Polynomial Methods in Learning and Statistics

Ankur Moitra, MIT

July 11, 2013
Outline

- Mixtures of Gaussians
  - Highlights: method of moments and the heat equation
  - based on [Kalai, Moitra, Valiant]
    (see also [Belkin and Sinha])

- Topic Models
  - Highlights: tensor methods and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- Nonnegative Matrix Factorization
  - Highlights: separability and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
Outline

- **Mixtures of Gaussians**
  - Highlights: *method of moments* and the heat equation
  - based on [Kalai, Moitra, Valiant]
    (see also [Belkin and Sinha])

- **Topic Models**
  - Highlights: *tensor methods* and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- **Nonnegative Matrix Factorization**
  - Highlights: *separability* and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
Pearson (1894) and the Naples Crabs

(figure due to Peter Macdonald)
Gaussian Mixture Models

\[ F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x) \]

In particular, with probability \( w_1 \) output a sample from \( F_1 \), otherwise output a sample from \( F_2 \)
Gaussian Mixture Models

\[ F(x) = w_1 F_1(x) + (1 - w_1) F_2(x) \text{, where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x) \]

In particular, with probability \( w_1 \) output a sample from \( F_1 \), otherwise output a sample from \( F_2 \)

five unknowns: \( w_1, \mu_1, \sigma_1, \mu_2, \sigma_2 \)
Gaussian Mixture Models

\[ F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x) \]

In particular, with probability \( w_1 \) output a sample from \( F_1 \), otherwise output a sample from \( F_2 \)

five unknowns: \( w_1, \mu_1, \sigma_1, \mu_2, \sigma_2 \)

Question

*Can we learn these parameters approximately, given enough random samples from \( F \)?*
Gaussian Mixture Models

\[ F(x) = w_1 F_1(x) + (1 - w_1) F_2(x), \text{ where } F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x) \]

In particular, with probability \( w_1 \) output a sample from \( F_1 \), otherwise output a sample from \( F_2 \)

five unknowns: \( w_1, \mu_1, \sigma_1, \mu_2, \sigma_2 \)

Question

*Can we learn these parameters approximately, given enough random samples from \( F \)?*

Pearson invented the **method of moments**, to attack this problem...
Pearson’s Sixth Moment Test

Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]
Pearson’s Sixth Moment Test

Claim

$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$

Let $E_{x \leftarrow F(x)}[x^r] = M_r(\theta)$
Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]

Let \( E_{x \leftarrow F(x)}[x^r] = M_r(\theta) \)
Pearson’s Sixth Moment Test

Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]

Let \( E_{x \leftarrow F(x)}[x^r] = M_r(\theta) \)

- Gather samples \( S \)
Pearson’s Sixth Moment Test

Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]

Let \( E_{x \leftarrow F(x)}[x^r] = M_r(\theta) \)

- Gather samples \( S \)
- Set \( \tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r \) for \( r = 1, 2, \ldots 6 \)
Pearson’s Sixth Moment Test

Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]

Let \[ E_{x \leftarrow F(x)}[x^r] = M_r(\theta) \]

- Gather samples \( S \)
- Set \( \tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r \) for \( r = 1, 2, \ldots 6 \)
- Compute simultaneous roots of \( \{ M_r(\theta) = \tilde{M}_r \}_{r=1,2,\ldots 5} \), select root \( \theta \) that is closest in sixth moment
Pearson’s Sixth Moment Test

Claim

\[ E_{x \leftarrow F(x)}[x^r] \text{ is a polynomial in } \theta = (w_1, \mu_1, \sigma_1, \mu_2, \sigma_2) \]

Let \( E_{x \leftarrow F(x)}[x^r] = M_r(\theta) \)

- Gather samples \( S \)
- Set \( \tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r \) for \( r = 1, 2, \ldots, 6 \)
- Compute simultaneous roots of \( \{ M_r(\theta) = \tilde{M}_r \}_{r=1,2,\ldots,5} \), select root \( \theta \) that is closest in **sixth** moment
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, computationally hard
- Teicher (1961): Identifiability through tails requires many samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM) stuck in local minima
- Dasgupta (1999) and many others: Clustering assumes almost non-overlapping components
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE)
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, *computationally hard*
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, **computationally hard**
- Teicher (1961): Identifiability through tails

Dempster, Laird, Rubin (1977): Expectation-Maximization (EM) stuck in local minima

Dasgupta (1999) and many others: Clustering assumes almost non-overlapping components
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, *computationally hard*
- Teicher (1961): Identifiability through tails requires *many* samples
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, \textit{computationally hard}
- Teicher (1961): Identifiability through tails requires many samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM)
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, *computationally hard*
- Teicher (1961): Identifiability through tails requires *many* samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM) stuck in *local minima*
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, computationally hard
- Teicher (1961): Identifiability through tails requires many samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM) stuck in local minima
- Dasgupta (1999) and many others: Clustering
A Conceptual History

- Pearson (1894): Method of Moments (no guarantees)
- Fisher (1912-1922): Maximum Likelihood Estimator (MLE) consistent and efficient in the limit, computationally hard
- Teicher (1961): Identifiability through tails requires many samples
- Dempster, Laird, Rubin (1977): Expectation-Maximization (EM) stuck in local minima
- Dasgupta (1999) and many others: Clustering assumes almost non-overlapping components
In summary, these approaches are heuristic, computationally intractable or make an assumption about the mixture.
In summary, these approaches are heuristic, computationally intractable or make an assumption about the mixture.

**Question**

*To learn the parameters to within an additive $\epsilon$, is there an algorithm whose sample complexity and running time are bounded by $(1/\epsilon)^C$?*
In summary, these approaches are heuristic, computationally intractable or make an assumption about the mixture

**Question**

*To learn the parameters to within an additive $\epsilon$, is there an algorithm whose sample complexity and running time are bounded by $(1/\epsilon)^C$?*

(Kalai, Moitra, Valiant):

- Given an $n$-dimensional mixture of two Gaussians, our algorithm requires $\text{poly}(n, \frac{1}{\epsilon})$ samples and running time to output a mixture that is $\epsilon$-close to the true parameters
In summary, these approaches are heuristic, computationally intractable or make an assumption about the mixture

**Question**

*To learn the parameters to within an additive $\epsilon$, is there an algorithm whose sample complexity and running time are bounded by $(1/\epsilon)^C$?*

(Kalai, Moitra, Valiant):

- Given an $n$-dimensional mixture of two Gaussians, our algorithm requires $\text{poly}(n, \frac{1}{\epsilon})$ samples and running time to output a mixture that is $\epsilon$-close to the true parameters
- Reduce to the one-dimensional case
In summary, these approaches are heuristic, computationally intractable or make an assumption about the mixture.

