Disentangling Gaussians

Ankur Moitra, MIT

November 6th, 2014 — Dean's Breakfast

Algorithmic Aspects of Machine Learning

Note: These are unpolished, incomplete course notes.

Developed for educational use at MIT and for publication through MIT OpenCourseware.

The Gaussian Distribution

The Gaussian distribution is defined as (μ = mean, σ^2 = variance):

$$\mathcal{N}(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central Limit Theorem: The sum of independent random variables $X_1, X_2, ..., X_s$ converges in distribution to a Gaussian:

$$\frac{1}{\sqrt{s}} \sum_{i=1}^{s} X_i \rightarrow_d \mathcal{N}(\mu, \sigma^2)$$

This distribution is ubiquitous — e.g. used to model height, velocities in an ideal gas, annual rainfall, ...

Karl Pearson (1894) and the Naples Crabs

(figure due to Peter Macdonald)

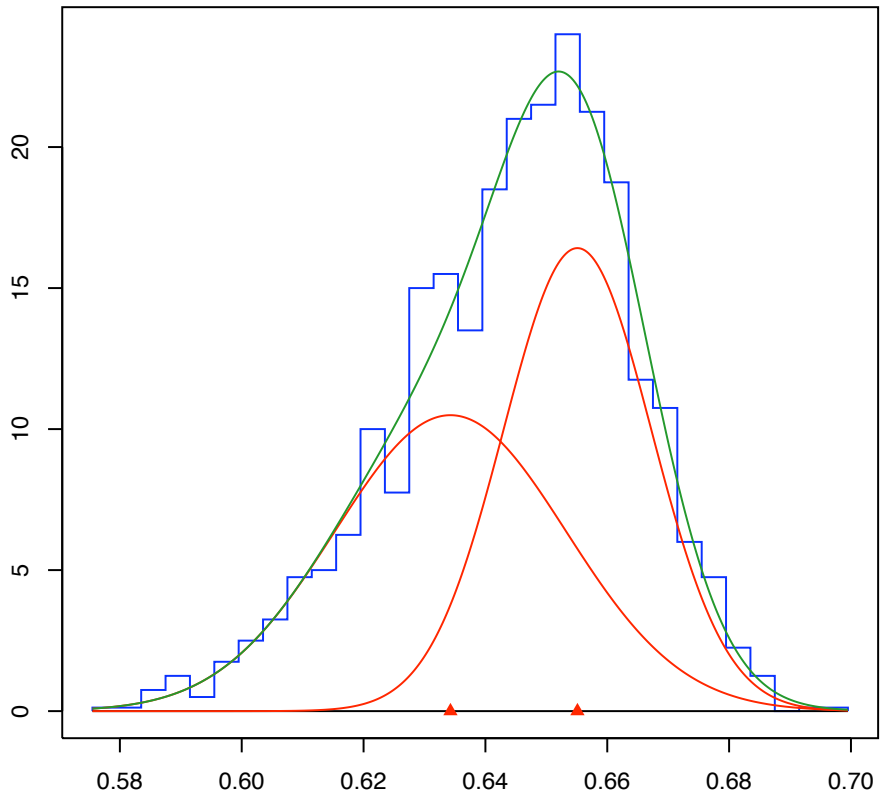


Image courtesy of Peter D. M. Macdonald. Used with permission.

Gaussian Mixture Models

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

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

five unknowns: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

Question

Given enough random samples from F, can we learn these parameters (approximately)?

Pearson invented the **method of moments**, to attack this problem...

Pearson's Sixth Moment Test

Claim

$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

In particular:

- $E_{x \leftarrow F(x)}[x] = w_1\mu_1 + (1 - w_1)\mu_2$
- $E_{x \leftarrow F(x)}[x^2] = w_1(\mu_1^2 + \sigma_1^2) + (1 - w_1)(\mu_2^2 + \sigma_2^2)$
- $E_{x \leftarrow F(x)}[x^3] = w_1(\mu_1^3 + 3\mu_1\sigma_1^2) + (1 - w_1)(\mu_2^3 + 3\mu_2\sigma_2^2)$
- ...

Pearson's Sixth Moment Test

Claim

$E_{x \leftarrow F(x)}[x^r]$ is a polynomial in $\theta = (w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$

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

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

---

Provable Guarantees?

In Contributions to the Mathematical Theory of Evolution (attributed to George Darwin):

> *"Given the probable error of every ordinate of a frequency curve, what are the probable errors of the elements of the two normal curves into which it may be dissected?"*

- Are the parameters of a mixture of two Gaussians uniquely determined by its moments?
- Are these polynomial equations robust to errors?

A View from Theoretical Computer Science

Suppose our goal is to **provably** learn the parameters of each component within an additive ϵ:

Goal

Output a mixture $\widehat{F} = \widehat{w}_1\widehat{F}_1 + \widehat{w}_2\widehat{F}_2$ *so that there is a permutation* $\pi : \{1,2\} \to \{1,2\}$ *and for* $i = \{1,2\}$

