Learning mixtures of Gaussians

Abbas Mehrabian

McGill University
IVADO Postdoctoral Fellow

30 January 2019

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

By KARL PEARSON, *University College, London.*

Communicated by Professor HENRICI, *F.R.*

Received October 18,—Read November 16, 1893.

[PLATES 1—5.]

CONTENTS.

(9.) The whole method may be illustrated by the following numerical example :—

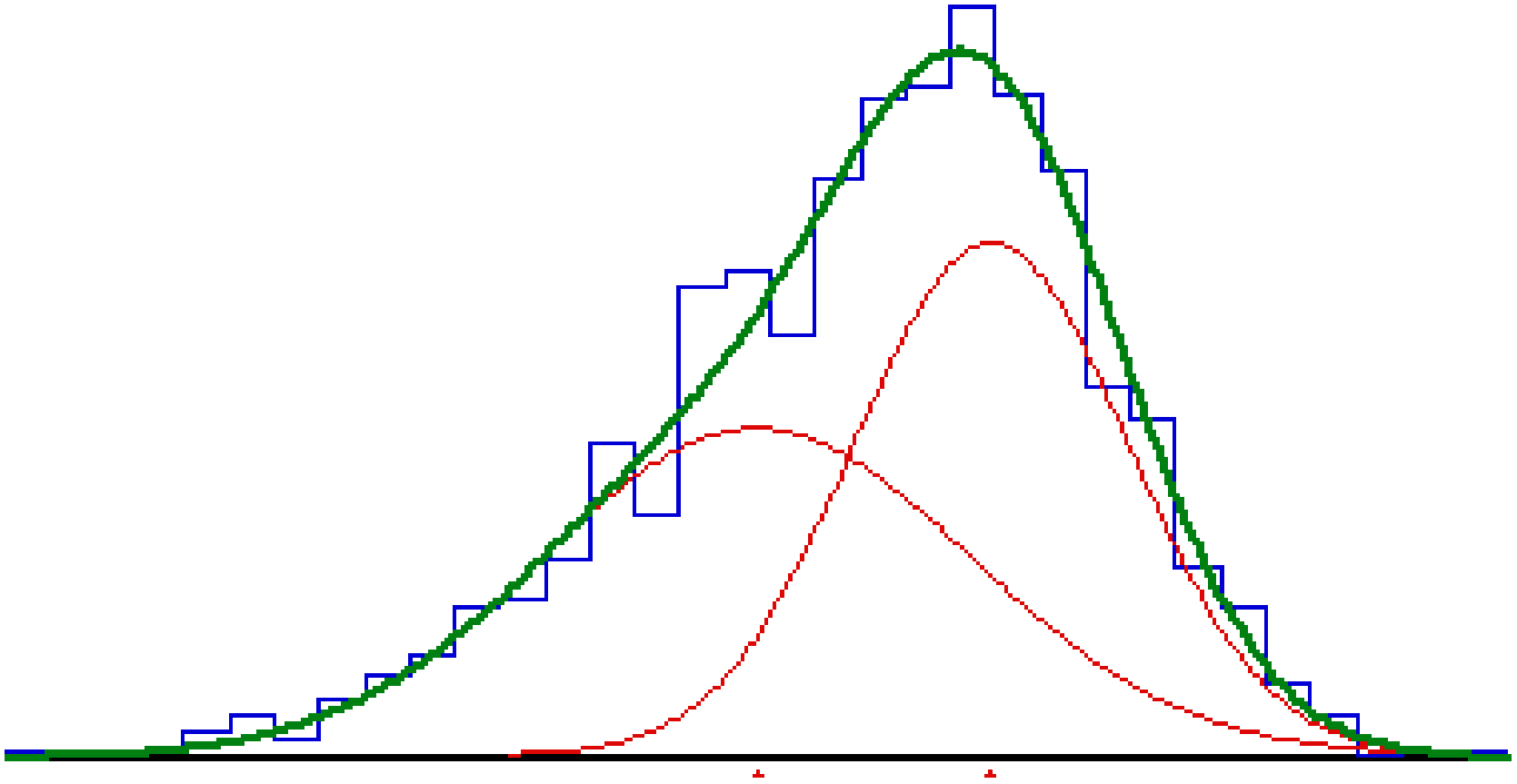
Breadth of "Forehead" of Crabs.—Professor W. F. R. Weldon has very kindly given me the following statistics from among his measurements on crabs. They are for 1000 individuals from Naples. The abscissæ of the curve are the ratio of "forehead" to body-length, and one unit of abscissa = ·004 of body-length. No. 1 of the abscissæ corresponds to ·580 — ·583 of body-length. The ordinates represent the number of individual crabs corresponding to each set of ratios of forehead to body-length. Thus there was one crab fell into the range ·580 — ·583, three fell into the range ·584 — ·587, five into the range ·588 — ·591, and so on. The average length of animals measured 35 millims., and measurements were recorded to ·1 millim.

Abscissæ.	Ordinates.	Abscissæ.	Ordinates.
1	1	16	74
2	3	17	84
3	5	18	86
4	2	19	96
5	7	20	85
6	10	21	75
7	13	22	47
8	19	23	43
9	20	24	24
10	25	25	19
11	40	26	9
12	31	27	5
13	60	28	0
14	62	29	1
15	54		

Observation: data is asymmetric.

Hypothesis: may be a mixture of two Gaussians.

Method: numerically estimated the parameters by matching the moments.



Plot by Peter D. M. Macdonald, McMaster University.

Learning mixtures of Gaussians in modern times

These days trying to fit data with mixtures of Gaussians is popular in data science.

Modern applications: high-dimensional data

Learning mixtures of Gaussians in modern times

These days trying to fit data with mixtures of Gaussians is popular in data science.

Modern applications: high-dimensional data

Why mixtures of Gaussians?

- ✓ fit some natural data well [Pearson 1894] [Redner and Walker 1984] [Brainard and Burmaster 1992]
- ✓ universal approximators
- ✓ clustering

High-dimensional Gaussians

Multivariate normal distribution:

$$\mathcal{N}_{\mu,\Sigma}(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right)}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} \qquad \text{for } x \in \mathbb{R}^d$$

$$X \sim \mathcal{N}_{\mu,\Sigma}\colon\ \mathbb{E}[X] = \mu \in \mathbb{R}^d, \mathbb{E}\left[(X-\mu)(X-\mu)^{\mathsf{T}}\right] = \Sigma \in \mathbb{R}^{d\times d}$$

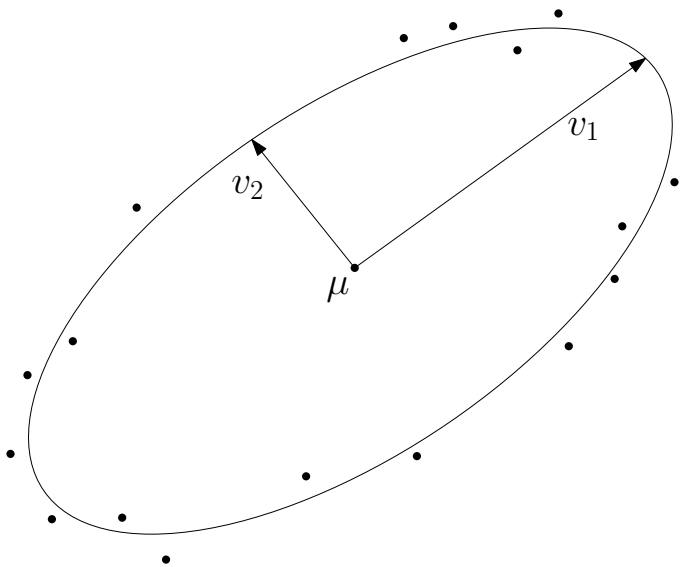
High-dimensional Gaussians

Multivariate normal distribution:

$$\mathcal{N}_{\mu,\Sigma}(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right)}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} \qquad \text{for } x \in \mathbb{R}^d$$

$$X \sim \mathcal{N}_{\mu,\Sigma}\colon\ \mathbb{E}[X] = \mu \in \mathbb{R}^d, \mathbb{E}\left[(X-\mu)(X-\mu)^{\mathsf{T}}\right] = \Sigma \in \mathbb{R}^{d\times d}$$

For large d, with-high-probability X lies in a thin ellipsoidal shell centred at μ with axes being eigenvectors of Σ