**Question**

*To learn the parameters to within an additive $\epsilon$, is there an algorithm whose sample complexity and running time are bounded by $(1/\epsilon)^C$?*

(Kalai, Moitra, Valiant):

- Given an $n$-dimensional mixture of two Gaussians, our algorithm requires $\text{poly}(n, 1/\epsilon)$ samples and running time to output a mixture that is $\epsilon$-close to the true parameters.
- Reduce to the one-dimensional case.
- Analyze Pearson’s sixth moment test (with noisy moments).
Analyzing the Method of Moments

Start with an easier question:

What if we are given the first six moments of the mixture, exactly? Does this uniquely determine the parameters of the mixture (up to a relabeling of the components)?

Do any two different mixtures $F$ and $\hat{F}$ differ on at least one of the first six moments?
Analyzing the Method of Moments

Start with an easier question:

**Question**

*What if we are given the first six moments of the mixture, exactly?*
Analyzing the Method of Moments

Start with an easier question:

**Question**

*What if we are given the first six moments of the mixture, exactly?*

Does this uniquely determine the parameters of the mixture? (up to a relabeling of the components)
Analyzing the Method of Moments

Start with an easier question:

**Question**

*What if we are given the first six moments of the mixture, exactly?*

Does this uniquely determine the parameters of the mixture?

(up to a relabeling of the components)

**Question**

*Do any two different mixtures $F$ and $\hat{F}$ differ on at least one of the first six moments?*
Method of Moments
Method of Moments

\[ f(x) = F(x) - \hat{F}(x) \]
Method of Moments

\[ f(x) = F(x) - \hat{F}(x) \]
Claim

*One of the first six moment of $F$, $\hat{F}$ is different!*
Claim

*One of the first six moment of $F$, $\hat{F}$ is different!*

Proof:

$$0 < \left| \int p(x)f(x)dx \right| = \left| \int_0^\infty \sum_{r=1}^{6} p_r x^r f(x)dx \right|$$
Claim

One of the first six moment of $F$, $\hat{F}$ is different!

Proof:

\[ 0 < \left| \int_x p(x)f(x)\,dx \right| = \left| \int_x \sum_{r=1}^{6} p_rx^r f(x)\,dx \right| \]
\[ \leq \sum_{r=1}^{6} |p_r| \left| \int_x x^r f(x)\,dx \right| \]
Claim

One of the first six moments of $F, \hat{F}$ is different!

Proof:

$$0 < \left| \int_x p(x)f(x)dx \right| = \left| \int_x \sum_{r=1}^{6} p_r x^r f(x)dx \right|$$

$$\leq \sum_{r=1}^{6} |p_r| \left| \int_x x^r f(x)dx \right|$$

$$= \sum_{r=1}^{6} |p_r| |M_r(F) - M_r(\hat{F})|$$
Claim

One of the first six moments of $F$, $\hat{F}$ is different!

Proof:

$$0 < \left| \int_x p(x)f(x)dx \right| = \left| \int_x \sum_{r=1}^6 p_r x^r f(x)dx \right|$$

$$\leq \sum_{r=1}^6 \left| p_r \right| \left| \int_x x^r f(x)dx \right|$$

$$= \sum_{r=1}^6 \left| p_r \right| \left| M_r(F) - M_r(\hat{F}) \right|$$

So $\exists r \in \{1, 2, \ldots, 6\}$ such that $\left| M_r(F) - M_r(\hat{F}) \right| > 0$
Proposition

If \( f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x) \) is not identically zero, \( f(x) \) has at most \( 2k - 2 \) zero crossings (\( \alpha_i \) can be negative).
Proposition

If \( f(x) = \sum_{i=1}^{k} \alpha_i N(\mu_i, \sigma_i^2, x) \) is not identically zero, \( f(x) \) has at most \( 2k - 2 \) zero crossings (\( \alpha_i \) can be negative).

Theorem (Hummel, Gidas)

Suppose \( f(x) : \mathbb{R} \to \mathbb{R} \) is analytic and has \( n \) zeros. Then \( f(x) \circ N(0, \sigma^2, x) \) has at most \( n \) zeros (for any \( \sigma^2 > 0 \)).
Proposition

If \( f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x) \) is not identically zero, \( f(x) \) has at most \( 2k - 2 \) zero crossings (\( \alpha_i \) can be negative).

Theorem (Hummel, Gidas)

Suppose \( f(x) : \mathbb{R} \rightarrow \mathbb{R} \) is analytic and has \( n \) zeros. Then \( f(x) \circ \mathcal{N}(0, \sigma^2, x) \) has at most \( n \) zeros (for any \( \sigma^2 > 0 \)).

Convolving by a Gaussian does not increase \( \# \) of zero crossings
Proposition

If \( f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x) \) is not identically zero, \( f(x) \) has at most \( 2k - 2 \) zero crossings (\( \alpha_i \) can be negative).

Theorem (Hummel, Gidas)

Suppose \( f(x) : \mathbb{R} \rightarrow \mathbb{R} \) is analytic and has \( n \) zeros. Then \( f(x) \circ \mathcal{N}(0, \sigma^2, x) \) has at most \( n \) zeros (for any \( \sigma^2 > 0 \)).

Convolving by a Gaussian does not increase \# of zero crossings

Fact

\[ \mathcal{N}(0, \sigma_1^2, x) \circ \mathcal{N}(0, \sigma_2^2, x) = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2, x) \]
Zero Crossings and the Heat Equation
Zero Crossings and the Heat Equation
Zero Crossings and the Heat Equation
Zero Crossings and the Heat Equation
Zero Crossings and the Heat Equation
Hence, the exact values of the first six moments determine the mixture parameters.
Hence, the **exact** values of the first six moments determine the mixture parameters

**An algebraic restatement:**

Let $\Gamma = \{\text{valid parameters}\}$ (in particular $w_i \in [0, 1]$, $\sigma_i \geq 0$)
Hence, the **exact** values of the first six moments determine the mixture parameters

**An algebraic restatement:**

Let \( \Gamma = \{ \text{valid parameters} \} \) (in particular \( w_i \in [0, 1], \sigma_i \geq 0 \))

**Claim**

Let \( \theta \) be the true parameters; then the variety

\[
\{ \theta' \in \Gamma | M_r(\theta') = M_r(\theta) \text{ for } r = 1, 2, \ldots, 6 \}
\]

contains only \((w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)\) and \((1 - w_1, \mu_2, \sigma_2, \mu_1, \sigma_1)\)
Hence, the **exact** values of the first six moments determine the mixture parameters

**An algebraic restatement:**

Let $\Gamma = \{\text{valid parameters}\}$ (in particular $w_i \in [0, 1]$, $\sigma_i \geq 0$)

**Claim**

*Let $\theta$ be the true parameters; then the variety*

$$\{\theta' \in \Gamma | M_r(\theta') = M_r(\theta) \text{ for } r = 1, 2, \ldots, 6\}$$

*contains only $(w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and $(1 - w_1, \mu_2, \sigma_2, \mu_1, \sigma_1)$*

*Are these equations stable, when we are given **noisy** estimates?*
Using deconvolution to isolate components, we show:

There are constants $c$, $C$ such that if $\epsilon < c$, the means and variances are bounded by $1/\epsilon$, the mixing weights are in $[\epsilon, 1-\epsilon]$ and $|M_r(\theta) - M_r(\theta')| \leq \epsilon C$ for $r = 1, 2, ..., 6$. Then there is a permutation $\pi$ such that $2\sum_{i=1}^{2}|w_i - w_{\pi(i)}| + |\mu_i - \mu_{\pi(i)}| + |\sigma_i^2 - \sigma_{\pi(i)}^2| \leq \epsilon$.

