Efficiently Learning Mixtures of Gaussians

Ankur Moitra, MIT

joint work with Adam Tauman Kalai and Gregory Valiant

May 11, 2010

What is a Mixture of Gaussians?

What is a Mixture of Gaussians?

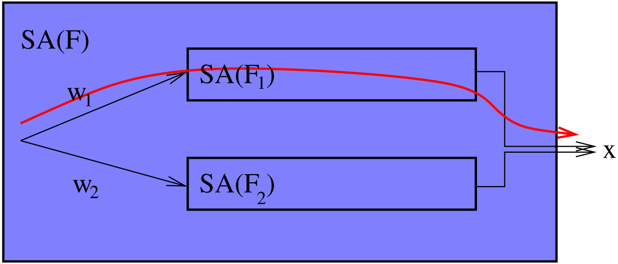
Distribution on $\Re^n$ ($w_1, w_2 \geq 0, w_1 + w_2 = 1$):

What is a Mixture of Gaussians?

Distribution on $\Re^n$ $(w_1, w_2 \geq 0, w_1 + w_2 = 1)$:

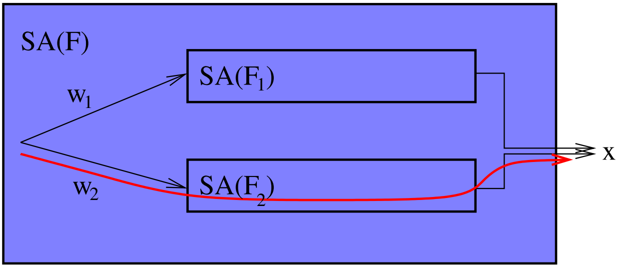
What is a Mixture of Gaussians?

Distribution on $\Re^n$ $(w_1, w_2 \geq 0, w_1 + w_2 = 1)$:



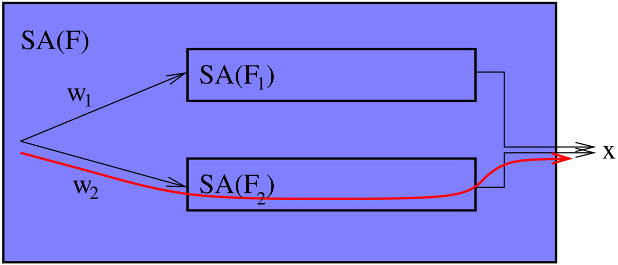
What is a Mixture of Gaussians?

Distribution on $\Re^n$ ($w_1, w_2 \geq 0, w_1 + w_2 = 1$):



What is a Mixture of Gaussians?

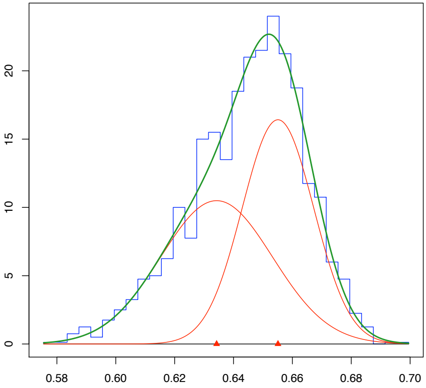
Distribution on $\Re^n$ $(w_1, w_2 \geq 0, w_1 + w_2 = 1)$:



$$F(x) = w_1 \mathcal{N}(\mu_1, \Sigma_1, x) + w_2 \mathcal{N}(\mu_2, \Sigma_2, x)$$

Pearson and the Naples Crabs

(figure due to Peter Macdonald)



Let $F(x) = w_1 F_1(x) + w_2 F_2(x)$, where $F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$

Let $F(x) = w_1 F_1(x) + w_2 F_2(x)$, where $F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$

Definition

We will refer to $E_{x \leftarrow F_i(x)}[x^r]$ as the r^{th}-raw moment of $F_i(x)$

Let $F(x) = w_1 F_1(x) + w_2 F_2(x)$, where $F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$

Definition

We will refer to $E_{x \leftarrow F_i(x)}[x^r]$ as the r^{th}-raw moment of $F_i(x)$

1. There are five unknown variables: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

Let $F(x) = w_1 F_1(x) + w_2 F_2(x)$, where $F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$

Definition

We will refer to $E_{x \leftarrow F_i(x)}[x^r]$ as the r^{th}-raw moment of $F_i(x)$

1. There are five unknown variables: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$
2. The r^{th}-raw moment of $F_i(x)$ is a polynomial in μ_i, σ_i

Let $F(x) = w_1 F_1(x) + w_2 F_2(x)$, where $F_i(x) = \mathcal{N}(\mu_i, \sigma_i^2, x)$

Definition

We will refer to $E_{x \leftarrow F_i(x)}[x^r]$ as the r^{th}-raw moment of $F_i(x)$

1. There are five unknown variables: $w_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$
2. The r^{th}-raw moment of $F_i(x)$ is a polynomial in μ_i, σ_i

Definition

Let $E_{x \leftarrow F_i(x)}[x^r] = M_r(\mu_i, \sigma_i^2)$

Question

What if we knew the r^{th}-raw moment of $F(x)$ **perfectly***?*

Question

What if we knew the r^{th}-raw moment of $F(x)$ **perfectly***?*

Each value yields a constraint:

$$E_{x \leftarrow F(x)}[x^r] = w_1 M_r(\mu_1, \sigma_1^2) + w_2 M_r(\mu_2, \sigma_2^2)$$

Question

What if we knew the r^{th}-raw moment of $F(x)$ **perfectly***?*

Each value yields a constraint:

$$E_{x \leftarrow F(x)}[x^r] = w_1 M_r(\mu_1, \sigma_1^2) + w_2 M_r(\mu_2, \sigma_2^2)$$

Definition

We will refer to $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$ as the empirical r^{th}-raw moment of $F(x)$

Pearson's Sixth Moment Test

Pearson's Sixth Moment Test

1. Compute the empirical r^{th}-raw moments $\tilde{M}_r$ for $r \in \{1, 2, ...6\}$

Pearson's Sixth Moment Test

1. Compute the empirical r^{th}-raw moments $\tilde{M}_r$ for $r \in \{1, 2, \ldots 6\}$
2. Find all simultaneous roots of

$$\{w_1 M_r(\mu_1, \sigma_1^2) + (1 - w_1) M_r(\mu_2, \sigma_2^2) = \tilde{M}_r\}_{r \in \{1,2,\ldots 5\}}$$

Pearson's Sixth Moment Test

1. Compute the empirical r^{th}-raw moments $\tilde{M}_r$ for $r \in \{1, 2, \ldots 6\}$
2. Find all simultaneous roots of

$$\{w_1 M_r(\mu_1, \sigma_1^2) + (1 - w_1) M_r(\mu_2, \sigma_2^2) = \tilde{M}_r\}_{r \in \{1,2,\ldots 5\}}$$

3. This yields a list of candidate parameters $\vec{\theta}^a, \vec{\theta}^b, \ldots$

Pearson's Sixth Moment Test

1. Compute the empirical r^{th}-raw moments $\tilde{M}_r$ for $r \in \{1, 2, ...6\}$
2. Find all simultaneous roots of

$$\{w_1 M_r(\mu_1, \sigma_1^2) + (1 - w_1) M_r(\mu_2, \sigma_2^2) = \tilde{M}_r\}_{r \in \{1,2,...5\}}$$

3. This yields a list of candidate parameters $\vec{\theta}^a, \vec{\theta}^b, ...$
4. Choose the candidate that is closest in sixth moment:

$$w_1 M_6(\mu_1, \sigma_1^2) + (1 - w_1) M_6(\mu_2, \sigma_2^2) \approx \tilde{M}_6$$

"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?" [Karl Pearson]

Question

How does noise in the empirical moments translate to noise in the derived parameters?

Gaussian Mixture Models

Applictions in physics, biology, geology, social sciences ...

Gaussian Mixture Models

Applictions in physics, biology, geology, social sciences ...

Goal

Estimate parameters in order to understand underlying process

Gaussian Mixture Models

Applictions in physics, biology, geology, social sciences ...

Goal

Estimate parameters in order to understand underlying process

Question

Can we **PROVABLY** *recover the parameters* **EFFICIENTLY***? (Dasgupta, 1999)*

Gaussian Mixture Models

Applictions in physics, biology, geology, social sciences ...

Goal

Estimate parameters in order to understand underlying process

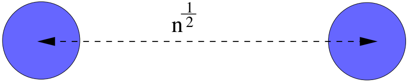
Question

Can we **PROVABLY** *recover the parameters* **EFFICIENTLY***? (Dasgupta, 1999)*

Definition

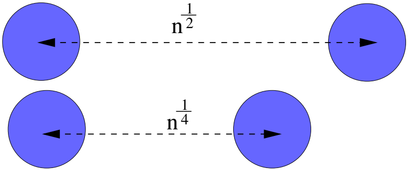
$D(f(x), g(x)) = \frac{1}{2}\|f(x) - g(x)\|_1$

A History of Learning Mixtures of Gaussians



[Dasgupta, 1999]

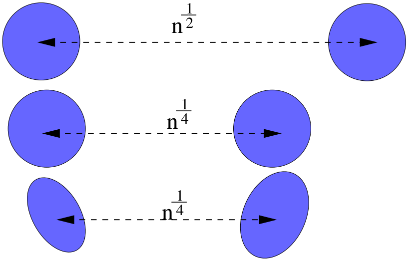
A History of Learning Mixtures of Gaussians



[Dasgupta, 1999]

[Dasgupta, Schulman, 2000]

A History of Learning Mixtures of Gaussians

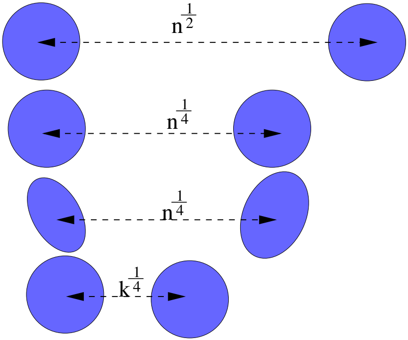


[Dasgupta, 1999]

[Dasgupta, Schulman, 2000]

[Arora, Kannan, 2001]

A History of Learning Mixtures of Gaussians



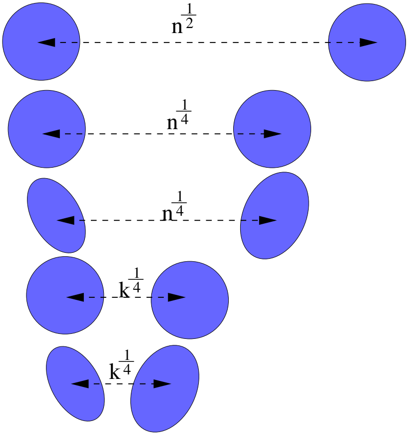
[Dasgupta, 1999]

[Dasgupta, Schulman, 2000]

[Arora, Kannan, 2001]

[Vempala, Wang, 2002]

A History of Learning Mixtures of Gaussians



[Dasgupta, 1999]

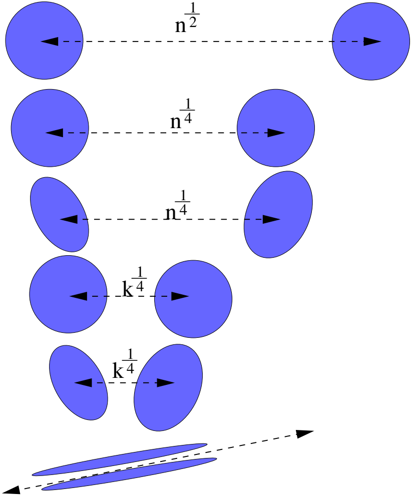
[Dasgupta, Schulman, 2000]

[Arora, Kannan, 2001]

[Vempala, Wang, 2002]

[Achlioptas, McSherry, 2005]

A History of Learning Mixtures of Gaussians



[Dasgupta, 1999]

[Dasgupta, Schulman, 2000]

[Arora, Kannan, 2001]

[Vempala, Wang, 2002]

[Achlioptas, McSherry, 2005]

[Brubaker, Vempala, 2008]

All previous results required $D(F_1, F_2) \approx 1$...

All previous results required $D(F_1, F_2) \approx 1$...

... because the results relied on **CLUSTERING**

All previous results required $D(F_1, F_2) \approx 1$...

... because the results relied on **CLUSTERING**

Question

Can we learn the parameters of the mixture without clustering?

All previous results required $D(F_1, F_2) \approx 1$...

... because the results relied on **CLUSTERING**

Question

Can we learn the parameters of the mixture without clustering?

Question

Can we learn the parameters when $D(F_1, F_2)$ is close to **ZERO**?

Goal

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

$$\left| w_i - \hat{w}_{\pi(i)} \right| \leq \epsilon, D(F_i, \hat{F}_{\pi(i)}) \leq \epsilon$$

Goal

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

$$|w_i - \hat{w}_{\pi(i)}| \leq \epsilon, D(F_i, \hat{F}_{\pi(i)}) \leq \epsilon$$

We will call such a mixture $\hat{F}$ ϵ-close to F.

Goal

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

$$|w_i - \hat{w}_{\pi(i)}| \le \epsilon, D(F_i, \hat{F}_{\pi(i)}) \le \epsilon$$

We will call such a mixture $\hat{F}$ ϵ-close to F.

