Learning probability distributions

Abbas Mehrabian

McGill University
IVADO Postdoctoral Fellow

29 November 2018

Co-authors: Hassan Ashtiani, Shai Ben-David, Luc Devroye, Nick Harvey, Chris Liaw, Yaniv Plan, and Tommy Reddad

An example of distribution learning

Generating random faces for computer games

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

An example of distribution learning

Generating random faces for computer games

- ✓ Training data consists of actual faces.
- ✓ A probability density function $\mathbf{P} : \mathbb{R}^d \to \mathbb{R}$ 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.

Distribution learning 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

Distribution learning task

also known as density estimation

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$, output a distribution $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

Distribution learning task

also known as density estimation

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$, output a distribution $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

- ✓ We assume $\mathbf{P}$ belongs to some known class $\mathcal{F}$ of distributions.
- ✓ We would like our algorithm to use as a small sample as possible.
- ✓ Closeness is measured by the total variation distance: $\mathrm{TV}(\mathbf{P}, \widehat{\mathbf{P}}) \coloneqq \sup_E |\mathbf{P}(E) - \widehat{\mathbf{P}}(E)| = \frac{1}{2}\int |p(x) - \widehat{p}(x)| \,\mathrm{d}x$

Distribution learning task

Our setup

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$ from a known class $\mathcal{F}$, output some $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

What is the smallest number of samples needed to guarantee $\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P}) \leq \varepsilon$ with probability 99%? $m_{\mathcal{F}}(\varepsilon)$.

Distribution learning task

Our setup

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$ from a known class $\mathcal{F}$, output some $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.

What is the smallest number of samples needed to guarantee $\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P}) \leq \varepsilon$ with probability 99%? $m_{\mathcal{F}}(\varepsilon)$.

Main problem

prove bounds for $m_{\mathcal{F}}(\varepsilon)$ for various classes $\mathcal{F}$.

Distribution learning task

Our setup

Given an i.i.d. sample generated from an unknown target distribution $\mathbf{P}$ from a known class $\mathcal{F}$, output some $\widehat{\mathbf{P}}$ that is close to $\mathbf{P}$.
What is the smallest number of samples needed to guarantee $\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P}) \leq \varepsilon$ with probability 99%? $m_{\mathcal{F}}(\varepsilon)$.

Main problem

prove bounds for $m_{\mathcal{F}}(\varepsilon)$ for various classes $\mathcal{F}$.

Often in statistics the problem is stated differently: given n samples from $\mathbf{P}$, how small can you make $\mathbb{E}\,\mathrm{TV}(\widehat{\mathbf{P}}, \mathbf{P})$?
The answer is called the minimax risk of $\mathcal{F}$.

A heuristic

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F} =$ Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} =$ Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 2/\varepsilon^2$
- ✓ $\mathcal{F} =$ d-dimensional Gaussians: $m_{\mathcal{F}}(\varepsilon) \leq Cd^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq C \log |\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01

A heuristic

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F} =$ Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} =$ Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 2/\varepsilon^2$
- ✓ $\mathcal{F} = d$-dimensional Gaussians: $m_{\mathcal{F}}(\varepsilon) \leq Cd^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq C \log |\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01

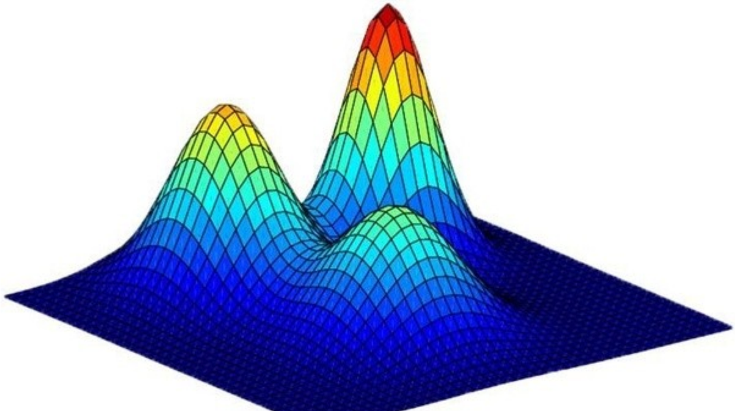
Main contribution: this heuristic also works for two more complicated classes: mixtures of multidimensional Gaussians, and the Ising model.