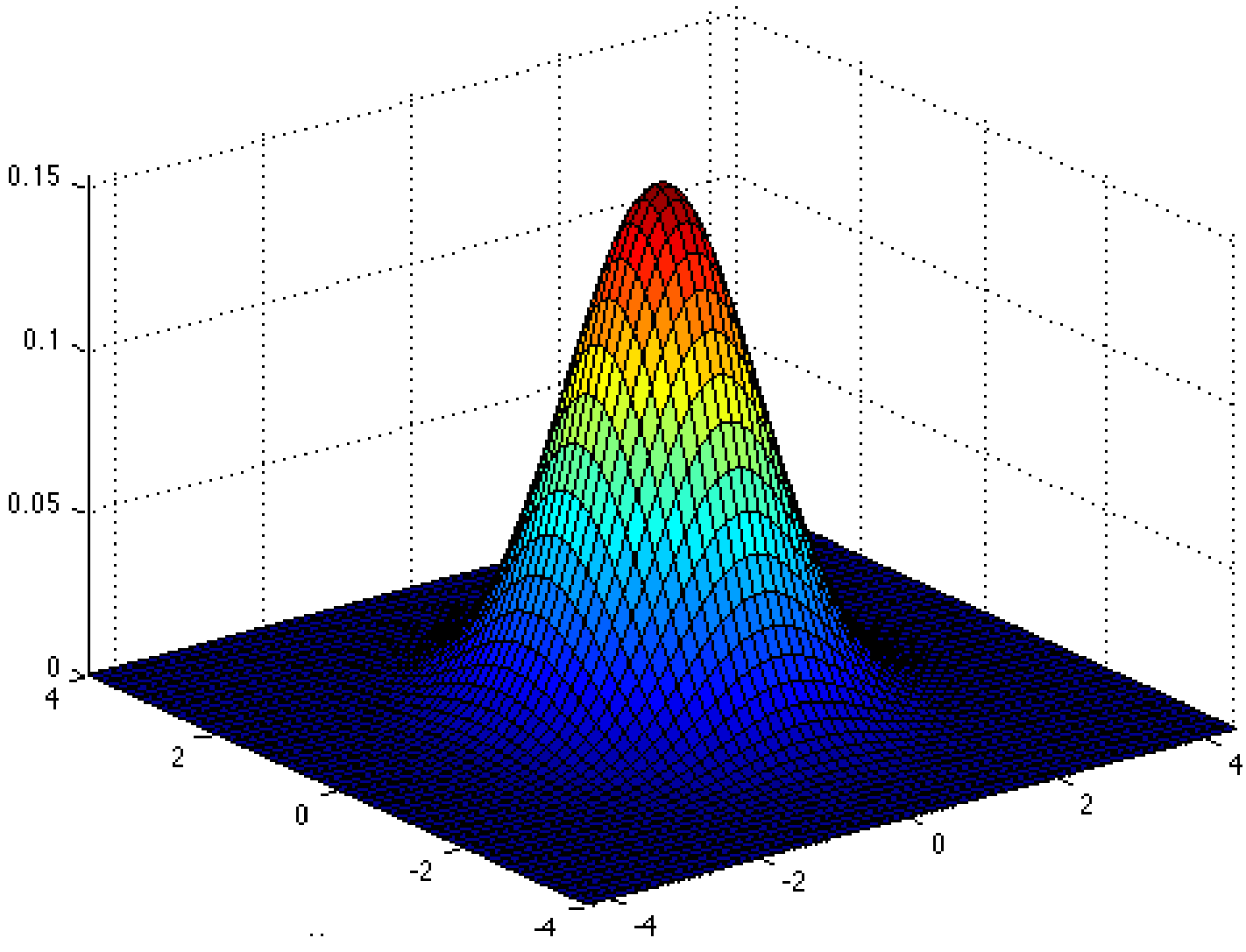
High-dimensional Gaussians

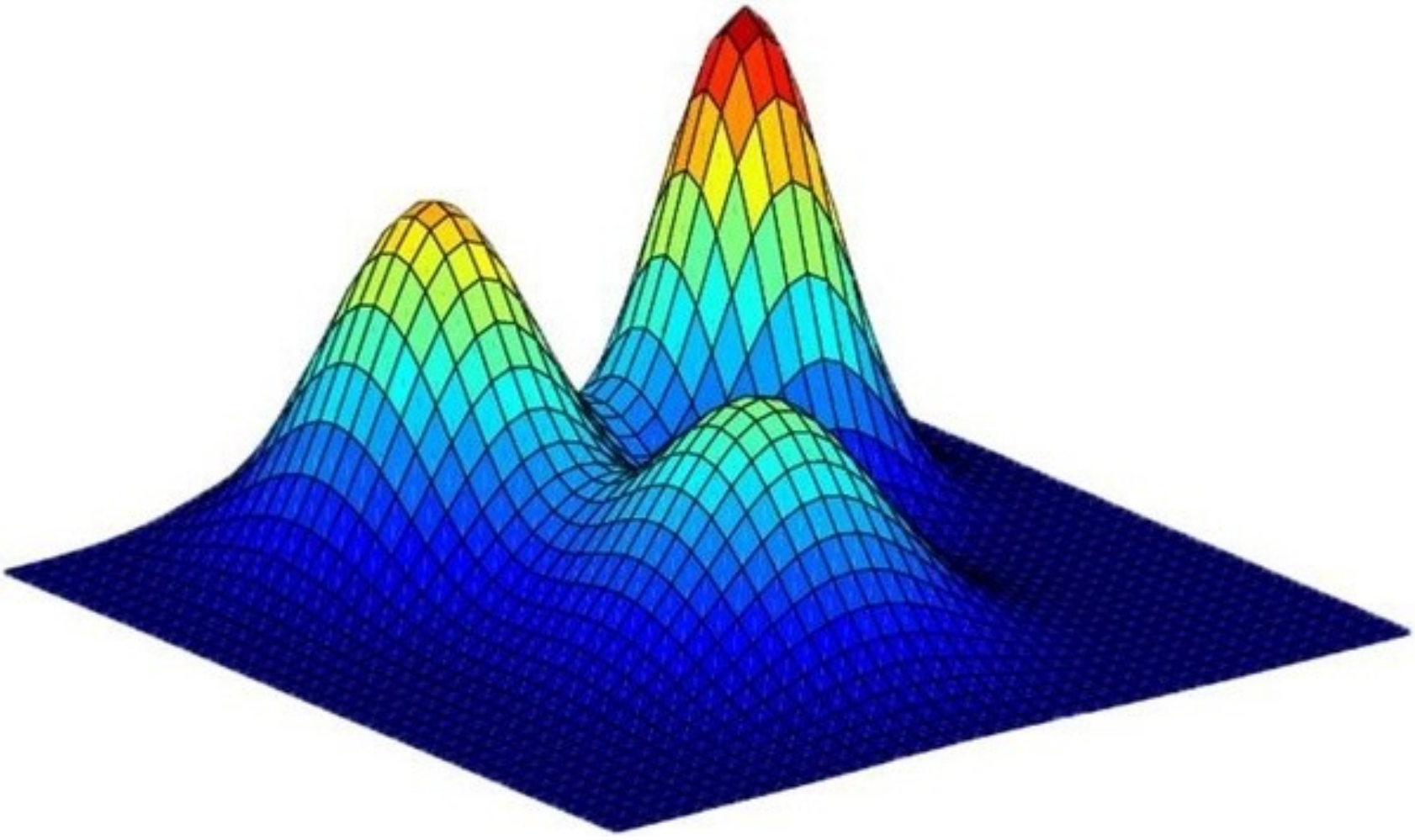
Multivariate normal distribution:

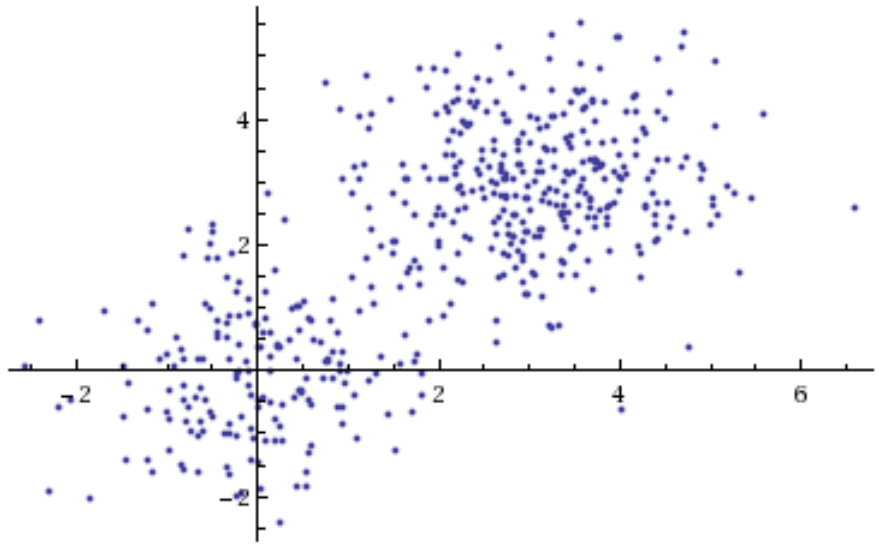
$$\mathcal{N}_{\mu,\Sigma}(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}\Sigma^{-1}(x-\mu)\right)}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} \qquad \text{for } x \in \mathbb{R}^d$$

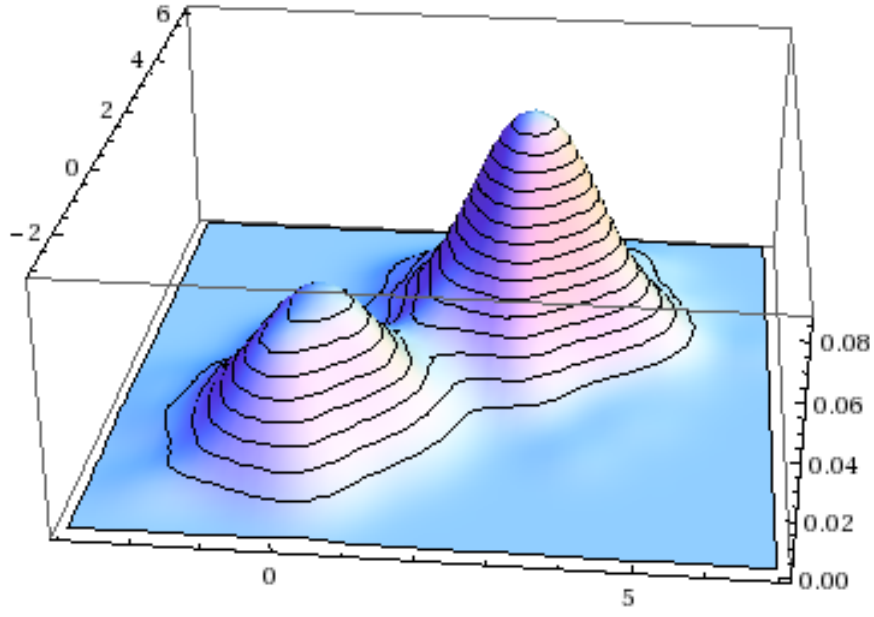
$X \sim \mathcal{N}_{\mu,\Sigma}$: $\mathbb{E}[X] = \mu \in \mathbb{R}^d, \mathbb{E}\left[(X-\mu)(X-\mu)^{\mathsf{T}}\right] = \Sigma \in \mathbb{R}^{d\times d}$

Mixture of k Gaussians in $\mathbb{R}^d$: $\sum_{i=1}^{k} w_i \mathcal{N}_{\mu_i,\Sigma_i}$
w_i are mixture weights (component weights), $w_i \geq 0$ and $\sum w_i = 1$
Parameters of the model: $(w_i, \mu_i, \Sigma_i)_{i=1}^{k}$: $\Theta(kd^2)$ parameters


0.15
0.1
0.05
0
4
2
0
-2
-4
-4
-2
0
2
4




4
2
−2
−2
2
4
6


6
4
2
0
−2
0
5
0.08
0.06
0.04
0.02
0.00

What does it mean to learn/estimate a mixture of Gaussians given data?

First answer: maximum likelihood estimation

Given samples $x_1, \ldots, x_n$, find parameters that maximize the likelihood:

$$\prod_{i=1}^{n} \left(\sum_{j=1}^{k} w_j \mathcal{N}_{\mu_j, \Sigma_j}(x_i) \right)$$

First answer: maximum likelihood estimation

Given samples $x_1, \ldots, x_n$, find parameters that maximize the likelihood:

$$\prod_{i=1}^{n} \left(\sum_{j=1}^{k} w_j \mathcal{N}_{\mu_j, \Sigma_j}(x_i) \right)$$

- ✓ Non-convex optimization problem, NP-hard [Arora and Kannan 2005]
- ✓ Widely used in practice: expectation-maximization (EM) an iterative algorithm [Dempster, Laird, and Rubin 1977]
- ✓ Convergence not well understood, very sensitive to initialization [Redner and Walker 1984]
- ✓ Almost-sure convergence to global maximum for $k = 2$, $w_1 = w_2 = 1/2$, fixed $\Sigma_1 = \Sigma_2$ [Xu, Hsu, Maleki 2016, Balakrishnan, Wainwright, Yu 2017 and Daskalakis, Tzamos, Zempetakis 2017]

Second answer: parameter estimation

Given samples $x_1, \ldots, x_n$ from some unknown mixture of Gaussians $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, find each of the parameters within threshold ε.