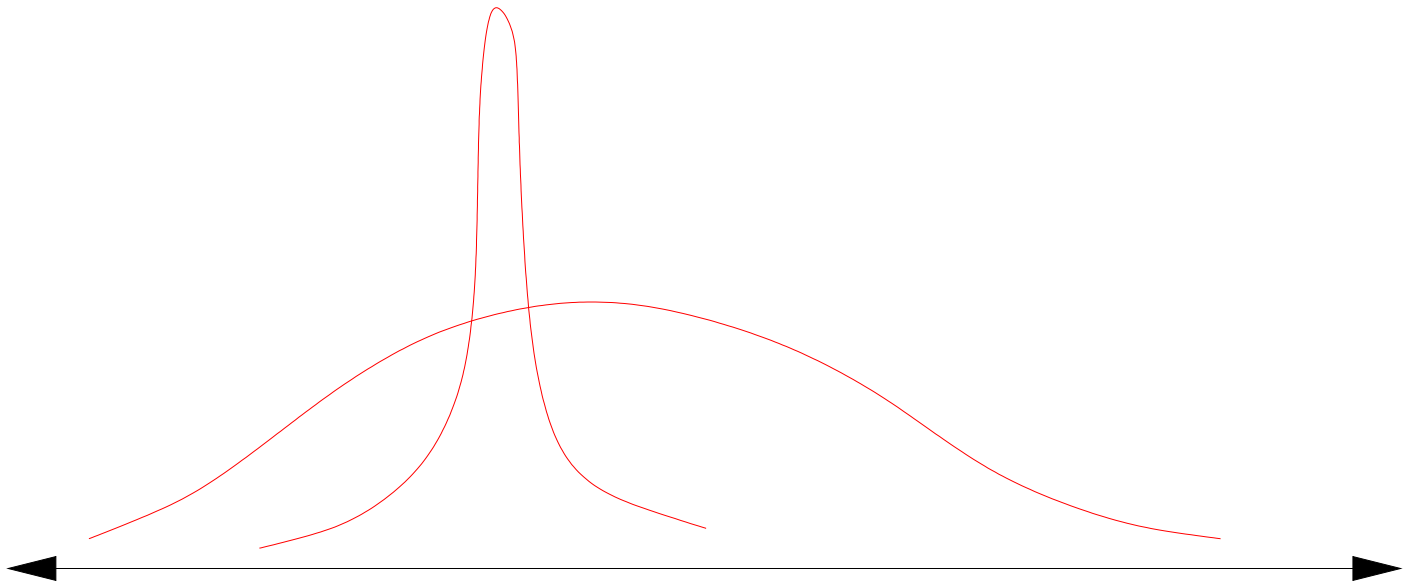
$$|w_i - \widehat{w}_{\pi(i)}|, |\mu_i - \widehat{\mu}_{\pi(i)}|, |\sigma_i^2 - \widehat{\sigma}^2_{\pi(i)}| \leq \epsilon$$

Is there an algorithm that takes poly$(1/\epsilon)$ samples and runs in time poly$(1/\epsilon)$?

A Conceptual History

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

Identifiability through the Tails

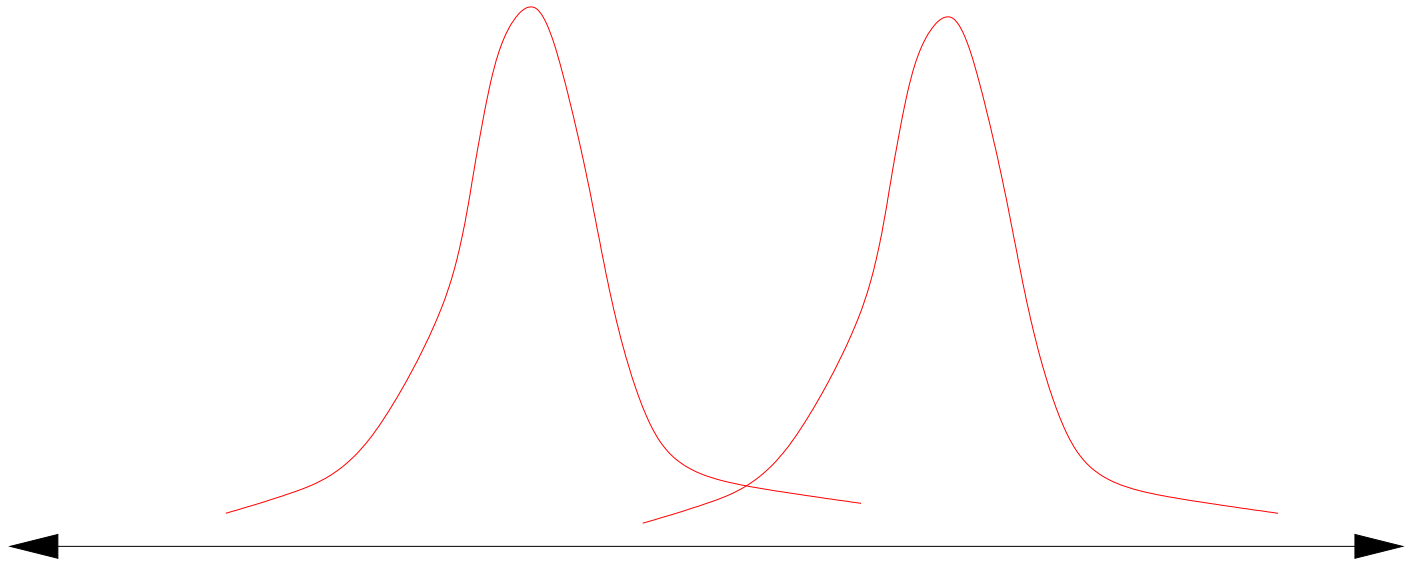


Approach: Find the parameters of the component with largest variance (it dominates the behavior of $F(x)$ at infinity); subtract it off and continue

A Conceptual History

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

Clustering Well-separated Mixtures



Approach: Cluster samples based on which component generated them; output the empirical mean and variance of each cluster

A Conceptual History

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

In summary, these approaches are heuristic, computationally intractable or make a separation assumption about the mixture

Question

What if the components overlap almost entirely?

