Tight Bounds for Learning a Mixture of Two Gaussians

Moritz Hardt      Eric Price

Google Research    UT Austin

2015-06-17
Problem

- Height distribution of American 20 year olds.
Height distribution of American 20 year olds.

- Male/female heights are very close to Gaussian distribution.
Problem

- Height distribution of American 20 year olds.
  - Male/female heights are very close to Gaussian distribution.
- Can we learn the average male and female heights from *unlabeled* population data?
Problem

- Height distribution of American 20 year olds.
  - Male/female heights are very close to Gaussian distribution.
- Can we learn the average male and female heights from *unlabeled* population data?
- How many samples to learn $\mu_1, \mu_2$ to $\pm \epsilon \sigma$?
Problem

- Height distribution of American 20 year olds.
  - Male/female heights are very close to Gaussian distribution.
- Can we learn the average male and female heights from *unlabeled* population data?
- How many samples to learn $\mu_1, \mu_2$ to $\pm \epsilon \sigma$?
- $d$-dimensional setting: also learn weight, shoe size, ...
III. Contributions to the Mathematical Theory of Evolution.

By Karl Pearson, University College, London.

Communicated by Professor Henrici, F.R.S.

Received October 18,—Read November 16, 1893.

[Plates 1—5.]

**CONTENTS.**

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>III. — Investigation of an Asymmetrical Frequency-Curve representing Mr. H. Thomson's measurements of the Carapace of Prawns. §§ 16–18</td>
<td>85–90</td>
</tr>
<tr>
<td>Table I. First Six Powers of First Thirty Natural Numbers</td>
<td>90–100</td>
</tr>
<tr>
<td>Table II. Ordinates of Normal Frequency-Curve</td>
<td>100–106</td>
</tr>
<tr>
<td>Note added February 10, 1894</td>
<td>106–107</td>
</tr>
<tr>
<td></td>
<td>107–110</td>
</tr>
</tbody>
</table>
Pearson’s naturalist buddy measured lots of crab body parts.
Pearson’s naturalist buddy measured lots of crab body parts.
Most lengths seemed to follow the “normal” distribution (a recently coined name)
Pearson’s naturalist buddy measured lots of crab body parts.

Most lengths seemed to follow the “normal” distribution (a recently coined name)

But the “forehead” size wasn’t symmetric.
Pearson’s naturalist buddy measured lots of crab body parts.

Most lengths seemed to follow the “normal” distribution (a recently coined name)

But the “forehead” size wasn’t symmetric.

Maybe there were actually two species of crabs?
More previous work

- Pearson 1894: proposed method for 2 Gaussians
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:

- Extended to general $k$ mixtures: Moitra-Valiant '10, Belkin-Sinha '10

Our result: tight upper and lower bounds for the sample complexity.
- For $k=2$ mixtures, arbitrary $d$ dimensions.
- Lower bound extends to larger $k$. 

Moritz Hardt, Eric Price (Google/UT)  
Tight Bounds for Learning a Mixture of Two Gaussians
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05

- Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10

Our result:
- Tight upper and lower bounds for the sample complexity.
  - For $k=2$ mixtures, arbitrary $d$ dimensions.
  - Lower bound extends to larger $k$.

Moritz Hardt, Eric Price (Google/UT)
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
  - Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
  - Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10
- The KMV polynomial is very large.
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
  - Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10
- The KMV polynomial is very large.
  - **Our result**: tight upper and lower bounds for the sample complexity.
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
  - Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10
- The KMV polynomial is very large.
  - **Our result**: tight upper and lower bounds for the sample complexity.
  - For $k = 2$ mixtures, arbitrary $d$ dimensions.
More previous work

- Pearson 1894: proposed method for 2 Gaussians
  - “Method of moments”
- Other empirical papers over the years:
  - Royce ’58, Gridgeman ’70, Gupta-Huang ’80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta ’99, DA ’00
  - Spectral methods: VW ’04, AK ’05, KSV ’05, AM ’05, VW ’05
  - Extended to general $k$ mixtures: Moitra-Valiant ’10, Belkin-Sinha ’10
- The KMV polynomial is very large.
  - **Our result**: tight upper and lower bounds for the sample complexity.
  - For $k = 2$ mixtures, arbitrary $d$ dimensions.
  - Lower bound extends to larger $k$. 

Moritz Hardt, Eric Price (Google/UT)  Tight Bounds for Learning a Mixture of Two Gaussians  2015-06-17 5 / 27
Learning the components vs. learning the sum

- It’s important that we want to learn the individual components:
Learning the components vs. learning the sum

- It’s important that we want to learn the individual components:
  - Male/female average heights, std. deviations.
Learning the components vs. learning the sum

- It’s important that we want to learn the individual components:
  - Male/female average heights, std. deviations.
- Getting $\epsilon$ approximation in TV norm to overall distribution takes $\tilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.
It’s important that we want to learn the individual components:

- Male/female average heights, std. deviations.

Getting $\epsilon$ approximation in TV norm to overall distribution takes $\tilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.

