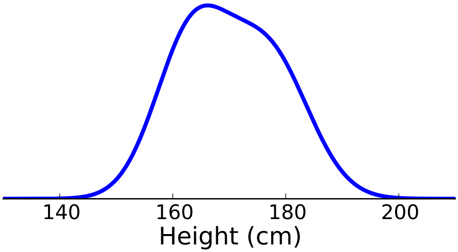
Sharp bounds for learning a mixture of two Gaussians

Moritz Hardt **Eric Price**

IBM Almaden

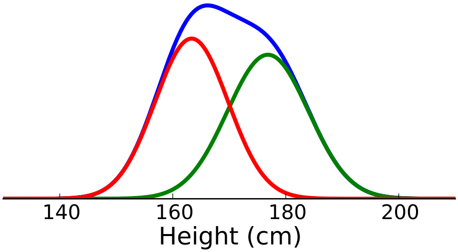
2014-05-28

Problem



- Height distribution of American 20 year olds.

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 μ_1, μ_2 to $\pm\epsilon\sigma$?

Gaussian Mixtures: Origins

III. *Contributions to the Mathematical Theory of Evolution.*

By KARL PEARSON, *University College, London.*

Communicated by Professor HENRICI, *F.R.S.*

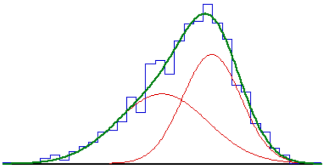
Received October 18,—Read November 16, 1893.

[PLATES 1—5.]

CONTENTS.

Gaussian Mixtures: Origins

Contributions to the Mathematical Theory of Evolution, Karl Pearson, 1894

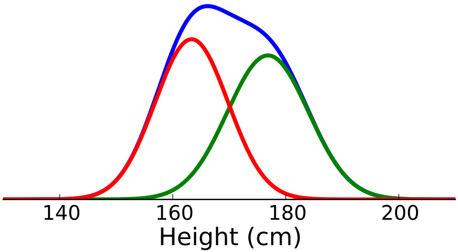


- Pearson's naturalist buddy measured lots of crab body parts.
- Most lengths seemed to follow the "normal" distribution (a recently coined name)
- But the "forehead" size wasn't symmetric.
- Maybe there were actually two species of crabs?

More previous work

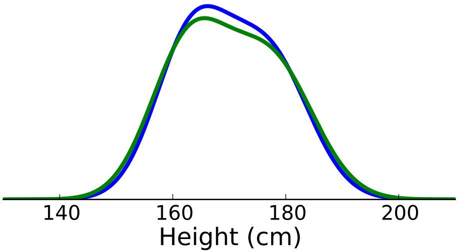
- Pearson 1894: proposed method for 2 Gaussians
  - "Method of moments"
- Other empirical papers over the years:
  - Royce '58, Gridgeman '70, Gupta-Huang '80
- Provable results assuming the components are well-separated:
  - Clustering: Dasgupta '99, DA '00
  - Spectral methods: VW '04, AK '05, KSV '05, AM '05, VW '05
- Kalai-Moitra-Valiant 2010: first general polynomial bound.
  - 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.

Learning the components vs. learning the sum



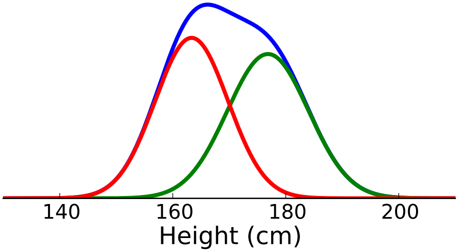
- It's important that we want to learn the individual components:

Learning the components vs. learning the sum



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

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 ϵ approximation in TV norm to overall distribution takes $\widetilde{\Theta}(1/\epsilon^2)$ samples from black box techniques.
  - Quite general: for any mixture of known unimodal distributions. [Chan, Diakonikolas, Servedio, Sun '13]