[**Kalai, Moitra, Valiant**] (studies n-dimensional version too):

- Reduce to the one-dimensional case
- Analyze Pearson's sixth moment test (with noisy estimates)

Our Results

Suppose $w_1 \in [\epsilon^{10}, 1-\epsilon^{10}]$ and $\int |F_1(x) - F_2(x)|dx \geq \epsilon^{10}$

Theorem (Kalai, Moitra, Valiant)

There is an algorithm that (with probability at least $1-\delta$) learns the parameters of F within an additive ϵ, and the running time and number of samples needed are poly$(\frac{1}{\epsilon}, \log \frac{1}{\delta})$.

Previously, the best known bound on the running time/sample complexity were exponential

See also **[Moitra, Valiant]** and **[Belkin, Sinha]** for mixtures of k Gaussians

Analyzing the Method of Moments

Let's start with an easier question:

Question

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

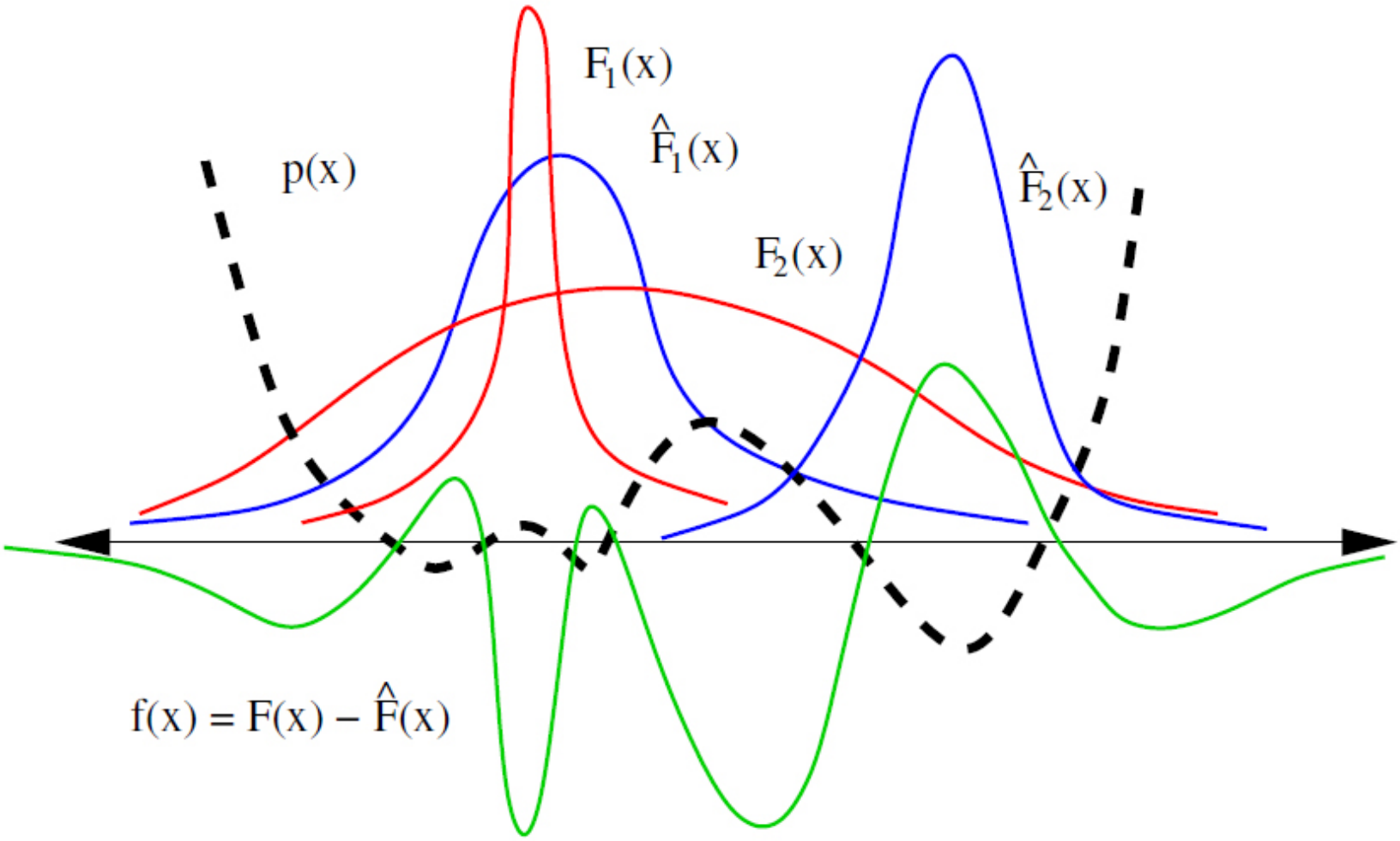
Does this uniquely determine the parameters of the mixture?

(up to a relabeling of the components)

Question

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

Method of Moments



Claim

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

Proof:

$$
\begin{aligned}
0 < \Big| \int_x p(x) f(x) dx \Big| &= \Big| \int_x \sum_{r=1}^{6} p_r x^r f(x) dx \Big| \\
&\leq \sum_{r=1}^{6} |p_r| \Big| \int_x x^r f(x) dx \Big| \\
&= \sum_{r=1}^{6} |p_r| |M_r(F) - M_r(\widehat{F})|
\end{aligned}
$$

So $\exists_{r \in \{1,2,\ldots,6\}}$ such that $|M_r(F) - M_r(\widehat{F})| > 0$

Our goal is to prove the following:

Proposition

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

.....

We will do it through properties of the **heat equation**

.....

The Heat Equation

Question

If the initial heat distribution on a one-dimensional infinite rod (κ) is $f(x) = f(x, 0)$ what is the heat distribution $f(x, t)$ at time t?