- Quite general: non-properly for any mixture of known unimodal distributions. [Chan, Diakonikolas, Servedio, Sun ’13]
Learning the components vs. learning the sum

- It’s important that we want to learn the individual components:
  - Male/female average heights, std. deviations.
- Getting $\epsilon$ approximation in TV norm to overall distribution takes $\tilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.
  - Quite general: non-properly for any mixture of known unimodal distributions. [Chan, Diakonikolas, Servedio, Sun ’13]
  - Proper learning: [Daskalakis-Kamath ’14]
Learning the components vs. learning the sum

- It’s important that we want to learn the individual components:
  - Male/female average heights, std. deviations.
- Getting $\epsilon$ approximation in TV norm to overall distribution takes $\tilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.
  - Quite general: non-properly for any mixture of known unimodal distributions. [Chan, Diakonikolas, Servedio, Sun ’13]
  - Proper learning: [Daskalakis-Kamath ’14]
  - But only in low dimensions.
It’s important that we want to learn the individual components:
- Male/female average heights, std. deviations.
- Getting $\epsilon$ approximation in TV norm to overall distribution takes $\tilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.
  - Quite general: non-properly for any mixture of known unimodal distributions. [Chan, Diakonikolas, Servedio, Sun ’13]
  - Proper learning: [Daskalakis-Kamath ’14]
  - But only in low dimensions.
  - Generic high-$d$ TV estimation algs use 1d parameter estimation.
Our result

- A variant of Pearson’s 1894 method is optimal!

Suppose we want means and variances to

\[ \epsilon \]

accuracy:

\[ \mu_i \to \pm \epsilon \]

\[ \sigma_i^2 \to \pm \epsilon^2 \sigma_i^2 \]

In one dimension:

\[ \Theta \left( \frac{1}{\epsilon_1^2} \right) \]

samples necessary and sufficient.

Previously: \( \frac{1}{\epsilon} \approx 300 \), no lower bound.

Moreover: algorithm is almost the same as Pearson (1894).

More precisely: if two gaussians are \( \alpha \) standard deviations apart,

getting \( \epsilon \alpha \) precision takes

\[ \Theta \left( \frac{1}{\alpha^2 \epsilon^2} \right) \] samples.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  
  $\mu_i$ to $\pm \epsilon \sigma_i^2$
  $\sigma_i^2$ to $\pm \epsilon^2 \sigma_i^2$

In one dimension: $\Theta(1/\epsilon^2)$ samples necessary and sufficient.

Previously: $1/\epsilon \approx 300$, no lower bound.

Moreover: algorithm is almost the same as Pearson (1894).

More precisely: if two gaussians are $\alpha$ standard deviations apart, getting $\epsilon \alpha$ precision takes $\Theta(1/\alpha^2 \epsilon^2)$ samples.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$

$\alpha$

More precisely: if two gaussians are $\alpha$ standard deviations apart, getting $\epsilon \alpha$ precision takes $\Theta(1/\epsilon \alpha^{1/2})$ samples.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$
  - $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$

In one dimension: $\Theta(1/\epsilon^2)$ samples necessary and sufficient.

Previously: $1/\epsilon \approx 300$, no lower bound.

Moreover: algorithm is almost the same as Pearson (1894).

More precisely: if two gaussians are $\alpha$ standard deviations apart, getting $\epsilon \alpha$ precision takes $\Theta(1/\alpha^2 \epsilon^2)$ samples.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$
  - $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$
- In one dimension: $\Theta(1/\epsilon^{12})$ samples necessary and sufficient.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$
  - $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$
- In one dimension: $\Theta(1/\epsilon^{12})$ samples necessary and sufficient.
  - Previously: $1/\epsilon \approx 300$, no lower bound.
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$
  - $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$
- In one dimension: $\Theta(1/\epsilon^{12})$ samples necessary and sufficient.
  - Previously: $1/\epsilon \approx 300$, no lower bound.
  - Moreover: algorithm is almost the same as Pearson (1894).
Our result

- A variant of Pearson’s 1894 method is optimal!
- Suppose we want means and variances to $\epsilon$ accuracy:
  - $\mu_i$ to $\pm \epsilon \sigma$
  - $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$
- In one dimension: $\Theta(1/\epsilon^{12})$ samples necessary and sufficient.
  - Previously: $1/\epsilon \approx 300$, no lower bound.
  - Moreover: algorithm is almost the same as Pearson (1894).

More precisely: if two gaussians are $\alpha$ standard deviations apart, getting $\epsilon \alpha$ precision takes $\Theta(\frac{1}{\alpha^{12} \epsilon^2})$ samples.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for *parameter distance*.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - “$\sigma^2$” is max variance in any coordinate.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - “$\sigma^2$” is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$. 

Caveat: assume $p_1, p_2$ are bounded away from zero throughout.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - “$\sigma^2$” is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - "$\sigma^2$" is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.
- Also gives an improved bound in TV error of each component:
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - $\sigma^2$ is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.
- Also gives an improved bound in TV error of each component:
  - If components overlap, then parameter distance $\approx$ TV.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - $\sigma^2$ is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.
- Also gives an improved bound in TV error of each component:
  - If components overlap, then parameter distance $\approx$ TV.
  - If components don’t overlap, then clustering is trivial.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12}\log d)$ samples for parameter distance.
  - "$\sigma^2$" is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm\epsilon^2\sigma^2$.
  - Useful when covariance matrix is sparse.
- Also gives an improved bound in TV error of each component:
  - If components overlap, then parameter distance $\approx$ TV.
  - If components don’t overlap, then clustering is trivial.
  - Straightforwardly gives $\tilde{O}(d^{30}/\epsilon^{36})$ samples.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - “$\sigma^2$” is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.
- Also gives an improved bound in TV error of each component:
  - If components overlap, then parameter distance $\approx$ TV.
  - If components don’t overlap, then clustering is trivial.
  - Straightforwardly gives $\tilde{O}(d^{30}/\epsilon^{36})$ samples.
  - Best known, but not the $\tilde{O}(d/\epsilon^c)$ we want.

Caveat: assume $p_1, p_2$ are bounded away from zero throughout.
Our result: higher dimensions

- In $d$ dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples for parameter distance.
  - "$\sigma^2$" is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm \epsilon^2 \sigma^2$.
  - Useful when covariance matrix is sparse.