Question

When can we hope to learn an ϵ*-close estimate?*

Question

What if $w_1 = 0$*?*

Question

What if $w_1 = 0$?

We never sample from F_1!

Question

What if $w_1 = 0$*?*

We never sample from F_1!

Question

What if $D(F_1, F_2) = 0$*?*

Question

What if $w_1 = 0$*?*

We never sample from F_1!

Question

What if $D(F_1, F_2) = 0$*?*

For any w_1, w_2, $F = w_1 F_1 + w_2 F_2$ is the same distribution!

Question

What if $w_1 = 0$*?*

We never sample from F_1!

Question

What if $D(F_1, F_2) = 0$*?*

For any w_1, w_2, $F = w_1 F_1 + w_2 F_2$ is the same distribution!

Definition

A mixture of Gaussians $F = w_1 F_1 + w_2 F_2$ is ϵ-statistically learnable if for $i = \{1, 2\}$, $w_i \geq \epsilon$ and $D(F_1, F_2) \geq \epsilon$.

Efficiently Learning Mixtures of Two Gaussians

Given oracle access to an ϵ-statistically learnable mixture of two Gaussians F:

Efficiently Learning Mixtures of Two Gaussians

Given oracle access to an ϵ-statistically learnable mixture of two Gaussians F:

Theorem (Kalai, M, Valiant)

There is an algorithm that (with probability at least $1-\delta$) learns a mixture of two Gaussians $\hat{F}$ that is an ϵ-close estimate to F, and the running time and data requirements are poly$(\frac{1}{\epsilon}, n, \frac{1}{\delta})$.

Efficiently Learning Mixtures of Two Gaussians

Given oracle access to an ϵ-statistically learnable mixture of two Gaussians F:

Theorem (Kalai, M, Valiant)

There is an algorithm that (with probability at least $1-\delta$) learns a mixture of two Gaussians $\hat{F}$ that is an ϵ-close estimate to F, and the running time and data requirements are poly$(\frac{1}{\epsilon}, n, \frac{1}{\delta})$.

Previously, even no inverse exponential estimator known for <u>univariate</u> mixtures of <u>two</u> Gaussians

Question

What about mixtures of k Gaussians?

Question

What about mixtures of k Gaussians?

Definition

A mixture of k Gaussians $F = \sum_i w_i F_i$ is ϵ-statistically learnable if for $i = \{1, 2, .., k\}$, $w_i \geq \epsilon$ and for all i, j $D(F_i, F_j) \geq \epsilon$.

Question

What about mixtures of k Gaussians?

Definition

A mixture of k Gaussians $F = \sum_i w_i F_i$ is ϵ-statistically learnable if for $i = \{1, 2, .., k\}$, $w_i \geq \epsilon$ and for all i, j $D(F_i, F_j) \geq \epsilon$.

Definition

An estimate $\hat{F} = \sum_i \hat{w}_i \hat{F}_i$ mixture of k Gaussians is ϵ-close to F if there is a permutation $\pi : \{1, 2, ..., k\} \to \{1, 2, ..., k\}$ and for $i = \{1, 2, ..., k\}$

$$|w_i - \hat{w}_{\pi(i)}| \leq \epsilon, D(F_i, \hat{F}_{\pi(i)}) \leq \epsilon$$

Efficiently Learning Mixtures of Gaussians

Given oracle access to an ϵ-statistically learnable mixture of k Gaussians F:

Efficiently Learning Mixtures of Gaussians

Given oracle access to an ϵ-statistically learnable mixture of k Gaussians F:

Theorem (M, Valiant)

There is an algorithm that (with probability at least $1 - \delta$) learns a mixture of k Gaussians $\hat{F}$ that is an ϵ-close estimate to F, and the running time and data requirements are poly$(\frac{1}{\epsilon}, n, \frac{1}{\delta})$.

Efficiently Learning Mixtures of Gaussians

Given oracle access to an ϵ-statistically learnable mixture of k Gaussians F:

Theorem (M, Valiant)

There is an algorithm that (with probability at least $1-\delta$) learns a mixture of k Gaussians $\hat{F}$ that is an ϵ-close estimate to F, and the running time and data requirements are poly$(\frac{1}{\epsilon}, n, \frac{1}{\delta})$.

The running time and sample complexity depends exponentially on k, but such a dependence is necessary!

Efficiently Learning Mixtures of Gaussians

Given oracle access to an ϵ-statistically learnable mixture of k Gaussians F:

Theorem (M, Valiant)

There is an algorithm that (with probability at least $1-\delta$) learns a mixture of k Gaussians $\hat{F}$ that is an ϵ-close estimate to F, and the running time and data requirements are poly$(\frac{1}{\epsilon}, n, \frac{1}{\delta})$.

The running time and sample complexity depends exponentially on k, but such a dependence is necessary!

Corollary: First polynomial time density estimation for mixtures of Gaussians **with no assumptions!**

Question

Can we give additive guarantees?

Question

Can we give additive guarantees?

Cannot give additive guarantees without defining an appropriate normalization

Question

Can we give additive guarantees?

Cannot give additive guarantees without defining an appropriate normalization

Definition

A distribution $F(x)$ is in isotropic position if

Question

Can we give additive guarantees?

Cannot give additive guarantees without defining an appropriate normalization

Definition

A distribution $F(x)$ is in isotropic position if

1. $E_{x \leftarrow F(x)}[x] = \vec{0}$

Question

Can we give additive guarantees?

Cannot give additive guarantees without defining an appropriate normalization

Definition

A distribution $F(x)$ is in isotropic position if

1. $E_{x \leftarrow F(x)}[x] = \vec{0}$
2. $E_{x \leftarrow F(x)}[(u^T x)^2] = 1$ for all $\|u\| = 1$

Question

Can we give additive guarantees?

Cannot give additive guarantees without defining an appropriate normalization

Definition

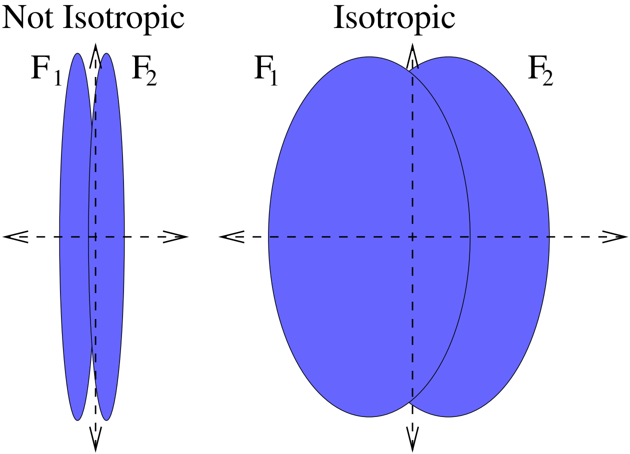
A distribution $F(x)$ is in isotropic position if

1. $E_{x \leftarrow F(x)}[x] = \vec{0}$
2. $E_{x \leftarrow F(x)}[(u^T x)^2] = 1$ for all $\|u\| = 1$

Fact

For any distribution $F(x)$ on $\Re^n$, there is an affine transformation T that places $F(x)$ in isotropic position

Isotropic Position



Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Output

$\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ *s.t.*

$$|w_i - \hat{w}_{\pi(i)}|, \|\mu_i - \hat{\mu}_{\pi(i)}\|, \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \le \epsilon$$

Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Output

$\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ *s.t.*

$$|w_i - \hat{w}_{\pi(i)}|, \|\mu_i - \hat{\mu}_{\pi(i)}\|, \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \leq \epsilon$$

Rough Idea

1. *Consider a series of projections down to one dimension*

Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Output

$\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ *s.t.*

$$|w_i - \hat{w}_{\pi(i)}|, \|\mu_i - \hat{\mu}_{\pi(i)}\|, \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \leq \epsilon$$

Rough Idea

1. *Consider a series of projections down to one dimension*
2. *Run a univariate learning algorithm*

Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Output

$\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ *s.t.*

$$|w_i - \hat{w}_{\pi(i)}|, \|\mu_i - \hat{\mu}_{\pi(i)}\|, \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \leq \epsilon$$

Rough Idea

1. *Consider a series of projections down to one dimension*
2. *Run a univariate learning algorithm*
3. *Use these estimates as constraints in a system of equations*

Given

Mixture F of two Gaussians, ϵ-statistically learnable, and in isotropic position

Output

$\hat{F} = \hat{w}_1 \hat{F}_1 + \hat{w}_2 \hat{F}_2$ *s.t.*

$$|w_i - \hat{w}_{\pi(i)}|, \|\mu_i - \hat{\mu}_{\pi(i)}\|, \|\Sigma_i - \hat{\Sigma}_{\pi(i)}\|_F \leq \epsilon$$

Rough Idea

1. *Consider a series of projections down to one dimension*
2. *Run a univariate learning algorithm*
3. *Use these estimates as constraints in a system of equations*
4. *Solve this system to obtain higher dimensional estimates*

Claim

$Proj_r[F_1] = \mathcal{N}(r^T\mu_1, r^T\Sigma_1 r, x)$

Claim

$Proj_r[F_1] = \mathcal{N}(r^T\mu_1, r^T\Sigma_1 r, x)$

Each univariate estimate yields an approximate linear constraint on the parameters

Claim

$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r, x)$

Each univariate estimate yields an approximate linear constraint on the parameters

Definition

$D_p(\mathcal{N}(\mu_1, \sigma_1^2), \mathcal{N}(\mu_2, \sigma_2^2)) = |\mu_1 - \mu_2| + |\sigma_1^2 - \sigma_2^2|$

Problem

What if we choose a direction r s.t. $D_p(Proj_r[F_1], Proj_r[F_2])$ is extremely small?

Problem

What if we choose a direction r s.t. $D_p(Proj_r[F_1], Proj_r[F_2])$ is extremely small?

Then we would need to run the univariate algorithm with extremely fine precision!

Problem

What if we choose a direction r s.t. $D_p(Proj_r[F_1], Proj_r[F_2])$ is extremely small?

Then we would need to run the univariate algorithm with extremely fine precision!

Isotropic Projection Lemma: With high probability, $D_p(Proj_r[F_1], Proj_r[F_2])$ is at least polynomially large

Problem

What if we choose a direction r s.t. $D_p(Proj_r[F_1], Proj_r[F_2])$ is extremely small?

Then we would need to run the univariate algorithm with extremely fine precision!

Isotropic Projection Lemma: With high probability, $D_p(Proj_r[F_1], Proj_r[F_2])$ is at least polynomially large

(i.e. at least $poly(\epsilon, \frac{1}{n})$)

Isotropic Projection Lemma

Suppose $F = w_1 F_1 + w_2 F_2$ is in isotropic position and is ϵ-statistically learnable:

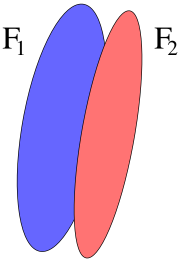
Isotropic Projection Lemma

Suppose $F = w_1 F_1 + w_2 F_2$ is in isotropic position and is ϵ-statistically learnable:

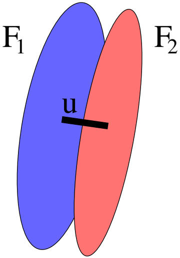
Lemma (Isotropic Projection Lemma)

With probability $\geq 1 - \delta$ *over a randomly chosen direction* r,
$D_p(Proj_r[F_1], Proj_r[F_2]) \geq \frac{\epsilon^5 \delta^2}{50 n^2} = \epsilon_3$.