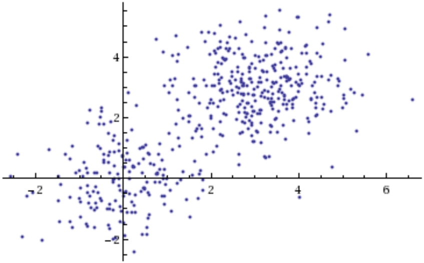
Mixtures of Gaussians

Mixtures of Gaussians

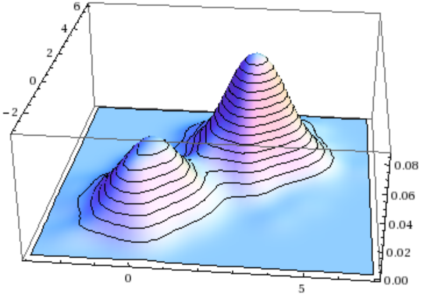
A mixture of k Gaussians in d dimensions has density $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)(x)$, where $w_i \geq 0$ and $\sum w_i = 1$.

$\mathcal{N}(\mu, \Sigma)(x)$ = density of a Gaussian with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$





-2
2
4
6
4
2
-2



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

Main result

mixtures of Gaussians

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

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ up to polylogarithmic factors.

Main result

mixtures of Gaussians

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

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ up to polylogarithmic factors.

Any density in $\mathcal{F}_{k,d}$ has form $\sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$, and Σ_i is $d \times d$, so has $\Theta(kd^2)$ parameters.

Proof of upper bound: compression

Definition

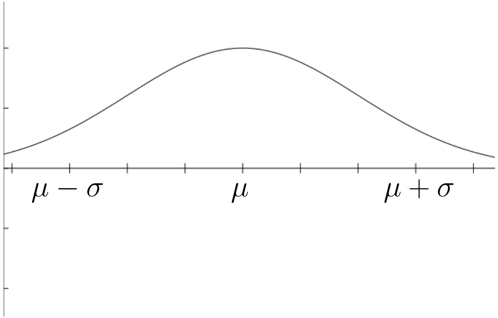
$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, there exist τ data points from which $\mathbf{P}$ can be reconstructed.

Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, there exist τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

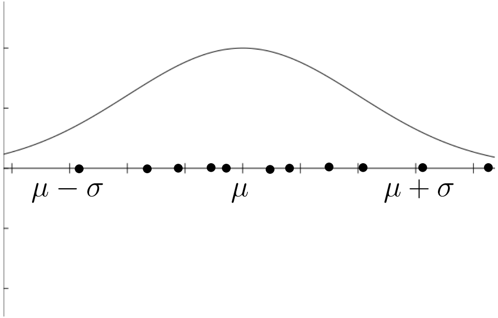


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

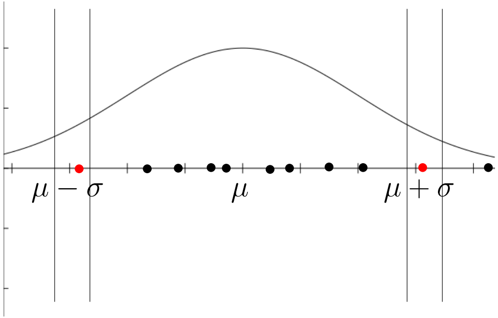


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

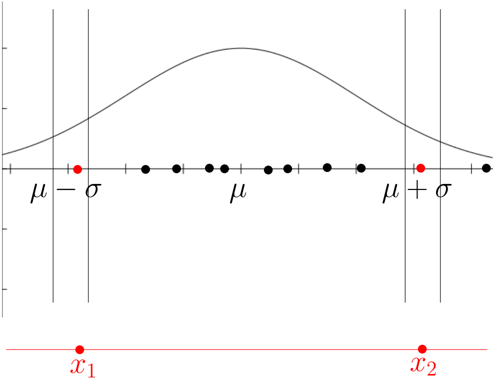


Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.