There is a probabilistic interpretation ($\sigma^2 = 2\kappa t$):

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

Alternatively, this is called a **convolution**:

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

The Key Facts

Theorem (Hummel, Gidas)

Suppose $f(x) : \mathbb{R} \to \mathbb{R}$ *is analytic and has N zeros. Then*

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

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

Convolving by a Gaussian does not increase # of zero crossings

Fact

$$\mathcal{N}(\mu_1, \sigma_1^2, x) * \mathcal{N}(\mu_2, \sigma_2^2, x) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2, x)$$

Recall, our goal is to prove the following:

Proposition

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

We will prove it by induction:

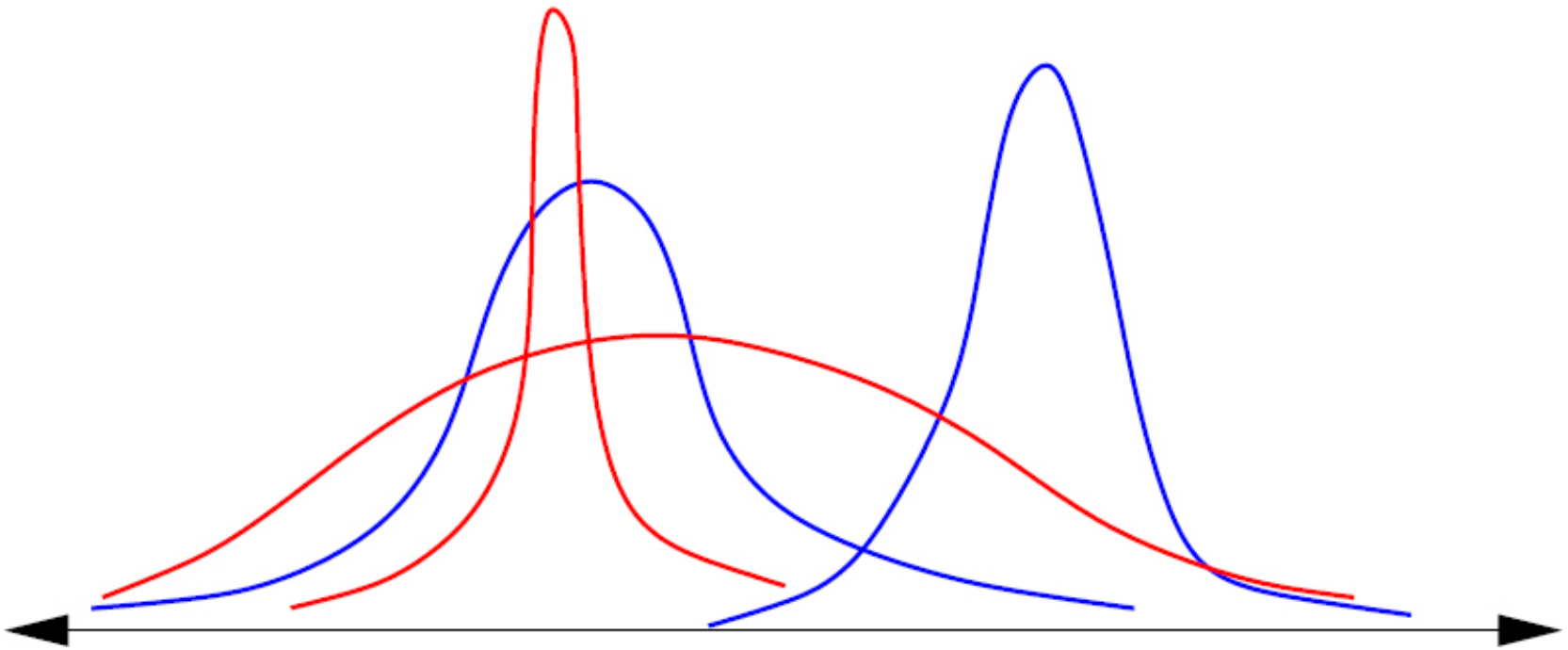
.....