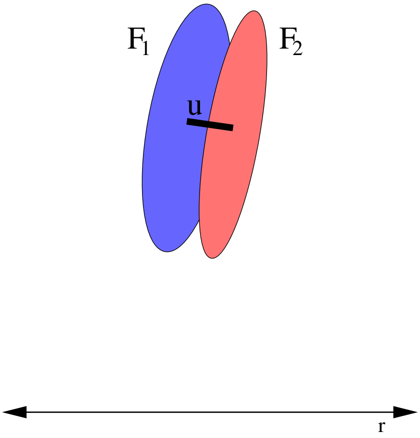
Isotropic Projection Lemma



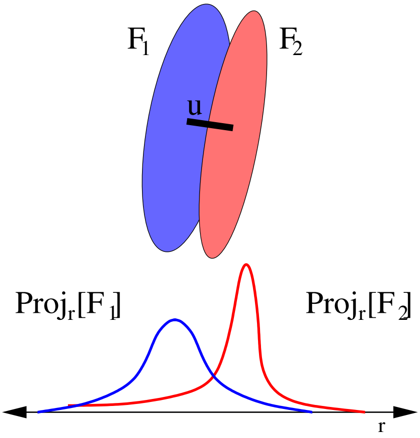
Isotropic Projection Lemma



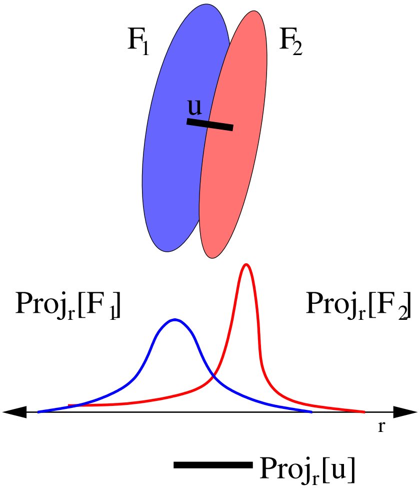
Isotropic Projection Lemma



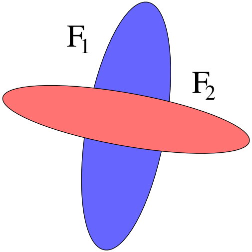
Isotropic Projection Lemma



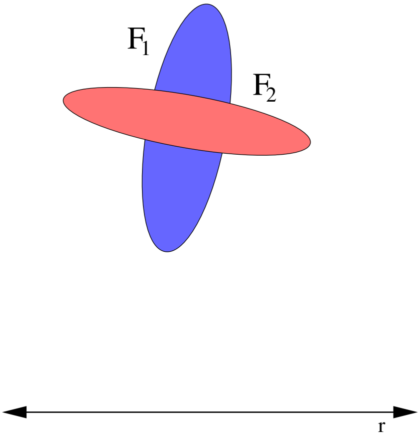
Isotropic Projection Lemma



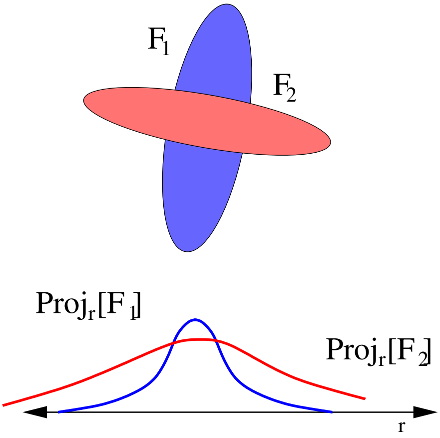
Isotropic Projection Lemma



Isotropic Projection Lemma



Isotropic Projection Lemma



Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

$\hat{F}_1^r, \hat{F}_1^s$ each yield constraints on multidimensional parameters of one Gaussian in F

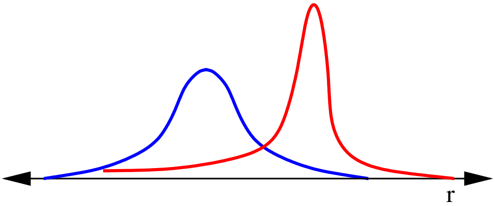
Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

$\hat{F}_1^r, \hat{F}_1^s$ each yield constraints on multidimensional parameters of one Gaussian in F

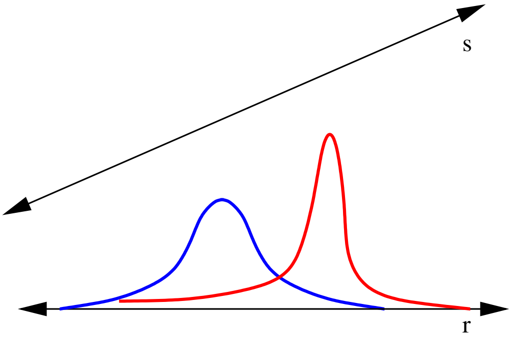
Problem

How do we know that they yield constraints on the **SAME** *Gaussian?*

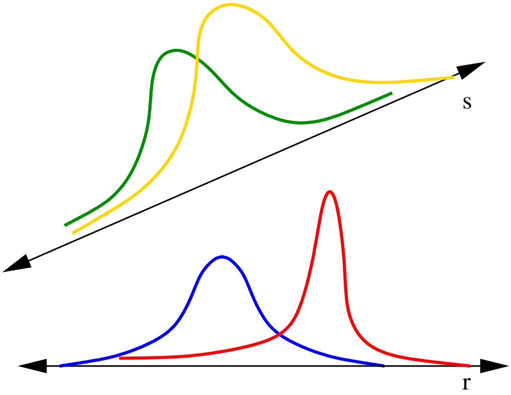
Searching Nearby Directions



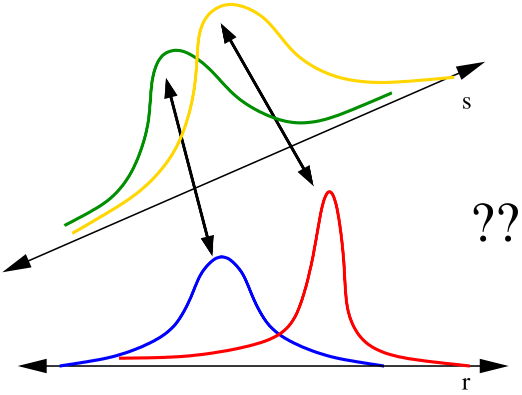
Searching Nearby Directions



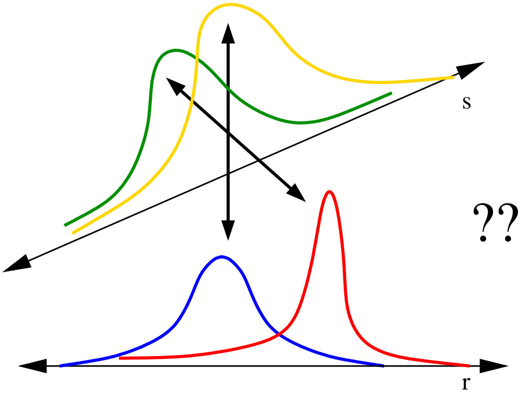
Searching Nearby Directions



Searching Nearby Directions



Searching Nearby Directions



Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

$\hat{F}_1^r, \hat{F}_1^s$ each yield constraints on multidimensional parameters of one Gaussian in F

Problem

How do we know that they yield constraints on the **SAME** *Gaussian?*

Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

$\hat{F}_1^r, \hat{F}_1^s$ each yield constraints on multidimensional parameters of one Gaussian in F

Problem

How do we know that they yield constraints on the **SAME** *Gaussian?*

Pairing Lemma: If we choose directions close enough, then pairing becomes easy

Suppose we learn estimates $\hat{F}^r, \hat{F}^s$ from directions r, s

$\hat{F}_1^r, \hat{F}_1^s$ each yield constraints on multidimensional parameters of one Gaussian in F

Problem

How do we know that they yield constraints on the **SAME** *Gaussian?*

Pairing Lemma: If we choose directions close enough, then pairing becomes easy

("close enough" depends on the **Isotropic Projection Lemma**)

Pairing Lemma

Suppose $\|r - s\| \leq \epsilon_2$ (for $\epsilon_2 << \epsilon_3$)

Pairing Lemma

Suppose $\|r - s\| \leq \epsilon_2$ (for $\epsilon_2 << \epsilon_3$)

We still assume $F = w_1 F_1 + w_2 F_2$ is in isotropic position and is ϵ-statistically learnable

Pairing Lemma

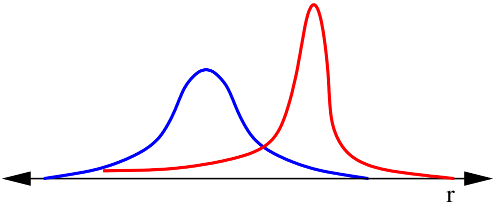
Suppose $\|r - s\| \leq \epsilon_2$ (for $\epsilon_2 << \epsilon_3$)

We still assume $F = w_1 F_1 + w_2 F_2$ is in isotropic position and is ϵ-statistically learnable

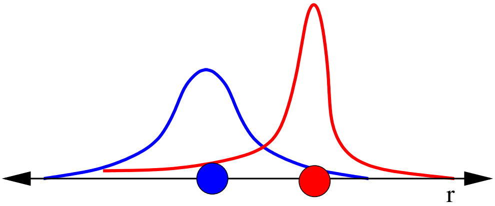
Lemma (Pairing Lemma)

$D_p(Proj_r[F_i], Proj_s[F_i]) \leq O(\frac{\epsilon_2}{\epsilon}) << \frac{\epsilon_3}{3}$

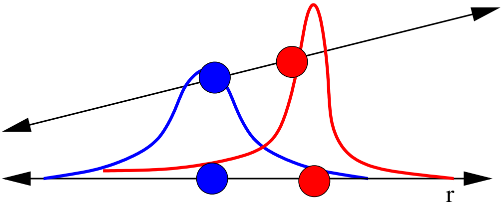
Searching Nearby Directions



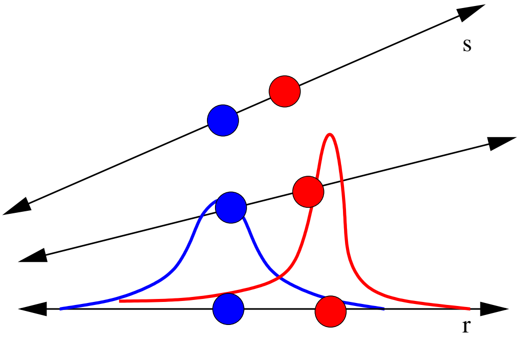
Searching Nearby Directions



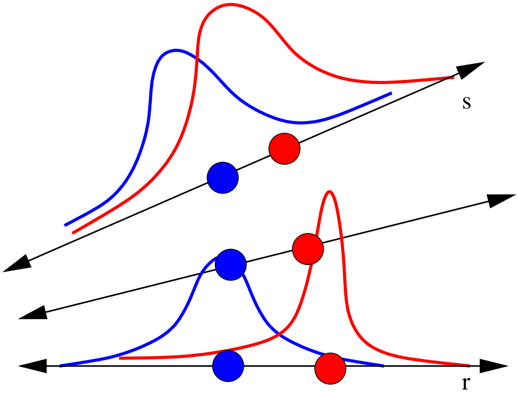
Searching Nearby Directions



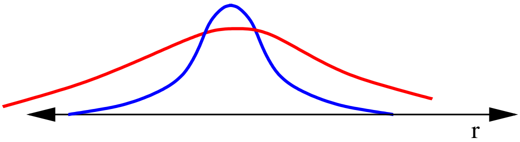
Searching Nearby Directions



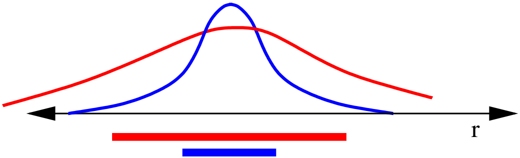
Searching Nearby Directions



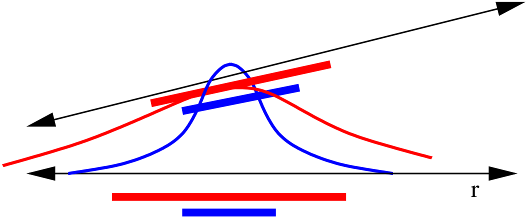
Searching Nearby Directions



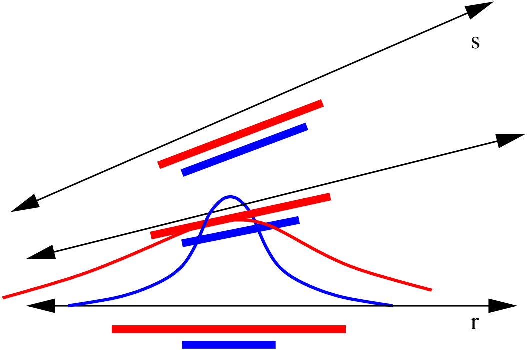
Searching Nearby Directions



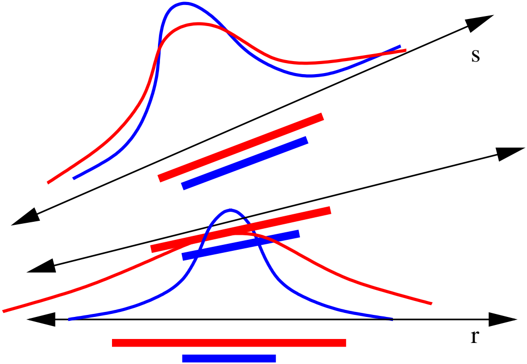
Searching Nearby Directions



Searching Nearby Directions



Searching Nearby Directions



Each univariate estimate yields a linear constraint on the parameters:

$$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r)$$

Each univariate estimate yields a linear constraint on the parameters:

$$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r)$$

Problem

What is the condition number of this system?

Each univariate estimate yields a linear constraint on the parameters:

$$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r)$$

Problem

What is the condition number of this system? (i.e. How do errors in univariate estimates translate to errors in multidimensional estimates?)

Each univariate estimate yields a linear constraint on the parameters:

$$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r)$$

Problem

What is the condition number of this system? (i.e. How do errors in univariate estimates translate to errors in multidimensional estimates?)