Hence, close enough estimates for the first six moments guarantee that the parameters are close too!
A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1 - \epsilon]$ and

$$|M_r(\theta) - M_r(\theta')| \leq \epsilon^C$$

for $r = 1, 2, \ldots 6$
A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1 - \epsilon]$ and

$$|M_r(\theta) - M_r(\theta')| \leq \epsilon^C$$

for $r = 1, 2, \ldots, 6$ then there is a permutation $\pi$ such that

$$\sum_{i=1}^{2} \left| w_i - w'_{\pi(i)} \right| + \left| \mu_i - \mu'_{\pi(i)} \right| + \left| \sigma_i^2 - \sigma'_{\pi(i)}^2 \right| \leq \epsilon$$
A Type of Condition Number

Using deconvolution to isolate components, we show:

There are constants $c, C$ such that if $\epsilon < c$, the means and variances are bounded by $\frac{1}{\epsilon}$, the mixing weights are in $[\epsilon, 1 - \epsilon]$ and

$$|M_r(\theta) - M_r(\theta')| \leq \epsilon^C$$

for $r = 1, 2, ... 6$ then there is a permutation $\pi$ such that

$$\sum_{i=1}^{2} |w_i - w_{\pi(i)}'| + |\mu_i - \mu_{\pi(i)}'| + |\sigma_i^2 - \sigma_{\pi(i)}'^2| \leq \epsilon$$

Hence, close enough estimates for the first six moments guarantee that the parameters are close too!
A Univariate Learning Algorithm

Our algorithm:

\[ \text{Take enough samples } S \text{ so that } \tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_r i \text{ is w.h.p. close to } M_r(\theta) \text{ for } r = 1, 2, \ldots, 6 \text{ (within an additive } \epsilon) \]

\[ \text{Compute } \theta' \text{ such that } M_r(\theta') \text{ is close to } \tilde{M}_r \text{ for } r = 1, 2, \ldots, 6 \text{ (within an additive } \epsilon) \]

And \( \theta' \) must be close to \( \theta \), because solutions to this system of polynomial equations are stable.
A Univariate Learning Algorithm

Our algorithm:

- Take enough samples $S$ so that $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2...6$
A Univariate Learning Algorithm

Our algorithm:

- Take enough samples $S$ so that $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2, ... 6$

  (within an additive $\frac{\epsilon C}{2}$)
A Univariate Learning Algorithm

Our algorithm:

- Take enough samples $S$ so that $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2...6$
  (within an additive $\frac{\epsilon C}{2}$)

- Compute $\theta'$ such that $M_r(\theta')$ is close to $\tilde{M}_r$ for $r = 1, 2...6$
A Univariate Learning Algorithm

Our algorithm:

- Take enough samples $S$ so that $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2, \ldots, 6$
  (within an additive $\frac{\epsilon C}{2}$)

- Compute $\theta'$ such that $M_r(\theta')$ is close to $\tilde{M}_r$ for $r = 1, 2, \ldots, 6$
  (within an additive $\frac{\epsilon C}{2}$)
A Univariate Learning Algorithm

Our algorithm:

- Take enough samples $S$ so that $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ is w.h.p. close to $M_r(\theta)$ for $r = 1, 2...6$
  (within an additive $\frac{\epsilon C}{2}$)

- Compute $\theta'$ such that $M_r(\theta')$ is close to $\tilde{M}_r$ for $r = 1, 2...6$
  (within an additive $\frac{\epsilon C}{2}$)

And $\theta'$ must be close to $\theta$, because solutions to this system of polynomial equations are stable
A More General Approach

(Belkin, Sinha)
(Belkin, Sinha): Consider a family of distributions $F(\theta)$

**Fact**

*If the moment generating function converges in a neighborhood around zero, then $M_r(\theta) = M_r(\theta')$ for all $r$ implies $F(\theta) = F(\theta')$.***
A More General Approach

(Belkin, Sinha): Consider a family of distributions $F(\theta)$

Fact

*If the moment generating function converges in a neighborhood around zero, then $M_r(\theta) = M_r(\theta')$ for all $r$ implies $F(\theta) = F(\theta')$.***

Definition

A family of distributions is a **polynomial family** if the above condition holds and furthermore $M_r(\theta)$ is a polynomial (for any $r$).
(Belkin, Sinha): Consider a family of distributions $F(\theta)$

**Fact**

If the moment generating function converges in a neighborhood around zero, then $M_r(\theta) = M_r(\theta')$ for all $r$ implies $F(\theta) = F(\theta')$.

**Definition**

A family of distributions is a **polynomial family** if the above condition holds and furthermore $M_r(\theta)$ is a polynomial (for any $r$).

e.g. a mixture of Gaussians
Definition

\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]
Definition

\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]

Consider the ideals \( I_1 = \langle Q_1 \rangle \subseteq I_2 = \langle Q_1, Q_2 \rangle \ldots \subseteq \mathbb{R}[\theta, \theta'] \)
Definition

\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]

Consider the ideals \( I_1 = \langle Q_1 \rangle \subseteq I_2 = \langle Q_1, Q_2 \rangle \cdots \subseteq \mathbb{R}[\theta, \theta'] \)

Fact

\( \mathbb{R}[\theta, \theta'] \) is a Noetherian Ring (Hilbert’s Basis Theorem)
Definition

\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]

Consider the ideals \( I_1 = \langle Q_1 \rangle \subseteq I_2 = \langle Q_1, Q_2 \rangle \ldots \subseteq \mathbb{R}[\theta, \theta'] \)

Fact

\( \mathbb{R}[\theta, \theta'] \) is a Noetherian Ring (Hilbert’s Basis Theorem)

Hence for some \( N \), \( I_N = I_{N+1} = \ldots \);
Definition

\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]

