Learning Arbitrary Statistical Mixtures of Discrete Distributions

Yuval Rabani - The Hebrew University of Jerusalem

joint works with J. Li, L.J. Schulman, C. Swamy

motivation: topic models

[Hofmann 1999, Papadimitriou et al. 2000]

Regard documents as "bags of words"

to generate a d-word document:

- draw d iid samples from a distribution p

k topic distributions $p_1, p_2, \ldots, p_k$

Pure documents: choose $p = p_i$ w/prob. w_i

mixed topic models

Each document is a mixture of topics.

to generate a d-word document:

- draw d iid samples from a distribution p

k topic distributions $p_1, p_2, \ldots, p_k$

probability measure θ on $\mathrm{conv}(p_1,\ldots,p_k)$

Choose $p \in \mathrm{conv}(p_1,\ldots,p_k)$ according to θ

<u>example</u>: latent Dirichlet allocation [Blei-Ng-Jordan 2003]

motivation: collaborative filtering

[Hofmann-Puzicha 1999, Kleinberg-Sandler 2004]

Purchase history of customers:

Customer has distribution p on purchases.

Purchases are drawn iid from p

p is chosen according to a probability measure θ on $\mathrm{conv}(p_1,\ldots,p_k)$

motivation: summary

<u>Data mining</u>: simple model for

- document features (LSI)
- customer taste (collaborative filtering)
- hyperlinks, citations (Kleinberg's HITS)
- observational studies (clinical, wildlife, ...)

<u>Properties</u>:

- a large number of possible features
- each specimen exhibits a few features
- population behaves "nicely"

learning the mixture model

known: dictionary $\{1,2,\ldots,n\}$

input: m samples of d-tuples from $\{1,2,\ldots,n\}$

How is a sample generated? -

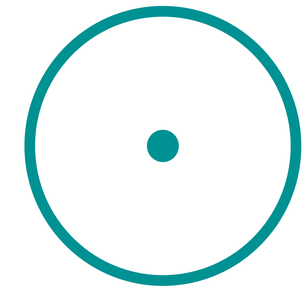
- pick p from θ (hidden from the observer)
- draw d items iid from p

goal: learn the model - θ

failure probability: a small constant δ

learning mixtures of Gaussians

k Gaussians in R^n



- Dasgupta (1999) $O(n^{1/2})$ sep.

  Dasgupta-Schulman (2000) $O(n^{1/4})$ sep.
- Arora-Kannan (2001) ellipsoidal
- Vempala-Wang (2002) $O(k^{1/4})$ sep.

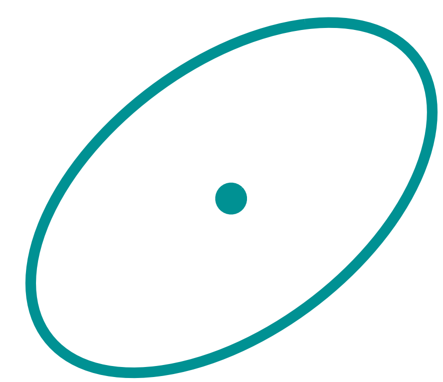
  Kannan-Salmasian-Vempala (2005)

  Achlioptas-McSherry (2005)

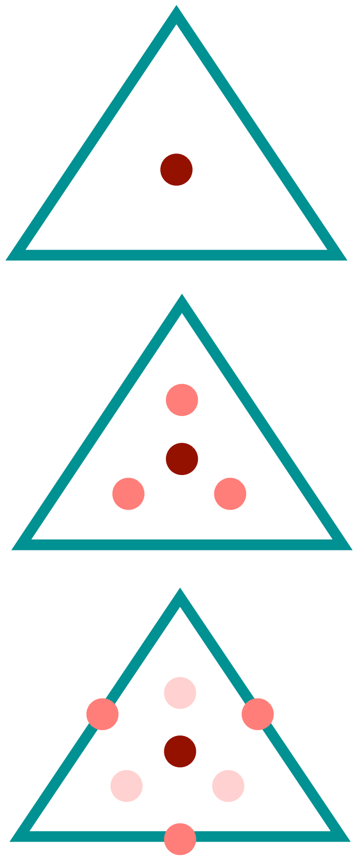
  Brubaker-Vempala (2008)
- Feldman et al. (2006) axis aligned

