Moreover, the algorithm works in higher dimensions too (still poly-time)

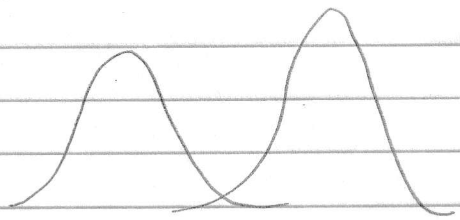
<u>Six Moments Suffice</u>

What if we are given the first six moments exactly? Does this determine the parameters?

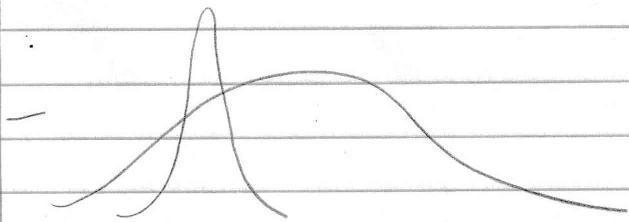
Alternatively:

Do any two different mixtures F and $\hat{F}$ necessarily differ on at least one of their first six moments?

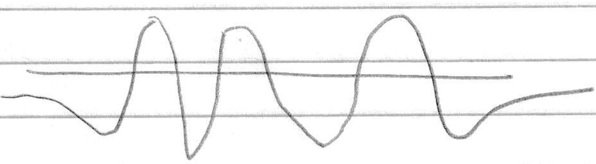
Consider the difference between their pdfs



$F(x)$



$\hat{F}(x)$



$f(x) = F(x) - \hat{F}(x)$

Lemma: If $f(x)$ has at most six zero crossings, then $F(x)$ and $\hat{F}(x)$ differ on one of their first six moments

Proof: We can find a degree ≤ 6 polynomial $p(x)$ that agrees in sign with $f(x)$.

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

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

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

So $\exists\, r \in \{1 \ldots 6\}$ s.t. $|M_r(F) - M_r(\hat{F})| > 0$ ☒

What remains is to show:

Proposition If $f(x) = \sum_{i=1}^{k} \alpha_i N(\mu_i, \sigma_i^2, x)$ is not

positive/negative

identically zero, then $f(x)$ has at most $2k-2$ zero crossings

Can check with $k=4 \Rightarrow$ six moments suffice

We'll use the heat equation.

Question: If the initial heat distribution on a 1-d infinite rod (x) is $f(x) = f(x, 0)$, what is the heat distribution at time t?

Probabilistic interpretation ($\sigma^2 = 2kt$)

$$f(x,t) = \mathbb{E}_{z \sim N(0,\sigma^2)} \left[f(x+z, 0) \right]$$

Alternatively this is a convolution

$$f(x,t) = \int_{z=-\infty}^{\infty} f(x+z) N(0, \sigma^2, z) \, dz$$

$$\triangleq f(x) * N(0, \sigma^2, x)$$

Theorem [Hummel, Gidas] Suppose $f(x): \mathbb{R} \to \mathbb{R}$ is analytic and has N zeros. Then

$$f(x) * N(0, \sigma^2, x)$$

has at most N zeros (for any $\sigma^2 > 0$)

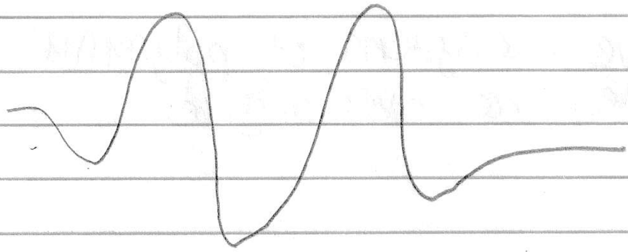
i.e. Convolving by a Gaussian / running the heat equation does not increase # of zeros

Last ingredient:

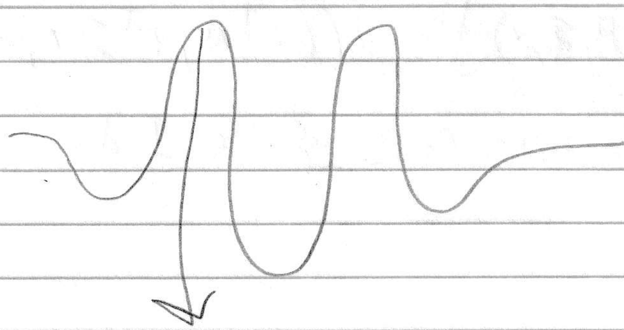
Fact: $N(\mu_1, \sigma_1^2, x) * N(\mu_2, \sigma_2^2, x) =$

$$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2, x)$$

has at most 4 ZC's. Hence



when we add back in the delta function, we add at most two new ZC's,



and convolving by $\mathcal{N}(0, \sigma_{min}^2, x)$ gets us back to the original linear combination, but does not increase the number of ZC's. $\boxtimes$

<u>Sampling Noise</u>

What if we only have estimates of the first six moments?