Proof of upper bound: compression

Definition

$\mathcal{F}$ admits τ-compression, if for any $\mathbf{P} \in \mathcal{F}$, you can find τ data points from which $\mathbf{P}$ can be reconstructed.

Example: 1 dimensional Gaussians admit 2-compression.

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

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

2. Compressing mixtures

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

3. Compression implies learnability

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

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

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) \leq (kd^2/\varepsilon^2) \times \mathrm{polylog}(kd/\varepsilon)$.

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

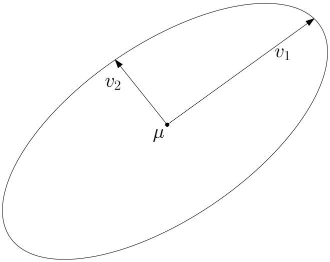
Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

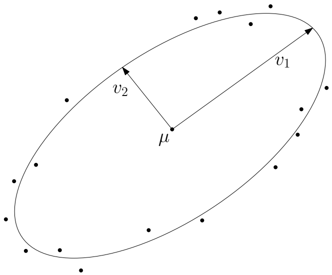


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

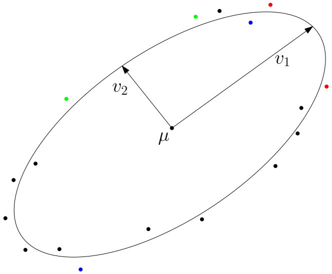


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

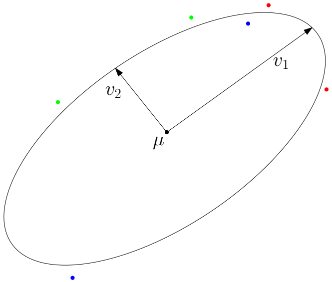


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^{\mathsf{T}} + v_2 v_2^{\mathsf{T}})$.

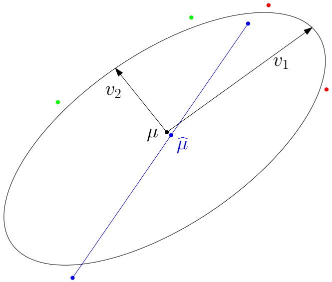


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\top + v_2 v_2^\top)$.

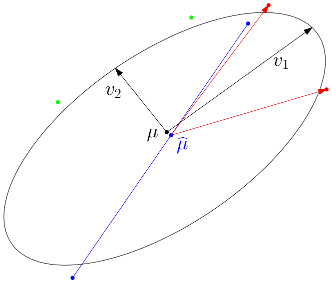


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

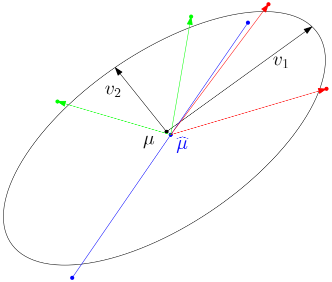


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.

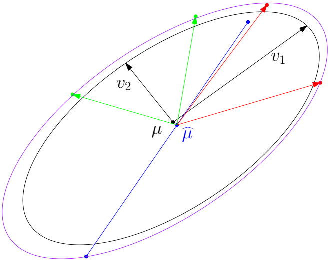


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\intercal + v_2 v_2^\intercal)$.

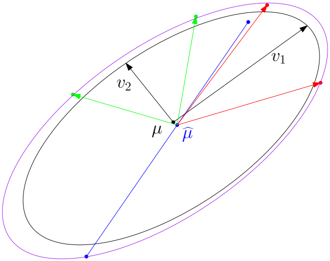


Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

Suppose $d = 2$, consider $\mathbf{P} = \mathcal{N}(\mu, \Sigma) = \mathcal{N}(\mu, v_1 v_1^\mathsf{T} + v_2 v_2^\mathsf{T})$.