Second answer: parameter estimation

Given samples $x_1, \ldots, x_n$ from some unknown mixture of Gaussians $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, find each of the parameters within threshold ε.

- ✓ Active area of research in theoretical computer science [Dasgupta 1999]
- ✓ Computational complexity: polynomial in d and $1/\varepsilon$ [Kalai, Moitra, Valiant 2010] (method of moments) [Belkin, Sinha 2010] (method of moments + tools from real algebraic geometry)
- ✓ Any algorithm has sample complexity exponential in the number of components [Moitra, Valiant 2010]

Third answer: density estimation

Given samples from an unknown mixture of Gaussians f,
output a density $\widehat{f}$ that is close to f with high probability, 99%.

Close in L^1 distance:

$$\left\|f-\widehat{f}\right\|_1 = \int_{\mathbb{R}^d} \left|f(x)-\widehat{f}(x)\right| \mathrm{d}x = 2\sup_{A\subseteq\mathbb{R}^d} \left|\int_A f - \int_A \widehat{f}\right|$$

Third answer: density estimation

Given samples from an unknown mixture of Gaussians f, output a density $\widehat{f}$ that is close to f with high probability, 99%.

Close in L^1 distance:
$\left\|f-\widehat{f}\right\|_1 = \int_{\mathbb{R}^d} \left|f(x)-\widehat{f}(x)\right| \mathrm{d}x = 2\sup_{A\subseteq\mathbb{R}^d} \left|\int_A f - \int_A \widehat{f}\right|$

In general, bounds for parameter estimation do not translate to bounds for density estimation for $d > 1$. Consider two zero-mean 2-dimensional Gaussians with

$$\Sigma_1 = \begin{pmatrix} 1 & -0.99 \\ -0.99 & 1 \end{pmatrix} \text{ and } \Sigma_2 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Then their parameters are close while L^1 distance is large.

Density estimation

Given samples from an unknown density f from some known family $\mathcal{C}$ of densities, output a density $\widehat{f}$ that is close to f.

REMARKS ON SOME NONPARAMETRIC ESTIMATES OF A DENSITY FUNCTION[1]

By Murray Rosenblatt[2]

University of Chicago

1956

On the Learnability of Discrete Distributions

(extended abstract)

1994

Michael Kearns
AT&T Bell Laboratories

Yishay Mansour
Tel-Aviv University

Dana Ron
Hebrew University

Ronitt Rubinfeld
Cornell University

Robert E. Schapire
AT&T Bell Laboratories

Linda Sellie
University of Chicago

Precise question we study today

Question

Let f be an unknown mixture of k Gaussians in $\mathbb{R}^d$. How many i.i.d. samples from f is needed to produce, with high probability, a density $\widehat{f}$ satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$?

Remarks:

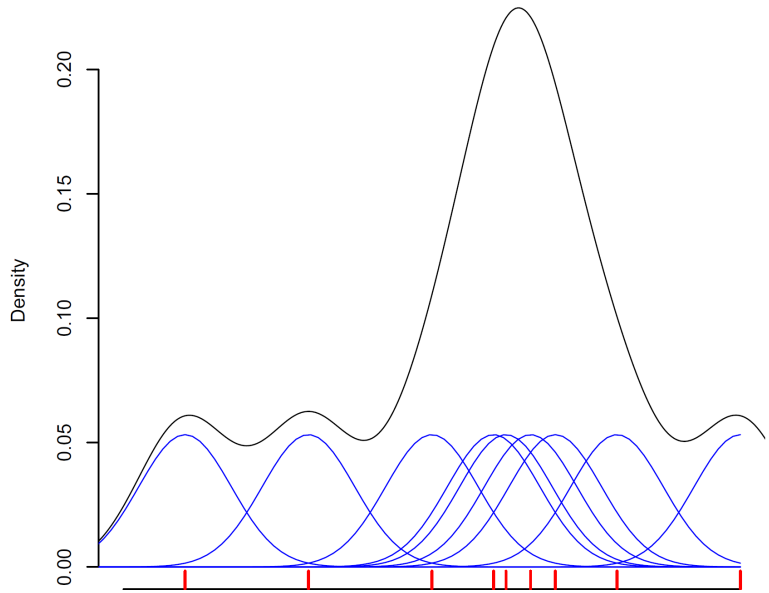
1. Algorithm knows k
2. Focus is on sample complexity
3. Equivalent formulation: given n samples from $f \in \mathcal{C}$, how small can you make $\mathbb{E}\left[\|f - \widehat{f}\|_1\right]$? Minimax risk of $\mathcal{C}$
   $n = g(\varepsilon) \Leftrightarrow \varepsilon = g^{-1}(n)$

Popular method in practice for density estimation

Kernel density estimation

Fix non-negative function $K : \mathbb{R}^d \to [0, \infty)$ with $\int K = 1$ (the kernel function). Given samples $x_1, \ldots, x_n \in \mathbb{R}^d$, output

$$\widehat{f}(x) := \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - x_i}{h}\right) \qquad \text{[Rosenblatt 1956]}$$



Plot by Patrick Breheny, University of Kentucky

Popular method in practice for density estimation

Kernel density estimation

Fix non-negative function $K : \mathbb{R}^d \to [0, \infty)$ with $\int K = 1$ (the kernel function). Given samples $x_1, \ldots, x_n \in \mathbb{R}^d$, output

$$\widehat{f}(x) := \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - x_i}{h}\right) \qquad \text{[Rosenblatt 1956]}$$

Tons of results showing fast convergence of $\widehat{f}$ to f in one dimension, assuming various boundedness/smoothness conditions on f [Devroye and Györfi 1985], [Ibragimov 1998], [Tsybakov 2008], [Jiang 2017]

Unfortunately, sample complexity is exponential in d.

Question

Let f be an unknown mixture of k Gaussians in $\mathbb{R}^d$. How many i.i.d. samples from f is needed to produce, with high probability, a density $\widehat{f}$ satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$?

$k = 1$: sample complexity $\leq O(d^2/\varepsilon^2)$

compute empirical mean and covariance, and use Gaussian concentration

$d = 1$: sample complexity $\leq O(k/\varepsilon^2)$

approximate by piecewise polynomials

[Chan, Diakonikolas, Servedio, Sun 2014]

Question

Let f be an unknown mixture of k Gaussians in $\mathbb{R}^d$. How many i.i.d. samples from f is needed to produce, with high probability, a density $\widehat{f}$ satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$?

$k = 1$: sample complexity $\leq O(d^2/\varepsilon^2)$
compute empirical mean and covariance, and use Gaussian concentration

$d = 1$: sample complexity $\leq O(k/\varepsilon^2)$
approximate by piecewise polynomials
[Chan, Diakonikolas, Servedio, Sun 2014]

Question : sample complexity $\leq$ number of parameters divided by ε^2? Indeed we will show
sample complexity $\leq kd^2/\varepsilon^2 \times \log^2(d)\log(k) = \widetilde{O}(kd^2/\varepsilon^2)$

Prior known results - 1

Definition

Given an i.i.d. sample from an unknown density $f \in \mathcal{C}$, output $\widehat{f}$ satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$ with high probability.

$m_{\mathcal{C}}(\varepsilon) =$ the smallest number of required samples.

k-mix$(\mathcal{C}) =$ class of distributions formed by taking k-mixtures of elements of $\mathcal{C}$

Prior known results - 1

Definition

Given an i.i.d. sample from an unknown density $f \in \mathcal{C}$, output $\widehat{f}$ satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$ with high probability.

$m_{\mathcal{C}}(\varepsilon) =$ the smallest number of required samples.

k-mix$(\mathcal{C}) =$ class of distributions formed by taking k-mixtures of elements of $\mathcal{C}$

Theorem (Ashtiani, Ben-david, M2017)

For any class $\mathcal{C}$, sample complexity for learning k-mix$(\mathcal{C}) \leq O(m_{\mathcal{C}}(\varepsilon) \times k \log k/\varepsilon^2)$

Corollary

Sample complexity for learning mixtures of Gaussians $\leq O((d^2/\varepsilon^2) \times k \log k/\varepsilon^2) = O(kd^2 \log(k)/\varepsilon^4)$

Prior known results - 2

$$Y(\mathcal{C}) = \left\{ \{x \in \mathbb{R}^d : f_1(x) > f_2(x)\} : f_1, f_2 \in \mathcal{C} \right\}$$

1. $m_{\mathcal{C}}(\varepsilon) \leq O(\text{VC-dim}(Y(\mathcal{C}))/\varepsilon^2)$ [Devroye and Lugosi 2001]

Prior known results - 2

$$Y(\mathcal{C}) = \left\{ \{x \in \mathbb{R}^d : f_1(x) > f_2(x)\} : f_1, f_2 \in \mathcal{C} \right\}$$

1. $m_{\mathcal{C}}(\varepsilon) \leq O(\text{VC-dim}(Y(\mathcal{C}))/\varepsilon^2)$ [Devroye and Lugosi 2001]
2. When $\mathcal{C}$ = mixtures of Gaussians, VC-dim$(Y(\mathcal{C})) \leq O(k^4 d^4)$ [Khovanskii 1991], [Karpinski and Macintyre 1997], [Anthony and Bartlett 1999]
3. Gives an upper bound of $O(k^4 d^4/\varepsilon^2)$ for the sample complexity of mixtures of Gaussians.

Lower bounds

Lower bounds?

Best known lower bound was $\Omega(kd/\varepsilon^2)$.

[Suresh, Orlitsky, Acharya, and Jafarpour 2014]

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Any algorithm that learns mixtures of Gaussians has sample complexity $\Omega(kd^2/\varepsilon^2)$.

Suffices to show a lower bound of $\Omega(d^2/\varepsilon^2)$ for learning a single Gaussian.

Lower bound proof

Suffices to show a lower bound of $\Omega(d^2/\varepsilon^2)$ for learning a single Gaussian.

General idea: a class of distribution has a large sample complexity if, there exist 'lots of distributions' in that class that are 'hard to distinguish' but are 'far in L^1 distance.'

[LeCam 1973], [Assoud 1983], [Hasminskii 1976]

Lower bound proof

Suffices to show a lower bound of $\Omega(d^2/\varepsilon^2)$ for learning a single Gaussian.