Start with $k = 3$ (at most 4 zero crossings),

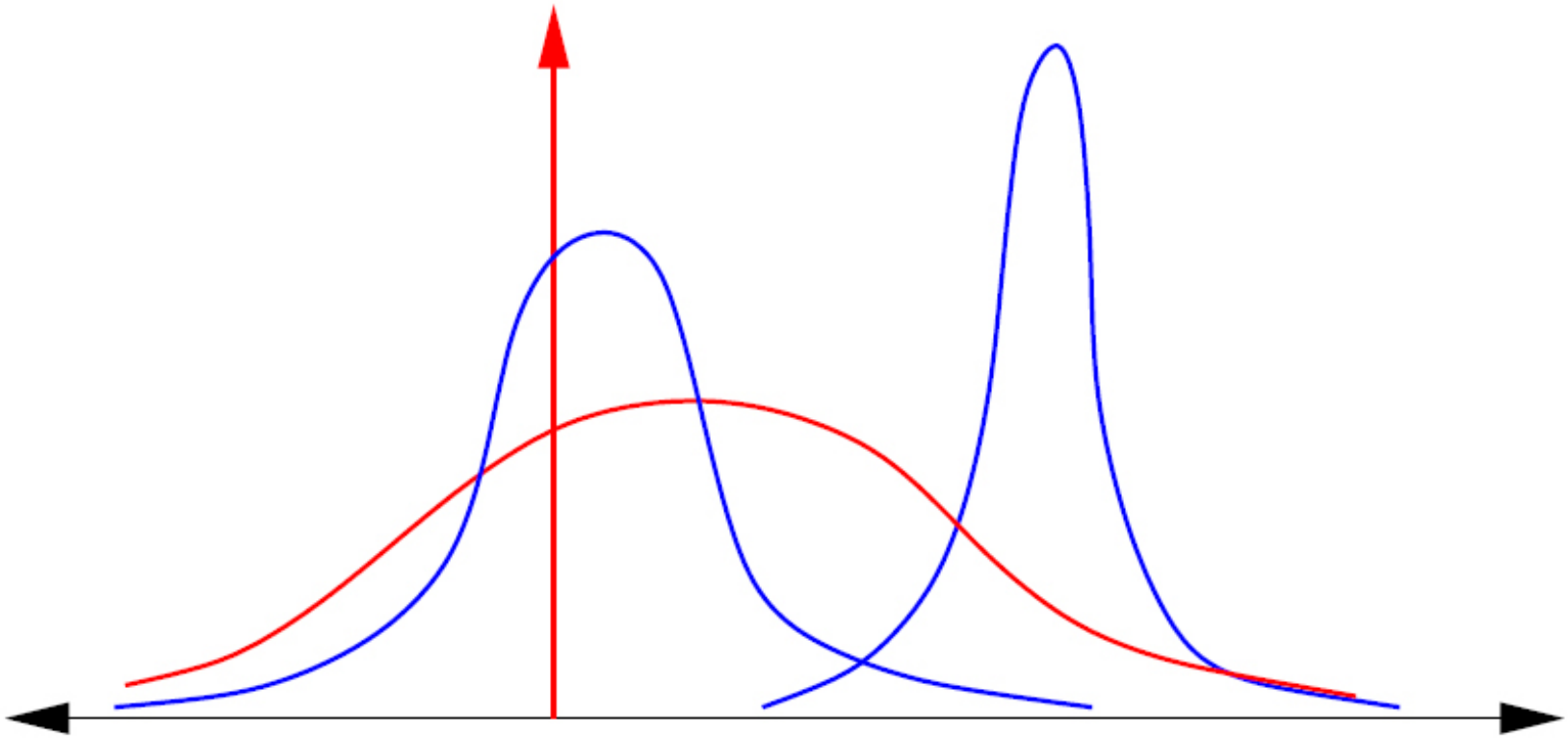
Let's prove it for $k = 4$ (at most 6 zero crossings)

.....

