STOC15

Learning Arbitrary Statistical Mixtures of Discrete Distributions

Jian Li (Tsinghua), Yuval Rabani(HUJI),
Leonard J. Schulman(Caltech), Chaitanya Swamy (Waterloo)

lijian83@mail.tsinghua.edu.cn

- **Problem Definition**
- Related Work
- Our Results
- The Coin Problem
- Higher Dimension
- Conclusion

Problem Definition

- $\Delta_n = \{x \in R_+^n \mid ||x||_1 = 1\}$
- So each point in Δ_n is a prob. distr. over [n]
- ϑ is a prob. distr. over Δ_n (unknown to us)

Mixture of discrete distributions

- Goal: learn ϑ (i.e., transportation distance in L_1 at most ϵ. $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \le \epsilon$)

Problem Definition

- $\Delta_n = \{x \in R_+^n |\ ||x||_1 = 1\}$
- So each point in Δ_n is a prob. distr. over [n]
- ϑ is a prob. distr. over Δ_n (unknown to us)
- Goal: learn ϑ (i.e., transportation distance in L_1 at most ϵ. $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \le \epsilon$)

- **A k-snapshot sample: (k: snapshot#)**
  - Take a sample point $x \sim \vartheta$ $(x \in \Delta_n)$ (we don't get to observe x directly)
  - Take k i.i.d. samples $s_1 s_2 \ldots s_k$ from x (we observe $s_1 s_2 \ldots s_k$, called a ***k*-snapshot sample**)
- **Question:**

How large the snapshot# k needs to be in order to learn ϑ??

How many k-snapshot samples do we need to learn ϑ ??

- Problem Definition
- **Related Work**
- Our Results
- The Coin Problem
- Higher Dimension
- Conclusion

Related Work

- Previous work
  - Mixture of Gaussians: a large body of work
    - Only need 1-snapshot samples
    - k-snapshot (k>1) is necessary for mixtures of discrete distributions
    - Learn the parameters
  - Topic Models
    - ϑ is a mixture of topics (each topic is a distribution of words)

    How a document is generated:

    - Sample a topic from $x \sim \vartheta \quad (x \in \Delta_n)$
    - Use x to generate a document of size k (a document is a k-snapshot sample)

Related Work

- Previous work
  - Mixture of Gaussians: a large body of work
  - Only need 1-snapshot samples
    - k-snapshot (k>1) is necessary for mixtures of discrete distribution

  - Topic Models
    - Various assumptions:
      - LSI, Separability [Papadimitriou,Raghavan,Tamaki,Vempala'00]
      - LDA [Blei, Ng, Jordan'03]
      - Anchor words [Arora,Ge,Moitra'12] (snapshot#=2)
      - Topic linear independent [Anandkumar, Foster, Hsu, Kakade, Liu'12] (snapshot#=O(1))
      - Several others
  - Collaborative Filtering
    - L1 condition number [Kleinberg, Sandler '08]

Transportation Distance

- Also known as earth mover distance, Rubinstein distance, Wasserstein distance
- $\text{Tran}(P, Q)$: Distance between two probability distributions P, Q

If we want to turn P to Q, the metric is the cost of the optimal transportation T (i.e., $\int ||x - T(x)||dP$)

E.g., in discrete case, it is the solution of the following LP:

$$\begin{aligned} \text{minimize } \textstyle\sum_{i,j} d(v_i, v_j) x_{ij} \quad \text{subject to } & \textstyle\sum_j x_{ij} = P(\{v_i\}), \ \forall i \in [n], \\ & \textstyle\sum_i x_{ij} = Q(\{v_j\}), \ \forall i \in [n], \\ & x_{ij} \in [0, 1] \ \forall i \in [n], j \in [n]. \end{aligned}$$

Transportation Distance

- Also known as earth mover distance, Rubinstein distance, Wasserstein distance
- $\text{Tran}_1(P, Q)$: Distance between two probability distributions P, Q

If we want to turn P to Q, the metric is the cost of the optimal transportation T (i.e., $\int ||x - T(x)||_1 \, dP$)

E.g., in discrete case, it is the solution of the following LP:

$$\text{minimize } \sum_{i,j} d(v_i, v_j) x_{ij} \quad \text{subject to } \sum_j x_{ij} = P(\{v_i\}), \ \forall i \in [n],$$
$$\sum_i x_{ij} = Q(\{v_j\}), \ \forall i \in [n],$$
$$x_{ij} \in [0, 1] \ \forall i \in [n], j \in [n].$$