- Also gives an improved bound in TV error of each component:
  - If components overlap, then parameter distance $\approx$ TV.
  - If components don’t overlap, then clustering is trivial.
  - Straightforwardly gives $\tilde{O}(d^{30}/\epsilon^{36})$ samples.
  - Best known, but not the $\tilde{O}(d/\epsilon^c)$ we want.

- Caveat: assume $p_1, p_2$ are bounded away from zero throughout.
Outline

1. Algorithm in One Dimension
Outline

1 Algorithm in One Dimension

2 Lower Bound
Outline

1. Algorithm in One Dimension

2. Lower Bound

3. Algorithm in $d$ Dimensions
Outline

1. Algorithm in One Dimension
2. Lower Bound
3. Algorithm in $d$ Dimensions
We want to learn five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, p_1, p_2$ with $p_1 + p_2 = 1$. 
We want to learn five parameters: \( \mu_1, \mu_2, \sigma_1, \sigma_2, p_1, p_2 \) with \( p_1 + p_2 = 1 \).

Moments give polynomial equations in parameters:

\[
M_1 := \mathbb{E}[x^1] = p_1 \mu_1 + p_2 \mu_2
\]
\[
M_2 := \mathbb{E}[x^2] = p_1 \mu_1^2 + p_2 \mu_2^2 + p_1 \sigma_1^2 + p_2 \sigma_2^2
\]
\[
M_3, M_4, M_5, M_6 = [...]
\]
We want to learn five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, p_1, p_2$ with $p_1 + p_2 = 1$.

Moments give polynomial equations in parameters:

- $M_1 := \mathbb{E}[x^1] = p_1 \mu_1 + p_2 \mu_2$
- $M_2 := \mathbb{E}[x^2] = p_1 \mu_1^2 + p_2 \mu_2^2 + p_1 \sigma_1^2 + p_2 \sigma_2^2$
- $M_3, M_4, M_5, M_6 = [...]$

Use our samples to estimate the moments.
We want to learn five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, p_1, p_2$ with $p_1 + p_2 = 1$.

Moments give polynomial equations in parameters:

\[
M_1 := \mathbb{E}[x^1] = p_1 \mu_1 + p_2 \mu_2
\]
\[
M_2 := \mathbb{E}[x^2] = p_1 \mu_1^2 + p_2 \mu_2^2 + p_1 \sigma_1^2 + p_2 \sigma_2^2
\]

$M_3, M_4, M_5, M_6 = [...]$

Use our samples to estimate the moments.

Solve the system of equations to find the parameters.
Method of Moments

Solving the system

- Start with five parameters.
Method of Moments

Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”

\[ M'_2 = M_2 - M_1 \] is independent of translation.

Analogously, can assume \( \min(\sigma_1, \sigma_2) = 0 \) by converting to “excess moments”

\[ X_4 = M_4 - 3M_2^2 \] is independent of adding \( N(0, \sigma^2) \).

“Excess kurtosis” coined by Pearson, appearing in every Wikipedia probability distribution infobox.

Leaves three free parameters.
Method of Moments
Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”
  - $M'_2 = M_2 - M_1^2$ is independent of translation.
Method of Moments

Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”
  - $M'_2 = M_2 - M_1^2$ is independent of translation.
- Analogously, can assume $\min(\sigma_1, \sigma_2) = 0$ by converting to “excess moments”
Method of Moments

Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”
  - $M'_2 = M_2 - M_1^2$ is independent of translation.
- Analogously, can assume $\min(\sigma_1, \sigma_2) = 0$ by converting to “excess moments”
  - $X_4 = M_4 - 3M_2^2$ is independent of adding $N(0, \sigma^2)$. 
Method of Moments

Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”
  - $M'_2 = M_2 - M_1^2$ is independent of translation.
- Analogously, can assume $\min(\sigma_1, \sigma_2) = 0$ by converting to “excess moments”
  - $X_4 = M_4 - 3M_2^2$ is independent of adding $N(0, \sigma^2)$.
  - “Excess kurtosis” coined by Pearson, appearing in every Wikipedia probability distribution infobox.
Method of Moments
Solving the system

- Start with five parameters.
- First, can assume mean zero:
  - Convert to “central moments”
  - $M'_2 = M_2 - M_1^2$ is independent of translation.
- Analogously, can assume $\min(\sigma_1, \sigma_2) = 0$ by converting to “excess moments”
  - $X_4 = M_4 - 3M_2^2$ is independent of adding $N(0, \sigma^2)$.
  - “Excess kurtosis” coined by Pearson, appearing in every Wikipedia probability distribution infobox.
- Leaves three free parameters.
Method of Moments: system of equations

- Convenient to reparameterize by

\[ \alpha = -\mu_1\mu_2, \beta = \mu_1 + \mu_2, \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1} \]
Method of Moments: system of equations

- Convenient to reparameterize by

\[ \alpha = -\mu_1 \mu_2, \beta = \mu_1 + \mu_2, \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1} \]

- Gives that

\[ X_3 = \alpha(\beta + 3\gamma) \]
\[ X_4 = \alpha(-2\alpha + \beta^2 + 6\beta\gamma + 3\gamma^2) \]
\[ X_5 = \alpha(\beta^3 - 8\alpha\beta + 10\beta^2\gamma + 15\gamma^2\beta - 20\alpha\gamma) \]
\[ X_6 = \alpha(16\alpha^2 - 12\alpha\beta^2 - 60\alpha\beta\gamma + \beta^4 + 15\beta^3\gamma + 45\beta^2\gamma^2 + 15\beta\gamma^3) \]
Method of Moments: system of equations