We show

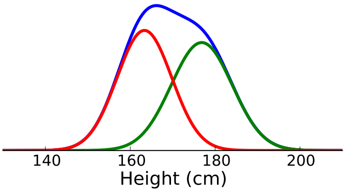
- Pearson's 1894 method can be extended to be optimal!
- Suppose we want means and variances to ϵ accuracy:
  - μ_i to $\pm\epsilon\sigma$
  - σ_i^2 to $\pm\epsilon^2\sigma^2$
- In one dimension: $\Theta(1/\epsilon^{12})$ samples *necessary* and *sufficient*.
  - Previously: $O(1/\epsilon^{300})$.
  - Moreover: algorithm is almost the same as Pearson (1894).
- In d dimensions, $\Theta(1/\epsilon^{12} \log d)$ samples *necessary* and *sufficient*.
  - "σ^2" is max variance in any coordinate.
  - Get each entry of covariance matrix to $\pm\epsilon^2\sigma^2$.
  - Previously: $O((d/\epsilon)^{300,000})$.
- Caveat: assume p_1, p_2 are bounded away from zero.

Outline

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

Outline

Method of 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:

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

- Use our samples to estimate the moments.
- Solve the system of equations to find the parameters.

Method of Moments

Solving the system

Parameters	$\lambda > 0$ rate, or inverse scale
Support	$x \in [0, \infty)$
pdf	$\lambda e^{-\lambda x}$
CDF	$1 - e^{-\lambda x}$
Mean	λ^{-1}
Median	$\lambda^{-1} \ln(2)$
Mode	0
Variance	λ^{-2}
Skewness	2
Ex. kurtosis	6
Entropy	$1 - \ln(\lambda)$
MGF	$\left(1 - \frac{t}{\lambda}\right)^{-1}$ for $t < \lambda$
CF	$\left(1 - \frac{it}{\lambda}\right)^{-1}$
Fisher information	λ^{-2}

- 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

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

- 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

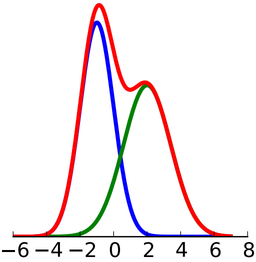
Pearson's Polynomial

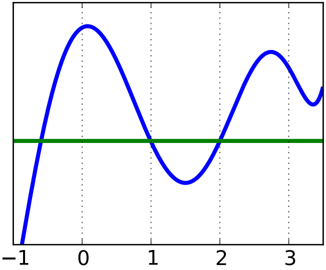
- Chug chug chug...
- Get a 9th degree polynomial in the excess moments X_3, X_4, X_5:

$$\begin{aligned} p(\alpha) = 8\alpha^9 &+ 28X_4\alpha^7 - 12X_3^2\alpha^6 + (24X_3X_5 + 30X_4^2)\alpha^5 \\ &+ (6X_5^2 - 148X_3^2X_4)\alpha^4 + (96X_3^4 - 36X_3X_4X_5 + 9X_4^3)\alpha^3 \\ &+ (24X_3^3X_5 + 21X_3^2X_4^2)\alpha^2 - 32X_3^4X_4\alpha + 8X_3^6 \\ = 0 \end{aligned}$$

- Easy to go from solutions α to mixtures μ_i, σ_i, p_i.

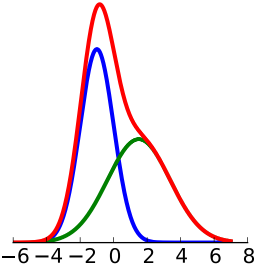
Pearson's Polynomial

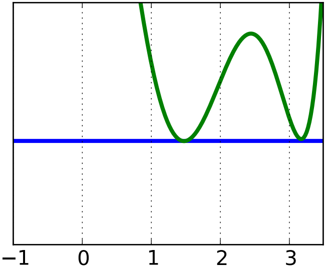




- Get a 9th degree polynomial in the excess moments X_3, X_4, X_5.
    - Positive roots correspond to mixtures that match on five moments.
    - Usually have two roots.
    - Pearson's proposal: choose candidate with closer 6th moment.
- Works because six moments uniquely identify mixture [KMV]
- How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