Consider the ideals \( I_1 = \langle Q_1 \rangle \subseteq I_2 = \langle Q_1, Q_2 \rangle \ldots \subseteq \mathbb{R}[\theta, \theta'] \)

Fact

\( \mathbb{R}[\theta, \theta'] \) is a Noetherian Ring (Hilbert’s Basis Theorem)

Hence for some \( N \), \( I_N = I_{N+1} = \ldots \); i.e.

\[ Q_{N+j}(\theta, \theta') = \sum_{r=1}^{N} \alpha(\theta, \theta') Q_r(\theta, \theta') \]
Definition
\[ Q_r(\theta, \theta') = M_r(\theta) - M_r(\theta') \]

Consider the ideals \( I_1 = \langle Q_1 \rangle \subseteq I_2 = \langle Q_1, Q_2 \rangle \ldots \subseteq \mathbb{R}[\theta, \theta'] \)

Fact
\( \mathbb{R}[\theta, \theta'] \) is a Noetherian Ring (Hilbert’s Basis Theorem)

Hence for some \( N \), \( I_N = I_{N+1} = \ldots \); i.e.

\[ Q_{N+j}(\theta, \theta') = \sum_{r=1}^{N} \alpha(\theta, \theta') Q_r(\theta, \theta') \]

\[ \sum_{i=1}^{N} |M_i(\theta) - M_i(\theta')| = 0 \Rightarrow \text{all moments are equal} \Rightarrow F(\theta) = F(\theta') \]
Question

If $\sum_{r=1}^{N} |M_r(\theta) - M_r(\theta')| < \delta$, for what $\epsilon(\delta)$ is $|\theta - \theta'| < \epsilon$?
Question

If \( \sum_{r=1}^{N} |M_r(\theta) - M_r(\theta')| < \delta \), for what \( \epsilon(\delta) \) is \( |\theta - \theta'| < \epsilon \)?

There is a notion of **condition number** for systems of polynomial equations:
Question

If \[ \sum_{r=1}^{N} |M_r(\theta) - M_r(\theta')| < \delta, \] for what \( \epsilon(\delta) \) is \( |\theta - \theta'| < \epsilon \)?

There is a notion of **condition number** for systems of polynomial equations:

**Claim**

We can take \( \epsilon \) to be fixed a polynomial in \( \delta \) (e.g. via quantifier elimination)
Question

If \( \sum_{r=1}^{N} |M_r(\theta) - M_r(\theta')| < \delta \), for what \( \epsilon(\delta) \) is \( |\theta - \theta'| < \epsilon \)?

There is a notion of **condition number** for systems of polynomial equations:

**Claim**

*We can take \( \epsilon \) to be fixed a polynomial in \( \delta \) (e.g. via quantifier elimination)*

(Belkin, Sinha): The method of moments learns the parameters of a polynomial family \( F(\theta) \) to within \( \epsilon \) in \((1/\epsilon)^C\) samples and time
Question

If $\sum_{r=1}^{N} |M_r(\theta) - M_r(\theta')| < \delta$, for what $\epsilon(\delta)$ is $|\theta - \theta'| < \epsilon$?

There is a notion of **condition number** for systems of polynomial equations:

Claim

*We can take $\epsilon$ to be fixed a polynomial in $\delta$ (e.g. via quantifier elimination)*

(Belkin, Sinha): The method of moments learns the parameters of a polynomial family $F(\theta)$ to within $\epsilon$ in $(1/\epsilon)^C$ samples and time

**Caveat:** This uses Hilbert’s Basis Theorem, hence no **effective** bound for number of moments (or $C$)
Outline

- Mixtures of Gaussians
  - Highlights: *method of moments* and the heat equation
  - based on [Kalai, Moitra, Valiant]
    (see also [Belkin and Sinha])

- Topic Models
  - Highlights: *tensor methods* and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- Nonnegative Matrix Factorization
  - Highlights: *separability* and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
Outline

- Mixtures of Gaussians
  - Highlights: *method of moments* and the heat equation
  - based on [Kalai, Moitra, Valiant]
    (see also [Belkin and Sinha])

- **Topic Models**
  - Highlights: *tensor methods* and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- Nonnegative Matrix Factorization
  - Highlights: *separability* and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
Topic Models

Large collection of articles, say from the New York Times:

newspaper articles
Topic Models

Large collection of articles, say from the New York Times:

newspaper articles

Question

*How can we automatically organize them by topic? (unsupervised learning)*
Large collection of articles, say from the New York Times:

newspaper articles

**Question**

*How can we automatically organize them by topic? (unsupervised learning)*

**Challenge:** Develop tools for automatic comprehension of data - e.g. newspaper articles, webpages, images, genetic sequences, user ratings...
Parceling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients’ skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between President Obama and Congress in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government in estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?
Parceling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients’ skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between President Obama and Congress in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government in estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?
Parceling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients’ skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between President Obama and Congress in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?
Parcelling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients’ skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between President Obama and Congress in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government in estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?

• Each document is a distribution on topics
Parcelling Out a Nest Egg, Without Emptying It

By PAUL SULLIVAN

What clients often forget are fixed costs — homes, cars, insurance — that must come down but take time to reduce, she said. Beyond that is her clients’ skittish approach to risk; putting all of their money in cash may make them feel safe, she said, but it probably will not support the lifestyle they want for decades.

A generational disconnect is at work here: most people plan to retire at 65, the retirement age established for Social Security in 1935, when the average life expectancy was 61. Today the average is over 80 for men and women with a college degree.

So the $5.12 million gift exemption — created in a compromise between Congress and President Obama in 2010 — presents the well-off with a decision laden with short- and long-term consequences. How much should they give heirs now — and thus avoid giving the government estate taxes later — while maintaining their lifestyle over a probably longer but still unpredictable remaining life span?

• Each document is a distribution on topics
• Each topic is a distribution on words
fixed stochastic

A W M
fixed stochastic

\[ A \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} W \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = M \]

document #1: (1.0, personal finance)
fixed stochastic

document #1: (1.0, personal finance)
fixed stochastic

\[
A \quad W = M
\]
fixed stochastic

A W M

= 

document #2: (0.5, baseball); (0.5, movie review)
fixed stochastic

A

W

M

document #2: (0.5, baseball); (0.5, movie review)
fixed stochastic

A W

≈

document #2: (0.5, baseball); (0.5, movie review)
Latent Dirichlet Allocation (Blei, Ng, Jordan)

fixed Dirichlet

document #2: (0.5, baseball); (0.5, movie review)
fixed

\[
\begin{align*}
&\text{document #2: (0.5, baseball); (0.5, movie review)}
\end{align*}
\]
Correlated Topic Model (Blei, Lafferty)

fixed Logistic Normal

A W M

≈

document #2: (0.5, baseball); (0.5, movie review)
fixed

A

W

\approx

M

document #2: (0.5, baseball); (0.5, movie review)
Pachinko Allocation Model (Li, McCallum)

fixed Multilevel DAG

document #2: (0.5, baseball); (0.5, movie review)
These models differ only in how $W$ is generated.