General idea: a class of distribution has a large sample complexity if, there exist 'lots of distributions' in that class that are 'hard to distinguish' but are 'far in L^1 distance.'
[LeCam 1973], [Assoud 1983], [Hasminskii 1976]

Using a lemma of [Hasminskii 1976], based on Fano's inequality in information theory (1952):
suffices to find $2^{\Omega(d^2)}$ Gaussians with
pairwise KL-divergence $\leq \varepsilon^2$ and pairwise L^1 distance $> \varepsilon$.

$$\mathrm{KL}(f_1 \parallel f_2) := \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \mathrm{d}x \qquad [Kullback - Leibler]$$

Lower bound proof

Need to build $2^{\Omega(d^2)}$ Gaussians with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise L^1 distance $> \varepsilon$.

We will use zero-mean Gaussians, so just need to specify the covariance matrices.

Lower bound proof

Need to build $2^{\Omega(d^2)}$ Gaussians with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise L^1 distance $> \varepsilon$.

We will use zero-mean Gaussians, so just need to specify the covariance matrices.

First construction [Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18]. Repeat 2^{d^2} times: start with an identity covariance matrix, then choose a random subspace of dimension $d/9$ and slightly increase the eigenvalues corresponding to this eigenspace: $\Sigma = I + \frac{\varepsilon}{\sqrt{d}} UU^\mathsf{T}$, with $U \in \mathbb{R}^{d \times d/9}$ orthonormal.

Then prove that with large probability, $1 - 2^{-\Omega(d^2)}$, any two of these have L^1 distance $> \varepsilon$. (KL is easy.)

Lower bound proof

Need to build $2^{\Omega(d^2)}$ Gaussians with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise L^1 distance $> \varepsilon$.

We will use zero-mean Gaussians, so just need to specify the covariance matrices.

Second construction (combinatorial) [Devroye,M,Reddad 2018]. For $d = 3$, consider the following inverse covariance matrices:

$$\begin{pmatrix} 0 & -\delta & -\delta \\ -\delta & 0 & -\delta \\ -\delta & -\delta & 0 \end{pmatrix}, \begin{pmatrix} 0 & \delta & \delta \\ \delta & 0 & -\delta \\ \delta & -\delta & 0 \end{pmatrix}, \begin{pmatrix} 0 & \delta & -\delta \\ \delta & 0 & \delta \\ -\delta & \delta & 0 \end{pmatrix}, \begin{pmatrix} 0 & -\delta & \delta \\ -\delta & 0 & \delta \\ \delta & \delta & 0 \end{pmatrix}$$

For general d, build $2^{d^2/10}$ inverse covariance matrices so that any two of them are different in at least $d^2/3$ coordinates (Gilbert-Varshamov bound in coding theory).

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Any algorithm that learns mixtures of Gaussians has sample complexity $\Omega(kd^2/\varepsilon^2)$.

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

Any algorithm that learns mixtures of Gaussians has sample complexity $\Omega(kd^2/\varepsilon^2)$.

Next: upper bound

Our upper bound

Recall that $m_{\mathcal{C}}(\varepsilon)$ is the sample complexity for learning a density from class $\mathcal{C}$.

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

If $\mathcal{C}$ = mixtures of k Gaussians in $\mathbb{R}^d$, then $m_{\mathcal{C}}(\varepsilon) = \widetilde{O}(kd^2/\varepsilon^2)$.

Covering number argument

Lemma (Yatracos 1985)

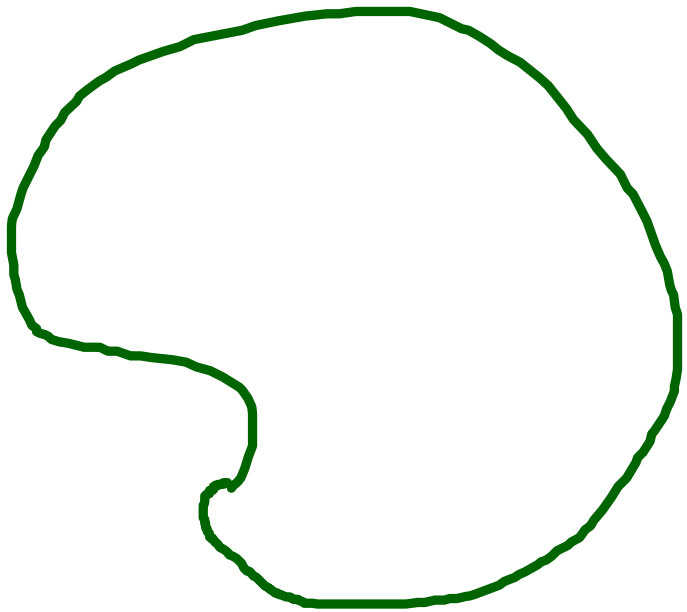
Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ *such that for any* $f \in \mathcal{C}$, *there exists some* i *with* $\|f - f_i\|_1 \leq \varepsilon$. *Then* $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.

Proved by a clever combination of Hoeffding's inequality and the union bound.

Covering number argument

Lemma (Yatracos 1985)

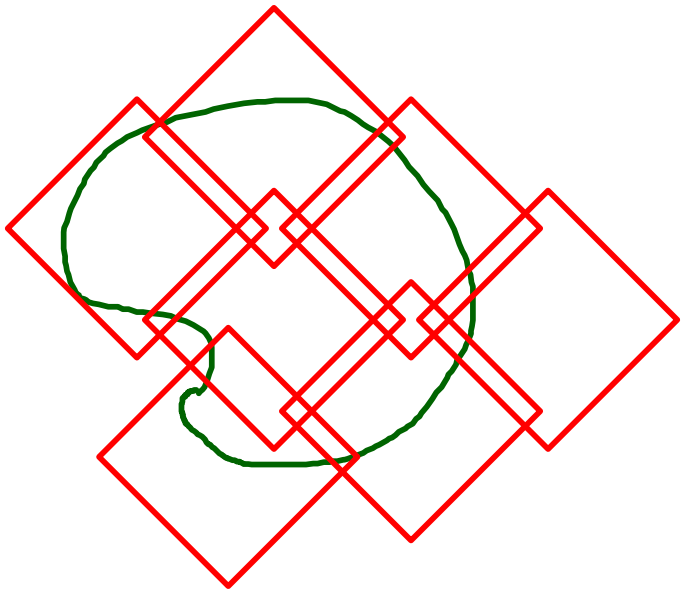
Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ *such that for any* $f \in \mathcal{C}$, *there exists some* i *with* $\|f - f_i\|_1 \leq \varepsilon$. *Then* $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.



Covering number argument

Lemma (Yatracos 1985)

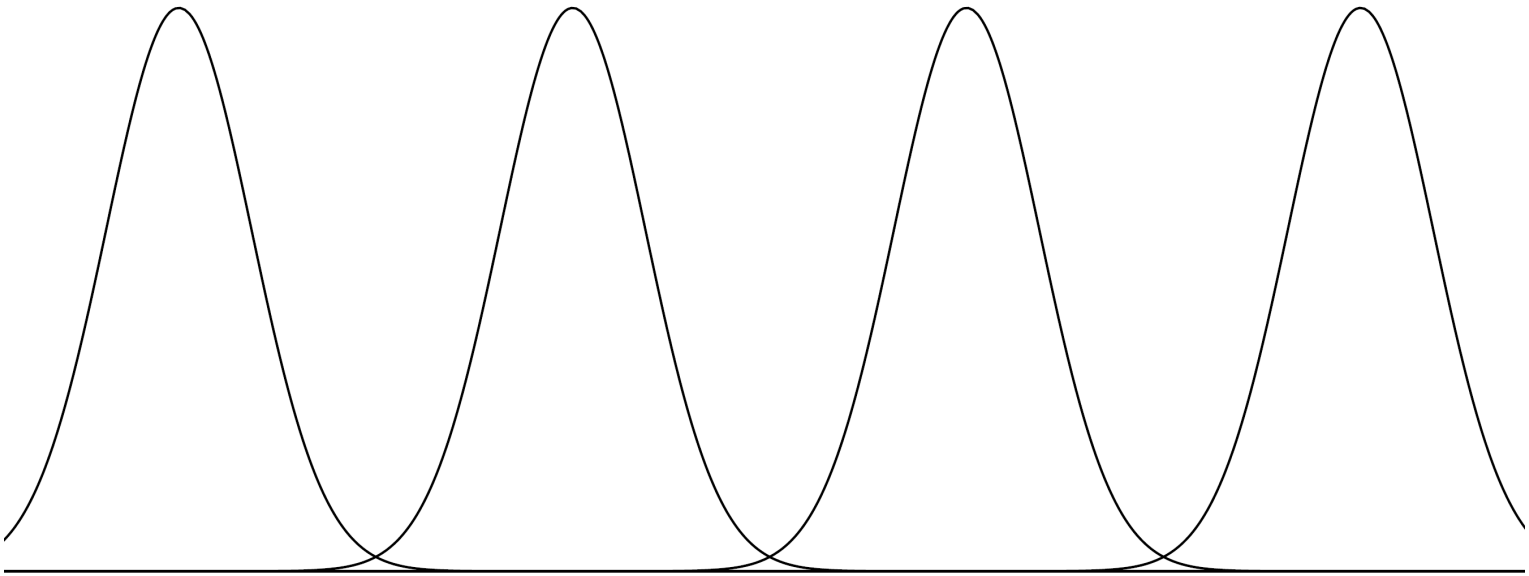
Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ *such that for any* $f \in \mathcal{C}$, *there exists some* i *with* $\|f - f_i\|_1 \leq \varepsilon$. *Then* $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.



A bound on the covering number of a distribution class bounds its sample complexity.