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.
  - Usually have two roots.
  - Pearson's proposal: choose candidate 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 α is the *unique* positive root of

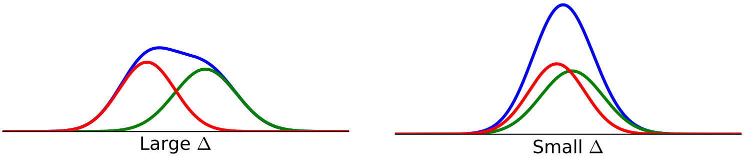
$$r(\alpha) := p_5(\alpha)^2 + p_6(\alpha)^2 = 0.$$

- Therefore $q(x) := r/(x-\alpha)^2$ has no positive roots.
- Would like that $q(x) \geq c > 0$ for all x and all mixtures α, β, γ.
  - Then for $|\widetilde{p}_5 - p_6|, |\widetilde{p}_6 - p_6| \leq \epsilon$,

$$|\alpha - \arg\min \widetilde{r}(x)| \leq \epsilon/\sqrt{c}.$$

  - Compactness: true for any closed and bounded region.
- Bounded:
  - For unbounded variables, dominating terms show $q \to \infty$.
- Closed:
  - Issue is that $x > 0$ isn't closed.
  - Can use X_3, X_4 to get an $O(1)$ approximation $\overline{\alpha}$ to α.
  - $x \in [\overline{\alpha}/10, \alpha]$ is closed.

Result



- Suppose the two components have means $\Delta\sigma$ apart.
- Then if we know M_i to $\pm\epsilon(\Delta\sigma)^i$, the algorithm recovers the means to $\pm\epsilon\Delta\sigma$.
- Therefore $O(\Delta^{-12}\epsilon^{-2})$ samples give an $\epsilon\Delta$ approximation.
  - 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

Algorithm in d dimensions

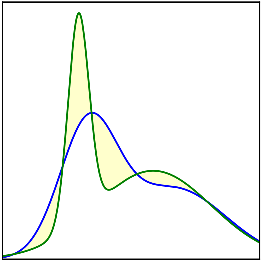
- Idea: project to lower dimensions.
- Look at individual coordinates: get $\{\mu_{1,i}, \mu_{2,i}\}$ to $\pm\epsilon\sigma$.
- How do we piece them together?
- Suppose we could solve $d = 2$:
  - Can match up $\{\mu_{1,i}, \mu_{2,i}\}$ with $\{\mu_{1,j}, \mu_{2,j}\}$.
- Solve $d = 2$:
  - Project $x \to \langle v, x \rangle$ for many random v.
  - For $\mu' \neq \mu$, will have $\langle \mu', v \rangle \neq \langle \mu', v \rangle$ with constant probability.
- So we solve d case with poly(d) calls to 1-dimensional case.
- Only loss is $\log(1/\delta) \to \log(d/\delta)$:

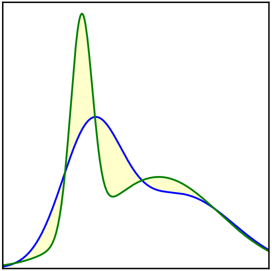
$$\Theta(1/\epsilon^{12} \log(d/\delta)) \text{ samples}$$

Outline

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:

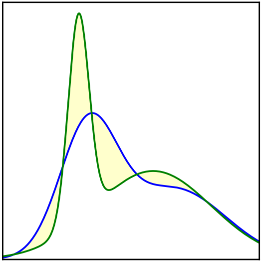


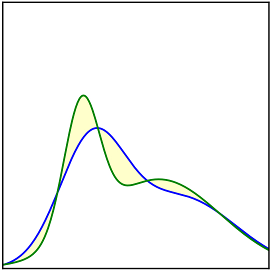


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

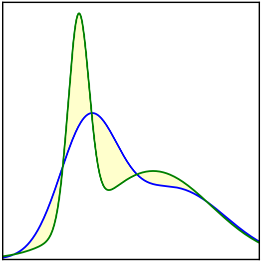


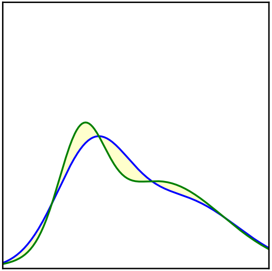


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

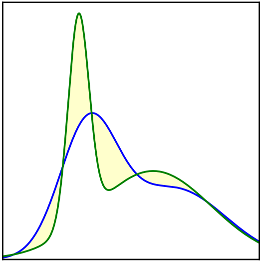


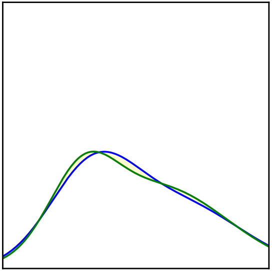


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

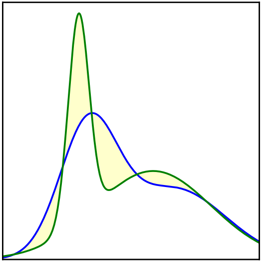


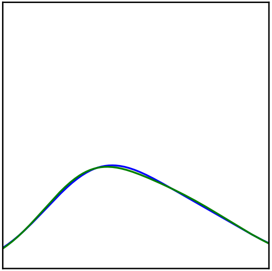


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

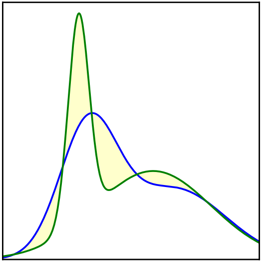


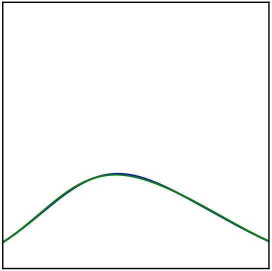


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

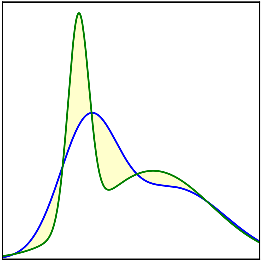


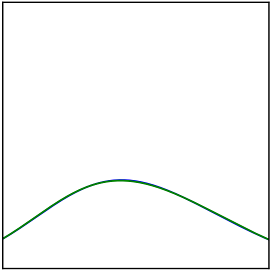


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

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:

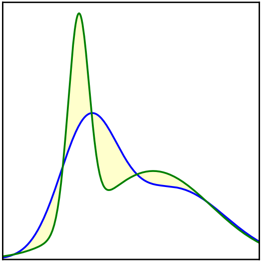


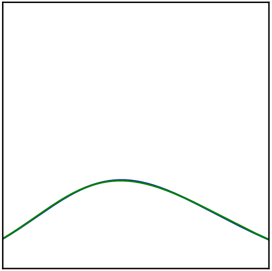


  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.

Lower bound in one dimension

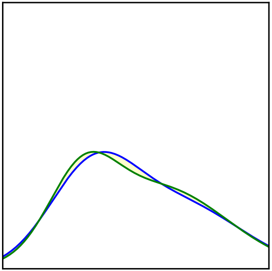
- The algorithm takes $O(\epsilon^{12})$ samples because it uses six moments
  - Necessary to get sixth moment to $\pm(\epsilon\sigma)^6$.
- Let F, F' be any two mixtures with five matching moments:





  - Constant means and variances.
  - Add $N(0, \sigma^2)$ to each mixture as σ grows.
- Claim: $\Omega(\sigma^{12})$ samples necessary to distinguish the distributions.

Lower bound in one dimension



- Two mixtures F, F' with $F \approx F'$.
- Have $\mathrm{TV}(F, F') \approx 1/\sigma^6$.
- 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
  - Sample complexity is $\Omega(1/H^2(F, F'))$
  - $H^2 \lesssim TV \lesssim H$, but often $H \approx TV$.