- Convenient to reparameterize by

\[ \alpha = -\mu_1 \mu_2, \quad \beta = \mu_1 + \mu_2, \quad \gamma = \frac{\sigma_2^2 - \sigma_1^2}{\mu_2 - \mu_1} \]

- Gives that

\[ X_3 = \alpha(\beta + 3\gamma) \]
\[ X_4 = \alpha(-2\alpha + \beta^2 + 6\beta \gamma + 3\gamma^2) \]
\[ X_5 = \alpha(\beta^3 - 8\alpha \beta + 10\beta^2 \gamma + 15\gamma^2 \beta - 20\alpha \gamma) \]
\[ X_6 = \alpha(16\alpha^2 - 12\alpha \beta^2 - 60\alpha \beta \gamma + \beta^4 + 15\beta^3 \gamma + 45\beta^2 \gamma^2 + 15\beta \gamma^3) \]

All my attempts to obtain a simpler set have failed... It is possible, however, that some other ... equations of a less complex kind may ultimately be found.

—Karl Pearson
Pearson’s Polynomial

- Chug chug chug...
Pearson’s Polynomial

- Chug chug chug...
- Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$:

$$p(\alpha) = 8\alpha^9 + 28X_4\alpha^7 - 12X_3^2\alpha^6 + (24X_3X_5 + 30X_4^2)\alpha^5$$
$$+ (6X_5^2 - 148X_3^2X_4)\alpha^4 + (96X_3^4 - 36X_3X_4X_5 + 9X_4^3)\alpha^3$$
$$+ (24X_3^3X_5 + 21X_3^2X_4^2)\alpha^2 - 32X_3^4X_4\alpha + 8X_3^6$$
$$= 0$$
Pearson’s Polynomial

- Chug chug chug...
- Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$:
  \[
  p(\alpha) = 8\alpha^9 + 28X_4\alpha^7 - 12X_3^2\alpha^6 + (24X_3X_5 + 30X_4^2)\alpha^5 \\
  + (6X_5^2 - 148X_3^2X_4)\alpha^4 + (96X_3^4 - 36X_3X_4X_5 + 9X_4^3)\alpha^3 \\
  + (24X_3^3X_5 + 21X_3^2X_4^2)\alpha^2 - 32X_3^4X_4\alpha + 8X_3^6 \\
  = 0
  \]
- Easy to go from solutions $\alpha = -\mu_1\mu_2$ to mixtures $\mu_i, \sigma_i, p_i$. 

Moritz Hardt, Eric Price  (Google/UT)  
Tight Bounds for Learning a Mixture of Two Gaussians  
2015-06-17 14 / 27
Get a 9th degree polynomial in the excess moments $X_3$, $X_4$, $X_5$. 
Pearson’s Polynomial

- Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.
  - Positive roots correspond to mixtures that match on five moments.
Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.

- Positive roots correspond to mixtures that match on five moments.
- Pearson’s proposal: choose root with closer 6th moment.
Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.  
- Positive roots correspond to mixtures that match on five moments.
- Pearson’s proposal: choose root with closer 6th moment.

Works because six moments uniquely identify mixture [KMV]
Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.
- Positive roots correspond to mixtures that match on five moments.
- Pearson’s proposal: choose root with closer 6th moment.

Works because six moments uniquely identify mixture [KMV]

How robust to moment estimation error?
Pearson’s Polynomial

- Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.
  - Positive roots correspond to mixtures that match on five moments.
  - Pearson’s proposal: choose root with closer 6th moment.
- Works because six moments uniquely identify mixture [KMV]
- How robust to moment estimation error?
  - Usually works well
Get a 9th degree polynomial in the excess moments $X_3, X_4, X_5$.
- Positive roots correspond to mixtures that match on five moments.
- Pearson’s proposal: choose root with closer 6th moment.

Works because six moments uniquely identify mixture [KMV]

How robust to moment estimation error?
- Usually works well
- Not when there’s a double root.
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$. 

Then $\alpha$ is the unique positive root of $r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0$.

How robust is the solution to perturbations of $X_3, \ldots, X_6$?

We know $q(x) := r/(x - \alpha)^2$ has no positive roots.

By compactness: $q(x) \geq c > 0$ for some constant $c$.

Therefore plugging in empirical moments $\tilde{X}_i$ to estimate polynomials $p_5, p_6$ is robust:

$|\tilde{p}_5 - p_5|, |\tilde{p}_6 - p_6| \leq \epsilon$.

Getting $\alpha$ lets us estimate means, variances.

Moritz Hardt, Eric Price (Google/UT)  
Tight Bounds for Learning a Mixture of Two Gaussians  
2015-06-17 16 / 27
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the *unique* positive root of

$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the *unique* positive root of
  \[ r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0. \]
- How robust is the solution to perturbations of $X_3, \ldots, X_6$?

\[ \text{Given approximations} \quad \left| \tilde{p}_5 - p_5 \right|, \left| \tilde{p}_6 - p_6 \right| \leq \epsilon, \]
\[ \left| \alpha - \arg \min \tilde{r}(x) \right| \lesssim \epsilon. \]
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the unique positive root of

$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$ 

- How robust is the solution to perturbations of $X_3, \ldots, X_6$?
- We know $q(x) := r/(x - \alpha)^2$ has no positive roots.
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the *unique* positive root of

$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$  

- How robust is the solution to perturbations of $X_3, \ldots, X_6$?
- We know $q(x) := r/(x - \alpha)^2$ has no positive roots.
- By compactness: $q(x) \geq c > 0$ for some constant $c$. 