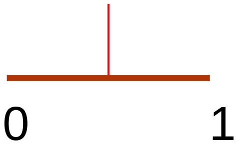
Our Results

- The Coin problem: 1-dimension
  - A mixture ϑ defined over [0,1]
  - If mixture ϑ is a k-spike distribution (k different coins)
    - Require k-snapshot **(k>1)** samples

- (H 0,T 1) w.p. 0.5 (H 1,T 0) w.p. 0.5

- (H 0.1, T 0.9) w.p. 0.5 (H 0.9, T 0.1) w.p. 0.5

- ……
- (H 0.5, T 0.5) w.p. 1



Our Results

The Coin problem: 1-dimension

- A mixture ϑ defined over [0,1]
- If mixture ϑ is a k-spike distribution, a lower bound is known
  - Require k-snapshot (k>1) samples
  - Lower bound : To guarantee $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \leq O(1/k)$ [Rabani,Schulman,Swamy'14]

    (1) (2k-1)-snapshot is necessary

    (2) We need $\exp(\Omega(k))$ (2k-1)-snapshot samples

Our Result:

- A nearly matching upper bound:

  $(k/\epsilon)^{O(k)} \log 1/\delta$ (2k-1)-snapshot samples suffice (w.p. $1-\delta$)

Our Results

The Coin problem: 1-dimension

- A mixture ϑ over [0,1]
- ϑ is arbitrary (may even be continuous)
  - Lower bound [Rabani,Schulman,Swamy'14]: Still applies. (rewrite a bit)
    - We can use K-snapshot samples.
    - We need $\exp(\Omega(K))$ K-snapshot samples to make $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \leq O(1/K)$
- Our Result
  - A nearly matching upper bound
    - Using $\exp(O(K))$ K-snapshot samples, we can recover ϑ s.t. $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \leq O(1/K)$

A tight tradeoff between K and transportation distance

Our Results

Higher Dimension

- A mixture ϑ over Δ_n
- Assumption: ϑ is a k-spike distribution (think k very small, k<<n)

Our result:

- Using poly(n) 1- and 2-snapshot samples and $(k/\epsilon)^{O(k^2)}$ (2k-1)-snapshot samples, we can obtain a mixture $\hat{\vartheta}$ s.t. $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \leq \epsilon$

L1 distance. Harder than L2

Our Results

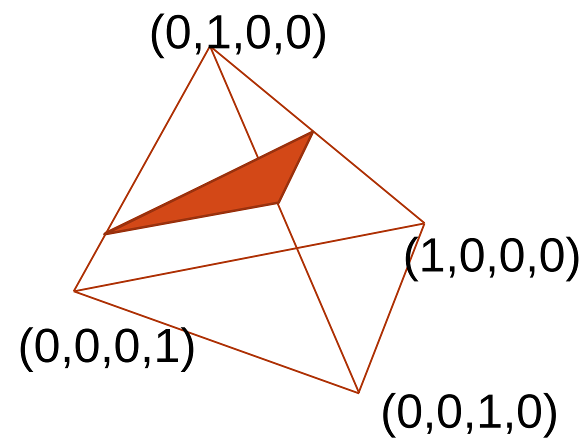
- Higher Dimension
- A mixture ϑ over Δ_n
- Assumption: ϑ is a k-spike distribution (think k very small, k<<n)

- Why L1 distance?
  - $P, Q \in \Delta^n \quad d_{TV}(P,Q) = ||P-Q||_1$
  - E.g., $\left(\frac{1}{n}, \dots, \frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right)$ and $\left(0, \dots, 0, \frac{2}{n}, \dots, \frac{2}{n}\right)$ are two very different distributions. But their L2 distance is small ($1/\sqrt{n}$)

Our Results

- Higher Dimension
- A mixture ϑ over Δ_n
- Assumption: ϑ is an **arbitrary** distribution

supported on a k-dim slice of Δ_n
(again think k<<n)



A 2-dim slice in Simplex Δ_4

Our result:

- Using poly(n) 1- and 2-snapshot samples, and $(k/\epsilon)^{O(k)}$ K-snapshot samples ($K = \text{poly}(k,\epsilon)$), we can obtain a mixture $\hat{\vartheta}$ s.t. $\text{Tran}_1(\vartheta, \hat{\vartheta}) \le \epsilon$