Recovery Lemma: Condition number is polynomially bounded

Each univariate estimate yields a linear constraint on the parameters:

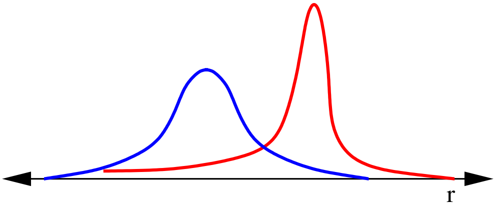
$$Proj_r[F_1] = \mathcal{N}(r^T \mu_1, r^T \Sigma_1 r)$$

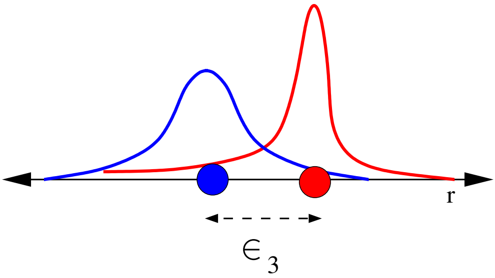
Problem

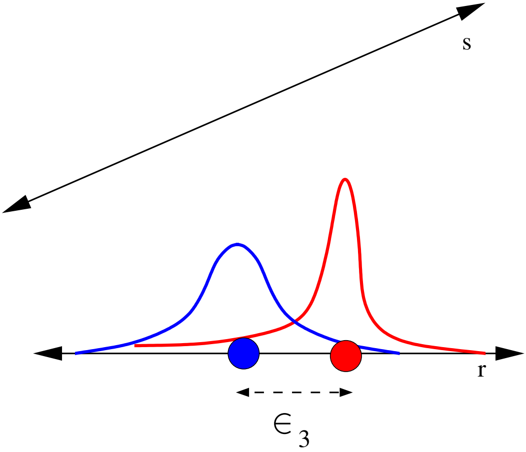
What is the condition number of this system? (i.e. How do errors in univariate estimates translate to errors in multidimensional estimates?)

Recovery Lemma: Condition number is polynomially bounded : $O(\frac{n}{\epsilon_2^2})$

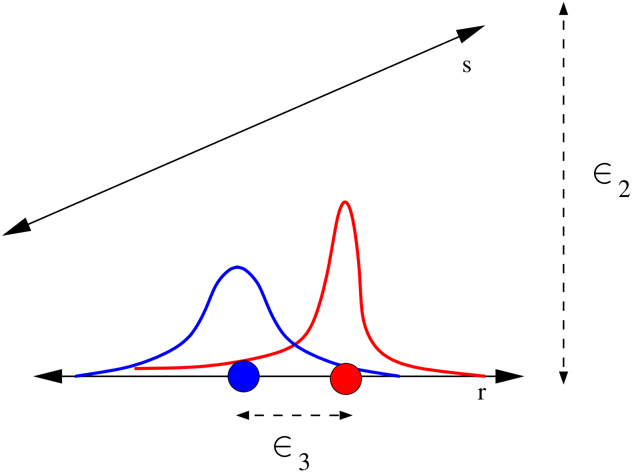

r


r

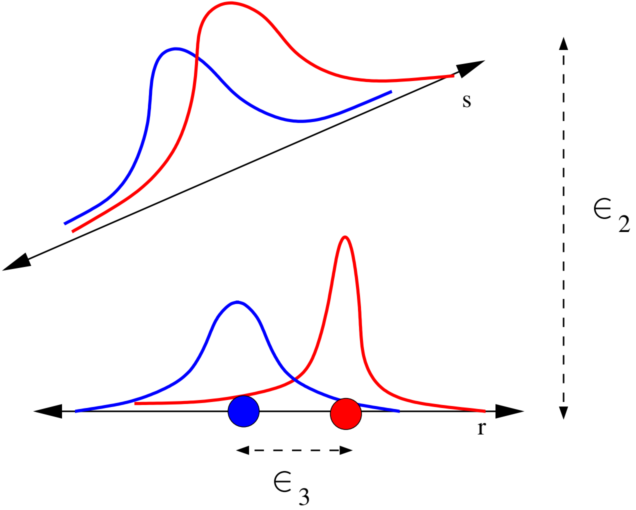

r
∈ 3



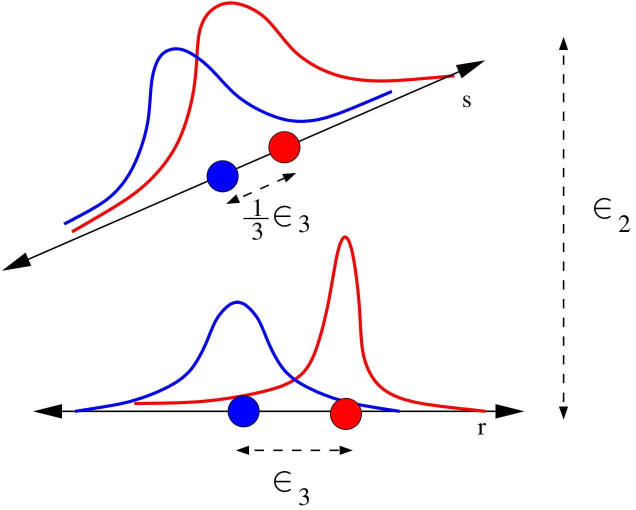
s
r
$\in_3$

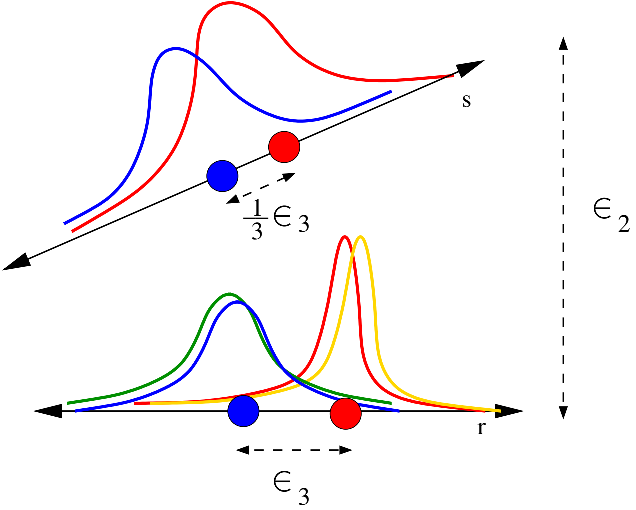


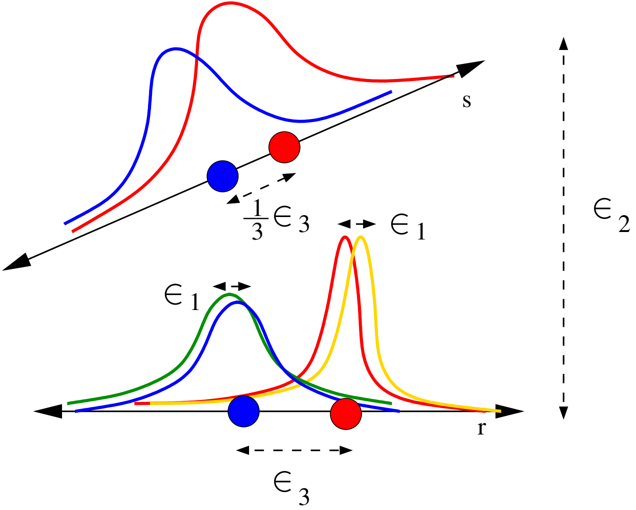
s
$\in_2$
r
$\in_3$



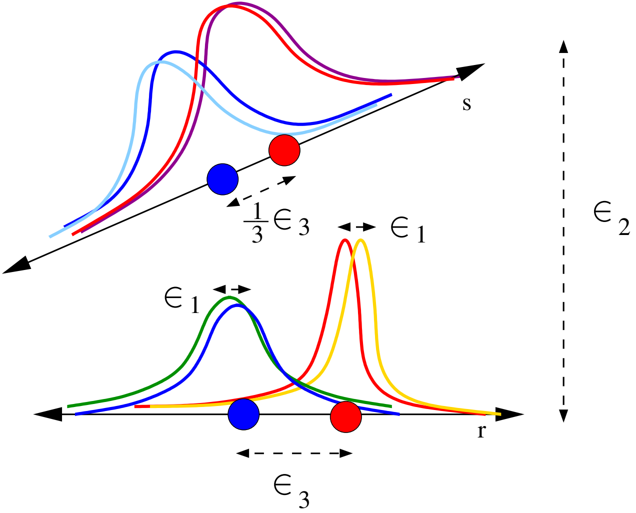
s
r
$\in_2$
$\in_3$


s
$\frac{1}{3}\in_3$
$\in_2$
r
$\in_3$

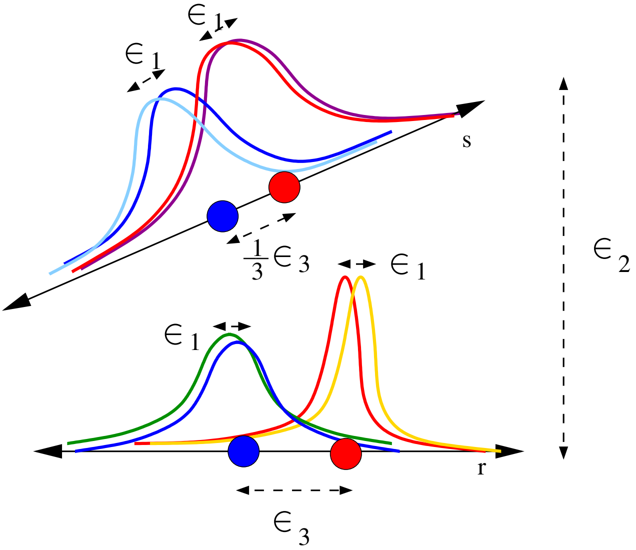

s
$\frac{1}{3}\in_3$
$\in_2$
r
$\in_3$



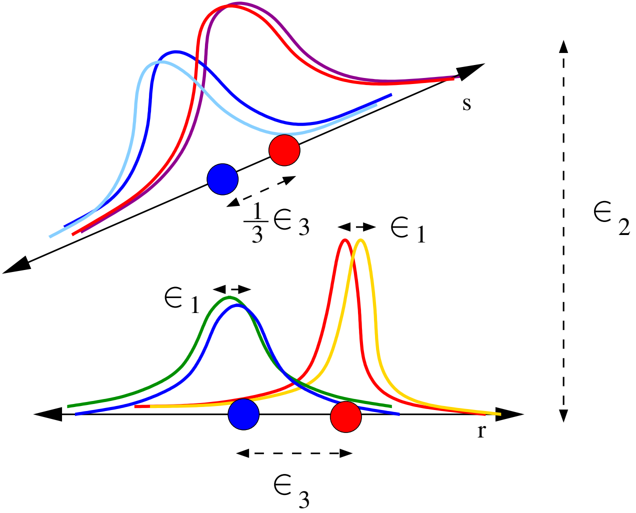
s
$\frac{1}{3}\epsilon_3$
ϵ_1
ϵ_2
ϵ_1
r
ϵ_3



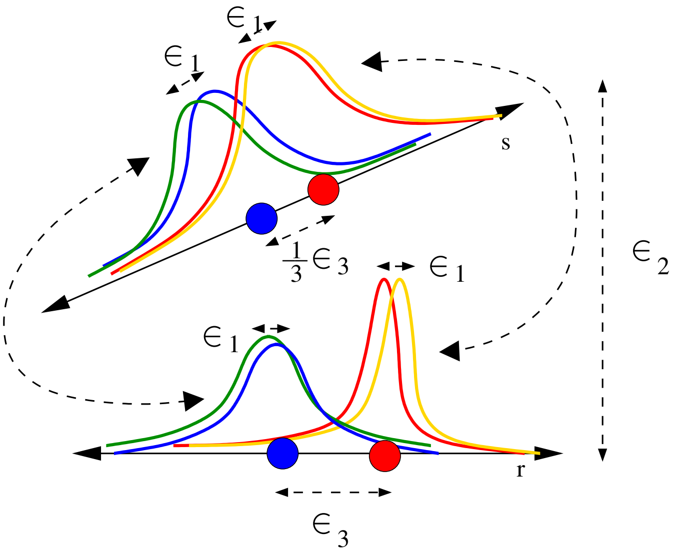
s
$\frac{1}{3}\in_3$
$\in_1$
$\in_2$
$\in_1$
r
$\in_3$



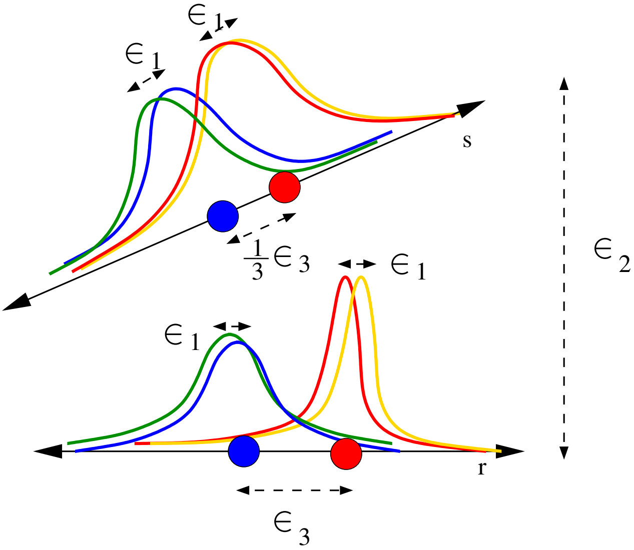
ϵ_1
ϵ_1
s
$\frac{1}{3}\epsilon_3$
ϵ_1
ϵ_2
ϵ_1
r
ϵ_3