Pachinko Allocation Model (Li, McCallum)

Multilevel DAG

Document #2: (0.5, baseball); (0.5, movie review)
- **Maximum Likelihood:** Find the parameters that maximize the likelihood of generating the observed data.
Algorithms

- **Maximum Likelihood**: Find the parameters that maximize the likelihood of generating the observed data.
  Hard to compute!
Algorithms

- **Maximum Likelihood**: Find the parameters that maximize the likelihood of generating the observed data.
  Hard to compute!

- **Spectral**: Compute the singular value decomposition of $\hat{M}$. [Papadimitriou et al], [Azar et al], ...

- **Tensors**: Use powerful tensor decompositions to recover $A$, when the topic model can be “diagonalized”. [Anandkumar et al]

- **Nonnegative Matrix Factorization**: When $A$ is separable, works for any topic model [Arora et al]
Algorithms

- **Maximum Likelihood**: Find the parameters that maximize the likelihood of generating the observed data. Hard to compute!

- **Spectral**: Compute the singular value decomposition of $\hat{M}$. [Papadimitriou et al], [Azar et al], ...
  But this only recovers the span of $A$
Algorithms

- **Maximum Likelihood**: Find the parameters that maximize the likelihood of generating the observed data. Hard to compute!

- **Spectral**: Compute the singular value decomposition of $\hat{M}$. [Papadimitriou et al], [Azar et al], ... But this only recovers the span of $A$

- **Tensors**: Use powerful tensor decompositions to recover $A$, when the topic model can be “diagonalized”. [Anandkumar et al]
Algorithms

- **Maximum Likelihood**: Find the parameters that maximize the likelihood of generating the observed data. Hard to compute!

- **Spectral**: Compute the singular value decomposition of \( \hat{M} \).
  [Papadimitriou et al], [Azar et al], ...
  But this only recovers the span of \( A \)

- **Tensors**: Use powerful tensor decompositions to recover \( A \), when the topic model can be “diagonalized”. [Anandkumar et al]

- **Nonnegative Matrix Factorization**: When \( A \) is separable, works for any topic model [Arora et al]
Chang’s Lemma

Given \( T = \sum_i u_i \otimes v_i \otimes w_i \)
Chang’s Lemma

Given $T = \sum_i u_i \otimes v_i \otimes w_i$

- Suppose that $\{u_i\}$, $\{v_i\}$ and $\{w_i\}$ are linearly independent

Hence we compute find $U$ and $V$ through eigen-decomposition (can also recover $W$) if diagonals of $D$ are distinct
Chang’s Lemma

Given $T = \sum_i u_i \otimes v_i \otimes w_i$

- Suppose that $\{u_i\}$, $\{v_i\}$ and $\{w_i\}$ are linearly independent
  (even better: well-conditioned)

Hence we compute find $U$ and $V$ through eigen-decomposition (can also recover $W$) if diagonals of $D_a D^{-1} b U^{-1}$ are distinct.
Chang’s Lemma

Given \( T = \sum_i u_i \otimes v_i \otimes w_i \)

- Suppose that \( \{u_i\}, \{v_i\} \) and \( \{w_i\} \) are linearly independent (even better: well-conditioned)
- Choose random unit vectors \( a, b \)
Chang’s Lemma

Given \( T = \sum_i u_i \otimes v_i \otimes w_i \)

- Suppose that \( \{u_i\}, \{v_i\} \) and \( \{w_i\} \) are linearly independent (even better: well-conditioned)
- Choose random unit vectors \( a, b \)
- Set \( T(\cdot, \cdot, a) = \sum_i (w_i^T a) u_i v_i^T = U D_a V^T \), similarly for \( T(\cdot, \cdot, b) \)
Chang’s Lemma

Given $T = \sum_i u_i \otimes v_i \otimes w_i$

- Suppose that $\{u_i\}, \{v_i\}$ and $\{w_i\}$ are linearly independent (even better: well-conditioned)
- Choose random unit vectors $a, b$
- Set $T(\cdot, \cdot, a) = \sum_i (w_i^T a) u_i v_i^T = UD_a V^T$, similarly for $T(\cdot, \cdot, b)$
- Then $T(\cdot, \cdot, a) (T(\cdot, \cdot, b))^{-1} = UD_a D_b^{-1} U^{-1}$, similarly for $V$
Chang’s Lemma

Given \( T = \sum_i u_i \otimes v_i \otimes w_i \)

- Suppose that \( \{u_i\}, \{v_i\} \) and \( \{w_i\} \) are linearly independent (even better: well-conditioned)
- Choose random unit vectors \( a, b \)
- Set \( T(\cdot, \cdot, a) = \sum_i (w_i^T a) u_i v_i^T = U D_a V^T \), similarly for \( T(\cdot, \cdot, b) \)
- Then \( T(\cdot, \cdot, a)(T(\cdot, \cdot, b))^{-1} = U D_a D_b^{-1} U^{-1} \), similarly for \( V \)

Hence we compute find \( U \) and \( V \) through eigen-decomposition (can also recover \( W \)) if diagonals of \( D_a D_b^{-1} \) are distinct
Applications

(Mossel, Roch): Applications to phylogenetic reconstruction and HMMs, when transition matrices are full-rank
Applications

(Mossel, Roch): Applications to phylogenetic reconstruction and HMMs, when transition matrices are full-rank

(Anandkumar, Hsu, Kakade): Applications to pure topic models
Applications

(Mossel, Roch): Applications to phylogenetic reconstruction and HMMs, when transition matrices are full-rank

(Anandkumar, Hsu, Kakade): Applications to pure topic models

Definition

Let \( T = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta, \text{word}_3 = \gamma] \) in a random document of length \( \geq 3 \).
Applications

(Mossel, Roch): Applications to phylogenetic reconstruction and HMMs, when transition matrices are full-rank

(Anandkumar, Hsu, Kakade): Applications to pure topic models

Definition

Let $T = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta, \text{word}_3 = \gamma]$ in a random document of length $\geq 3$.

$$T = \sum_i Pr[\text{topic} = i] A_i \otimes A_i \otimes A_i,$$

where $A_i = Pr[\text{word}|\text{topic} = i]$
Applications

(Mossel, Roch): Applications to phylogenetic reconstruction and HMMs, when transition matrices are full-rank

(Anandkumar, Hsu, Kakade): Applications to pure topic models

Definition
Let $T = \Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta, \text{word}_3 = \gamma]$ in a random document of length $\geq 3$.

$$T = \sum_i \Pr[\text{topic} = i] A_i \otimes A_i \otimes A_i,$$ where $A_i = \Pr[\text{word}|\text{topic} = i]$.