---

Given approximations $|\tilde{p}_5 - p_5|, |\tilde{p}_6 - p_6| \leq \epsilon$,

$$|\alpha - \arg \min \tilde{r}(x)| \lesssim \epsilon.$$  

---

Getting $\alpha$ lets us estimate means, variances.
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the *unique* positive root of

$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$ 

- How robust is the solution to perturbations of $X_3, \ldots, X_6$?
- We know $q(x) := r/(x - \alpha)^2$ has no positive roots.
- By compactness: $q(x) \geq c > 0$ for some constant $c$.
- Therefore plugging in empirical moments $\tilde{X}_i$ to estimate polynomials $p_5, p_6$ is robust:
Making it robust in all cases

- Can create another ninth degree polynomial $p_6$ from $X_3, X_4, X_5, X_6$.
- Then $\alpha$ is the *unique* positive root of

$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$  

- How robust is the solution to perturbations of $X_3, \ldots, X_6$?
- We know $q(x) := r/(x - \alpha)^2$ has no positive roots.
- By compactness: $q(x) \geq c > 0$ for some constant $c$.
- Therefore plugging in empirical moments $\tilde{X}_i$ to estimate polynomials $p_5, p_6$ is robust:
  - Given approximations $|\tilde{p}_5 - p_5|, |\tilde{p}_6 - p_6| \leq \epsilon$,

$$|\alpha - \arg \min \tilde{r}(x)| \lesssim \epsilon.$$
Making it robust in all cases

- Can create another ninth degree polynomial \( p_6 \) from \( X_3, X_4, X_5, X_6 \).
- Then \( \alpha \) is the unique positive root of

\[
r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.
\]

- How robust is the solution to perturbations of \( X_3, \ldots, X_6 \)?
- We know \( q(x) := r/(x - \alpha)^2 \) has no positive roots.
- By compactness: \( q(x) \geq c > 0 \) for some constant \( c \).
- Therefore plugging in empirical moments \( \tilde{X}_i \) to estimate polynomials \( p_5, p_6 \) is robust:
  - Given approximations \( |\tilde{p}_5 - p_5|, |\tilde{p}_6 - p_6| \leq \epsilon \),
    \[
    |\alpha - \arg \min \tilde{r}(x)| \lesssim \epsilon.
    \]
  - Getting \( \alpha \) lets us estimate means, variances.
Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.
Result

- Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.
- Getting the $\tilde{p}_i$ to $O(\epsilon)$ requires getting the first six moments to $\pm O(\epsilon)$.
Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.

Getting the $\tilde{p}_i$ to $O(\epsilon)$ requires getting the first six moments to $\pm O(\epsilon)$.

If the variance is $\sigma^2$, then $M_i$ has variance $O(\sigma^{2i})$. 
Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.

Getting the $\tilde{p}_i$ to $O(\epsilon)$ requires getting the first six moments to $\pm O(\epsilon)$.

If the variance is $\sigma^2$, then $M_i$ has variance $O(\sigma^{2i})$.

Thus $O(\sigma^{12}/\epsilon^2)$ samples to learn the $\mu_i$ to $\pm \epsilon$. 
• Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.
• Getting the $\tilde{p}_i$ to $O(\epsilon)$ requires getting the first six moments to $\pm O(\epsilon)$.
• If the variance is $\sigma^2$, then $M_i$ has variance $O(\sigma^{2i})$.
• Thus $O(\sigma^{12}/\epsilon^2)$ samples to learn the $\mu_i$ to $\pm \epsilon$.
  ▶ If components are $\Omega(1)$ standard deviations apart, $O(1/\epsilon^2)$ samples suffice.
Result

- Scale so the excess moments are $O(1)$: $\mu_i$ are $\pm O(1)$.
- Getting the $\tilde{p}_i$ to $O(\epsilon)$ requires getting the first six moments to $\pm O(\epsilon)$.
- If the variance is $\sigma^2$, then $M_i$ has variance $O(\sigma^{2i})$.
- Thus $O(\sigma^{12}/\epsilon^2)$ samples to learn the $\mu_i$ to $\pm \epsilon$.
  - If components are $\Omega(1)$ standard deviations apart, $O(1/\epsilon^2)$ samples suffice.
  - In general, $O(1/\epsilon^{12})$ samples suffice to get $\epsilon \sigma$ accuracy.
Outline

1. Algorithm in One Dimension

2. Lower Bound

3. Algorithm in $d$ Dimensions
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments.
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$. 
Lower bound in one dimension

The algorithm takes \( O(\epsilon^{-12}) \) samples because it uses six moments
  
  - Necessary to get sixth moment to \( \pm (\epsilon \sigma)^6 \).

Let \( F, F' \) be any two mixtures with five matching moments:

- Constant means and variances.
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm(\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 

![Graphs showing the distributions of two mixtures](image-url)
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm(\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 

---

Let $F, F'$ be any two mixtures with five matching moments:

- Constant means and variances.
- Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 

Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 

![Graph](image-url)
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$. 
Lower bound in one dimension

- The algorithm takes $O(\epsilon^{-12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm (\epsilon \sigma)^6$.
- Let $F, F'$ be any two mixtures with five matching moments:
  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture for growing $\sigma$.
- Claim: $\Omega(\sigma^{12})$ samples necessary to distinguish the distributions.
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$. 

\[
\int \left( \sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx
\]
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
- Improve using squared Hellinger distance.
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
- Improve using squared Hellinger distance.
  - $H^2(P, Q) := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$

$⇒ H^2 \approx \text{TV} \approx H$, but often $H \approx TV$. 