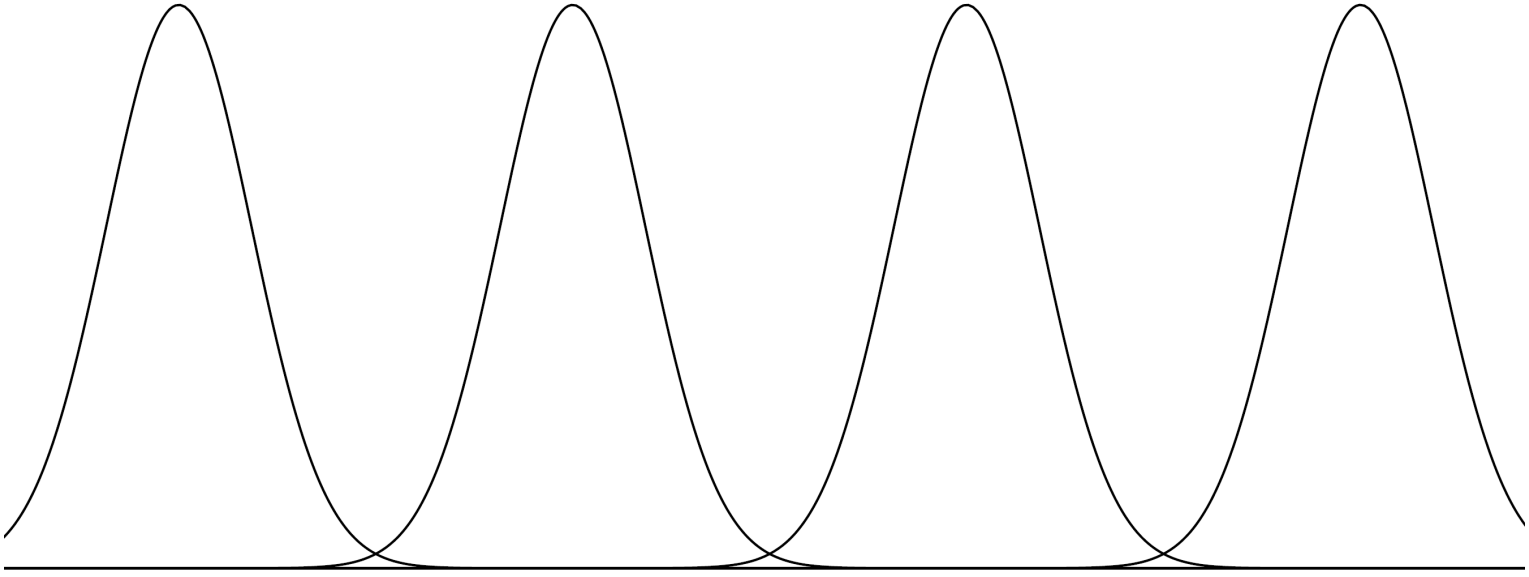
Gaussians are not bounded

Unfortunately, Gaussian distributions have infinite covering number, even if the mean is bounded.



Gaussians are not bounded

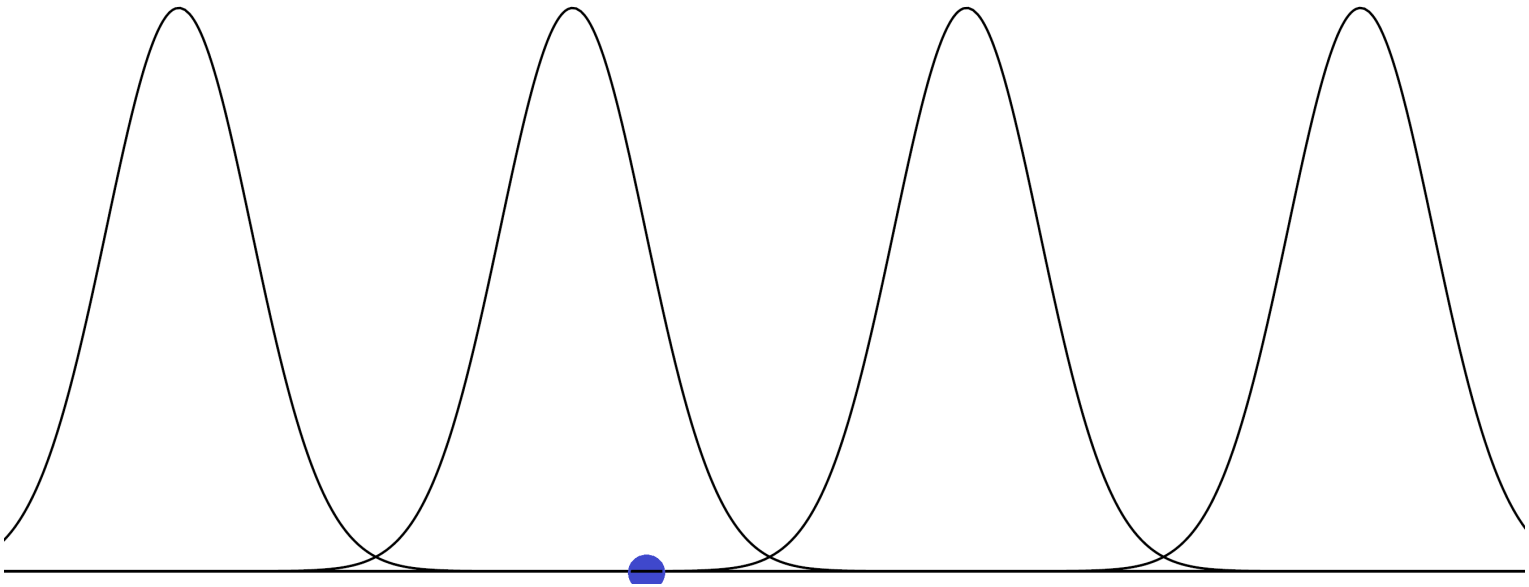
Unfortunately, Gaussian distributions have infinite covering number, even if the mean is bounded.



Our novel idea to solve this: Use some of the data to reduce the search space significantly. To formalize this idea, we introduce the notion of compression.

Gaussians are not bounded

Unfortunately, Gaussian distributions have infinite covering number, even if the mean is bounded.



Our novel idea to solve this: Use some of the data to reduce the search space significantly. To formalize this idea, we introduce the notion of compression.

Idea of compression

If class $\mathcal{C}$ has a covering of size 2^t, any distribution in this covering can be encoded using t bits.

Thus, given any $f \in \mathcal{C}$, there are t bits from which one can construct a distribution close to f. We say the essence of each $f \in \mathcal{C}$ can (almost) be compressed into t bits.

We generalize this and use some i.i.d. samples from $f \in \mathcal{C}$ in addition to these bits to compress f.

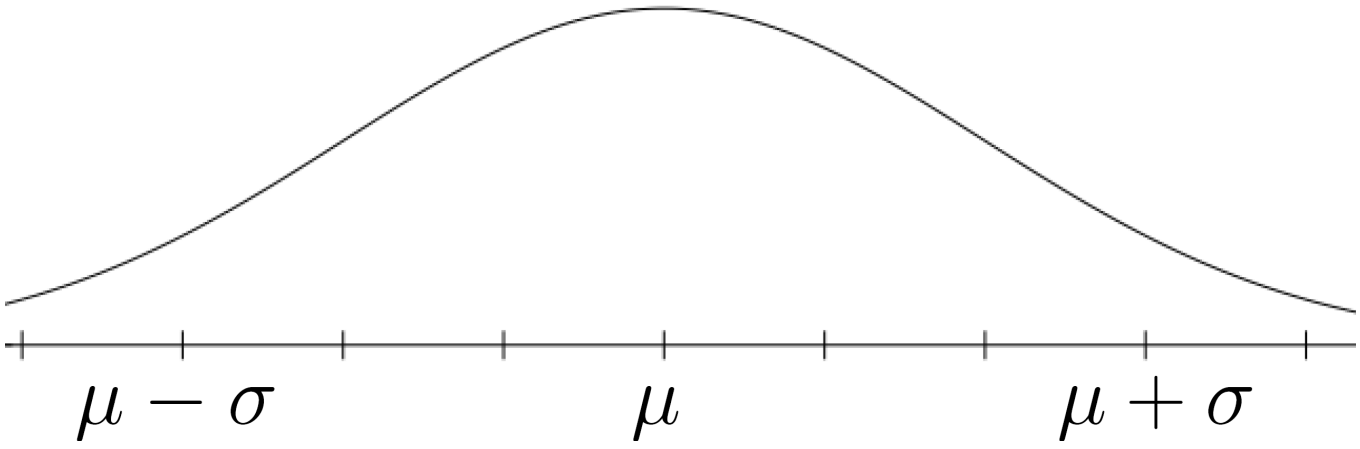
Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$ and ε, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exists a sequence of $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed with $\|f - \widehat{f}\|_1 \leq \varepsilon$.

Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$ and ε, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exists a sequence of $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed with $\|f - \widehat{f}\|_1 \leq \varepsilon$.

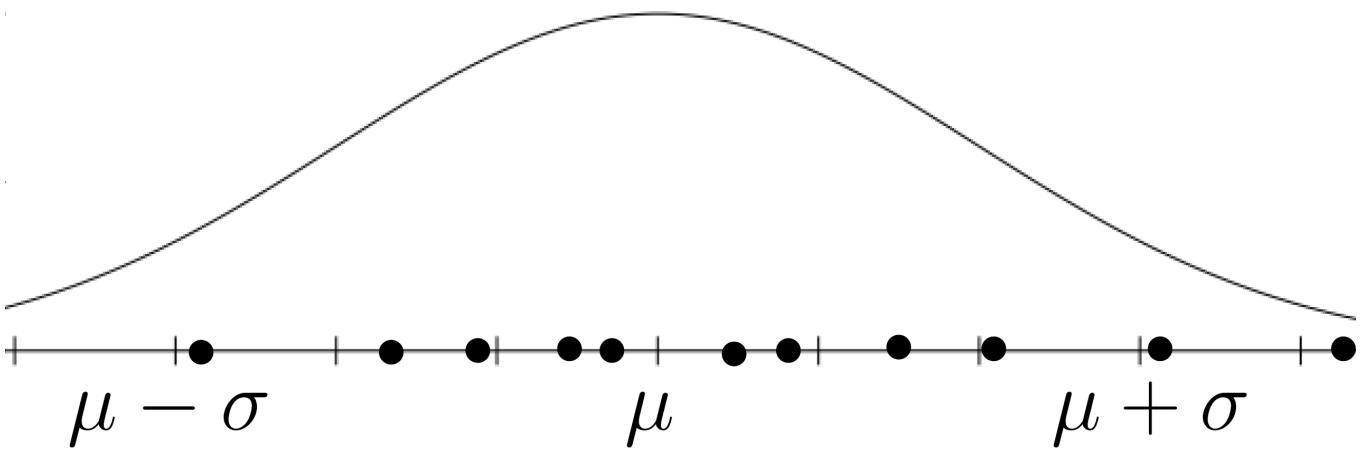
One dimensional Gaussians admit $(100/\varepsilon, 2)$-compression.



Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$ and ε, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exists a sequence of $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed with $\|f - \widehat{f}\|_1 \leq \varepsilon$.