Hence we can recover $A$ if it is full rank and each document is about only one topic.
(Anandkumar, Foster, Hsu, Kakade, Liu): What about more general topic models? (e.g. LDA)
(Anandkumar, Foster, Hsu, Kakade, Liu): What about more general topic models? (e.g. LDA)

The challenge is $T$ has a more complicated form: ($D$ is not necessarily diagonal)
(Anandkumar, Foster, Hsu, Kakade, Liu): What about more general topic models? (e.g. LDA)

The challenge is $T$ has a more complicated form: ($D$ is not necessarily diagonal)

**Definition**

A Tucker decomposition of $T$ is a set of $n \times r$ matrices $U$, $V$, $W$ and an $r \times r$ matrix $D$, such that

$$T = \sum_{i,j,k \in [r]} D_{i,j,k} U_i \otimes V_j \otimes W_k$$
(Anandkumar, Foster, Hsu, Kakade, Liu): What about more general topic models? (e.g. LDA)

The challenge is $T$ has a more complicated form: ($D$ is not necessarily diagonal)

Definition

A Tucker decomposition of $T$ is a set of $n \times r$ matrices $U, V, W$ and an $r \times r$ matrix $D$, such that

$$T = \sum_{i,j,k \in [r]} D_{i,j,k} U_i \otimes V_j \otimes W_k$$

In our setting, $U = V = W = A$, but $D$ corresponds to **moments** of the Dirichlet distribution
Let $\mu = Pr[\text{word}_1 = \alpha]$ and $M = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta]$
Let $\mu = Pr[\text{word}_1 = \alpha]$ and $M = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta]$

**Observation**

*We can form new tensors, e.g. $\mu \otimes \mu \otimes \mu$ or $M \otimes \mu$*
Let $\mu = Pr[\text{word}_1 = \alpha]$ and $M = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta]$

Observation

*We can form new tensors, e.g. $\mu \otimes \mu \otimes \mu$ or $M \otimes \mu$*

Claim

*Given Tucker decompositions $T$ and $T'$ with same $U, V, W$

$$T - T' = \sum_{i,j,k \in [r]} (D_{i,j,k} - D'_{i,j,k}) U_i \otimes V_j \otimes W_k$$

(Anandkumar, Foster, Hsu, Kakade, Liu): The formula $T + 2\mu \otimes \mu \otimes \mu - M \otimes \mu$ (all three ways) diagonalizes the decomposition, and hence we can recover $A$ if it is full rank for LDA topic models!
Let $\mu = Pr[\text{word}_1 = \alpha]$ and $M = Pr[\text{word}_1 = \alpha, \text{word}_2 = \beta]$

**Observation**

*We can form new tensors, e.g. $\mu \otimes \mu \otimes \mu$ or $M \otimes \mu$*

**Claim**

*Given Tucker decompositions $T$ and $T'$ with same $U, V, W$*

$$T - T' = \sum_{i,j,k \in [r]} (D_{i,j,k} - D'_{i,j,k}) U_i \otimes V_j \otimes W_k$$

(Anandkumar, Foster, Hsu, Kakade, Liu): The formula

$$T + 2\mu \otimes \mu \otimes \mu - M \otimes \mu$$  (all three ways)

**diagonalizes** the decomposition
Let \( \mu = Pr[word_1 = \alpha] \) and \( M = Pr[word_1 = \alpha, word_2 = \beta] \)

**Observation**

We can form new tensors, e.g. \( \mu \otimes \mu \otimes \mu \) or \( M \otimes \mu \)

**Claim**

Given Tucker decompositions \( T \) and \( T' \) with same \( U, V, W \)

\[
T - T' = \sum_{i,j,k \in [r]} (D_{i,j,k} - D'_{i,j,k}) U_i \otimes V_j \otimes W_k
\]

(Anandkumar, Foster, Hsu, Kakade, Liu): The formula

\[
T + 2\mu \otimes \mu \otimes \mu - M \otimes \mu \text{ (all three ways)}
\]

**diagonalizes** the decomposition, and hence we can recover \( A \) if it is full rank for LDA topic models!
Outline

- **Mixtures of Gaussians**
  - Highlights: *method of moments* and the heat equation
  - based on [Kalai, Moitra, Valiant] (see also [Belkin and Sinha])

- **Topic Models**
  - Highlights: *tensor methods* and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- **Nonnegative Matrix Factorization**
  - Highlights: *separability* and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
Outline

- Mixtures of Gaussians
  - Highlights: *method of moments* and the heat equation
  - based on [Kalai, Moitra, Valiant]
    (see also [Belkin and Sinha])

- Topic Models
  - Highlights: *tensor methods* and Chang’s Lemma
  - based on [Anandkumar, Foster, Hsu, Kakade and Liu]

- **Nonnegative Matrix Factorization**
  - Highlights: *separability* and more general topic models
  - based on [Arora, Ge, Kannan, Moitra]
\[ n \times m \times M = A \imes \text{inner-dimension} \]
\text{rank} \quad M = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = A

\text{inner-dimension}
\[ \text{rank} \quad \begin{array}{c} \text{non-negative} \\ \text{inner-dimension} \\ \text{non-negative} \end{array} \]

\[
\begin{array}{c}
A \\
M
\end{array} =
\begin{array}{c}
A \\
W
\end{array}
\]

\[
\begin{array}{c}
m \\
n
\end{array}
\]
A non-negative rank

\[ m = \text{inner-dimension} \]

\[ W \]

\[ \text{non-negative} \]

\[ n = \text{inner-dimension} \]

\[ M \]

\[ \text{non-negative} \]

\[ A \]

\[ \text{inner-dimension} \]
Local Search: Given $A$, compute $W$, compute $A$, ....

Known to fail on worst-case inputs (stuck in local minima)

Highly sensitive to cost function, regularization, update procedure

Question (theoretical)
Is there an algorithm that (provably) works on all inputs?

[Arora, Ge, Kannan, Moitra]: There is an $(nm)$ time algorithm

but improving this to $(nm o(r))$ would imply a subexponential time algorithm for 3-SAT

Question
Can we do better for natural instances?
Local Search: Given $A$, compute $W$, compute $A$, ....

- Known to fail on worst-case inputs (stuck in local minima)

[Arora, Ge, Kannan, Moitra]: There is an $(nm)$ time algorithm but improving this to $(nm)^{o(r)}$ would imply a subexponential time algorithm for 3-SAT.

Question: Can we do better for natural instances?
Local Search: Given $A$, compute $W$, compute $A$, ....