Moritz Hardt, Eric Price (Google/UT)  Tight Bounds for Learning a Mixture of Two Gaussians  2015-06-17  20 / 27
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
- Improve using *squared Hellinger distance*.
  - $H^2(P, Q) := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$
  - $H^2$ is subadditive on product measures:
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
- Improve using squared Hellinger distance.
  - $H^2(P, Q) := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$
  - $H^2$ is subadditive on product measures:
    - $H^2((x_1, \ldots, x_m), (x'_1, \ldots, x'_m)) \leq m H^2(x, x')$. 
Lower bound in one dimension

- Two mixtures $F, F'$ with $F \approx F'$.
- Have $\text{TV}(F, F') \approx 1/\sigma^6$.
- Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.
- Improve using squared Hellinger distance.
  - $H^2(P, Q) := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$
  - $H^2$ is subadditive on product measures:
    - $H^2((x_1, \ldots, x_m), (x'_1, \ldots, x'_m)) \leq mH^2(x, x')$.
  - Sample complexity is $\Omega(1/H^2(F, F'))$
Lower bound in one dimension

Two mixtures $F, F'$ with $F \approx F'$.

Have $\text{TV}(F, F') \approx 1/\sigma^6$.

Shows $\Omega(\sigma^6)$ samples, $O(\sigma^{12})$ samples.

Improve using squared Hellinger distance.

\[
H^2(P, Q) := \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx
\]

$H^2$ is subadditive on product measures:

\[
H^2((x_1, \ldots, x_m), (x'_1, \ldots, x'_m)) \leq mH^2(x, x').
\]

Sample complexity is $\Omega(1/H^2(F, F'))$

$H^2 \lapprox TV \lapprox H$, but often $H \approx TV$. 

Bounding the Hellinger distance: general idea

Definition

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \]
Bounding the Hellinger distance: general idea

Definition

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx \]
Bounding the Hellinger distance: general idea

**Definition**

\[
H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx = 1 - \int \sqrt{p(x)q(x)} \, dx
\]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]
Bounding the Hellinger distance: general idea

**Definition**

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx \]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[ H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)p(x)} dx \]
Bounding the Hellinger distance: general idea

**Definition**

\[
H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx
\]

- If \(q(x) = (1 + \Delta(x))p(x)\) for some small \(\Delta\), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)p(x)} dx
\]

\[
= 1 - \mathbb{E}_{x \sim p} [\sqrt{1 + \Delta(x)}]
\]
Bounding the Hellinger distance: general idea

Definition

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx \]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)}p(x) dx
= 1 - \mathbb{E}_{x \sim p} [\sqrt{1 + \Delta(x)}]
= 1 - \mathbb{E}_{x \sim p} [1 + \Delta(x)/2 - O(\Delta^2(x))]\]
Bounding the Hellinger distance: general idea

Definition

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx \]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)p(x)} dx \\
= 1 - \mathbb{E}_{x \sim p} [\sqrt{1 + \Delta(x)}] \\
= 1 - \mathbb{E}_{x \sim p} [1 + \Delta(x)/2 - O(\Delta^2(x))] 
\]
Bounding the Hellinger distance: general idea

**Definition**

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx = 1 - \int \sqrt{p(x)q(x)} \, dx \]

If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)}p(x) \, dx \\
= 1 - \mathbb{E}_{x \sim p}[\sqrt{1 + \Delta(x)}] \\
= 1 - \mathbb{E}_{x \sim p}[1 + \Delta(x)/2 - O(\Delta^2(x))] \\
\int q(x)-p(x)=0
\]
Bounding the Hellinger distance: general idea

**Definition**

\[
H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = 1 - \int \sqrt{p(x)q(x)} dx
\]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)}p(x) dx \\
= 1 - \mathbb{E}_{x \sim p}[\sqrt{1 + \Delta(x)}] \\
= 1 - \mathbb{E}_{x \sim p}[1 + \frac{\Delta(x)}{2} - O(\Delta^2(x))] \\
\leq \mathbb{E}_{x \sim p}[\Delta^2(x)]
\]
Bounding the Hellinger distance: general idea

**Definition**

\[ H^2(P, Q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \, dx = 1 - \int \sqrt{p(x)q(x)} \, dx \]

- If \( q(x) = (1 + \Delta(x))p(x) \) for some small \( \Delta \), then [Pollard ’00]

\[
H^2(p, q) = 1 - \int \sqrt{1 + \Delta(x)}p(x) \, dx
= 1 - \mathbb{E}_{x \sim p}[\sqrt{1 + \Delta(x)}]
= 1 - \mathbb{E}_{x \sim p}[1 + \Delta(x)/2 - O(\Delta^2(x))]
\]

\[
\lesssim \mathbb{E}_{x \sim p}[\Delta^2(x)]
\]

- Compare to \( TV(p, q) = \frac{1}{2} \mathbb{E}_{x \sim p}[|\Delta(x)|] \)
Bounding the Hellinger distance: our setting

**Lemma**

Let \( F, F' \) be two subgaussian distributions with \( k \) matching moments and constant parameters. Then for \( G, G' = F + N(0, \sigma^2), F' + N(0, \sigma^2) \),

\[
H^2(G, G') \lesssim \frac{1}{\sigma^{2k+2}}.
\]
Bounding the Hellinger distance: our setting

Lemma

Let $F, F'$ be two subgaussian distributions with $k$ matching moments and constant parameters. Then for $G, G' = F + N(0, \sigma^2), F' + N(0, \sigma^2)$, $H^2(G, G') \lesssim 1/\sigma^{2k+2}$.

- Power series expansion of $\mathbb{E}[\Delta^2] = \mathbb{E} \left[ \left( \frac{G'(x) - G(x)}{G(x)} \right)^2 \right]$. 