  Kalai-Moitra-Valiant (2010) k=2 general

  $\left.\begin{array}{l}\text{Moitra-Valiant (2010)}\\ \text{Belkin-Sinha (2010)}\end{array}\right\}$ general



Gaussians vs. pure topic models



- single view vs. multi-view samples:
  Gaussians - learnable using single view
  topic models - require multi-view:

- sample info. vs. model size:
  Gaussians - n vs. $k \cdot n^2$
  topic models - d vs. $k \cdot n$

- multi-view topic models =
  power distributions on $\{1,2,\dots,n\}^d$
  n is large, d is small.

common techniques

- spectral decomposition
- random projection
- method of moments

back to topic models

m, n are large
d, k are small

minimize:
\# samples m
aperture (# views) d
running time (in terms of m, d, n, k)

trivial information theoretic bounds:

- if $m \cdot d = o(n)$ then the error could be ≫
  (we don't see most of the dictionary)
- if $d = \Omega(n \log n)$ then most samples give an accurate estimate of their p.

some notation

constituents matrix: $P = (p_1, p_2, \ldots, p_k)$

mean: $\mu = \int p \, d\theta$ $(w_1 p_1 + w_2 p_2 + \cdots + w_k p_k)$

pairwise distrib.: $M = \int p p^\dagger \, d\theta$ $(w_1 p_1 p_1^\dagger + w_2 p_2 p_2^\dagger + \cdots + w_k p_k p_k^\dagger)$

variance: $V = M - \mu\mu^\dagger$

---

i^{th} largest (left) singular value: $\sigma_i(A)$

i^{th} largest (real) eigenvalue: $\lambda_i(A)$

condition number: $\kappa(A) = \sigma_1(A) \,/\, \sigma_{rank(A)}(A)$

min. variation distance $\zeta_1 = \sqrt{n} \cdot \min\{\|p_i - p_j\|_2 : i \neq j\}$

min. non-zero eigenvalue $\zeta_2 = \sqrt{n \cdot \lambda_{rank(V)}(V)}$

spreading parameter: $\zeta = \max\{\zeta_1, \zeta_2\}$

pure mixtures

Anandkumar-Hsu-Kakade (2012)

assumption: P is full-rank ($\mathrm{rank}(P) = k$)

aperture: $d = 3$

$\max_j \|p_j - \acute{p}_j\|_2$

guarantee: w.h.p. L_2 error $\varepsilon \cdot \max_i \|p_i\|_2$

sample size:

alg. A: $m = k^c / (\sigma_k(P))^8 (\lambda_k(M))^4 \varepsilon^2$

alg. B: $m = k^c n (\kappa(P))^8 / (\zeta_1)^2 (\lambda_k(M))^2 \varepsilon^2$

R.-Schulman-Swamy (2014)

tight

<u>no assumptions</u>

$\max_j \|p_j - \acute{p}_j\|_1$

aperture: $d = 2k-1$

guarantee: w.h.p. L_1 error ε (weights, too)

<u>sample size:</u>

$m = k^c\, n \log^c n / \varepsilon^6 + \exp(k^2 \log(k/\zeta\varepsilon))$

(1st term uses $d \leq 2$)

exp(k) needed for d=O(k)

comparison with [AHK12]:

to get L_1 error ε they (might) need (for constant ζ):

alg. A: $m = k^c\, n^8 / \varepsilon^2$

alg. B: $m = k^c\, n^3 / \varepsilon^2$

Li-R.-Schulman-Swamy (2015)

no assumptions

aperture: $d = 2k-1$

guarantee: w.h.p. L_1 error ε

sample size:

$m = k^4 n^3 \log n / \varepsilon^6 + \exp(k^2 \log(k/\varepsilon))$

(1st term uses $d \le 2$)

comparison with previous results:

the sample size does not depend on ζ

mixed mixtures

Arora-Ge-Moitra (2012)

assumption: $p_1, p_2, \ldots, p_k$ are ρ-separable
(each p_i has an item w/prob. $\geq \rho$ that has 0 probability in the other p_j-s)

guarantee: w.h.p. L_∞ error ε (L_1 error $\varepsilon \cdot n$)
sometimes (e.g., LDA) also θ recovered

sample size: $m = k^c \log n / \varepsilon^2 \rho^6 d$

technique: nonnegative matrix factorization [Arora-Ge-Kannan-Moitra 2012]