Bounding the Hellinger distance: general idea

Definition

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

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

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

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 Δ, then [Pollard '00]

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

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 Δ, then [Pollard '00]

$$\begin{aligned} H^2(p,q) &= 1 - \int \sqrt{1+\Delta(x)}p(x)dx \\ &= 1 - \mathop{\mathbb{E}}_{x\sim p}[\sqrt{1+\Delta(x)}] \\ &= 1 - \mathop{\mathbb{E}}_{x\sim p}[1 + \underbrace{\Delta(x)}_{\int q(x)-p(x)=0}/2 - O(\Delta^2(x))] \end{aligned}$$

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 Δ, then [Pollard '00]

$$\begin{aligned}
H^2(p,q) &= 1 - \int \sqrt{1+\Delta(x)}p(x)dx \\
&= 1 - \mathop{\mathbb{E}}_{x\sim p}[\sqrt{1+\Delta(x)}] \\
&= 1 - \mathop{\mathbb{E}}_{x\sim p}[1 + \underbrace{\Delta(x)}_{\int q(x)-p(x)=0}/2 - O(\Delta^2(x))] \\
&\lesssim \mathop{\mathbb{E}}_{x\sim p}[\Delta^2(x)]
\end{aligned}$$

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 Δ, then [Pollard '00]

$$\begin{aligned} H^2(p,q) &= 1 - \int \sqrt{1+\Delta(x)}p(x)dx \\ &= 1 - \mathop{\mathbb{E}}_{x\sim p}[\sqrt{1+\Delta(x)}] \\ &= 1 - \mathop{\mathbb{E}}_{x\sim p}[1 + \underbrace{\Delta(x)}_{\int q(x)-p(x)=0}/2 - O(\Delta^2(x))] \\ &\lesssim \mathop{\mathbb{E}}_{x\sim p}[\Delta^2(x)] \end{aligned}$$

- 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 $G, G' = F + N(0, \sigma^2), F' + N(0, \sigma^2)$,

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

- Can show both G', G are within $O(1)$ of $N(0, \sigma^2)$ over $[-\sigma^2, \sigma^2]$.
- We have that

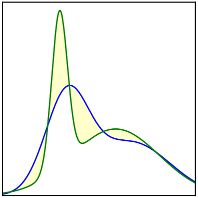
$$\begin{aligned}\Delta(x) \eqsim \frac{G'(x) - G(x)}{\nu(x)} &= \int \frac{\nu(x-t)}{\nu(x)}(F'(t) - F(t))dt \\ &\lesssim \int \sum_{d=0}^{\infty} \left(\frac{1 + x/\sigma}{\sigma\sqrt{d}}\right)^d t^d (F'(t) - F(t))dt \\ &\lesssim \sum_{d=k+1}^{\infty} \left(\frac{1 + x/\sigma}{\sigma}\right)^d \lesssim \left(\frac{1 + x/\sigma}{\sigma}\right)^{k+1}\end{aligned}$$

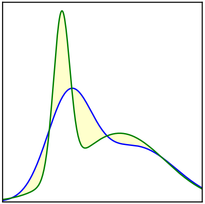
so

$$H^2(G, G') \leq \mathop{\mathbb{E}}_{x \sim G}[\Delta(x)^2] \lesssim 1/\sigma^{2k+2}$$

Lower bound in one dimension

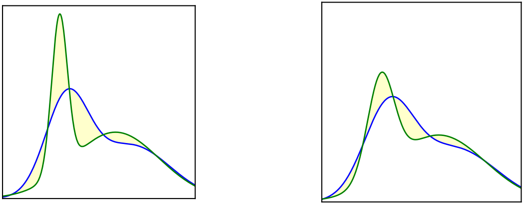
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.





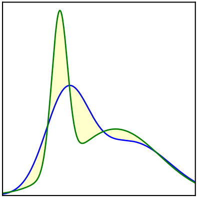
Lower bound in one dimension

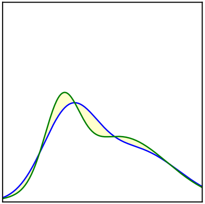
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.