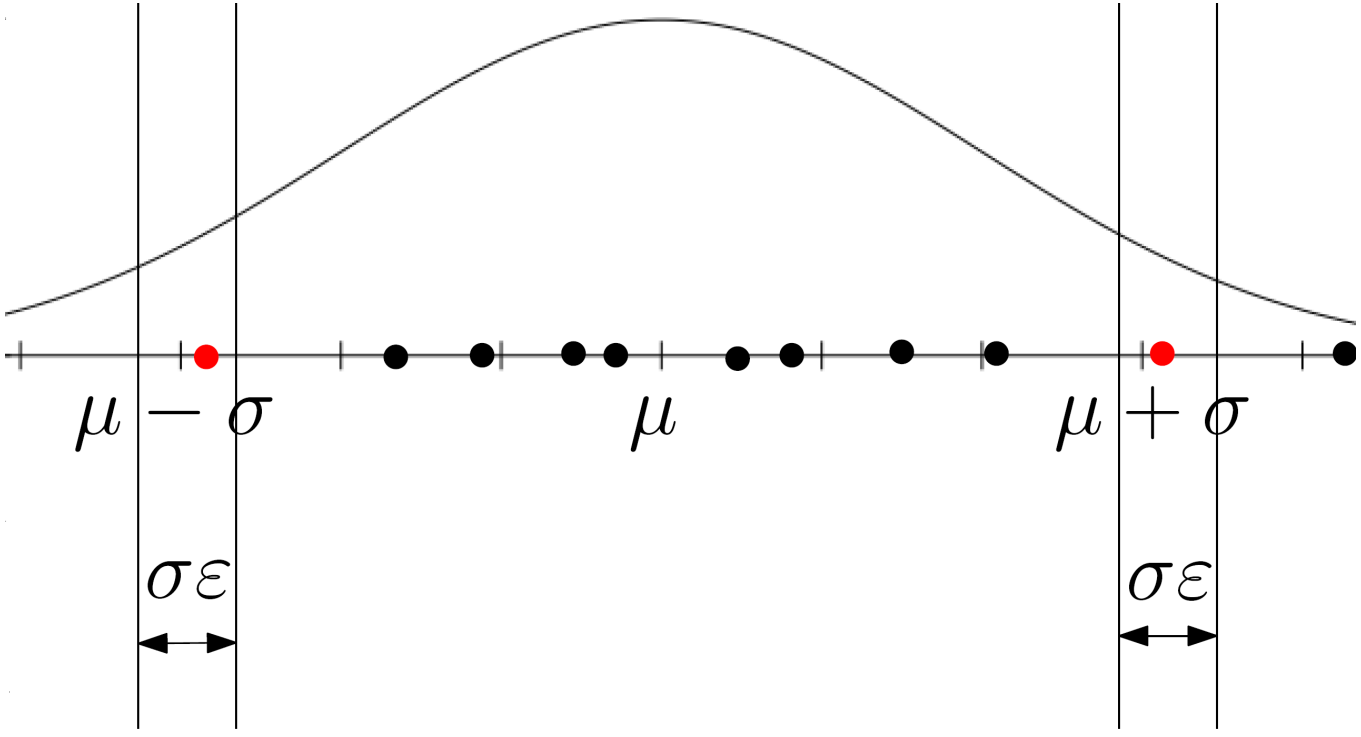
One dimensional Gaussians admit $(100/\varepsilon, 2)$-compression.



Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$ and ε, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exists a sequence of $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed with $\|f - \widehat{f}\|_1 \leq \varepsilon$.

One dimensional Gaussians admit $(100/\varepsilon, 2)$-compression.



Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$ and ε, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exists a sequence of $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed with $\|f - \widehat{f}\|_1 \leq \varepsilon$.

One dimensional Gaussians admit $(100/\varepsilon, 2)$-compression.

$$\widehat{\mu} = \frac{x_1+x_2}{2}$$
$$\widehat{\sigma} = \frac{|x_1-x_2|}{2}$$

Compression implies learnability

Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exist $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$.

Lemma (compression implies learnability)

If $\mathcal{C}$ admits (n, τ)-compression, $m_{\mathcal{C}}(\varepsilon) = O\left(n + \frac{\tau \log n}{\varepsilon^2}\right)$.

Compression implies learnability

Definition (compression)

Class $\mathcal{C}$ admits $(n(\varepsilon), \tau(\varepsilon))$-compression if, for any $f \in \mathcal{C}$, after $n(\varepsilon)$ i.i.d. samples from f are generated, with high probability there exist $\tau(\varepsilon)$ of the samples and $\tau(\varepsilon)$ bits from which $\widehat{f}$ can be constructed satisfying $\|f - \widehat{f}\|_1 \leq \varepsilon$.

Lemma (compression implies learnability)

If $\mathcal{C}$ admits (n, τ)-compression, $m_{\mathcal{C}}(\varepsilon) = O\left(n + \frac{\tau \log n}{\varepsilon^2}\right)$.

Learning algorithm: generate sample of size n from unknown f. Try to reconstruct f by considering all n^τ sample sequences of length τ and all 2^τ bit sequences of length τ. We obtain $M = n^\tau 2^\tau$ densities, and one of them is ε-close to f. Use Yatracos' learning algorithm which needs $O(\log(M)/\varepsilon^2)$ more samples.
Total samples $= O(n + \frac{\log(M)}{\varepsilon^2}) = O(n + \frac{\tau \log(n)}{\varepsilon^2})$.

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d\log(d), d^2\log^2(d/\varepsilon))$-compression.

2. Compressing mixtures

If $\mathcal{C}$ admits (n,τ)-compression, then k-mix$(\mathcal{C})$ admits
$(nk\log(k), k\tau + k\log k)$-compression.

3. Compression implies learnability

If $\mathcal{C}$ admits (n,τ)-compression, $m_{\mathcal{C}}(\varepsilon) = O(n + \tau\log\tau/\varepsilon^2)$.

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan'18)

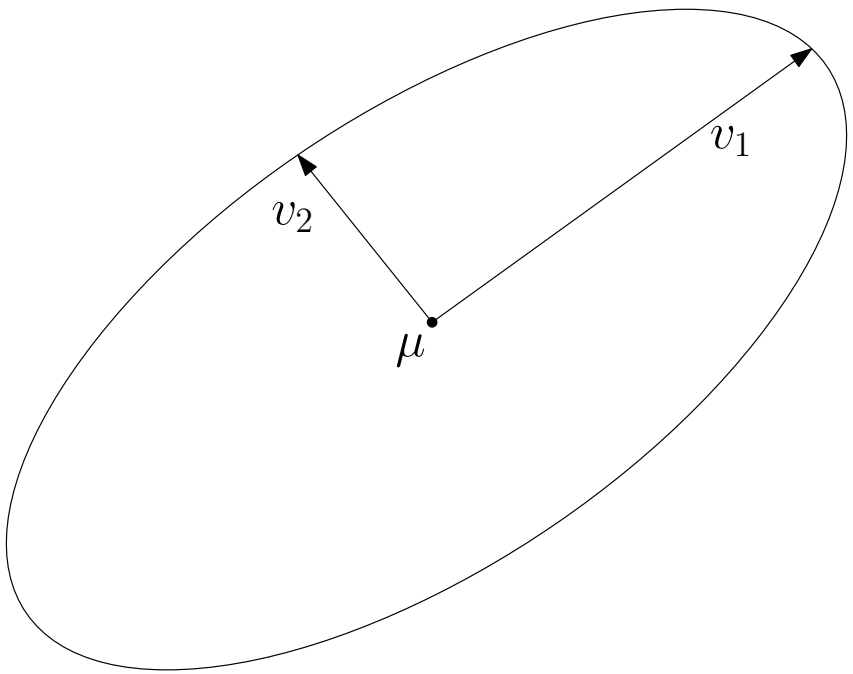
If $\mathcal{C}$ is mixtures of k Gaussians in $\mathbb{R}^d$ then
$m_{\mathcal{C}}(\varepsilon) = \widetilde{O}(kd^2/\varepsilon^2)$.

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d \log(d), d^2 \log^2(d/\varepsilon))$-compression.

$$\mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^T + v_2 v_2^T).$$

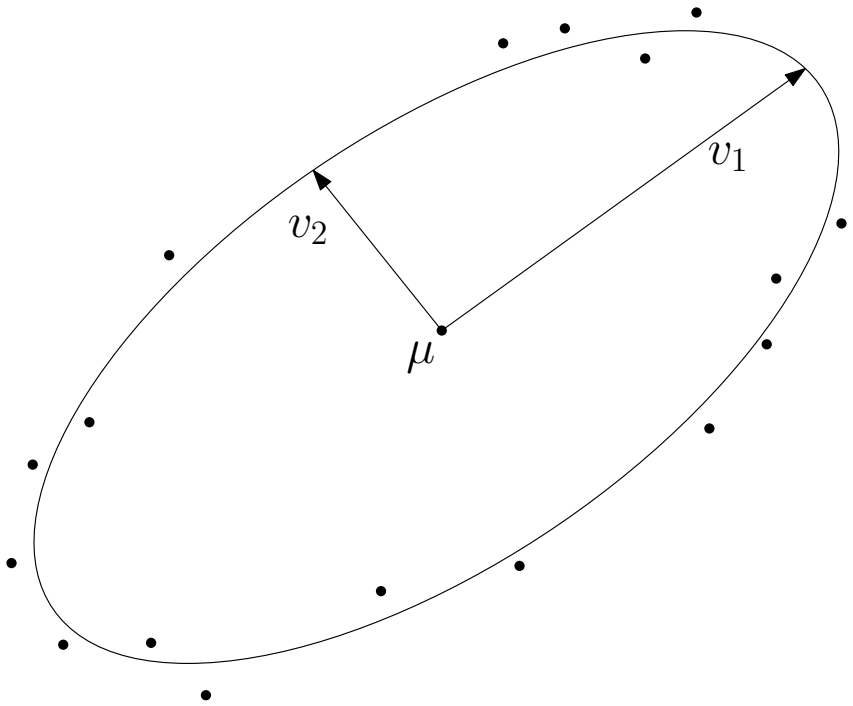


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d \log(d), d^2 \log^2(d/\varepsilon))$-compression.

$$\mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^T + v_2 v_2^T).$$

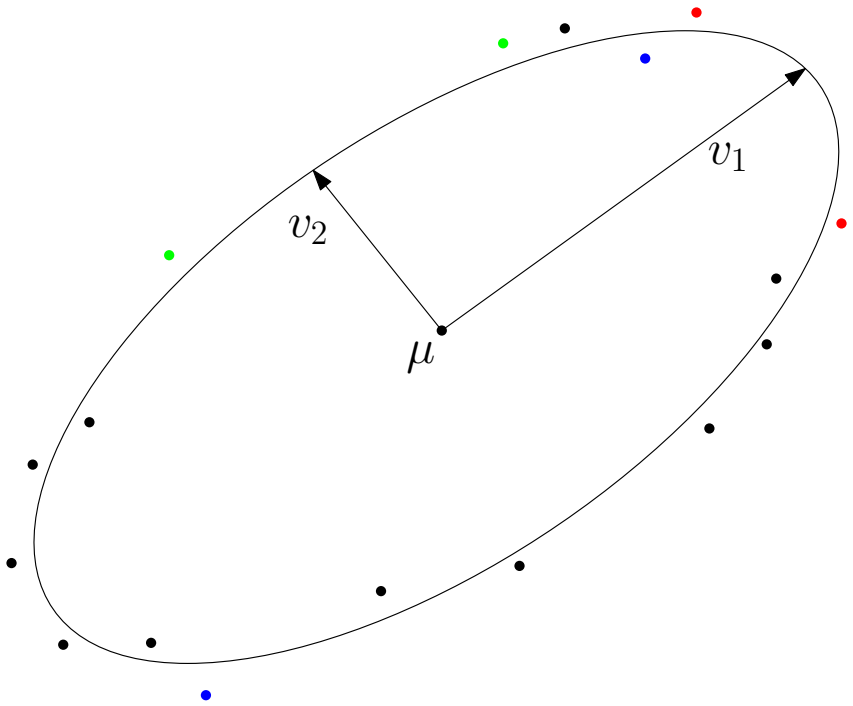


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d \log(d), d^2 \log^2(d/\varepsilon))$-compression.