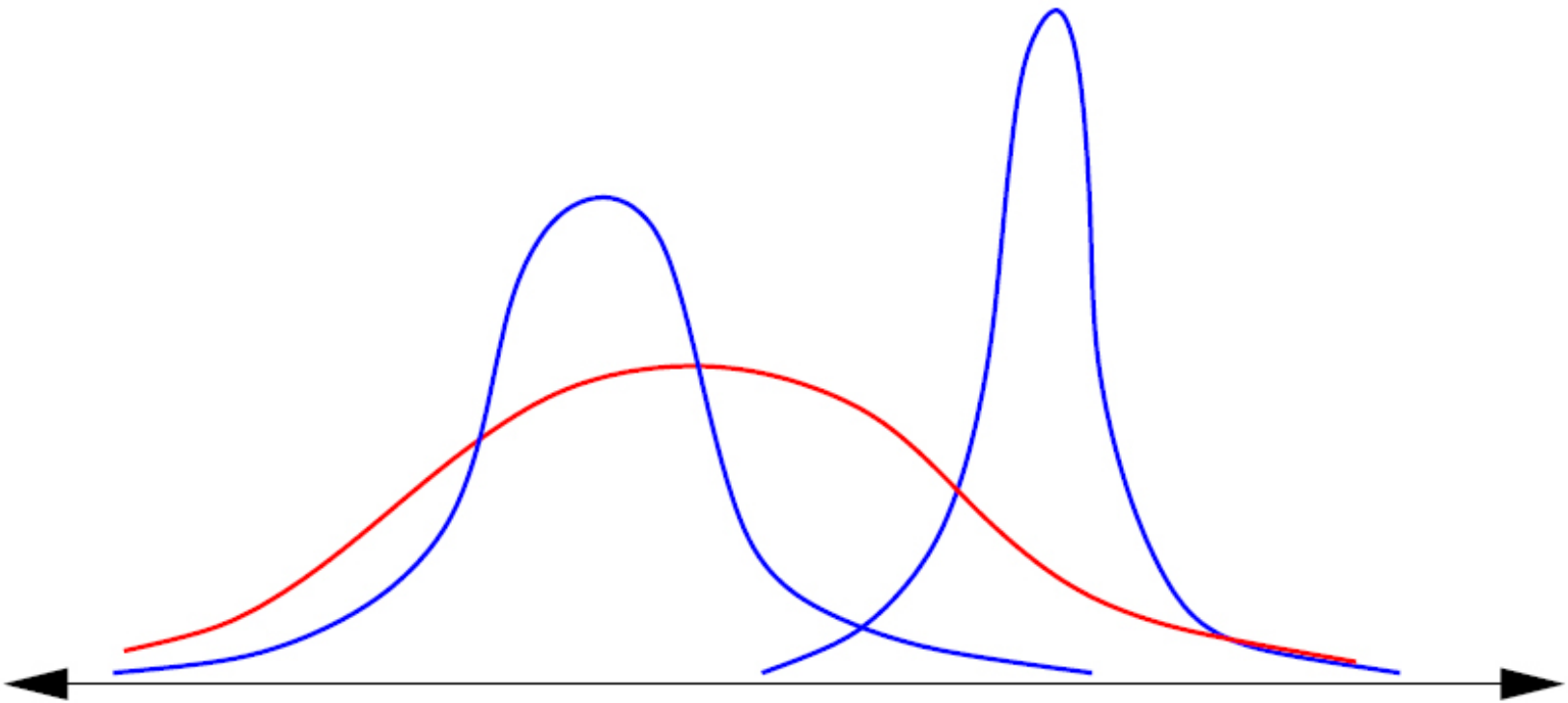
Bounding the Number of Zero Crossings



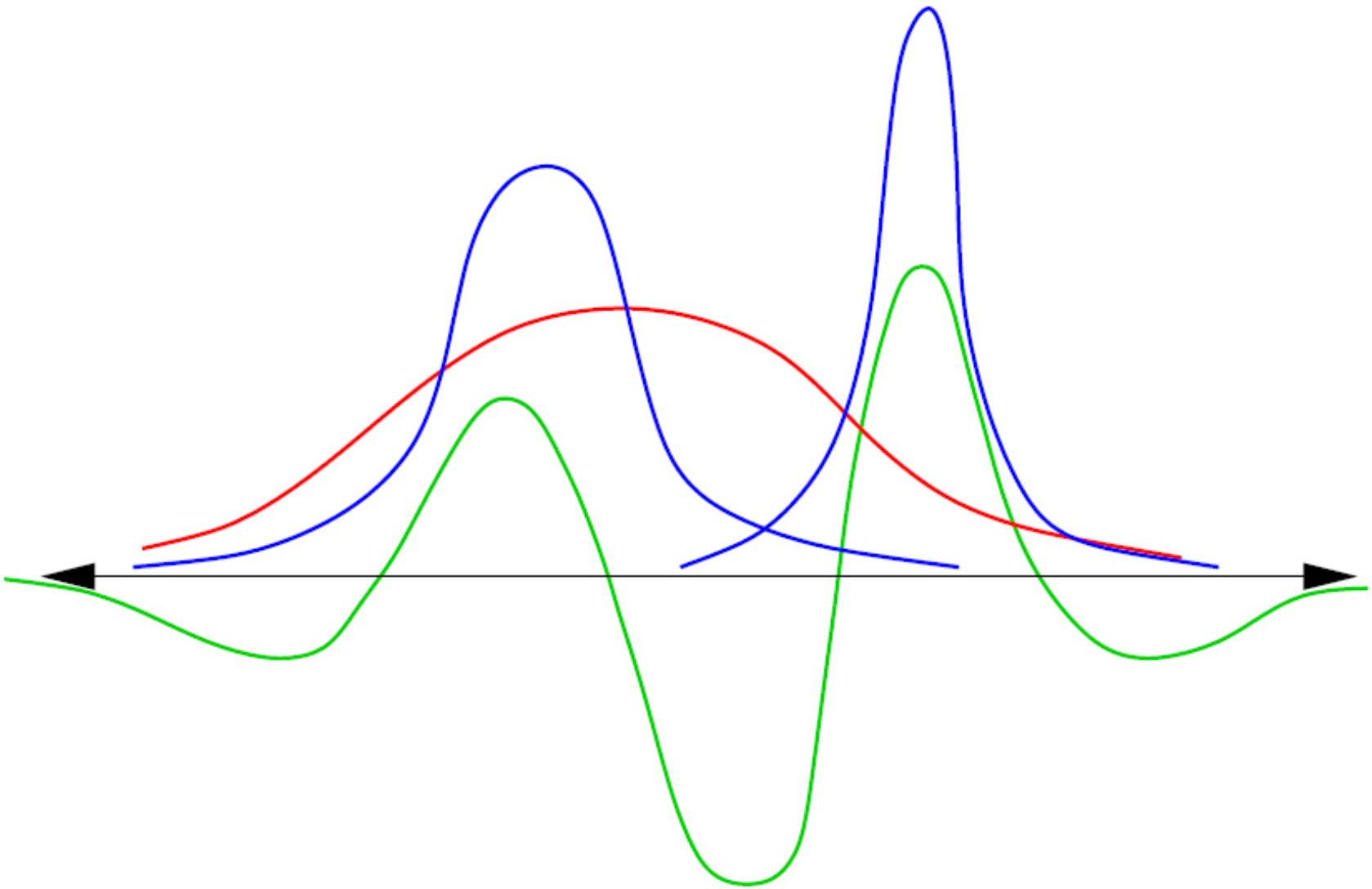
Bounding the Number of Zero Crossings



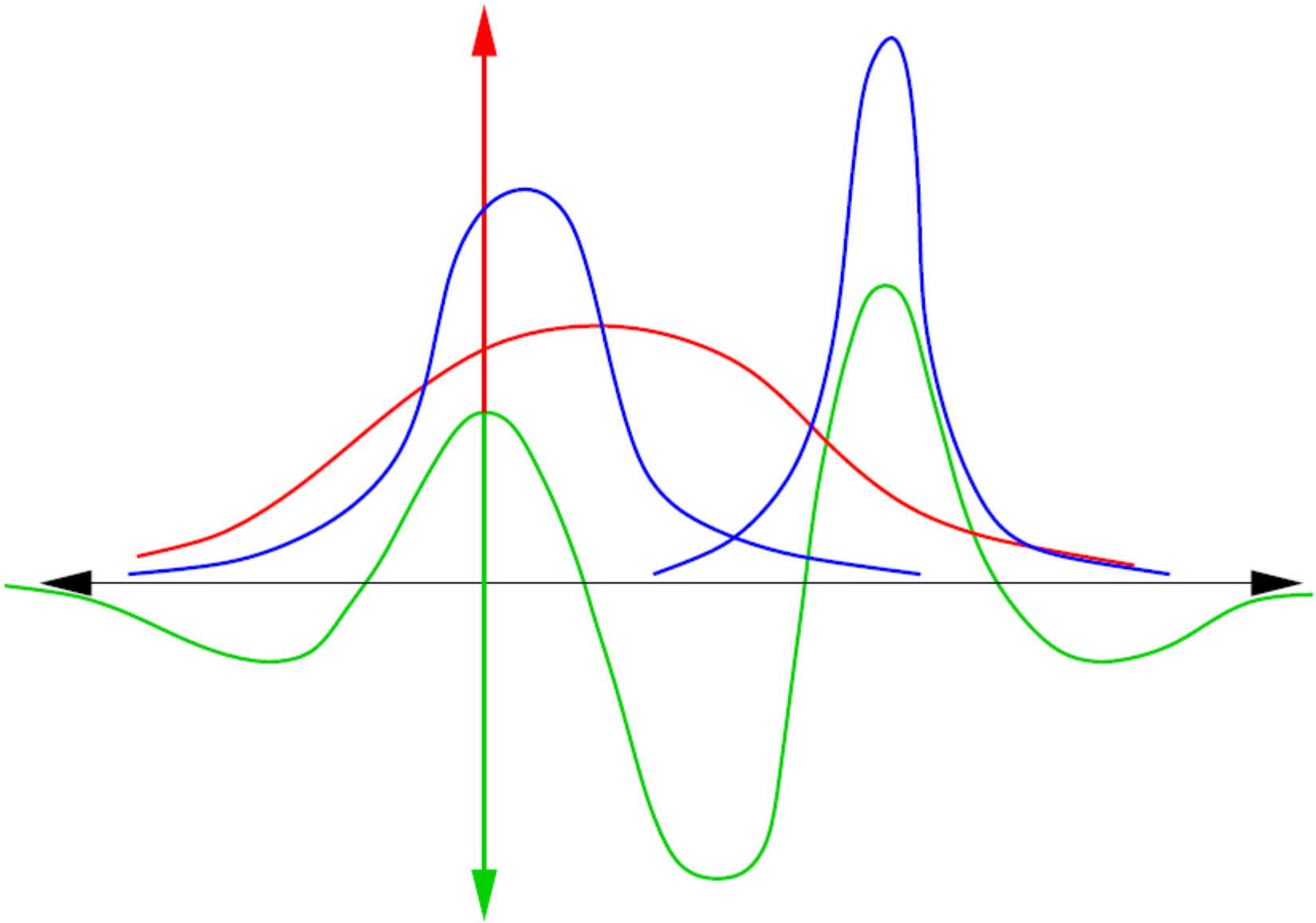
Bounding the Number of Zero Crossings