Lower bound in one dimension

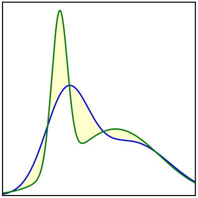
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

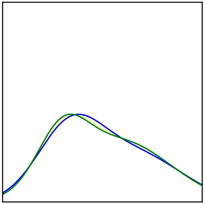




Lower bound in one dimension

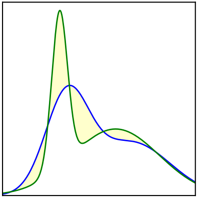
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

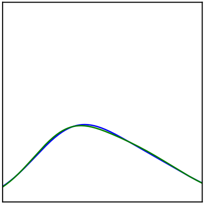




Lower bound in one dimension

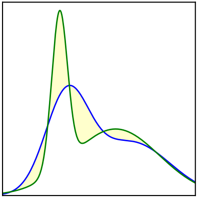
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

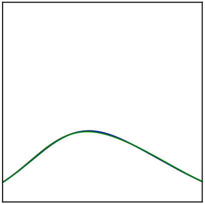




Lower bound in one dimension

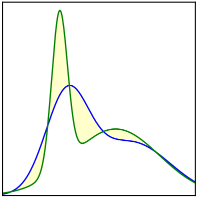
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.

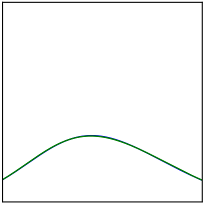




Lower bound in one dimension

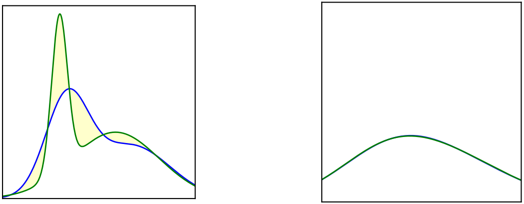
- Add $N(0, \sigma^2)$ to two mixtures with five matching moments.





Lower bound in one dimension

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



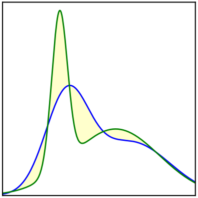
- For

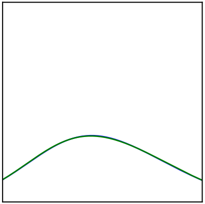
$$G = \frac{1}{2}N(-1, 1+\sigma^2) + \frac{1}{2}N(1, 2+\sigma^2)$$
$$G' \approx 0.297N(-1.226, 0.610+\sigma^2) + 0.703N(0.517, 2.396+\sigma^2)$$

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

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

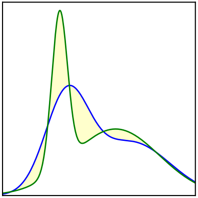
$$G' \approx 0.297N(-1.226, 0.610+\sigma^2) + 0.703N(0.517, 2.396+\sigma^2)$$

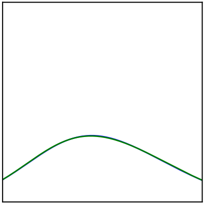
have $H^2(G, G') \lesssim 1/\sigma^{12}$.

- 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.





- For

$$G = \frac{1}{2}N(-1, 1+\sigma^2) + \frac{1}{2}N(1, 2+\sigma^2)$$
$$G' \approx 0.297N(-1.226, 0.610+\sigma^2) + 0.703N(0.517, 2.396+\sigma^2)$$

  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.

Recap and open questions

- Our result:
  - $\Theta(\epsilon^{-12} \log d)$ samples necessary and sufficient to estimate μ_i to $\pm\epsilon\sigma$, σ_i^2 to $\pm\epsilon^2\sigma^2$.
  - If the means have $\Delta\sigma$ separation, just $O(\epsilon^{-2}\Delta^{-12})$ for $\epsilon\Delta\sigma$ accuracy.
- Extend to $k > 2$?
  - Lower bound extends, so $\Omega(\epsilon^{-6k})$.
  - Do we really care about finding an $O(\epsilon^{-18})$ algorithm?
  - Solving the system of equations gets nasty.
- Automated way of figuring out whether solution to system of polynomial equations is robust?