$$\mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^T + v_2 v_2^T).$$

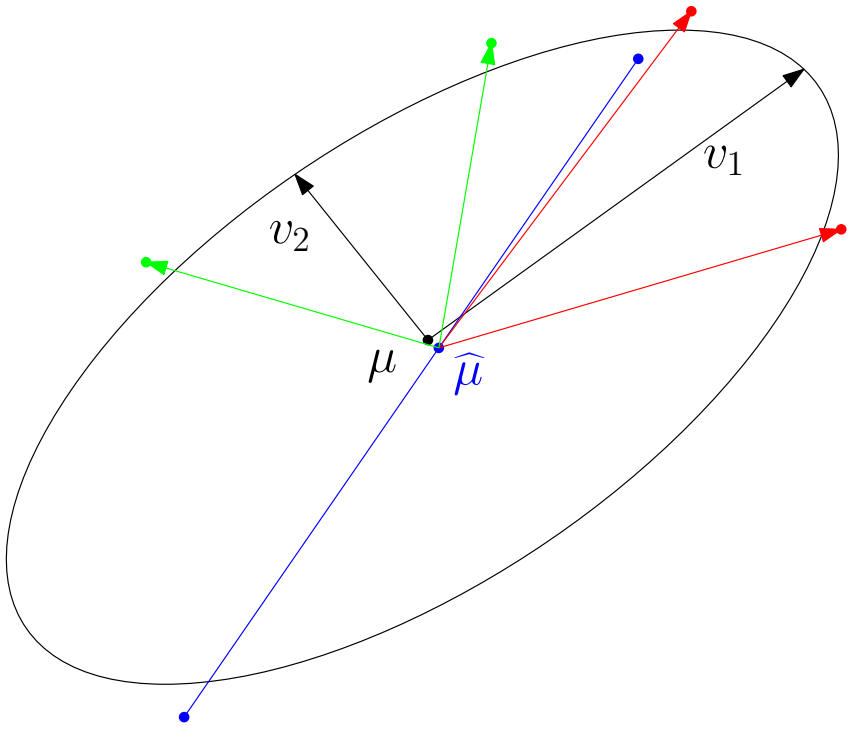


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d \log(d), d^2 \log^2(d/\varepsilon))$-compression.

$$\mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^T + v_2 v_2^T).$$



Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

d-dimensional Gaussians admit
$O(d \log(d), d^2 \log^2(d/\varepsilon))$-compression.

In general, use $d \log^2(d/\varepsilon)$ data points to encode the mean, and $d \log^2(d/\varepsilon)$ data points for each eigenvector.

Lemma (Litvak, Pajor, Rudelson, Tomczak-Jaegermann 2005)

If we take $O(d \log d)$ samples from $\mathcal{N}(0, I_d)$, their convex hull with high probability contains $\frac{1}{20} B_2^d$

Main result

Theorem (Ashtiani, Ben-David, Harvey, Liaw, M, Plan, NeurIPS 2018 (NIPS 2018))

Given $\widetilde{O}(kd^2/\varepsilon^2)$ samples from an unknown mixture of k Gaussians in d dimensions, we can output a density that is ε-close in L^1 to the underlying density with high probability. Moreover, any algorithm achieving this task requires at least $\Omega(kd^2/\varepsilon^2)$ many samples.

improve previous upper bounds of $\widetilde{O}(kd^2/\varepsilon^4)$ and $O(k^4d^4/\varepsilon^2)$, and the lower bound of $\Omega(kd/\varepsilon^2)$.

Upper bound. a novel technique for distribution learning based on compressions, high-dimensional geometry + Yatracos' algorithm.

Lower bound. a packing argument, Fano's inequality.

Conclusion

Distribution learning has rich connections with

- ✓ probability theory and statistics (concentration)
- ✓ machine learning (Vapnik-Chervonenkis dimension)
- ✓ optimization (expectation-maximization algorithm)
- ✓ information theory (Fano's inequality)
- ✓ coding theory (Gilbert-Varshamov bound)
- ✓ high-dimensional geometry
  (Litvak-Pajor-Rudelson-Tomczak-Jaegermann's result)

My research

2009–11 Graph theory (University of Waterloo)

2012–15 Random graphs with applications to network science (Waterloo)

2015–16 Randomized algorithms with applications to network science/distributed computing (e.g. gossip and load balancing protocols) (University of British Columbia) worked with Omer Angel, Yaniv Plan, Nick Harvey, Chris Liaw

2017–19 Theoretical machine learning (UC Berkeley, McGill)

Current interests: distribution learning and online learning.

Research agenda – 1

Distribution learning problem

Given data, estimate the underlying distribution.

Central problem in statistics.

Fundamental problem in unsupervised machine learning: anomaly detection, generative models, PCA, clustering.

My research goal

Design sample-efficient, time-efficient algorithms with mathematical guarantees that are robust against noise.

Research agenda – 2

Online learning problem

Data to the learning problem is not revealed at once, but is received sequentially. Can we compete against batch algorithm that sees all the data at once?

Connections with game theory, probability theory and optimization.

At the core of reinforcement learning, very active area of machine learning research, many applications.

My research goal

Design fast algorithms with provably small regret for various settings.

Thanks to my co-authors

Hassan Ashtiani
(McMaster)

Shai Ben-David
(Waterloo)

Luc Devroye
(McGill)

Nick Harvey
(UBC)

Chris Liaw
(UBC)

Yaniv Plan
(UBC)

Tommy Reddad
(McGill)

Research direction 1

What is the sample complexity for learning a class $\mathcal{C}$?

- ✓ Relate this to some notion of dimension of the class?
- ✓ Apply the compression idea to other classes?

- ✓ Probabilistic graphical models, e.g. the Ising model [Devroye, M, Reddad'18]
- ✓ Distributions generated by neural networks

Picture taken from the work of Karras, Aila, Laine, and Lehtinen 2017

Research direction 2: computational complexity

Which classes are learnable in polynomial time?

Polynomial time algorithm for mixtures of Gaussians?

Exists for mixtures of spherical Gaussians.
[Suresh, Orlitsky, Acharya, and Jafarpour 2014]

Research direction 2: computational complexity

Which classes are learnable in polynomial time?

Polynomial time algorithm for mixtures of Gaussians?

Exists for mixtures of spherical Gaussians.
[Suresh, Orlitsky, Acharya, and Jafarpour 2014]

Research direction 3: robustness

Design learners that are robust against noisy data.

Our algorithm works in agnostic learning.

What if a small fraction of the data is corrupted in an adversarial way?

Research direction 4: online learning

What if data is not revealed at once, but is received in an online manner? Can we compete against a batch algorithm that sees all the data at once?

Research direction 5: model selection

Can we learn the class $\mathcal{C}$ itself from data?

What if the number of Gaussian components, k, is not known?

Popular method in practice for density estimation

Kernel density estimation–continued

Table 4.2 *Sample size required (accurate to about 3 significant figures) to ensure that the relative mean square error at zero is less than 0.1, when estimating a standard multivariate normal density using a normal kernel and the window width that minimizes the mean square error at zero*

Dimensionality	*Required sample size*
1	4
2	19
3	67
4	223
5	768
6	2 790
7	10 700
8	43 700
9	187 000
10	842 000

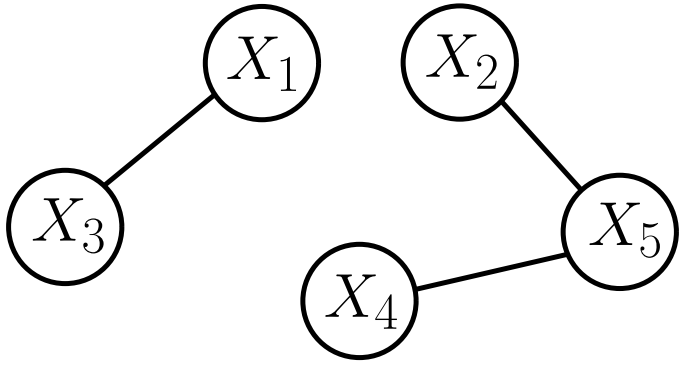
Silverman 1998, Density estimation for Statistics and Data Analysis

VC-dimension

For a family $\mathcal{Y}$ of subsets of X, the VC-dimension of $\mathcal{Y}$ is the size of the largest set $A \subseteq X$, such that for any $B \subseteq A$ there exists some $Y \in \mathcal{Y}$ with $Y \cap A = B$.

Interesting classes - 1

Probabilistic graphical models



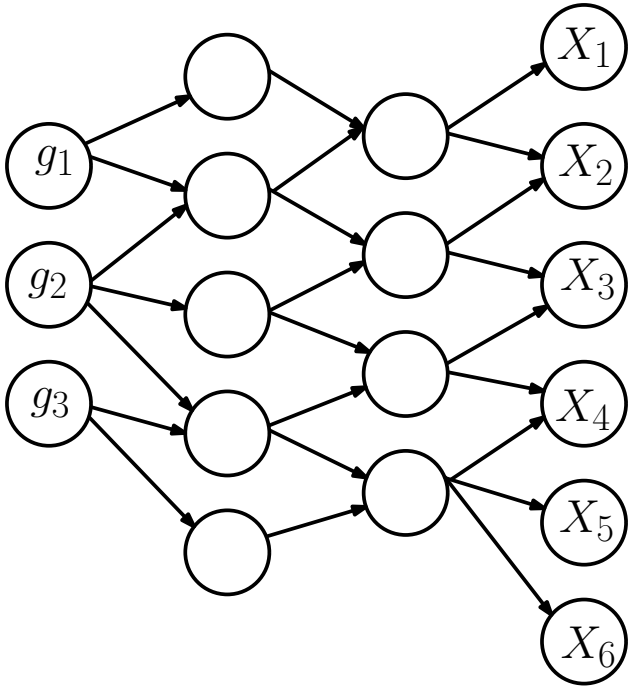
Example (The Ising model). Each $X_i \in \{-1, +1\}$ and

$$\mathbb{P}[X_1 = x_1, \ldots, X_d = x_d] \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j\right)$$

Theorem (Devroye, M, Reddad'18)

Let $\mathcal{I}_G$ = Ising models on G. Then, $m_{\mathcal{I}_G}(\varepsilon) = \Theta(|E(G)|/\varepsilon^2)$.