- Known to fail on worst-case inputs (stuck in local minima)
- Highly sensitive to cost function, regularization, update procedure

[Question (theoretical)]

Is there an algorithm that (provably) works on all inputs?

[Arora, Ge, Kannan, Moitra]: There is an $(nm)O(r^2)$ time algorithm but improving this to $(nm)o(r)$ would imply a subexponential time algorithm for 3-SAT

[Question]

Can we do better for natural instances?
Local Search: Given $A$, compute $W$, compute $A$, ....

- Known to fail on worst-case inputs (stuck in local minima)
- Highly sensitive to cost function, regularization, update procedure

**Question (theoretical)**

*Is there an algorithm that (provably) works on all inputs?*
Local Search: Given $A$, compute $W$, compute $A$, ....

- Known to fail on worst-case inputs (stuck in local minima)
- Highly sensitive to cost function, regularization, update procedure

**Question (theoretical)**

*Is there an algorithm that (provably) works on all inputs?*

[Arora, Ge, Kannan, Moitra]: There is an $(nm)^{O(r^2)}$ time algorithm but improving this to $(nm)^{o(r)}$ would imply a subexponential time algorithm for 3-SAT
Local Search: Given $A$, compute $W$, compute $A$, ....

- Known to fail on worst-case inputs (stuck in local minima)
- Highly sensitive to cost function, regularization, update procedure

**Question (theoretical)**

*Is there an algorithm that (provably) works on all inputs?*

[Arora, Ge, Kannan, Moitra]: There is an $(nm)^{O(r^2)}$ time algorithm but improving this to $(nm)^{o(r)}$ would imply a subexponential time algorithm for 3-SAT

**Question**

*Can we do better for natural instances?*
If an anchor word occurs then the document is at least partially about the topic.
<table>
<thead>
<tr>
<th>topics (r)</th>
<th>personal finance</th>
</tr>
</thead>
</table>

If an **anchor word** occurs then the document is at least partially about the topic.
If an anchor word occurs then the document is at least partially about the topic.
If an anchor word occurs then the document is at least partially about the topic.
If an anchor word occurs then the document is at least partially about the topic.
If an **anchor word** occurs then the document is at least partially about the topic.
If an anchor word occurs then the document is at least partially about the topic.
If an **anchor word** occurs then the document is at least partially about the topic.
If an **anchor word** occurs then the document is at least partially about the topic.
If an **anchor word** occurs then the document is at least partially about the topic.
If an **anchor word** occurs then the document is at least partially about the topic.

A is **p-separable** if each topic has an anchor word that occurs with probability $\geq p$. 

<table>
<thead>
<tr>
<th>topics (r)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>movie reviews</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>words (m)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>bunt</td>
<td></td>
</tr>
<tr>
<td>401k</td>
<td></td>
</tr>
<tr>
<td>oscar-winning</td>
<td></td>
</tr>
</tbody>
</table>
Question

How do anchor words help?

Observation

If $A$ is separable, then rows of $W$ appear as (scaled) rows of $M$, we just need to find the anchor words!

Question

How can we find the anchor words?

Anchor words are extreme points; can be found by linear programming (or a combinatorial distance-based algorithm)
A
Question

*How do anchor words help?*
Using Anchor Words

Question

*How do anchor words help?*

Observation

*If A is separable, then rows of \( W \) appear as (scaled) rows of \( M \), we just need to find the anchor words!*
Using Anchor Words

Question

How do anchor words help?

Observation

If $A$ is separable, then rows of $W$ appear as (scaled) rows of $M$, we just need to find the anchor words!

Question

How can we find the anchor words?
A

W

= M
Deleting a word changes the convex hull. It is an anchor word.
Using Anchor Words

Question

*How do anchor words help?*

Observation

*If $A$ is separable, then rows of $W$ appear as (scaled) rows of $M$, we just need to find the anchor words!*

Question

*How can we find the anchor words?*
Using Anchor Words

Question

*How do anchor words help?*

Observation

*If A is separable, then rows of W appear as (scaled) rows of M, we just need to find the anchor words!*

Question

*How can we find the anchor words?*

Anchor words are extreme points; can be found by linear programming (or a combinatorial distance-based algorithm)
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

Find the anchor words (linear programming):

If a word cannot be written as a convex combination of the other words, it is an anchor word.
Paste these vectors in as rows of $W$.

Find the nonnegative $A$ so that $AW \approx M$ (convex programming).

Claim: The following greedy algorithm works too: repeatedly find the word furthest from the span of the ones we have found so far!
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

- Find the anchor words (linear programming):

  If a word cannot be written as a convex combination of the other words, it is an anchor word. Paste these vectors in as rows of $W$. Find the nonnegative $A$ so that $AW \approx M$ (convex programming).
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

- Find the anchor words (linear programming):
  If a word cannot be written as a convex combination of the other words, it is an anchor word.
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

- Find the anchor words (linear programming):
  
  If a word cannot be written as a convex combination of the other words, it is an anchor word

- Paste these vectors in as rows of $W$
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

- Find the anchor words (linear programming):
  If a word cannot be written as a convex combination of the other words, it is an anchor word
- Paste these vectors in as rows of $W$
- Find the nonnegative $A$ so that $AW \approx M$ (convex programming)
An Algorithm for NMF

(Arora, Ge, Kannan, Moitra):

- Find the anchor words (linear programming):
  
  If a word cannot be written as a convex combination of the other words, it is an anchor word

- Paste these vectors in as rows of $W$

- Find the nonnegative $A$ so that $AW \approx M$ (convex programming)

Claim

The following greedy algorithm works too: repeatedly find the word furthest from the span of the ones we have found so far!
Question

What if documents are short; can we still find $A$?
Back to Topic Models

Question

*What if documents are short; can we still find $A$?*

Crucial observation: We can work with the Gram matrix (define next) to find the anchor words.
Gram Matrix

$\hat{M} \hat{M}^T$
Gram Matrix

\[ \hat{M} \hat{M}^T \]
Gram Matrix

\[ \hat{M} \hat{M}^T \rightarrow E[MM^T] \]

\[ AA^T \]

\[ WW^T \]
Gram Matrix

\[ \hat{M} \hat{M}^T \rightarrow E[M M^T] = A E[WW^T] A^T \]
Gram Matrix

\( \hat{M} \hat{M}^T \rightarrow E[M M^T] \equiv A E[WW^T] A^T \rightarrow A A^T \)
Gram Matrix

\[ \hat{M} \hat{M}^T \quad \Rightarrow \quad E[MM^T] = A E[WW^T] A^T \quad \Rightarrow \quad ARA^T \]

nonnegative
Gram Matrix

\[ \hat{M} \hat{M}^T \]

\[ E[M M^T] \equiv A E[WW^T] A^T \rightarrow ARA^T \]

nonnegative

separable!
Anchor words are extreme rows of the Gram matrix!
Question

*What if documents are short; can we still find A?*

Crucial observation: We can work with the Gram matrix (define next) to find the anchor words
Question

*What if documents are short; can we still find A?*

Crucial observation: We can work with the Gram matrix (define next) to find the anchor words

Question

*How can we use the anchor words to find the rest of A?*
Question

What if documents are **short**; can we still find $A$?