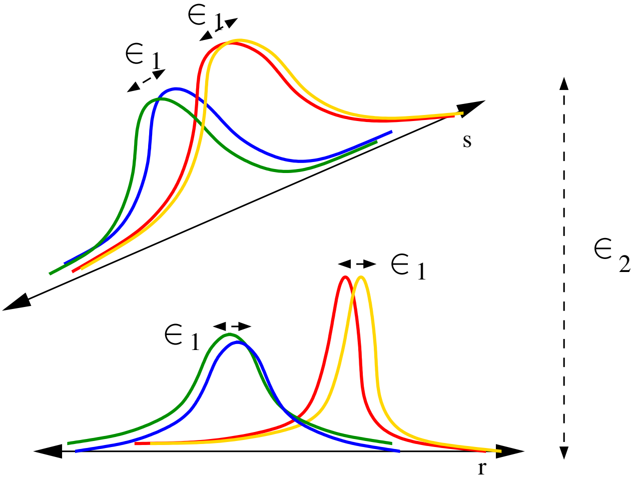
s
$\frac{1}{3}\in_3$
$\in_1$
$\in_2$
$\in_1$
r
$\in_3$



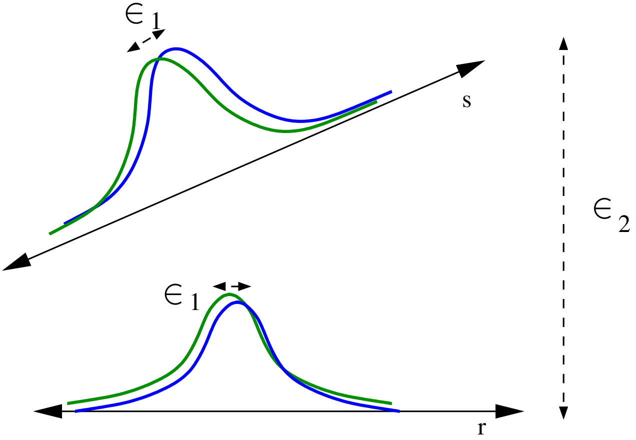
$\in_1$
$\in_1$
s
$\frac{1}{3}\in_3$
$\in_1$
$\in_2$
$\in_1$
r
$\in_3$



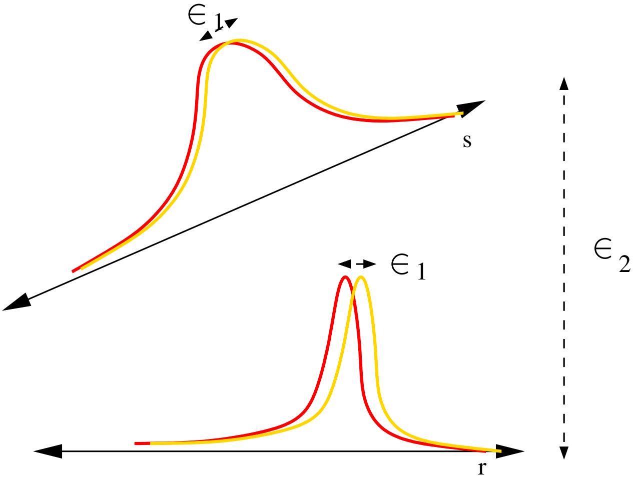
ϵ_1
ϵ_1
s
$\frac{1}{3}\epsilon_3$
ϵ_1
ϵ_2
ϵ_1
r
ϵ_3



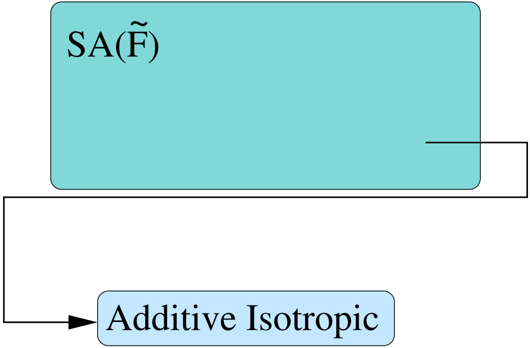
∈ 1
∈ 1
s
∈ 2
∈ 1
∈ 1
r



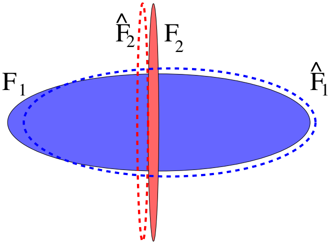
$\in_1$
s
$\in_2$
$\in_1$
r



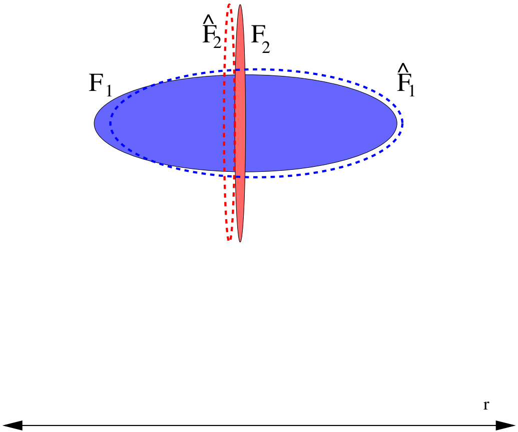
$\in_1$
s
$\in_1$
$\in_2$
r


SA($\tilde{F}$)
Additive Isotropic

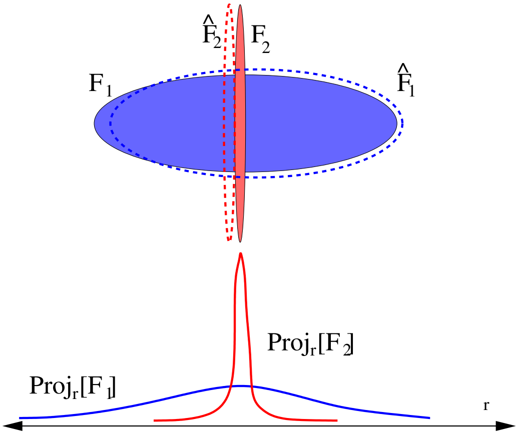
Using Additive Estimates to Cluster



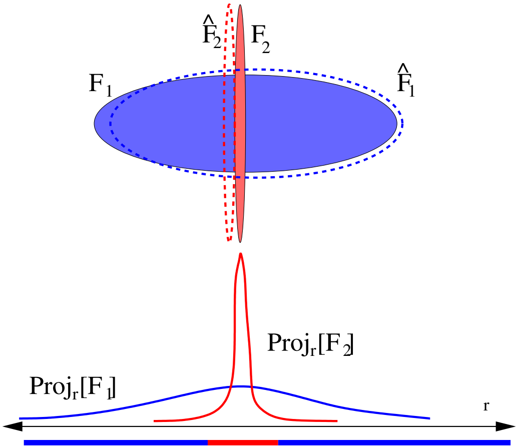
Using Additive Estimates to Cluster



Using Additive Estimates to Cluster

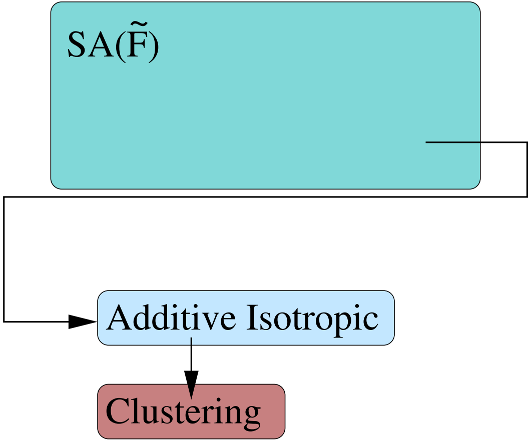


Using Additive Estimates to Cluster

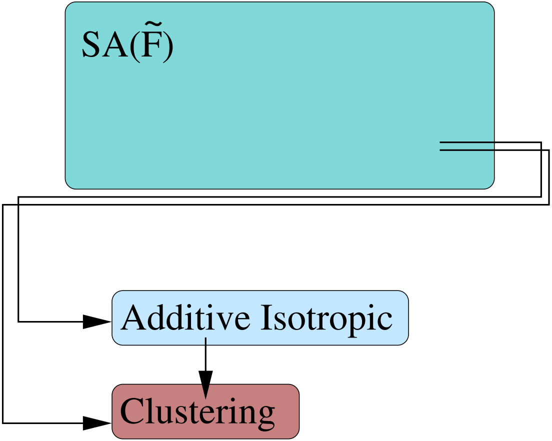


SA($\tilde{F}$)

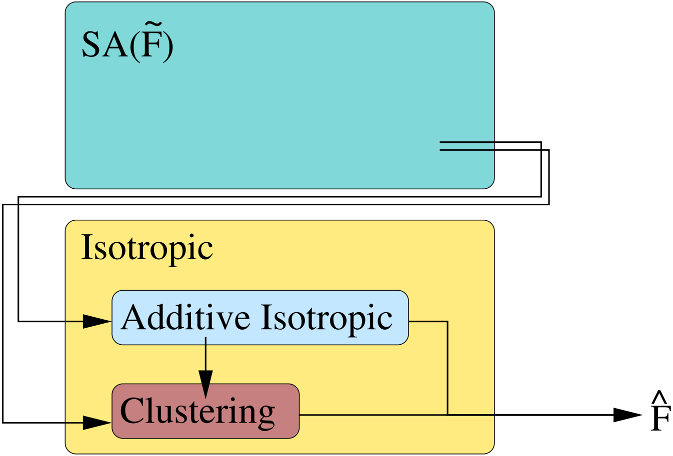
Additive Isotropic

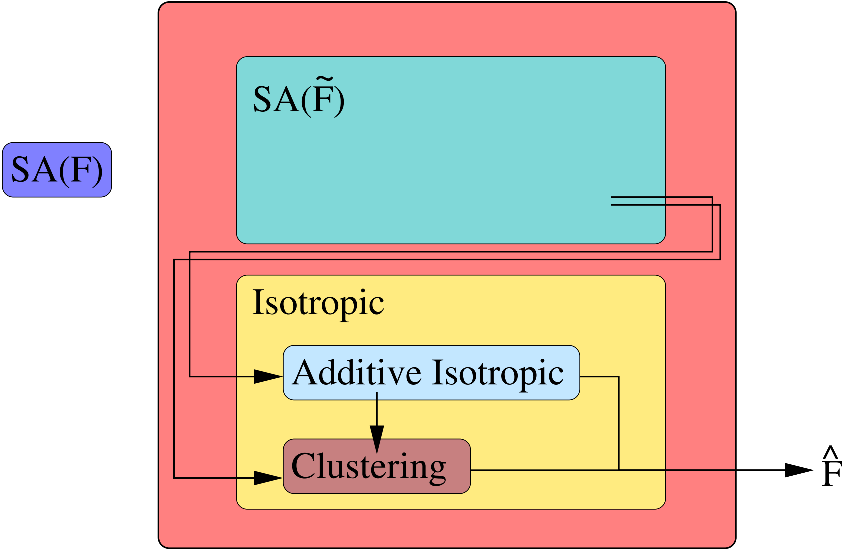


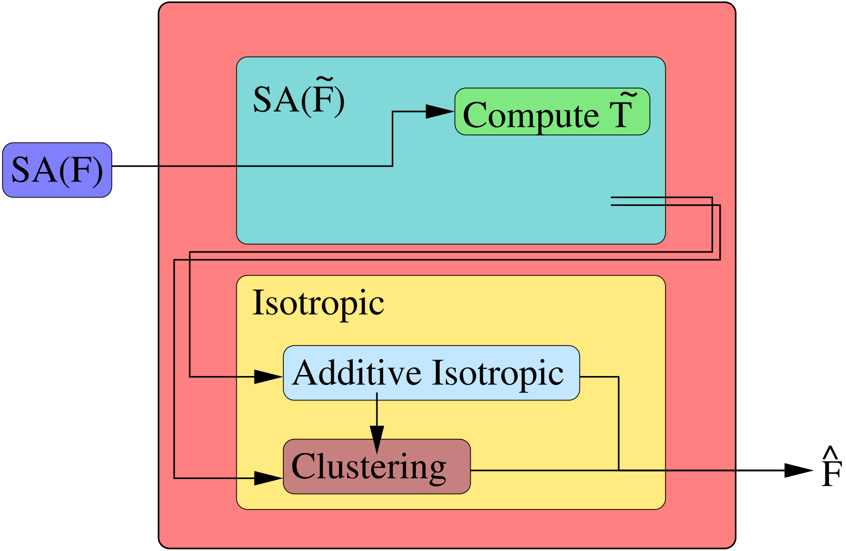
SA($\tilde{F}$)
Additive Isotropic
Clustering