For $d > 2$, use $d \log d$ data points to 'encode' the mean, and $d \log d$ data points for each eigenvector.

Proof of upper bound: compression

2. Compressing mixtures

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

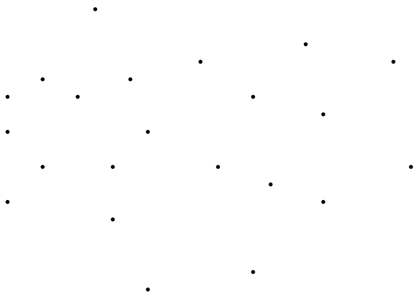
Let $\mathbf{P} = \frac{1}{3}P_1 + \frac{1}{3}P_2 + \frac{1}{3}P_3$, where each P_i is 2-compressible.

Proof of upper bound: compression

2. Compressing mixtures

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

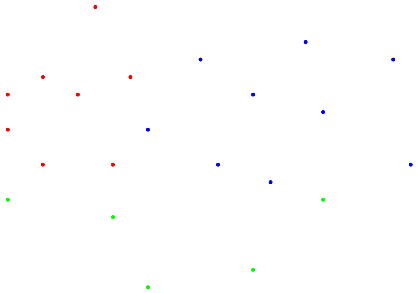


Proof of upper bound: compression

2. Compressing mixtures

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

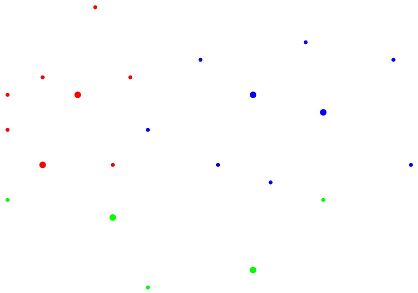


Proof of upper bound: compression

2. Compressing mixtures

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

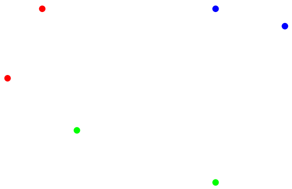


Proof of upper bound: compression

2. Compressing mixtures

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

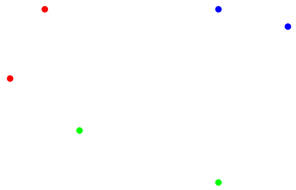


Proof of upper bound: compression

2. Compressing mixtures

If $\mathcal{F}$ admits τ-compression, then k-mix$(\mathcal{F})$ admits $(k\tau + k\log k)$-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 $\widehat{\mathbf{P}} = \frac{1}{3}\widehat{P_1} + \frac{1}{3}\widehat{P_2} + \frac{1}{3}\widehat{P_3}$

Proof of upper bound: compression

3. Compression implies learnability

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

First, generate a sample of size $m = \mathrm{poly}(\tau)$.
Try to reconstruct the distribution by considering all $\binom{m}{\tau}$ subsets of size τ (we know one of them is correct).

Proof of upper bound: compression

3. Compression implies learnability

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

First, generate a sample of size $m = \mathrm{poly}(\tau)$.
Try to reconstruct the distribution by considering all $\binom{m}{\tau}$ subsets of size τ (we know one of them is correct).

Theorem (Devroye and Lugosi'01)

Given a finite set $\mathcal{C}$ of candidates, given $\log(|\mathcal{C}|)/\epsilon^2$ *additional samples from the target distribution, we can find the candidate that is closest to the target.*

In our case, $|C| = \binom{m}{\tau} \leq m^\tau$, hence total sample complexity $< \tau \log(m)/\varepsilon^2$.