- Problem Definition
- Related Work
- Our Results
- **The Coin Problem**
- Higher Dimension
- Conclusion

The Coin Problem

- A (even continuous) mixture ϑ of coins
- Consider a K-snapshot sample

$$\Pr[\text{exactly } i \text{ heads}] = \int \binom{K}{i} x^i (1-x)^{K-i} \mathrm{d}\vartheta = \int B_{i,K}(x) \mathrm{d}\vartheta$$

Bernstein Polynomial

$$\mathsf{fq}(\vartheta) = \{\Pr[\text{exactly } 0 \text{ heads}], \Pr[\text{exactly } 1 \text{ heads}], \ldots, \Pr[\text{exactly } K \text{ heads}]\}$$

Using $\kappa^{-2}\log(K/\delta)$ samples, we can obtain $\|\widetilde{\mathsf{fq}} - \mathsf{fq}(\vartheta)\| \leq \kappa$

The Coin Problem

- A simple but useful lemma:

For any two distributions P and Q on $[0,1]$,

$$\mathrm{Tran}(P,Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda).$$

C and λ satisfy the following statement:
For any $f \in 1\text{-}\mathsf{Lip}[0,1]$,

$$f = \sum_i c_i B_{i,K} \pm O(\lambda) \quad \text{where } c_0, \ldots, c_K \in [-C, C]$$

Pf based on the Dual formulation (Kantorovich&Rubinstein)

$$\mathrm{Tran}(P,Q) = \sup \left\{ \left| \int f \mathrm{d}(P-Q) \right| : f \in 1\text{-}\mathsf{Lip} \right\}.$$

$|f(x) - f(y)| \le ||x - y||$

The Coin Problem

$$\mathrm{Tran}(P, Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda).$$

- If we want to make $\mathrm{Tran}(P, Q) \le \epsilon$

  need
  $$\begin{cases} \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 \le O(\epsilon / C) \\ \lambda = O(\epsilon) \end{cases}$$

Require $\mathrm{poly}(C/\epsilon)$ samples

The Coin Problem

$$\mathrm{Tran}(P,Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda).$$

- If we want to make $\mathrm{Tran}(P,Q) \le \epsilon$

  need
  $$\begin{cases} \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 \le O(\epsilon/C) \\ \lambda = O(\epsilon) \end{cases}$$

Require $\mathrm{poly}(C/\epsilon)$ samples

What C and λ can we achieve??

$f \in 1\text{-Lip}[0,1] \quad f = \sum c_i B_{i,K} \pm O(\lambda) \quad$ where $c_0, \ldots, c_K \in [-C, C]$

WELL KNOWN in approximation theory (e.g., Rivlin03):

$$\left\| f - \sum_{i=0}^{K} f(i/K) B_{i,K}(x) \right\|_\infty \le O(1/\sqrt{K})$$

Bernstein polynomial approximation

So, with $\mathrm{poly}(K)$ K-snapshot samples, $\mathrm{Tran} = O(1/\sqrt{K})$

The Coin Problem

$$\mathrm{Tran}(P,Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda).$$

- If we want to make $\mathrm{Tran}(P,Q) \le \epsilon$

  need $\begin{cases} \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 \le O(\epsilon/C) \\ \lambda = O(\epsilon) \end{cases}$

Require $\mathrm{poly}(C/\epsilon)$ samples

What C and λ can we achieve??

$f \in$ 1-Lip$[0,1]$ $\quad f = \sum c_i B_{i,K} \pm O(\lambda)$ $\quad$ where $c_0, \ldots, c_K \in [-C, C]$

Jackson's theorem:

$$f(x) = \sum_{i=0}^{K} t_i T_i(x) \pm O(1/K) \qquad |t_i| \le \mathrm{poly}(K)$$