Anandkumar-Foster-Hsu-Kakade-Liu (2012)

assumption: $\mathrm{rank}(P) = k$ and θ is Dirichlet

aperture: $d = 3$

guarantee: w.h.p. L_2 error ε

sample size: $m = k^c / (\sigma_k(P))^6 \varepsilon^2$

Li-R.-Schulman-Swamy (2015)

$d > 1/\varepsilon$ needed

no assumptions

aperture: $d \gg k^{11} / \varepsilon^{10}$

guarantee: w.h.p. L_1-transportation cost ε

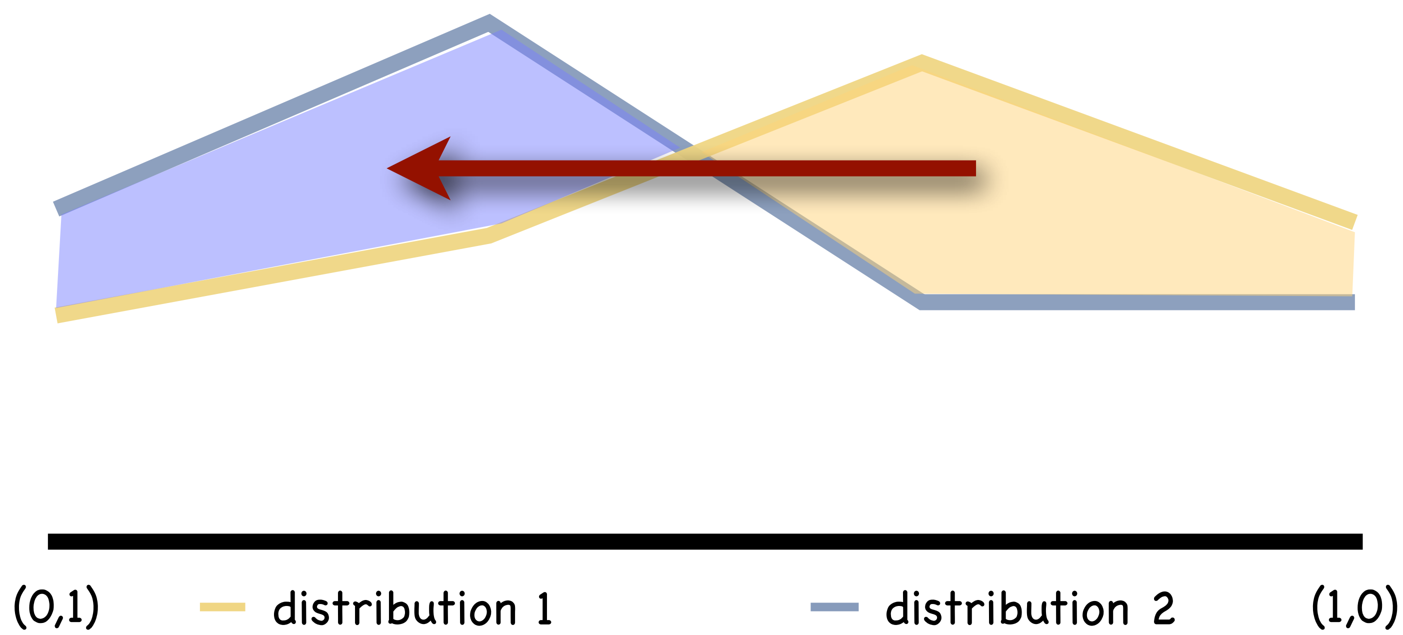
sample size:

$m = k^4 n^3 \log n / \varepsilon^6 + \exp(k \log(k/\varepsilon))$

(1st term uses $d \leq 2$)

about the proofs

transportation distance



in general:

$$\mathrm{Tran}(\eta,\theta) = \inf\{\textstyle\int \|p-q\|_1 \, d\varphi : \varphi \text{ has marginals } \eta,\theta\}$$

$$= \sup\{|\textstyle\int f \, d(\eta-\theta)| : f \text{ is 1-Lipschitz}\}$$

Kantorovich-Rubinstein duality

one-dimensional problem

goal: learn a prob. distribution θ on $[0,1]$

sample: pick $p \in [0,1]$ by θ (p hidden)
toss a p-biased coin d times
the alg. sees the d-tuple in $\{0,1\}^d$

repeated sampling gives ≈ the mean
$F_i = F_i(\theta)$ of $B_{i,d}(p) = \binom{d}{i} p^i (1-p)^{d-i}$
for all $i = 0, 1, \ldots, d$

distributions with similar first d moments

<u>Lemma</u>: if $\forall$ 1-Lip. function on $[0,1]$ is $\pm\gamma$ a linear combination of $B_{i,d}$-s with coefficients $\in [-C,+C]$, then

$$\mathrm{Tran}(\eta,\theta) \leq C \cdot \|F(\eta)-F(\theta)\|_1 + \gamma$$

<u>proof</u>: Kantorovich-Rubinstein duality + triangle inequality.

A bound on the error

thm: for $C = O(1)$ we can get $\gamma = O(1/\sqrt{d})$

$\Rightarrow$ poly(d) sample, $O(1/\sqrt{d})$ error.

Jackson's thm: if f is 1-Lip. on $[-1,+1]$ then

$\exists$ degree-d polynomial q such that

$\|f-q\|_\infty = O(1/d)$.

uses Chebyshev polynomials $\Rightarrow$ $C = d^c \cdot 2^d$

$\Rightarrow$ exp(d) sample, $O(1/d)$ error.

the algorithm

- get a good estimate F' of the frequency moments F (we want $||F'-F||_\infty < 1/d^c 2^d$)
- partition $[0,1]$ into $d^c 2^d$ segments; put $b_{i,j} = E[B_{i,d}]$ in segment j.
- solve a linear system to get a piecewise constant probability measure η with $\Sigma_j\, b_{i,j}\, \eta_j = F_i' \pm 1/d^c 2^d$, $\forall i$
- (notice that $F_i(\eta) \approx \Sigma_j\, b_{i,j}\, \eta_j \approx F_i(\theta) \pm 1/d^c 2^d$)

k spikes

θ has finite support of size k

$d = 2k-1$

$F(\theta)$ = vector of the first d moments of θ

Lemma: $\forall$two k-spike distributions η,θ,

$\|F(\eta) - F(\theta)\|_2 \geq (\mathrm{Tran}(\eta,\theta) / k)^{O(k)}$

(in general, $|F_i(\eta) - F_i(\theta)| \leq i \cdot \mathrm{Tran}(\eta,\theta)$)

higher dimensions

W.l.o.g. the mixture model is <u>isotropic</u>:

$$\forall i \in \{1,2,\ldots,n\},\ 1/2n \leq \mu_i \leq 2/n$$

$\Rightarrow$ L_1 and L_2 norms are $\approx$ isometric and
$\exists$basis $b_1, b_2, \ldots, b_{K'}$ for span(P) with bounded entries

$\Rightarrow$ span(P) learnable
from empirical pairwise distribution, using Vu (2005)

Project samples onto $b_1', b_2', \ldots, b_{K'}'$ or $\approx$span(P).

<u>Notice</u>: $\langle p, b_i \rangle = E[(b_i)_s : s \sim p_j]$

Compute a model that matches $\approx$ the projections.

a multidimensional version of Jackson's thm

thm (Yudin): if $f: B^k(R) \to \mathbb{C}$ is 1-Lip. then $\exists c_z$ $\forall z \in \mathbb{Z}^k \cap B^k(R)$ with $|c_z| \leq \exp(k)$ such that $\forall x \in B^k(R)$, $|f(x) - \sum c_z \cdot e^{i\langle z,x\rangle}| = O(k/R)$

$\Rightarrow$ $\mathrm{Tran}(\eta,\theta)$ is bounded by $\sup_b \mathrm{Tran}(\langle b,\eta\rangle,\langle b,\theta\rangle)$

concluding remarks

- better bounds for mixed documents?
  under what conditions?
- learning from sparse samples?