<u>definition</u> Let Θ be the set of valid parameters, i.e. $w_i \in [0,1]$ and $\sigma_i^2 \geq 0$

what we just showed is:

$$\{\hat{\theta} \in \Theta \mid M_r(\hat{\theta}) = M_r(\theta) \text{ for } r = 1 \text{ to } 6\}$$

the only solutions are $\theta = (\omega_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ and the relabeling $\theta' = (1-\omega_1, \mu_2, \sigma_2^2, \mu_1, \sigma_1^2)$

Are these equations stable, when given noisy estimates?

Again using deconvolution, can show

Proposition $\exists$ constants c, C s.t. if $\varepsilon < c$, the means and variances $\leq \frac{1}{\varepsilon}$, and mixing weights are in $[\varepsilon, 1-\varepsilon]$ and

$$|M_r(\theta) - M_r(\hat{\theta})| \leq \varepsilon^C \quad r = 1 \text{ to } 6$$

then there is a permutation π s.t.

$$\sum_{i=1}^{2} |\omega_i - \hat{\omega}_{\pi(i)}| + |\mu_i - \hat{\mu}_{\pi(i)}| + |\sigma_i^2 - \hat{\sigma}_{\pi(i)}^2| \leq \varepsilon$$

Hence close enough estimates for the first six moments guarantee close parameters too

This is called polynomial identifiability ← reduction from high d to 1-d

A View from Algebraic Geometry

Following Belkin and Sinha, we'll give a more general framework definition: A class of distributions $F(\Theta)$ is a polynomial family if $\forall r$

$$M_r(\theta) \triangleq \mathop{\mathbb{E}}_{x \sim F(\theta)} [x^r]$$

is a polynomial in $\theta = (\theta_1, \ldots, \theta_k)$

e.g. GMMs, mixtures of uniform, exponential, Poisson, gamma

definition: the moment generating function (mgf) is defined as

$$f(t) = \sum_{n=0}^{\infty} \mathbb{E}[x^n] \frac{t^n}{n!}$$

Fact: If the mgf converges in a neighborhood of zero then

$$\{M_r(\theta) = M_r(\hat{\theta}) \ \forall r\} \Rightarrow F(\theta) = F(\hat{\theta})$$

i.e. the infinite sequence of moments determines the density function

Now we'll need some notions/tools from algebraic geometry

definition Given a ring R, an ideal I generated by $g_1, \ldots g_n \in R$ is

$$\langle g_1, \ldots g_n \rangle \triangleq I = \left\{ \sum r_i g_i \mid \forall i \ r_i \in R \right\}$$

Moreover:

<u>definition</u>: A <u>Noetherian ring</u> is a ring s.t. for any sequence of ideals

$$I_1 \subseteq I_2 \subseteq \cdots$$

$\exists N$ s.t. $I_N = I_{N+1} = I_{N+2} \cdots$

And our main tool:

Theorem [Hilbert's Basis theorem] If R is a Noetherian ring, then $R[X]$ is too

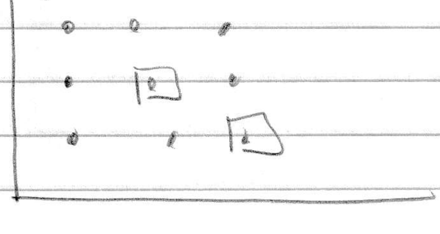
polynomials w/ coeffs in R

Let's pause for some intuition

<u>Lemma</u>: [Dickson] For any subset $K \subseteq \mathbb{N} \times \mathbb{N}$ there are a finite number of <u>minimal</u> elements

i.e. $(a,b) \in K$ s.t. $\nexists (a',b') \in K$ with $a' \leq a$, $b' \leq b$ and one inequality is strict

we can visualize this



containing only
If I is an ideal ~~generated by~~ polynomials of the form

$$x^a y^b$$

then we know I is Noetherian, and hence has a finite list of generators. This is what Dickson's lemma tells us too

Warning: An ideal I over the ring $\mathbb{R}[x,y]$ is Noetherian, but there is no effective bound of how many generators we need

Theorem [Belkin, Sinha] Let $F(\Theta)$ be a polynomial family. If the mgf converges in a neighborhood of zero the $\exists N$ s.t

$$F(\theta) = F(\hat{\theta}) \iff M_r(\theta) = M_r(\hat{\theta}) \text{ for } r = 1 \text{ to } N$$

Proof: Let $\varphi_r(\theta, \hat{\theta}) = M_r(\theta) - M_r(\hat{\theta})$

Now let

$$I_1 = \langle \varphi_1 \rangle, \; I_2 = \langle \varphi_1, \varphi_2 \rangle, \text{ etc}$$