Bounding the Hellinger distance: our setting

Lemma

Let $F, F'$ be two subgaussian distributions with $k$ matching moments and constant parameters. Then for $G, G' = F + N(0, \sigma^2), F' + N(0, \sigma^2)$,

$$H^2(G, G') \lesssim 1/\sigma^{2k+2}.$$  

- Power series expansion of $\mathbb{E}[\Delta^2] = \mathbb{E} \left[ \left( \frac{G'(x) - G(x)}{G(x)} \right)^2 \right]$.

- Matching moments make the first $k$ terms zero.
Bounding the Hellinger distance: our setting

Lemma

Let $F, F'$ be two subgaussian distributions with $k$ matching moments and constant parameters. Then for $G, G' = F + N(0, \sigma^2), F' + N(0, \sigma^2)$,

$$H^2(G, G') \lesssim \frac{1}{\sigma^{2k+2}}.$$ 

- Power series expansion of $\mathbb{E}[\Delta^2] = \mathbb{E} \left[ \left( \frac{G'(x) - G(x)}{G(x)} \right)^2 \right].$
- Matching moments make the first $k$ terms zero.
- Leaves $(1/\sigma^{k+1})^2$ as largest remaining term.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

For $G = \frac{1}{2} N(-\frac{1}{2}, \frac{1}{2} + \sigma^2) + \frac{1}{2} N(\frac{1}{2}, 2 + \sigma^2)$

$G' \approx 0.297 N(-1.226, 0.610 + \sigma^2) + 0.703 N(0.517, 2.396 + \sigma^2)$

$H_2(G, G') \ll \frac{1}{\sigma_{12}}$

Therefore distinguishing $G$ from $G'$ takes $\Omega(\sigma_{12})$ samples.

Cannot learn either means to $\pm \epsilon\sigma$ or variance to $\pm \epsilon^2\sigma^2$ with $o(\frac{1}{\epsilon^{12}})$ samples.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

\[ G = \frac{1}{2} N(-1, 1 + \sigma^2) + \frac{1}{2} N(1, 2 + \sigma^2) \]

\[ G' \approx 0.297 N(-1.226, 0.610 + \sigma^2) + 0.703 N(0.517, 2.396 + \sigma^2) \]

Have $H_2(G, G') \lessapprox 1/\sigma^{12}$.

Therefore distinguishing $G$ from $G'$ takes $\Omega(\sigma^{12})$ samples.

Cannot learn either means to $\pm \epsilon \sigma$ or variance to $\pm \epsilon^2 \sigma^2$ with $o(1/\epsilon^{12})$ samples.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.


Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

For

$$G = \frac{1}{2} N(-1, 1 + \sigma^2) + \frac{1}{2} N(1, 2 + \sigma^2)$$

$$G' \approx 0.297 N(-1.226, 0.610 + \sigma^2) + 0.703 N(0.517, 2.396 + \sigma^2)$$

have $H^2(G, G') \lesssim 1/\sigma^{12}$. 

Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

For

$$G = \frac{1}{2} N(-1, 1 + \sigma^2) + \frac{1}{2} N(1, 2 + \sigma^2)$$

$$G' \approx 0.297 N(-1.226, 0.610 + \sigma^2) + 0.703 N(0.517, 2.396 + \sigma^2)$$

have $H^2(G, G') \lesssim 1/\sigma^{12}$.
- Therefore distinguishing $G$ from $G'$ takes $\Omega(\sigma^{12})$ samples.
Lower bound in one dimension

- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

\[ G = \frac{1}{2} N(-1, 1 + \sigma^2) + \frac{1}{2} N(1, 2 + \sigma^2) \]

\[ G' \approx 0.297 N(-1.226, 0.610 + \sigma^2) + 0.703 N(0.517, 2.396 + \sigma^2) \]

have $H^2(G, G') \lesssim 1/\sigma^{12}$.

Therefore distinguishing $G$ from $G'$ takes $\Omega(\sigma^{12})$ samples.

Cannot learn either means to $\pm \epsilon \sigma$ or variance to $\pm \epsilon^2 \sigma^2$ with $o(1/\epsilon^{12})$ samples.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
- Solution must solve all $d$ instances.

Each instance has Hellinger distance $O\left(\epsilon^{1/2}\right)$. Therefore $\Omega\left(\epsilon^{-1/2} \log \left(\frac{d}{\delta}\right)\right)$ samples are necessary to succeed with probability $1 - \delta$:

▶ Each set of $\epsilon^{-1/2}$ samples has a constant chance of giving no information about each coordinate.
▶ With $o\left(\epsilon^{-1/2} \log d\right)$ samples, some coordinate will be independent of all the samples.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
- Solution must solve all $d$ instances.
- Each instance has Hellinger distance $O(\epsilon^{12})$. 

Therefore $\Omega(\epsilon^{-12} \log (d/\delta))$ samples are necessary to succeed with probability $1 - \delta$. 

- Each set of $\epsilon^{-12}$ samples has a constant chance of giving no information about each coordinate.
- With $o(\epsilon^{-12} \log d)$ samples, some coordinate will be independent of all the samples.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
- Solution must solve all $d$ instances.
- Each instance has Hellinger distance $O(\epsilon^{12})$.

Therefore $\Omega(\epsilon^{-12} \log(d/\delta))$ samples are necessary to succeed with probability $1 - \delta$:
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
- Solution must solve all $d$ instances.
- Each instance has Hellinger distance $O(\epsilon^{12})$.