Interesting classes - 2



[Karras, Aila, Laine, and Lehtinen 2017]

Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

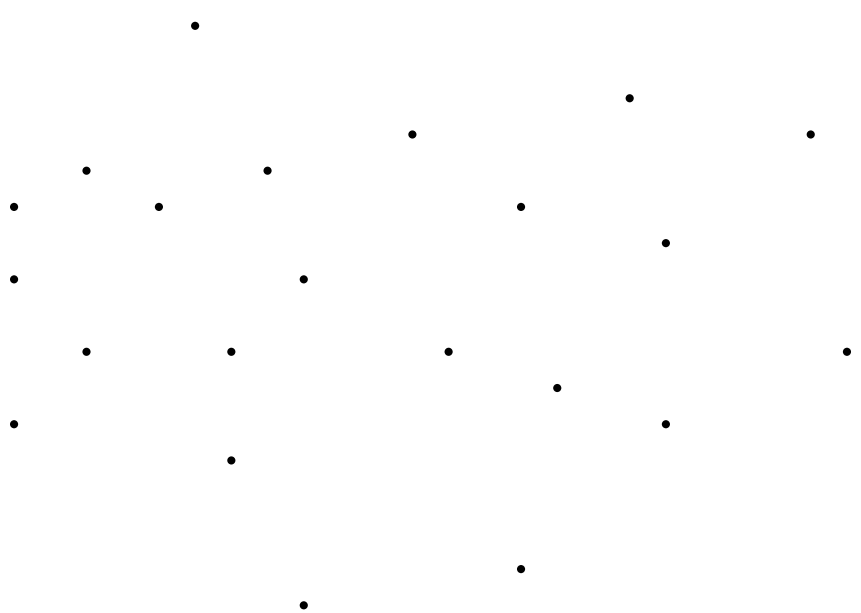
Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

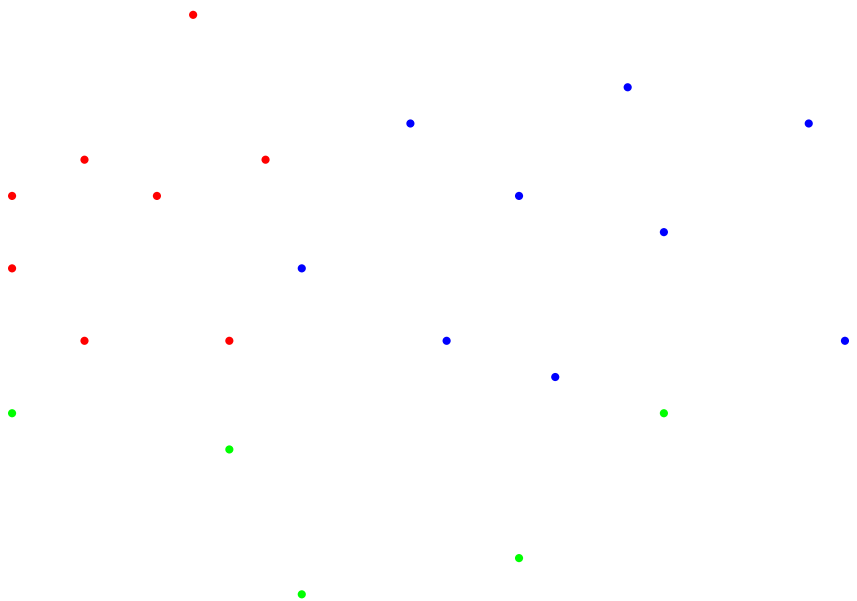


Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

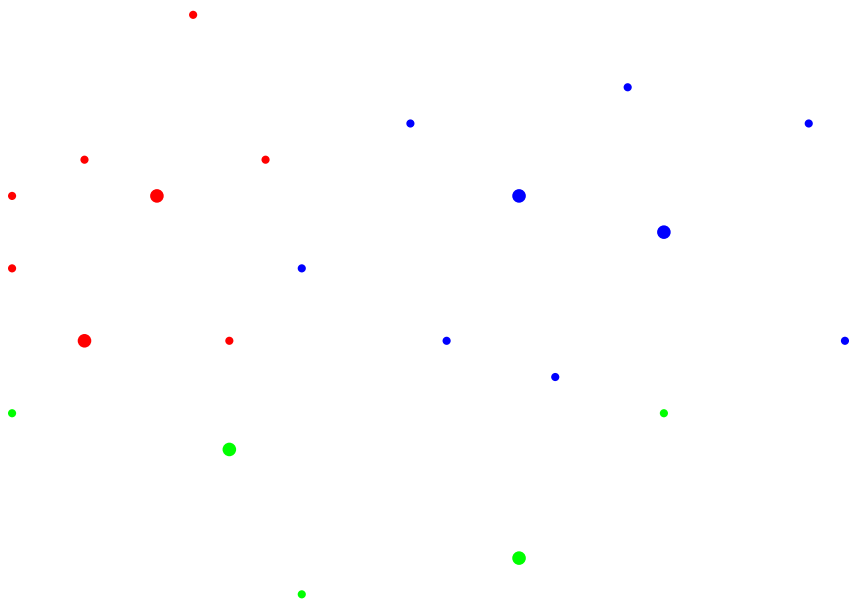


Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

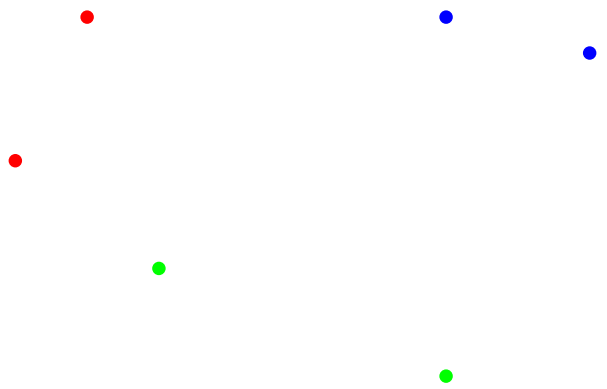


Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

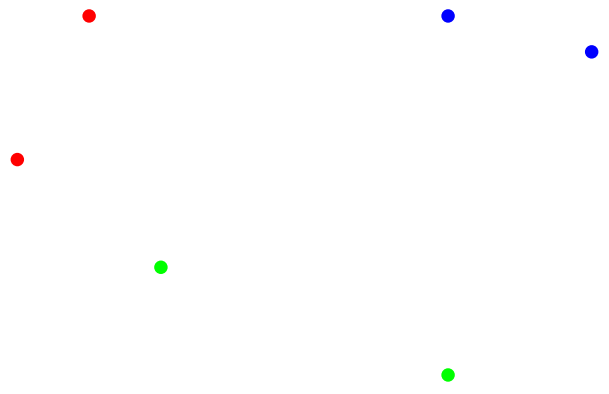


Proof of mixture lemma

Compressing mixtures

If $\mathcal{C}$ admits (n, τ)-compression, then k-mix$(\mathcal{C})$ admits $(nk\log(k), k\tau + k\log k)$-compression.

Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.



Let $\widehat{\mathbf{P}} = \frac{1}{3}\widehat{P_1} + \frac{1}{3}\widehat{P_2} + \frac{1}{3}\widehat{P_3}$

Lemma (Yatracos 1985)

Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ *such that for any* $f \in \mathcal{C}$, *there exists some* i *with* $\|f - f_i\|_1 \leq \varepsilon/5$. *Then* $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.

Lemma (Yatracos 1985)

Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ such that for any $f \in \mathcal{C}$, there exists some i with $\|f - f_i\|_1 \leq \varepsilon/5$. Then $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.

Let $Y := \left\{\{x : f_i(x) > f_j(x)\} \text{ for } i, j = 1, \ldots, m\right\}$ and let S be an i.i.d. sample of size $50\log(M)/\varepsilon^2$ from f. For density f, let $f(A) := \int_A f$. $|S \cap A| \sim \text{binomial}(|S|, f(A))$. By Hoeffding [1963] and a union bound over $A \in Y$,

$$\text{err}(f) := \sup_{A \in Y} \left| f(A) - \frac{|S \cap A|}{|S|} \right| \leq \varepsilon/5$$

with probability $1 - 2M^2 \exp(-|S|\varepsilon^2/25) \geq 99\%$.
Thus there exists some i with $\text{err}(f_i) \leq 2\varepsilon/5$.
So $\min_j \text{err}(f_j) \leq 2\varepsilon/5$, and it can be shown that the argmin here is within L1 distance ε of f.

Lemma (Yatracos 1985)

Suppose there exist $f_1, \ldots, f_M \in \mathcal{C}$ such that for any $f \in \mathcal{C}$, there exists some i with $\|f - f_i\|_1 \leq \varepsilon/5$. Then $m_{\mathcal{C}}(\varepsilon) = O(\log(M)/\varepsilon^2)$.

Let $Y := \left\{ \{x : f_i(x) > f_j(x)\} \text{ for } i, j = 1, \ldots, m \right\}$ and let S be an i.i.d. sample of size $50 \log(M)/\varepsilon^2$ from f. For density f, let $f(A) := \int_A f$. Output

$$\min_{j=1,\ldots,M} \sup_{A \in Y} \left| f_j(A) - \frac{|S \cap A|}{|S|} \right|$$

An application of density estimation

detecting breast cancer

- ✓ Training data consists of normal (non-cancerous) X-ray images.
- ✓ A probability density function $f : \mathbb{R}^d \to \mathbb{R}$ is learned from the data.
- ✓ When a new input x is presented, a high value for $f(x)$ indicates a normal image, while a low value indicates a novel input, which might be characteristic of an abnormality.

[Tarassenko, Hayton, Cerneaz, Brady 1995: Novelty detection for the identification of masses in mammograms]

An example of density estimation

Generating random faces for computer games

- ✓ Training data consists of actual faces.
- ✓ A probability density function $f : \mathbb{R}^d \to \mathbb{R}$ is learned from the data.
- ✓ New random faces are generated using the learned distribution.

An example of density estimation

Generating random faces for computer games

- ✓ Training data consists of actual faces.
- ✓ A probability density function $f : \mathbb{R}^d \to \mathbb{R}$ is learned from the data.
- ✓ New random faces are generated using the learned distribution.

A popular approach: generative adversarial networks (GANs), based on deep neural networks.

Density estimation in action

Top: generated images using generative adversarial networks

Bottom: a small part of the training data

Picture from Karras, Aila, Laine, and Lehtinen (NVIDIA and Aalto University), October 2017