By Hilbert's basis theorem, $\exists N$ s.t

$$I_N = I_{N+j} \quad \forall j \geq 0$$

Thus we have

$$\varphi_{N+j}(\theta, \hat{\theta}) = \sum_{i=1}^{N} p_{ij}(\theta, \hat{\theta}) \, \varphi_i(\theta, \hat{\theta})$$

for some polynomials $p_{ij} \in \mathbb{R}[\theta, \hat{\theta}]$. Thus if

$$M_r(\theta) = M_r(\hat{\theta}) \text{ for } r = 1 \text{ to } N$$

we have $\Phi_r(\theta, \hat{\theta}) = 0$ for $r = 1$ to N which implies

$$\Phi_{N+j}(\theta, \hat{\theta}) = 0 \text{ too!}$$

Now from the fact, we get that $F(\theta) = F(\hat{\theta})$ $\boxtimes$

So we know finitely many moments suffice, but don't have an effective bound

Belkin and Sinha also gave a stability analysis through quantifier elimination

def: A set S is semialgebraic if $\exists$ polynomials s.t

$$S = \{x_1, \dots x_d \mid \forall i \; p_i(x_1, \dots, x_d) \geq \text{ or } = \text{ or } \leq 0\}$$

or if S is the finite union or intersection of such sets

① we can define the projection i.e.

$$T = \{x_1, \dots x_{d-1} \mid \exists x_d \text{ s.t. } (x_1, \dots x_d) \in S\}$$

theorem: [Tarski] the projection of a semialgebraic set is also semi-algebraic

Notice, if you only use polynomial equations, you might need polynomial inequalities for the proj. ②

Fact: The complement $\bar{S}$ of a semialgebraic set is also semialgebraic.

①

Corollary: The set $U = \{x_1, \dots x_{d-1} \mid \forall x_d \ (x_1, \dots x_d) \in S\}$ is also semialgebraic

Proof: We can write

$$\bar{U} = \{x_1, \dots x_{d-1} \mid \exists x_d \ (x_1, \dots x_d) \notin S\}$$

Thus we have

$$S \overset{\text{Fact}}{\Rightarrow} \bar{S} \overset{\text{Tarski}}{\Rightarrow} \bar{U} \overset{\text{Fact}}{\Rightarrow} U$$

②

are all semialgebraic □

Now let's prove stability via quantifier elimination

Consider the set $H(\varepsilon,\delta)$ defined to be the set of all ε and δ s.t.

$$\left.\begin{array}{c}\forall\, \theta, \hat{\theta} \quad |M_r(\theta) - M_r(\hat{\theta})| \leq \delta \text{ for } r = 1 \text{ to } 6 \\ \Downarrow \\ d_p(\theta, \hat{\theta}) \leq \varepsilon \end{array}\right\} (P)$$

<u>Claim</u>: $H(\varepsilon,\delta)$ is a semialgebraic set

<u>Proof</u>: Consider the tuple $(\varepsilon, \delta, \theta, \hat{\theta})$ that satisfy the predicate (P). Equivalently

$$\{|M_r(\theta) - M_r(\hat{\theta})| \leq \delta \text{ for } r=1 \text{ to } 6\} \cup \{d_p(\theta,\hat{\theta}) \leq \varepsilon\}$$

Thus it's semialgebraic

Now applying quantifier elimination to get rid of the $\forall$ completes the proof. $\boxed{}$

<u>Theorem</u> [Belkin, Sinha] $\exists$ constants C_1, C_2 and s s.t. if $\delta \leq 1/C_1$ the $\varepsilon(\delta) \leq C_2\, \delta^{1/s}$

[sketch]

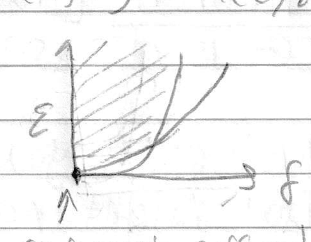
<u>Proof</u>: Since $H(\varepsilon,\delta)$ is semialgebraic,

need to show $\varepsilon^*(\delta) > 0$ $\forall \delta > 0$ "Proof by wiggling"

$$\varepsilon^*(\delta) = \text{smallest } \varepsilon \text{ s.t. } (\varepsilon,\delta) \in H(\varepsilon,\delta)$$

$$\varepsilon^*(\delta) \geq \text{poly}(\delta)$$

for sufficiently small δ. $\boxed{}$



six moments suffice!

Robust Statistics

Basic estimation problem, but we'll go in a new direction:

Given samples from a 1-d Gaussian $\mathcal{N}(\mu, \sigma^2)$, can we estimate its parameters?

Of course! Use:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \quad , \quad \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

These are examples of <u>maximum likelihood</u> paradigm (Ronald Fisher 1912-1922)

(1) <u>consistent</u>: converges to true parameters as $N \to \infty$ under tame conditions

(2) <u>asymptotically normal</u>: has smallest variance among all unbiased estimators

<u>Main Question</u>: But what if the samples are only approximately Gaussian?

<u>definition</u>: In the <u>strong contamination model</u>:

(1) m samples are drawn iid from $P \in \mathcal{D}$

known class of distributions