Therefore $\Omega(\epsilon^{-12} \log(d/\delta))$ samples are necessary to succeed with probability $1 - \delta$:

- Each set of $\epsilon^{-12}$ samples has a constant chance of giving no information about each coordinate.
Lower bound in $d$ dimensions

- Trivial based on the Hellinger distance bound.
- Place the “hard” instance independently in all $d$ coordinates.
- Solution must solve all $d$ instances.
- Each instance has Hellinger distance $O(\epsilon^{12})$.
- Therefore $\Omega(\epsilon^{-12} \log(d/\delta))$ samples are necessary to succeed with probability $1 - \delta$:
  - Each set of $\epsilon^{-12}$ samples has a constant chance of giving no information about each coordinate.
  - With $o(\epsilon^{-12} \log d)$ samples, some coordinate will be independent of all the samples.
Outline

1. Algorithm in One Dimension
2. Lower Bound
3. Algorithm in $d$ Dimensions
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)

...
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$: 

  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$.
  - $(\mu_i, \mu_j)$ usually has a significantly different projection from $(\mu_i, \mu'_j)$.
  - Thus we can piece them together by solving the $O(d^2)$ one-dimensional problems.
  - For covariances: reduce to $d = 4$, so $O(d^4)$ one-dimensional problems.
  - Only loss is $\log(1/\delta) \to \log(d/\delta)$:
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu'_j$?
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu'_j$?
  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$. 
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu'_j$?
  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$.
  - $(\mu_i, \mu_j)$ usually has a significantly different projection from $(\mu_i, \mu'_j)$.
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu'_j$?
  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$.
  - $(\mu_i, \mu_j)$ usually has a significantly different projection from $(\mu_i, \mu'_j)$.
- Thus we can piece them together by solving the $O(d^2)$ one dimensional problems.
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu'_j$?
  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$.
  - $(\mu_i, \mu_j)$ usually has a significantly different projection from $(\mu_i, \mu'_j)$.

Thus we can piece them together by solving the $O(d^2)$ one dimensional problems.

For covariances: reduce to $d = 4$, so $O(d^4)$ one dimensional problems.
Algorithm in $d$ dimensions

- Want to learn average male/female height, weight, shoe size, ...
  - (And covariance matrix)
- Look at individual attributes to get all these.
- Just need to know: is the taller group also heavier or lighter?
- Suffices to consider $d = 2$:
  - Does $\mu_i$ go with $\mu_j$ or $\mu_j'$?
  - Project onto a random direction $e_i \sin \theta + e_j \cos \theta$.
  - $(\mu_i, \mu_j)$ usually has a significantly different projection from $(\mu_i, \mu_j')$.
- Thus we can piece them together by solving the $O(d^2)$ one dimensional problems.
- For covariances: reduce to $d = 4$, so $O(d^4)$ one dimensional problems.
- Only loss is $\log(1/\delta) \rightarrow \log(d/\delta)$:
  $$\Theta(1/\epsilon^{12} \log(d/\delta))$$ samples
Recap and open questions

Our result:

- $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
Recap and open questions

Our result:

- $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
- If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.
Recap and open questions

- **Our result:**
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.
- **Extend to $k > 2$?**
Recap and open questions

- **Our result:**
  - \( \Theta(\epsilon^{-12} \log d) \) samples necessary and sufficient to estimate \( \mu_i \) to within \( \pm \epsilon \sigma \), \( \sigma_i^2 \) to within \( \pm \epsilon^2 \sigma^2 \).
  - If the means have \( \alpha \sigma \) separation, just \( O(\epsilon^{-2} \alpha^{-12}) \) for \( \epsilon \alpha \sigma \) accuracy.

- **Extend to \( k > 2 \)?**
  - Lower bound extends, at least to \( \Omega(\epsilon^{-6k-2}) \).
Recap and open questions

- Our result:
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.

- Extend to $k > 2$?
  - Lower bound extends, at least to $\Omega(\epsilon^{-6k-2})$.
  - Do we really care about finding an $O(\epsilon^{-22})$ algorithm?
Recap and open questions

- **Our result:**
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.

- **Extend to $k > 2$?**
  - Lower bound extends, at least to $\Omega(\epsilon^{-6k^{-2}})$.
  - Do we really care about finding an $O(\epsilon^{-22})$ algorithm?
  - Solving the system of equations gets nasty.
Recap and open questions

- Our result:
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.

- Extend to $k > 2$?
  - Lower bound extends, at least to $\Omega(\epsilon^{-6k-2})$.
  - Do we really care about finding an $O(\epsilon^{-22})$ algorithm?
  - Solving the system of equations gets nasty.
  - [Next talk: Ge-Huang-Kakade avoid this for *smoothed* instances]
Recap and open questions

- Our result:
  - $\Theta(\epsilon^{−12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.

- Extend to $k > 2$?
  - Lower bound extends, at least to $\Omega(\epsilon^{-6k−2})$.
  - Do we really care about finding an $O(\epsilon^{-22})$ algorithm?
  - Solving the system of equations gets nasty.
  - [Next talk: Ge-Huang-Kakade avoid this for smoothed instances]

- Automated way of figuring out whether solution to system of polynomial equations is robust?
Recap and open questions

- Our result:
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate $\mu_i$ to $\pm \epsilon \sigma$, $\sigma_i^2$ to $\pm \epsilon^2 \sigma^2$.
  - If the means have $\alpha \sigma$ separation, just $O(\epsilon^{-2} \alpha^{-12})$ for $\epsilon \alpha \sigma$ accuracy.

- Extend to $k > 2$?
  - Lower bound extends, at least to $\Omega(\epsilon^{-6k-2})$.
  - Do we really care about finding an $O(\epsilon^{-22})$ algorithm?
  - Solving the system of equations gets nasty.
  - [Next talk: Ge-Huang-Kakade avoid this for smoothed instances]

- Automated way of figuring out whether solution to system of polynomial equations is robust?

- TV estimation in $d$ dimensions with $d/\epsilon^c$ rather than $d^{30}/\epsilon^c$?