hat{p}_i^*)) - (f(p_i) - P(p_i))$$

yields to polynomial approximation

Properties

- Implicit polynomial approximation: we did not construct any explicit polynomial in our estimator
- Universal estimation: a single estimator works for multiple functionals
- Too good to be true!?

Open Questions

- How "universal" is our estimator for general functionals?
- Can our estimator be applied to 2D functionals, e.g., the ℓ_1 distance $\|P - Q\|_1$ and the KL divergence $D(P\|Q)$?
- Why are polynomials so special in the unlabeled distribution learning problem? Can we match other symmetric functionals instead of moments?

Concluding Remarks

- It requires $n \gg \frac{S}{\ln S}$ samples for unlabeled distribution learning, while $n \gg S$ samples are required for the labeled one
- The natural estimator (MLE) is strictly suboptimal
- Beautiful duality
  - moment matching in both upper and lower bounds
  - Wasserstein distance argument in both upper and lower bounds
- The plug-in approach of the previous estimator is universal for functional estimation