Crucial observation: We can work with the Gram matrix (define next) to find the anchor words.

Question

*How can we use the anchor words to find the rest of $A$?*

The posterior distribution $Pr[topic|word]$ is supported on just one topic, for an anchor word.
Question

What if documents are short; can we still find $A$?

Crucial observation: We can work with the Gram matrix (define next) to find the anchor words

Question

How can we use the anchor words to find the rest of $A$?

The posterior distribution $Pr[topic|word]$ is supported on just one topic, for an anchor word.

We will find $Pr[topic|word]$ for all the other words...
points are now (normalized) rows of $\hat{M} \hat{M}^T$
points are now (normalized)
rows of $\hat{M} \hat{M}^T$
points are now (normalized) rows of $\hat{M} \hat{M}^T$

word #3: (0.5, anchor #2); (0.5, anchor #3)
points are now (normalized) rows of $\hat{M} \hat{M}^T$

word #3: (0.5, anchor #2); (0.5, anchor #3)

Pr[topic|word #3]: (0.5, topic #2); (0.5, topic #3)
points are now (normalized) rows of $\hat{M} \hat{M}^T$

what we have:

**$Pr[\text{topic} | \text{word}]$**

word #3: (0.5, anchor #2); (0.5, anchor #3)

$Pr[\text{topic} | \text{word} \#3]$: (0.5, topic #2); (0.5, topic #3)
points are now (normalized) rows of $\hat{M} \hat{M}^T$

what we have: 

A 

what we want: 

Pr[topic|word] 

Pr[topic|word #3]: (0.5, topic #2); (0.5, topic #3)

word #3: (0.5, anchor #2); (0.5, anchor #3)

Pr[topic|word #3]: (0.5, topic #2); (0.5, topic #3)
points are now (normalized) rows of $\mathbf{M} \mathbf{M}^T$

what we have: 

$A$

what we want: 

$\Pr[\text{word}|\text{topic}]$

$\Pr[\text{topic}|\text{word}]$

Bayes Rule

what we want:

$\Pr[\text{word}|\text{topic}]$

word #3: (0.5, anchor #2); (0.5, anchor #3)

Pr[topic|word #3]: (0.5, topic #2); (0.5, topic #3)
An Algorithm for Topic Models

(Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, Zhu):

Form the Gram matrix and find the anchor words. Write each word as a convex combination of the anchor words to find $\Pr[\text{topic} | \text{word}]$. Compute $A$ from Bayes' Rule: $\Pr[\text{word} | \text{topic}] = 1$. This algorithm provably works for any topic model (LDA, CTM, PAM, ...) provided $A$ is separable and $R$ is non-singular!
An Algorithm for Topic Models

(Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, Zhu):

- Form the Gram matrix and find the anchor words

\[ \text{Compute } \mathbf{A} \text{ from Bayes' Rule: } \Pr[\text{word} | \text{topic}] = \frac{1}{1} \]

This algorithm provably works for any topic model (LDA, CTM, PAM, ...) provided \( \mathbf{A} \) is separable and \( \mathbf{R} \) is non-singular!
An Algorithm for Topic Models

(Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, Zhu):

- Form the Gram matrix and find the anchor words
- Write each word as a convex combination of the anchor words to find $Pr[\text{topic}|\text{word}]$
An Algorithm for Topic Models

(Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, Zhu):

- Form the Gram matrix and find the anchor words
- Write each word as a convex combination of the anchor words to find $Pr[\text{topic} | \text{word}]$
- Compute $A$ from Bayes’ Rule:

$$Pr[\text{word} | \text{topic}] = \frac{Pr[\text{topic} | \text{word}]Pr[\text{word}]}{\sum_{\text{word}'} Pr[\text{topic} | \text{word}']Pr[\text{word}']}$$
An Algorithm for Topic Models

(Arora, Ge, Halpern, Mimno, Moitra, Sontag, Wu, Zhu):

- Form the Gram matrix and find the anchor words
- Write each word as a convex combination of the anchor words to find $Pr[\text{topic}|\text{word}]$
- Compute $A$ from Bayes’ Rule:

$$Pr[\text{word}|\text{topic}] = \frac{Pr[\text{topic}|\text{word}]Pr[\text{word}]}{\sum_{\text{word}'} Pr[\text{topic}|\text{word}']\Pr[\text{word}']}$$

This algorithm provably works for any topic model (LDA, CTM, PAM, ...) provided $A$ is separable and $R$ is non-singular!
Epilogue

Open Question
Are there better algorithms for general GMMs?

There are powerful uniqueness theorems for tensors (beyond Chang's Lemma):

Open Question
Is there an algorithmic proof of Kruskal's Theorem?

Other uses of the polynomial method?

(Moitra, Saks): Applications to inverse problems, population recovery
(Hsu, Kakade): Improved algorithm for spherical GMMs
(Hsu, Kakade): Improved algorithm for spherical GMMs

Open Question

Are there better algorithm for general GMMs?
(Hsu, Kakade): Improved algorithm for spherical GMMs

Open Question

Are there better algorithm for general GMMs?

There are powerful uniqueness theorems for tensors (beyond Chang’s Lemma):
(Hsu, Kakade): Improved algorithm for \textit{spherical} GMMs

**Open Question**

\textit{Are there better algorithm for general GMMs?}

There are powerful uniqueness theorems for tensors (beyond Chang’s Lemma):

**Open Question**

\textit{Is there an algorithmic proof of Kruskal’s Theorem?}
Epilogue

(Hsu, Kakade): Improved algorithm for \textit{spherical} GMMs

\textbf{Open Question}

Are there better algorithm for general GMMs?

There are powerful uniqueness theorems for tensors (beyond Chang’s Lemma):

\textbf{Open Question}

\textit{Is there an algorithmic proof of Kruskal’s Theorem?}

Other uses of the polynomial method?
Epilogue

(Hsu, Kakade): Improved algorithm for **spherical** GMMs

**Open Question**

*Are there better algorithm for general GMMs?*

There are powerful uniqueness theorems for tensors (beyond Chang’s Lemma):

**Open Question**

*Is there an algorithmic proof of Kruskal’s Theorem?*

Other uses of the polynomial method? (Moitra, Saks): Applications to inverse problems, population recovery
Thanks!