Chebyshev polynomials

By a change of basis $\{B_{i,K}\} \to \{T_i\}$

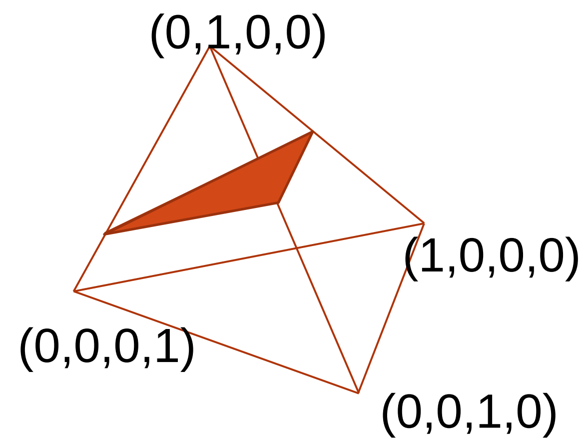
with $\exp(K)$ K-snapshot samples, Tran $= O(1/K)$

- Problem Definition
- Related Work
- Our Results
- The Coin Problem
- **Higher Dimension**
- Conclusion

High Dimensional Case

- A mixture ϑ over Δ_n
- ϑ is a k-spike distribution over

a k-dim slice A of Δ_n ($k<<n$)



A 2-dim slice in Simplex Δ_4

Outline:

- Step 1: Reduce the learning problem from n-dim to k-dim
  (we don't want the snapshot# depends on n)
- Step 2: Learn the projected mixture in the k-dim subspace
  (require $\text{Tran}_2 \leq \epsilon$, snapshot# depends only on k, ϵ)
- Step 3: Project back to Δ_n

High Dimensional Case

Step 1: From n-dim to k-dim

- Existing approach: apply SVD/PCA/Eigen decomposition to the 2-moment matrix, then take the subspace spanned by the first few eigenvectors
- Does NOT work!

High Dimensional Case

Step 1: From *n*-dim to *k*-dim

- Existing approach: apply SVD/PCA/Eigen decomposition to the 2-moment matrix, then take the subspace spanned by the first few eigenvectors
- Does NOT work!

Reason: we want $\mathrm{Tran}_1(\vartheta, \hat{\vartheta}) \leq \epsilon$ (L1 metric)

- L1 is not rotationally invariant. So it may happen (in the subspace) that

  $\|a-b\|_1 = O(\sqrt{n})\|a-b\|_2$ in some directions

  but $\|a-b\|_1 = O(1)\|a-b\|_2$ in some other directions

Implication: in the reduced k-dim learning problem, we have to be very accurate in some directions (only by making snapshot# depend on n)

High Dimensional Case

- Step 1: From n-dim to k-dim

- What we do:

  Find a k'-dim ($k' < k$) subspace B where the L1-ball is **almost spherical**, and the supporting slice A is <u>close</u> to B

  in L1 metric

High Dimensional Case

Step 1: From *n*-dim to *k*-dim

(sketch)

1. Put ϑ in an isotropic position: $r_i = \int x_i \mathrm{d}\vartheta \in [1/2n, 2/n]$
   (by deleting and splitting letters)
2. Compute the John Ellipsoid for a polytope $\mathcal{P} = \mathcal{H} \cap \mathrm{Span}(A)$
   take the first few (normalized) principle axes, where
   $\mathcal{H} = [-C/n, C/n]^n$

High Dimensional Case

Step 2: Learn the projected mixture in the *k*-dim subspace (sketch)

(1) project to a net of 1-dim directions

(2) Learn the 1-d projections

(3) Assemble the 1-d projections using LP

Similar to a Geometric Tomography question.
Analysis uses Fourier decomposition and a
multidimension version of Jackson theorem

- Problem Definition
- Related Work
- Our Results
- The Coin Problem
- Higher Dimension
- **Conclusion**

Conclusion

- Algorithms for learning mixtures of discrete distributions
- No assumption (on independence, conditional number etc.). Worst case analysis
- Tradeoff: Snapshot#, Tran, #samples
- Transportation distance

Thanks

lijian83@mail.tsinghua.edu.cn

More on Transportation Distance

- Def: $\mathrm{Tran}(P,Q) = \inf_{T} \|x - T(x)\| \mathrm{d}x$

  where T is a transportation from P to Q
- The Dual formulation (Kantorovich&Rubinstein)

$$\mathrm{Tran}(P,Q) = \sup \left\{ \left| \int f \mathrm{d}(P-Q) \right| : f \in \text{1-Lip} \right\}.$$