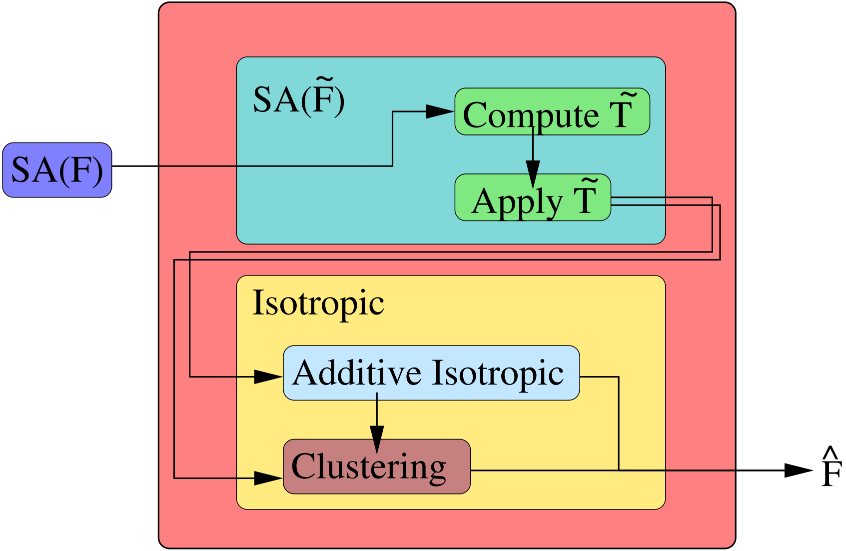
SA(F̃)
Additive Isotropic
Clustering

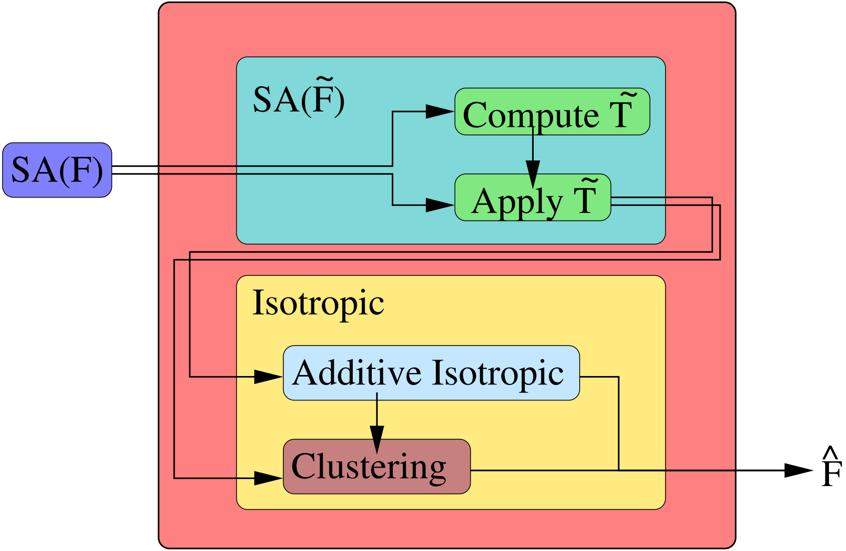

SA($\tilde{F}$)
Isotropic
Additive Isotropic
Clustering
$\hat{F}$


SA(F)
SA($\tilde{F}$)
Isotropic
Additive Isotropic
Clustering
$\hat{F}$



SA(F)
SA($\tilde{F}$)
Compute $\tilde{T}$
Isotropic
Additive Isotropic
Clustering
$\hat{F}$


SA(F)
SA($\tilde{F}$)
Compute $\tilde{T}$
Apply $\tilde{T}$
Isotropic
Additive Isotropic
Clustering
$\hat{F}$


SA(F)
SA($\tilde{F}$)
Compute $\tilde{T}$
Apply $\tilde{T}$
Isotropic
Additive Isotropic
Clustering
$\hat{F}$

Question

Can we learn an additive approximation in one dimension?

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

Additionally, each parameter is bounded:

Claim

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

Additionally, each parameter is bounded:

Claim

1. $w_1, w_2 \in [\epsilon, 1]$

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

Additionally, each parameter is bounded:

Claim

1. $w_1, w_2 \in [\epsilon, 1]$
2. $|\mu_1|, |\mu_2| \leq \frac{1}{\sqrt{\epsilon}}$

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

Additionally, each parameter is bounded:

Claim

1. $w_1, w_2 \in [\epsilon, 1]$
2. $|\mu_1|, |\mu_2| \leq \frac{1}{\sqrt{\epsilon}}$
3. $\sigma_1^2, \sigma_2^2 \leq \frac{1}{\epsilon}$

Question

Can we learn an additive approximation in one dimension? How many free parameters are there?

$$\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1$$

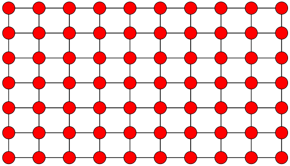
Additionally, each parameter is bounded:

Claim

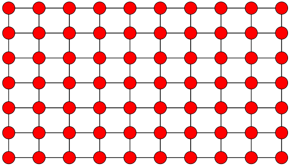
1. $w_1, w_2 \in [\epsilon, 1]$
2. $|\mu_1|, |\mu_2| \leq \frac{1}{\sqrt{\epsilon}}$
3. $\sigma_1^2, \sigma_2^2 \leq \frac{1}{\epsilon}$

In this case, we call the parameters ϵ-bounded

So we can use a grid search over $\hat{\mu}_1 \times \hat{\sigma}_1^2 \times \hat{\mu}_2 \times \hat{\sigma}_2^2 \times \hat{w}_1$



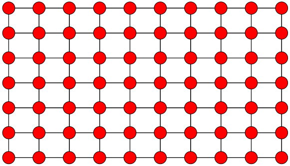
So we can use a grid search over $\hat{\mu}_1 \times \hat{\sigma}_1^2 \times \hat{\mu}_2 \times \hat{\sigma}_2^2 \times \hat{w}_1$



Question

How do we test if a candidate set of parameters is accurate?

So we can use a grid search over $\hat{\mu}_1 \times \hat{\sigma}_1^2 \times \hat{\mu}_2 \times \hat{\sigma}_2^2 \times \hat{w}_1$

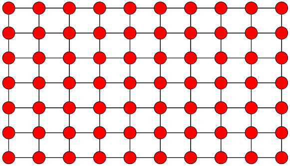


Question

How do we test if a candidate set of parameters is accurate?

1. Compute empirical moments $r = \{1, 2, ...6\}$: $\tilde{M}_r = \frac{1}{|S|}\sum_{i \in S} x_i^r$

So we can use a grid search over $\hat{\mu}_1 \times \hat{\sigma}_1^2 \times \hat{\mu}_2 \times \hat{\sigma}_2^2 \times \hat{w}_1$

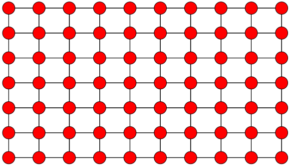


Question

How do we test if a candidate set of parameters is accurate?

1. Compute empirical moments $r = \{1, 2, ...6\}$: $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$
2. Compute the analytical moments $M_r(\hat{F}) = E_{x \leftarrow \hat{F}}[x^r]$ where $\hat{F} = \hat{w}_1 \mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1^2, x) + \hat{w}_2 \mathcal{N}(\hat{\mu}_2, \hat{\sigma}_2^2, x)$ for $r \in \{1, 2, ..., 6\}$

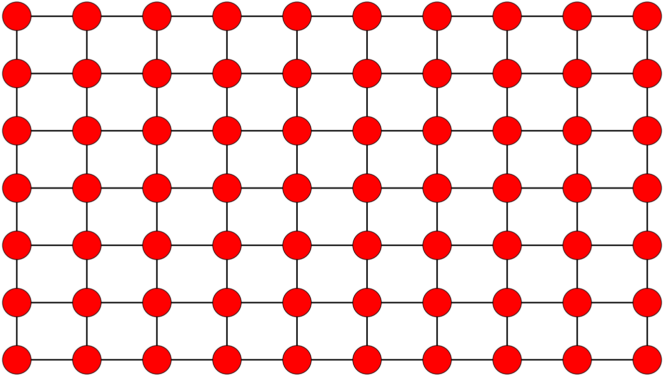
So we can use a grid search over $\hat{\mu}_1 \times \hat{\sigma}_1^2 \times \hat{\mu}_2 \times \hat{\sigma}_2^2 \times \hat{w}_1$

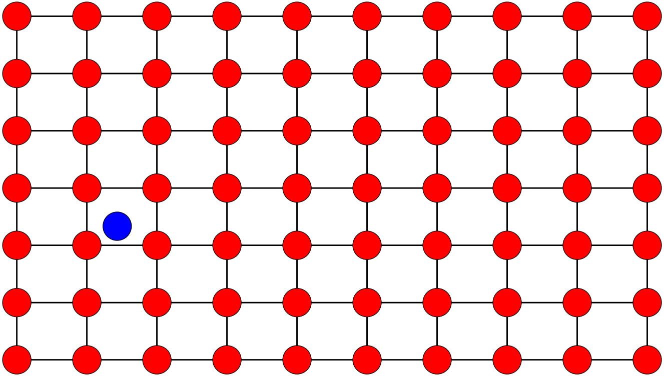


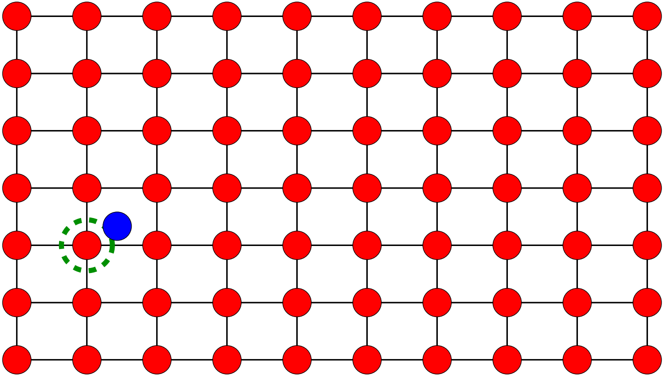
Question

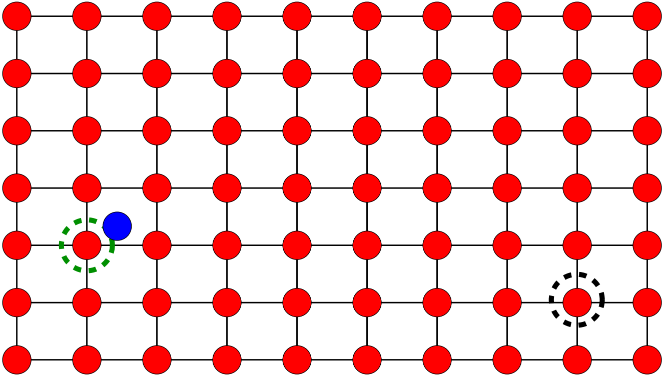
How do we test if a candidate set of parameters is accurate?

1. Compute empirical moments $r = \{1, 2, \ldots 6\}$: $\tilde{M}_r = \frac{1}{|S|} \sum_{i \in S} x_i^r$
2. Compute the analytical moments $M_r(\hat{F}) = E_{x \leftarrow \hat{F}}[x^r]$ where $\hat{F} = \hat{w}_1 \mathcal{N}(\hat{\mu}_1, \hat{\sigma}_1^2, x) + \hat{w}_2 \mathcal{N}(\hat{\mu}_2, \hat{\sigma}_2^2, x)$ for $r \in \{1, 2, \ldots, 6\}$
3. Accept if $\tilde{M}_r \approx M_r(\hat{F})$ for all $r \in \{1, 2, \ldots, 6\}$









Definition

The pair $F, \hat{F}$ ϵ-standard if

1. the parameters of $F, \hat{F}$ are ϵ-bounded

Definition

The pair $F, \hat{F}$ ϵ-standard if

1. the parameters of $F, \hat{F}$ are ϵ-bounded
2. $D_p(F_1, F_2), D_p(\hat{F}_1, \hat{F}_2) \geq \epsilon$

Definition

The pair $F, \hat{F}$ ϵ-standard if

1. the parameters of $F, \hat{F}$ are ϵ-bounded
2. $D_p(F_1, F_2), D_p(\hat{F}_1, \hat{F}_2) \geq \epsilon$
3. $\epsilon \leq \min_\pi \sum_i \Big(|w_i - \hat{w}_{\pi(i)}| + D_p(F_i, \hat{F}_{\pi(i)}) \Big)$

Definition

The pair $F, \hat{F}$ ϵ-standard if

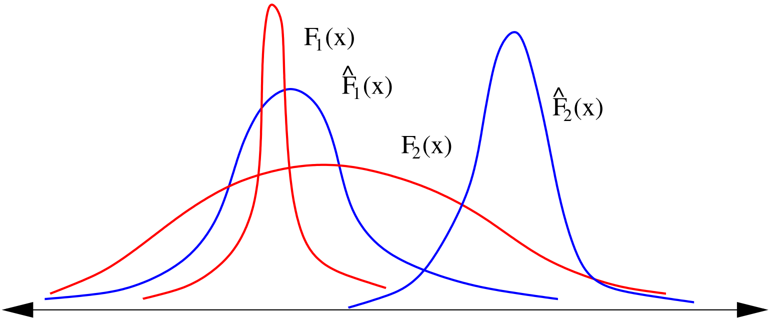
1. the parameters of $F, \hat{F}$ are ϵ-bounded
2. $D_p(F_1, F_2), D_p(\hat{F}_1, \hat{F}_2) \geq \epsilon$
3. $\epsilon \leq \min_\pi \sum_i \left(|w_i - \hat{w}_{\pi(i)}| + D_p(F_i, \hat{F}_{\pi(i)}) \right)$