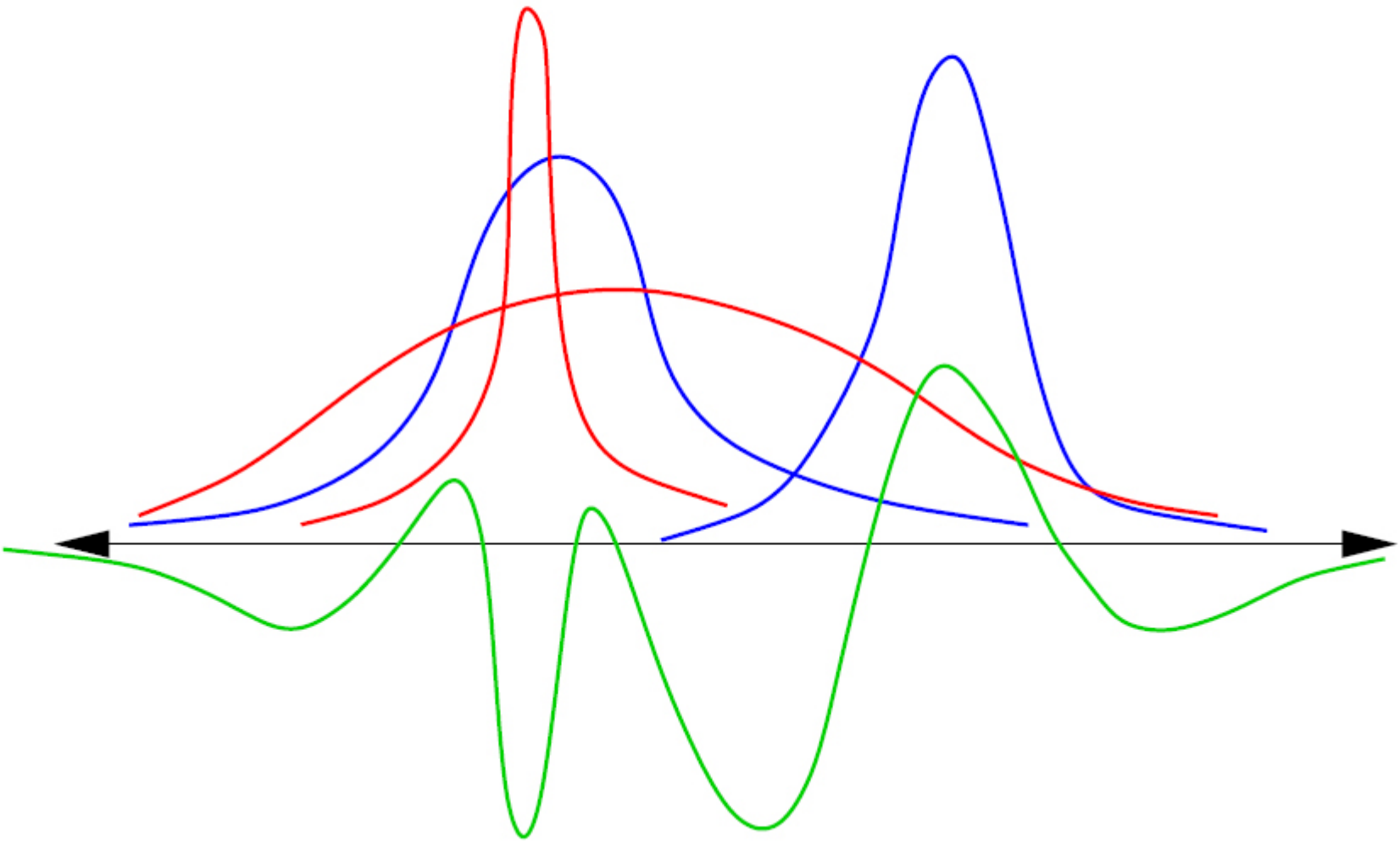
Bounding the Number of Zero Crossings



Bounding the Number of Zero Crossings



Bounding the Number of Zero Crossings



Hence, the **exact** values of the first six moments determine the mixture parameters!