$|f(x) - f(y)| \leq ||x - y||$

More on Transportation Distance

- Def: $\mathrm{Tran}(P,Q) = \inf_{T} \|x - T(x)\| \mathrm{d}x$

  where T is a transportation from P to Q

- The Dual formulation (Kantorovich&Rubinstein)

$$\mathrm{Tran}(P,Q) = \sup \left\{ \left| \int f \mathrm{d}(P-Q) \right| : f \in \text{1-Lip} \right\}.$$

$|f(x) - f(y)| \le ||x - y||$

If P, Q are finite supported discrete distributions, the above is simply the LP-duality

Primal: minimize $\sum_{i,j} d(v_i, v_j) x_{ij}$ subject to $\sum_j x_{ij} = P(\{v_i\}),\ \forall i \in [n],$

$\sum_i x_{ij} = Q(\{v_j\}),\ \forall i \in [n],$

$x_{ij} \in [0,1]\ \forall i \in [n], j \in [n].$

Dual: maximize $\sum_i f_i(P(\{v_i\}) - Q(\{v_i\})),$ subject to $f_i - f_j \le d(v_i, v_j)\ \forall i \in [n], j \in [n].$

The Coin Problem

- A simple but useful lemma:

For any two distributions P and Q on $[0,1]$,

$$\mathrm{Tran}(P,Q) \le C \cdot \| \mathsf{fq}(P) - \mathsf{fq}(Q) \|_1 + O(\lambda).$$

C and λ satisfy the following statement:
For any $f \in 1\text{-}\mathsf{Lip}[0,1]$,

$$f = \sum_i c_i B_{i,K} \pm O(\lambda) \qquad \text{where } c_0, \ldots, c_K \in [-C, C]$$

Pf sketch:

$$\begin{aligned}
\left| \int f \mathrm{d}(P-Q) \right| &= \left| \sum_{i=0}^{K} c_i \int B_{i,K} \, \mathrm{d}(P-Q) \right| + O(\lambda) \\
&= \left| \sum_{i=0}^{K} c_i (\mathsf{fq}_i(P) - \mathsf{fq}_i(Q)) \right| + O(\lambda) \\
&\le C \cdot \|\mathsf{fq}(P) - \mathsf{fq}(Q)\|_1 + O(\lambda).
\end{aligned}$$

This holds for any 1-Lip function f.
So the lemma follows from the dual formulation

High Dimensional Case

Step 1: From *n*-dim to *k*-dim

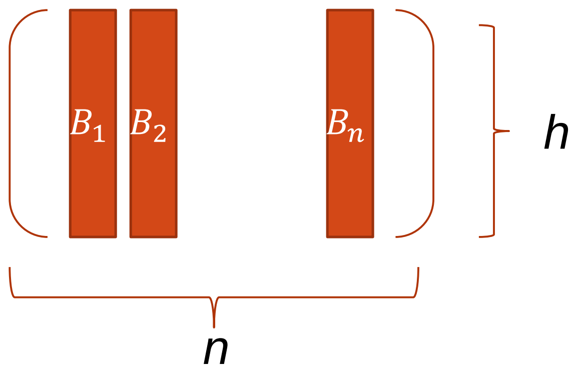
1. Put ϑ in an isotropic position: $r_i = \int x_i \mathrm{d}\vartheta \in [1/2n, 2/n]$
   (by deleting and splitting letters)
2. Consider $\mathcal{H} = [-C/n, C/n]^n$ and the polytope $\mathcal{P} = \mathcal{H} \cap \mathrm{Span}(A)$
   (C only depends on k and ϵ)
3. Compute the John Ellipsoid $\mathcal{E} \subseteq \mathcal{P} \subseteq \sqrt{k}\mathcal{E}$ with axes $\{e_1, \ldots, e_k\}$
4. Take the first few (normalized) principle axes

$$B = \left\{ b_i = \frac{e_i}{\|e_i\|_2} : \|e_i\|_2 \geq \frac{\epsilon}{\sqrt{n}} \right\}$$

High Dimensional Case

Step 2: Learn the projected mixture in the k-dim subspace

$B=$



For a K-snapshot sample $\mathbf{s} = \{s_1, \ldots, s_K\}, s_i \in [n]$,

let $u(\mathbf{s}) = \sum_{k=1..K} B_{s_k}$

Suppose we take N samples $\mathbf{s_1}, \ldots, \mathbf{s_N}$

The learnt project measure is the empirical measure

$$\frac{1}{N}\sum_{i=1}^{N} \delta(B^T u(\mathbf{s}_i))$$

Delta func