Theorem

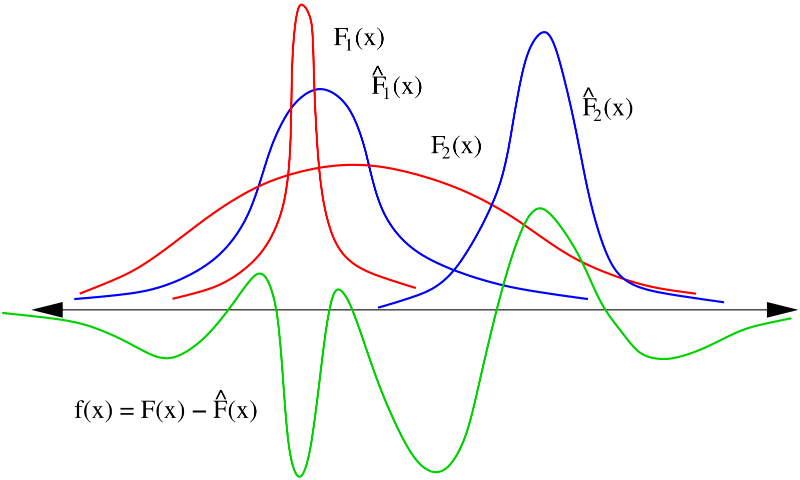
There is a constant $c > 0$ such that, for any any $\epsilon < c$ and any ϵ-standard $F, \hat{F}$,

$$\max_{r \in \{1,2,\ldots,6\}} |M_r(F) - M_r(\hat{F})| \geq \epsilon^{67}$$

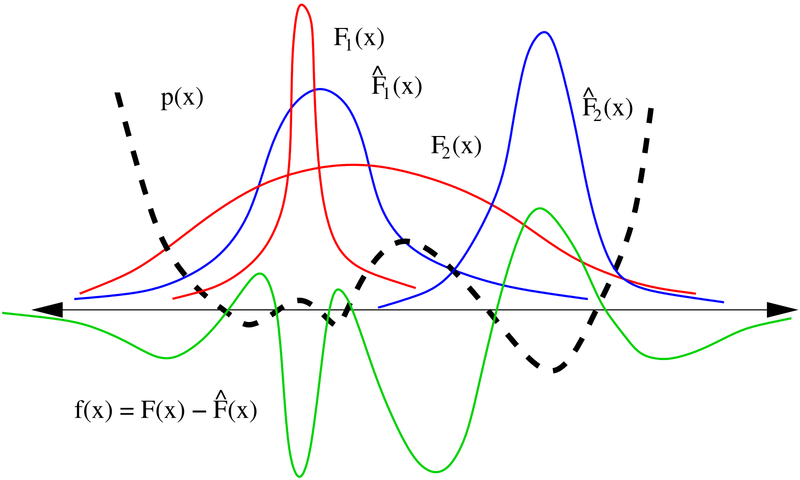
Method of Moments



Method of Moments



Method of Moments



Question

Why does this imply one of the first six moment of $F, \hat{F}$ is different?

$$0 < \left| \int_x p(x) f(x) dx \right|$$

Question

Why does this imply one of the first six moment of $F, \hat{F}$ is different?

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

Question

Why does this imply one of the first six moment of $F, \hat{F}$ is different?

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

Question

Why does this imply one of the first six moment of $F, \hat{F}$ is different?

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

So $\exists_{r \in \{1,2,\ldots,6\}}$ s.t. $|M_r(F) - M_r(\hat{F})| > 0$

Proposition

Let $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *be a linear combination of* k *Gaussians (*α_i *can be negative). Then if* $f(x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings.*

Proposition

Let $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *be a linear combination of* k *Gaussians* (α_i *can be negative). Then if* $f(x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings.*

Theorem (Hummel, Gidas)

Given $f(x) : \Re \rightarrow \Re$, *that is analytic and has* n *zeros, then for any* $\sigma^2 > 0$, *the function* $g(x) = f(x) \circ \mathcal{N}(0, \sigma^2, x)$ *has at most* n *zeros.*

Proposition

Let $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *be a linear combination of* k *Gaussians (*α_i *can be negative). Then if* $f(x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings.*

Theorem (Hummel, Gidas)

Given $f(x) : \Re \rightarrow \Re$*, that is analytic and has* n *zeros, then for any* $\sigma^2 > 0$*, the function* $g(x) = f(x) \circ \mathcal{N}(0, \sigma^2, x)$ *has at most* n *zeros.*

Convolving by a Gaussian does not increase the number of zero crossings!

Proposition

Let $f(x) = \sum_{i=1}^{k} \alpha_i \mathcal{N}(\mu_i, \sigma_i^2, x)$ *be a linear combination of* k *Gaussians* (α_i *can be negative). Then if* $f(x)$ *is not identically zero,* $f(x)$ *has at most* $2k - 2$ *zero crossings.*

Theorem (Hummel, Gidas)

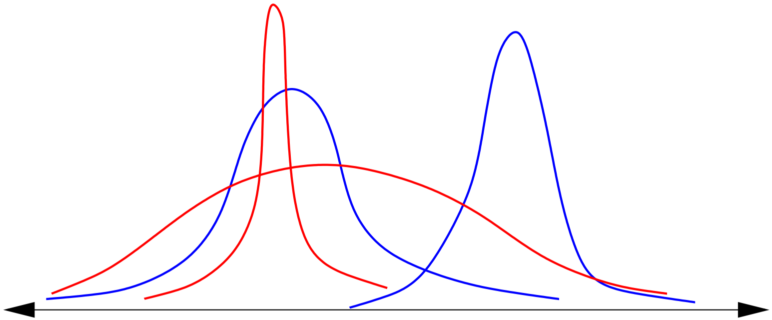
Given $f(x) : \Re \to \Re$*, that is analytic and has* n *zeros, then for any* $\sigma^2 > 0$*, the function* $g(x) = f(x) \circ \mathcal{N}(0, \sigma^2, x)$ *has at most* n *zeros.*

Convolving by a Gaussian does not increase the number of zero crossings!

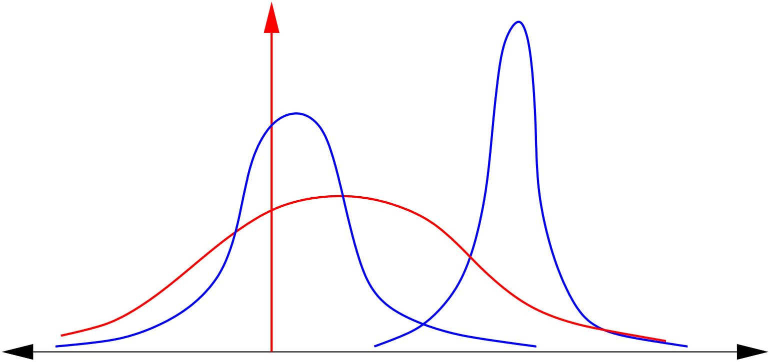
Fact

$\mathcal{N}(0, \sigma_1^2, x) \circ \mathcal{N}(0, \sigma_2^2, x) = \mathcal{N}(0, \sigma_1^2 + \sigma_2^2, x)$