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

Claim

Let θ be the true parameters; then the only solutions to

$$\left\{\widehat{\theta} \in \Theta | M_r(\widehat{\theta}) = M_r(\theta) \text{ for } r = 1, 2, ...6\right\}$$

are $(w_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and the relabeling $(1 - w_1, \mu_2, \sigma_2, \mu_1, \sigma_1)$

Are these equations stable, when we are given **noisy** estimates?

A Univariate Learning Algorithm

Our algorithm:

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

  (within an additive $\frac{\epsilon^C}{2}$)

- Compute $\widehat{\theta}$ such that $M_r(\widehat{\theta})$ is close to $\widetilde{M}_r$ for $r = 1, 2...6$

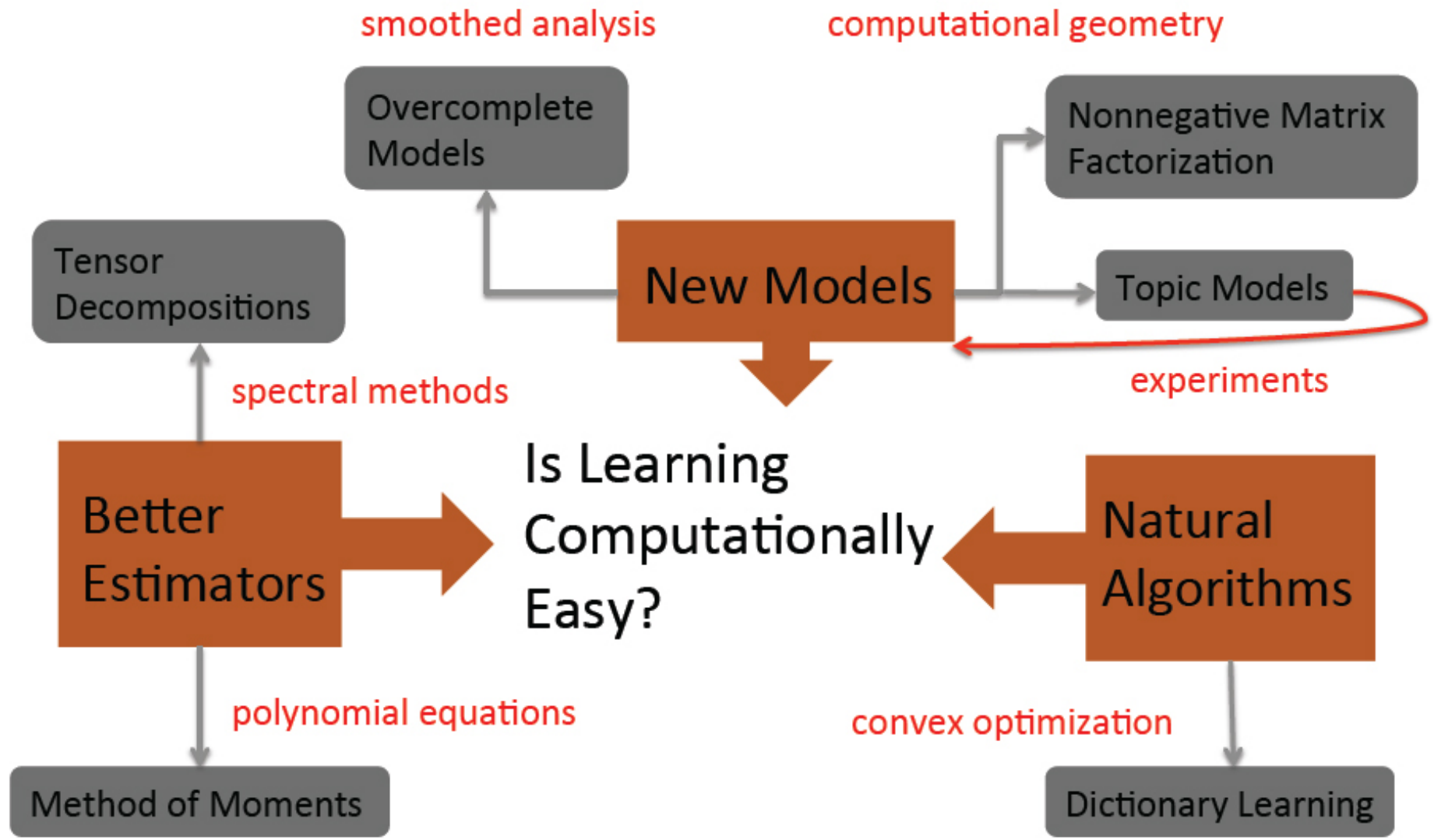
Summary and Discussion

- Here we gave the first efficient algorithms for learning mixtures of Gaussians with provably minimal assumptions

Key words: method of moments, polynomials, heat equation

- **Computational intractability** is everywhere in machine learning/statistics
- Currently, most approaches are **heuristic** and have no provable guarantees
- Can we design new algorithms for some of the fundamental problems in these fields?

My Work



e.g. mixtures of Gaussians

18.409 Algorithmic Aspects of Machine Learning

Spring 2015