Proof of upper bound: compression

1. Compressing d-dimensional Gaussians

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

2. Compressing mixtures

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

3. Compression implies learnability

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

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

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then, $m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{O}(kd^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then, $m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

Fano's lemma

Suppose there exist $f_1, \dots, f_M \in \mathcal{F}$ with

$$\mathrm{KL}(f_i \parallel f_j) = O(\varepsilon^2) \text{ and } \mathrm{TV}(f_i, f_j) = \Omega(\varepsilon) \qquad \forall i \neq j \in [M].$$

Then $m_{\mathcal{F}}(\varepsilon) = \Omega(\log M/\varepsilon^2)$.

$$\mathrm{KL}(f_1 \parallel f_2) := \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \mathrm{d}x$$

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then, $m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

Fano's lemma

Suppose there exist $f_1, \ldots, f_M \in \mathcal{F}$ with

$$\mathrm{KL}(f_i \parallel f_j) = O(\varepsilon^2) \text{ and } \mathrm{TV}(f_i, f_j) = \Omega(\varepsilon) \qquad \forall i \neq j \in [M].$$

Then $m_{\mathcal{F}}(\varepsilon) = \Omega(\log M/\varepsilon^2)$.

$$\mathrm{KL}(f_1 \parallel f_2) := \int f_1(x) \log \frac{f_1(x)}{f_2(x)} \mathrm{d}x$$

To apply this lemma, we need to build 2^{d^2} Gaussian distributions, with pairwise KL-divergence $\leq \varepsilon^2$, pairwise TV distance $\geq \varepsilon$.

Proof of lower bound via Fano's inequality

Need to build 2^{d^2} Gaussian distributions with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise TV distance $\geq \varepsilon$.
We will use zero-mean Gaussians, so just need to specify the covariance matrices.

Proof of lower bound via Fano's inequality

Need to build 2^{d^2} Gaussian distributions with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise TV distance $\geq \varepsilon$.
We will use zero-mean Gaussians, so just need to specify the covariance matrices.

First construction (geometric). 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: $\Sigma = I + \frac{\varepsilon}{\sqrt{d}} UU^\top$, with $U \in \mathbb{R}^{d \times d/9}$ orthonormal.
Then prove that with large probability, any two of these have TV distance $\geq \varepsilon$.

Proof of lower bound via Fano's inequality

Need to build 2^{d^2} Gaussian distributions with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise TV distance $\geq \varepsilon$.
We will use zero-mean Gaussians, so just need to specify the covariance matrices.

Second construction (combinatorial). 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.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d} = d$-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

It is easy to lift this to the class of mixtures, proving
$m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Omega}(kd^2/\varepsilon^2)$.

Proof of lower bound: Fano's inequality

Main lemma

Let $\mathcal{F}_{1,d}$ = d-dimensional Gaussians. Then,
$m_{\mathcal{F}_{1,d}}(\varepsilon) = \widetilde{\Omega}(d^2/\varepsilon^2)$.

It is easy to lift this to the class of mixtures, proving
$m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Omega}(kd^2/\varepsilon^2)$.

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

Let $\mathcal{F}_{k,d}$ = mixtures of k Gaussians in d dimensions. Then,
$m_{\mathcal{F}_{k,d}}(\varepsilon) = kd^2/\varepsilon^2$ *up to polylogarithmic factors.*

The Ising model

The Ising model

Definition

For a graph G on d vertices, and edge weights $\{w_{i,j}\}_{ij \in E(G)}$, the Ising model with parameters $\{w_{i,j}\}_{ij \in E(G)}$ is supported on $\{-1, +1\}^d$ and has probability mass function

$$p_{\mathrm{w}}(x_1, \ldots, x_d) \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j\right)$$

Number of parameters $= |E(G)|$.

The Ising model

Definition

For a graph G on d vertices, and edge weights $\{w_{i,j}\}_{ij \in E(G)}$, the Ising model with parameters $\{w_{i,j}\}_{ij \in E(G)}$ is supported on $\{-1,+1\}^d$ and has probability mass function

$$p_{\mathrm{w}}(x_1,\ldots,x_d) \propto \exp\left(\sum_{ij \in E(G)} w_{i,j} x_i x_j\right)$$

Number of parameters $= |E(G)|$.