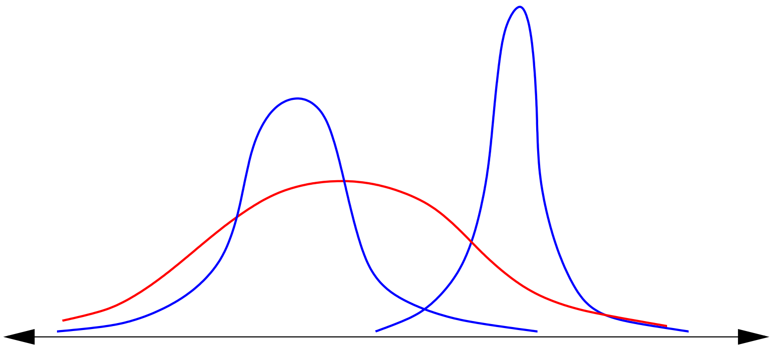
Zero Crossings and the Heat Equation



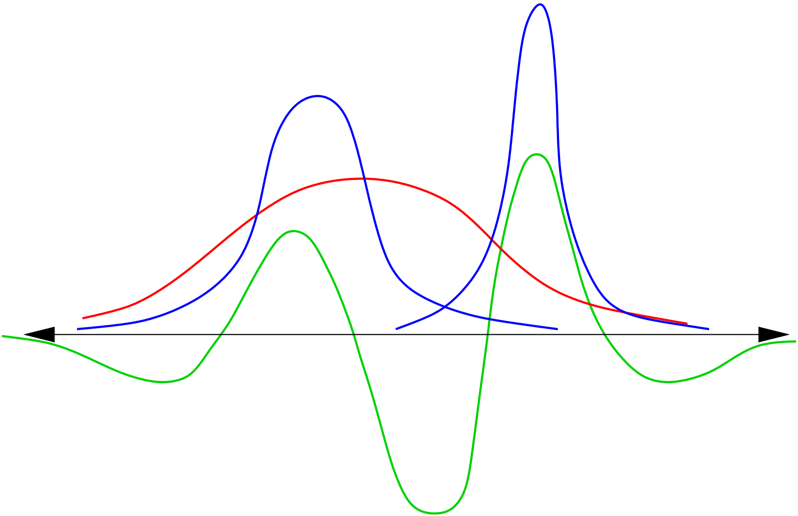
Zero Crossings and the Heat Equation



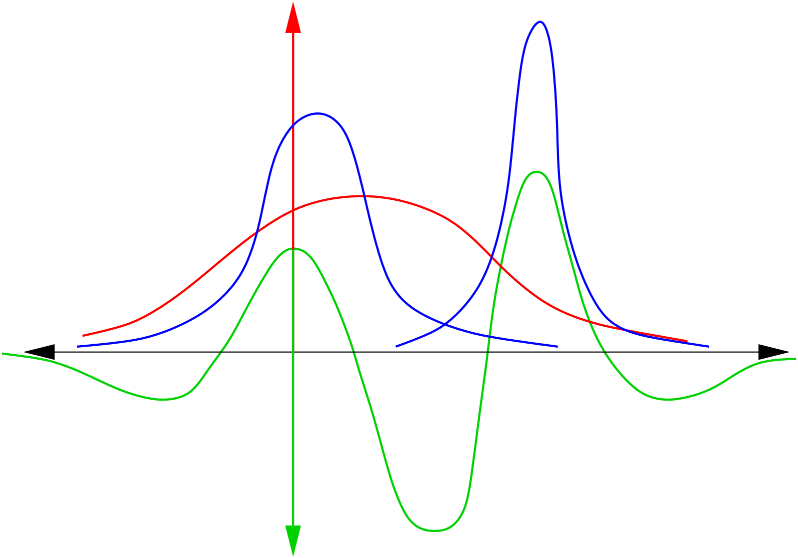
Zero Crossings and the Heat Equation



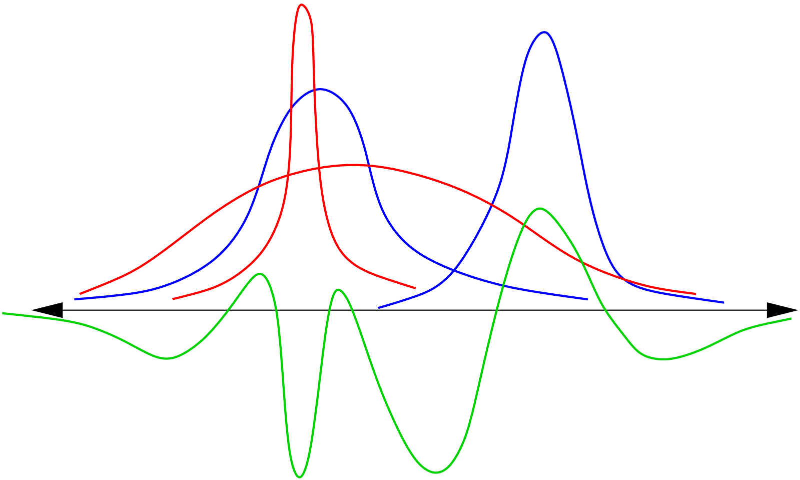
Zero Crossings and the Heat Equation

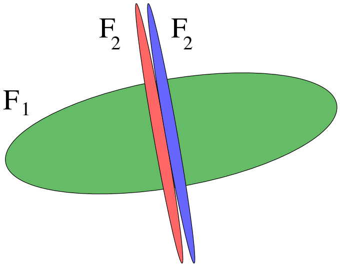


Zero Crossings and the Heat Equation

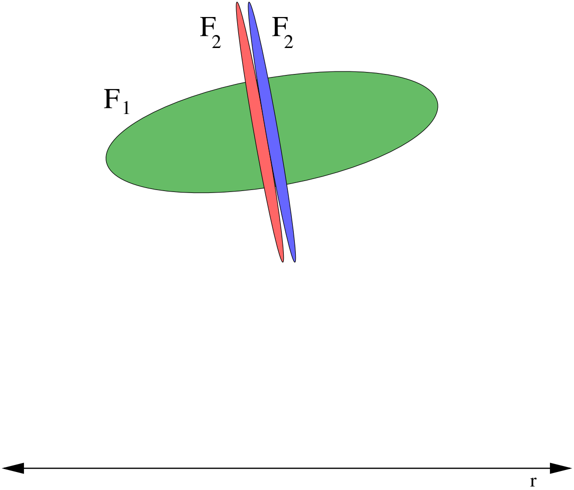


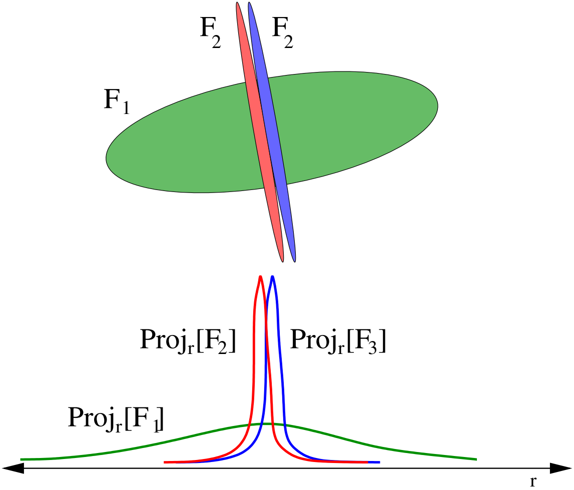
Zero Crossings and the Heat Equation



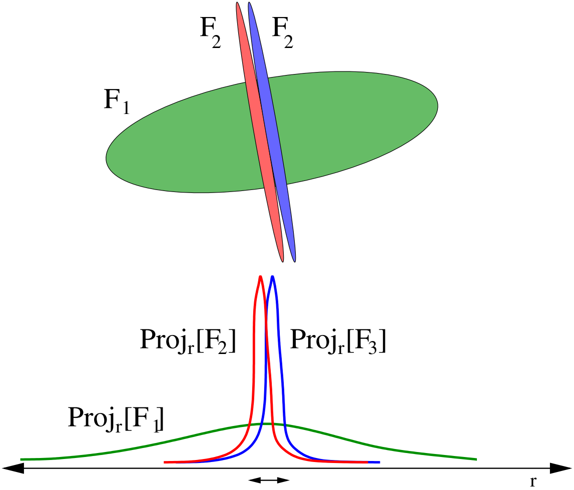


F_2
F_2
F_1

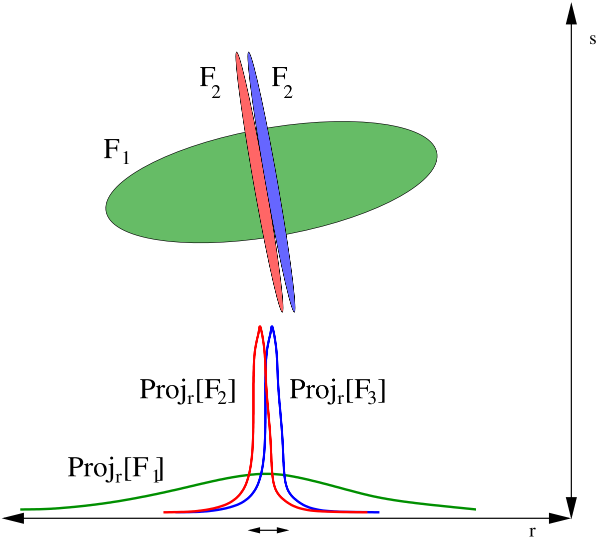

F_2
F_2
F_1
r

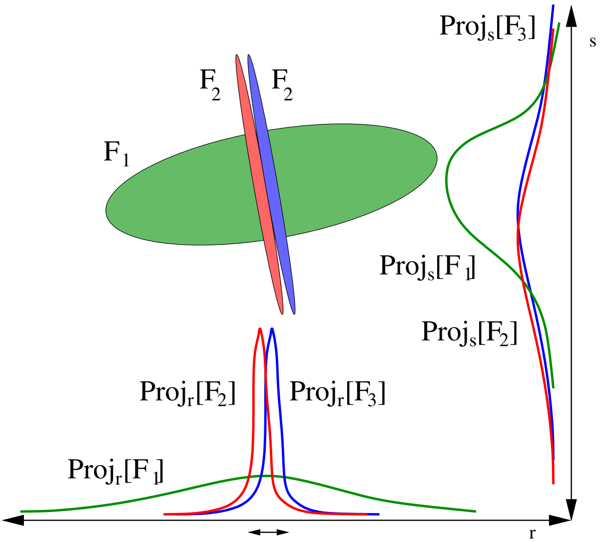


F_2
F_2
F_1
$\mathrm{Proj}_r[F_2]$
$\mathrm{Proj}_r[F_3]$
$\mathrm{Proj}_r[F_1]$
r

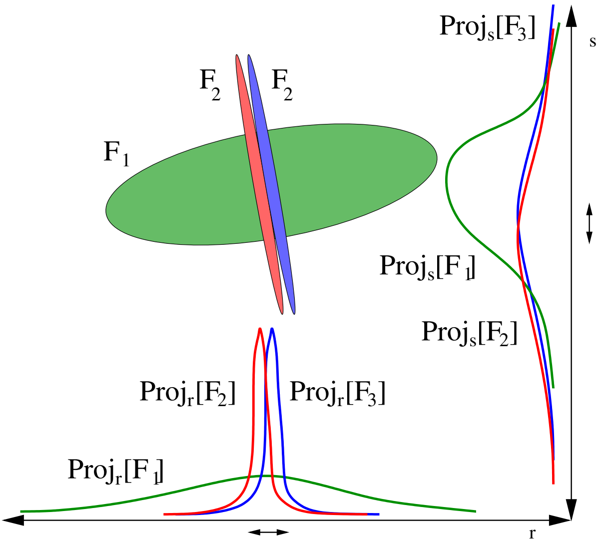


F_2
F_2
F_1
$\mathrm{Proj}_r[F_2]$
$\mathrm{Proj}_r[F_3]$
$\mathrm{Proj}_r[F_1]$
r


s
F2
F2
F1
Proj_r[F2]
Proj_r[F3]
Proj_r[F1]
r



Proj$_s$[F$_3$]
s
F$_2$
F$_2$
F$_1$
Proj$_s$[F$_1$]
Proj$_s$[F$_2$]
Proj$_r$[F$_2$]
Proj$_r$[F$_3$]
Proj$_r$[F$_1$]
r



Proj$_s$[F$_3$]
s
F$_2$
F$_2$
F$_1$
Proj$_s$[F$_1$]
Proj$_s$[F$_2$]
Proj$_r$[F$_2$]
Proj$_r$[F$_3$]
Proj$_r$[F$_1$]
r

Generalized Isotropic Projection Lemma

Lemma (Generalized Isotropic Projection Lemma)

With probability $\geq 1 - \delta$ *over a randomly chosen direction* r*, for all* $i \neq j$*,* $D_p(Proj_r[F_i], Proj_r[F_j]) \geq \epsilon_3$.

Generalized Isotropic Projection Lemma

Lemma (Generalized Isotropic Projection Lemma)

With probability $\geq 1 - \delta$ *over a randomly chosen direction* r*, for all* $i \neq j$*,* $D_p(Proj_r[F_i], Proj_r[F_j]) \geq \epsilon_3$.

FALSE!

Generalized Isotropic Projection Lemma

Lemma (Generalized Isotropic Projection Lemma)

With probability $\geq 1 - \delta$ *over a randomly chosen direction* r, *for all* $i \neq j$, $D_p(Proj_r[F_i], Proj_r[F_j]) \geq \epsilon_3$.

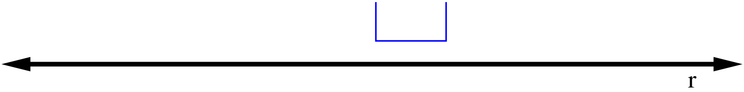
FALSE!

Lemma (Generalized Isotropic Projection Lemma)

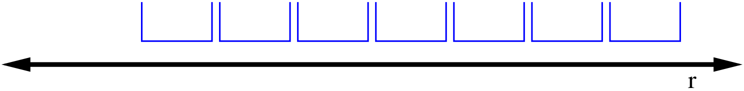
With probability $\geq 1 - \delta$ *over a randomly chosen direction* r, ~~*for all*~~ *there exists* $i \neq j$, $D_p(Proj_r[F_i], Proj_r[F_j]) \geq \epsilon_3$.

Windows

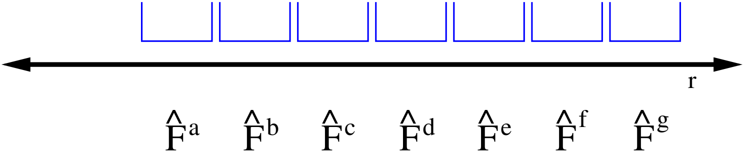
Windows



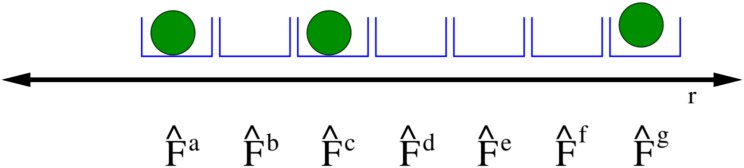
Windows



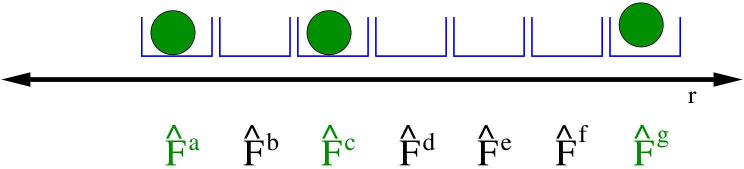
Windows



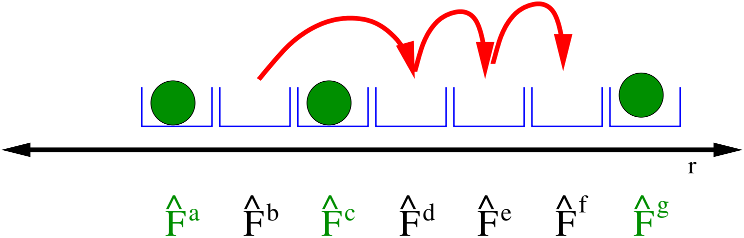
Windows



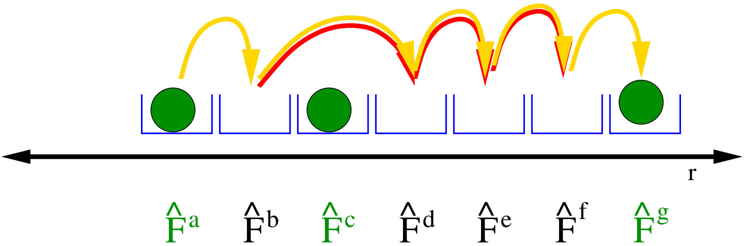
Windows



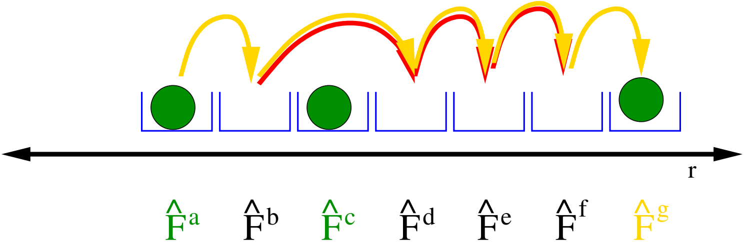
Windows



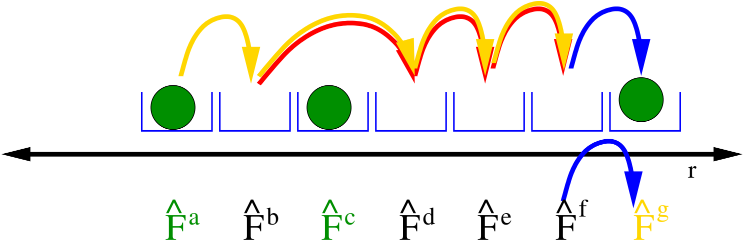
Windows



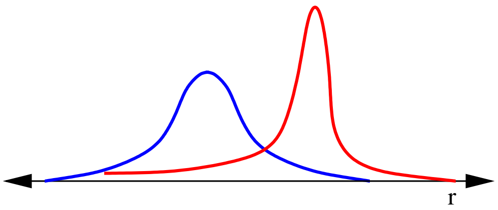
Windows



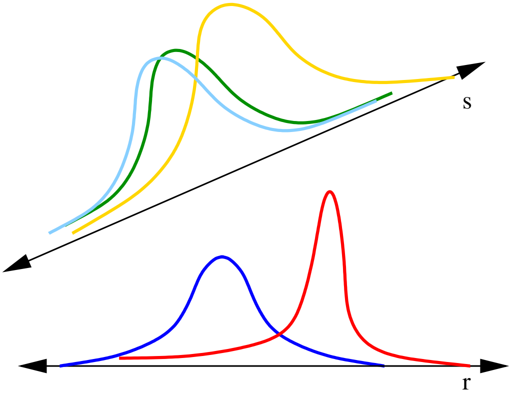
Windows



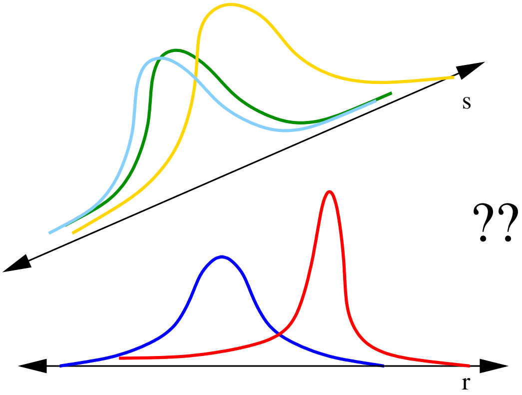
Pairing Lemma, Part 2?



Pairing Lemma, Part 2?



Pairing Lemma, Part 2?



Thanks!