Theorem (Devroye, M, Reddad'18)

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

Proof of lower bound: Fano's inequality

Lower bound proof uses Fano's inequality again.
Need to build $2^{|E(G)|}$ Ising models with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise TV distance $\geq \varepsilon$.

Proof of lower bound: Fano's inequality

Lower bound proof uses Fano's inequality again.
Need to build $2^{|E(G)|}$ Ising models with pairwise KL-divergence $\leq \varepsilon^2$ and pairwise TV distance $\geq \varepsilon$.

For $d = 3$ and G the complete graph, consider the following weight matrices W:

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

For a general interaction graph G, build $2^{|E(G)|/5}$ weight matrices so that any two of them are different in at least $|E(G)|/6$ coordinates.

Proof of upper bound

For class $\mathcal{F}$ of densities defined over X, consider the Yatracos set system:

$$A_{\mathcal{F}} \coloneqq \{S \subseteq X : \exists p_1, p_2 \in \mathcal{F} \text{ s. t. } S = \{x \in X : p_1(x) > p_2(x)\}\}$$

Proof of upper bound

For class $\mathcal{F}$ of densities defined over X, consider the Yatracos set system:

$$A_{\mathcal{F}} \coloneqq \{S \subseteq X : \exists p_1, p_2 \in \mathcal{F} \text{ s. t. } S = \{x \in X : p_1(x) > p_2(x)\}\}$$

Devroye and Lugosi'01 proved $m_{\mathcal{F}}(\varepsilon) \leq C \cdot \text{VC-dim}(A_{\mathcal{F}})/\varepsilon^2$.

Proof of upper bound

For class $\mathcal{F}$ of densities defined over X, consider the Yatracos set system:

$$A_{\mathcal{F}} \coloneqq \{S \subseteq X : \exists p_1, p_2 \in \mathcal{F} \text{ s. t. } S = \{x \in X : p_1(x) > p_2(x)\}\}$$

Devroye and Lugosi'01 proved $m_{\mathcal{F}}(\varepsilon) \le C \cdot \text{VC-dim}(A_{\mathcal{F}})/\varepsilon^2$.

If $\mathcal{F}$ is the class of Ising models on G, standard techniques give $\text{VC-dim}(A_{\mathcal{F}}) \le |E(G)| + 1$, whence $m_{\mathcal{F}}(\varepsilon) \le C(|E(G)| + 1)/\varepsilon^2$.

Theorem (Devroye, M, Reddad'18)

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

Recap

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

Example

- ✓ $\mathcal{F} =$ Bernoulli distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} =$ Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp 1/\varepsilon^2$
- ✓ $\mathcal{F} = d$-dimensional Gaussian distributions: $m_{\mathcal{F}}(\varepsilon) \asymp d^2/\varepsilon^2$
- ✓ Finite $\mathcal{F}$: $m_{\mathcal{F}}(\varepsilon) \leq 9 \log |\mathcal{F}|/\varepsilon^2$ Devroye-Lugosi'01
- ✓ $\mathcal{F}_{k,d} =$ mixture of k Gaussians in d dimensions: $m_{\mathcal{F}_{k,d}}(\varepsilon) = \widetilde{\Theta}(kd^2/\varepsilon^2)$.
- ✓ $\mathcal{I}_G =$ Ising models on G: $m_{\mathcal{I}_G}(\varepsilon) \asymp |E(G)|/\varepsilon^2$.

Questions

$\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \asymp$ number of free parameters in $\mathcal{F}$ in 'natural representation'

1. Does the heuristic works for other classes? For example, other exponential families, graphical models, distributions generated by neural networks?
2. $\varepsilon^2 m_{\mathcal{F}}(\varepsilon) \leq$ smallest compression size of $\mathcal{F}$. Is the converse true?
3. Can we use $\varepsilon^2 m_{\mathcal{F}}(\varepsilon)$ as a natural definition of 'dimension' for class $\mathcal{F}$? Are there connections with other dimensions?
4. What about computational complexity?