Settling the Sample Complexity of GMMs via Compression Schemes

Shai Ben-David
(Waterloo)

Nick Harvey
(UBC)

Chris Liaw
(UBC)

Abbas Mehrabian
(McGill)

Yaniv Plan
(UBC)

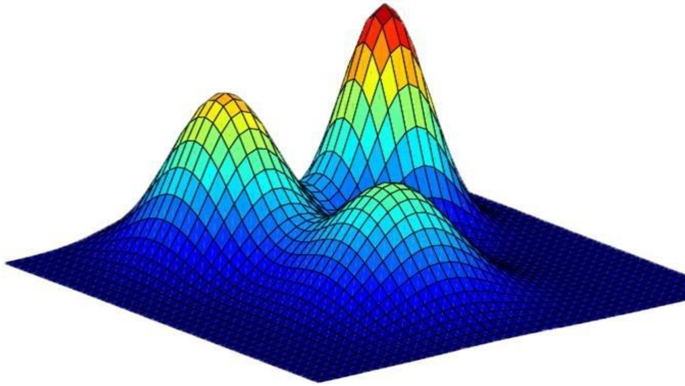
Hassan Ashtiani
McMaster & Vector

Density estimation

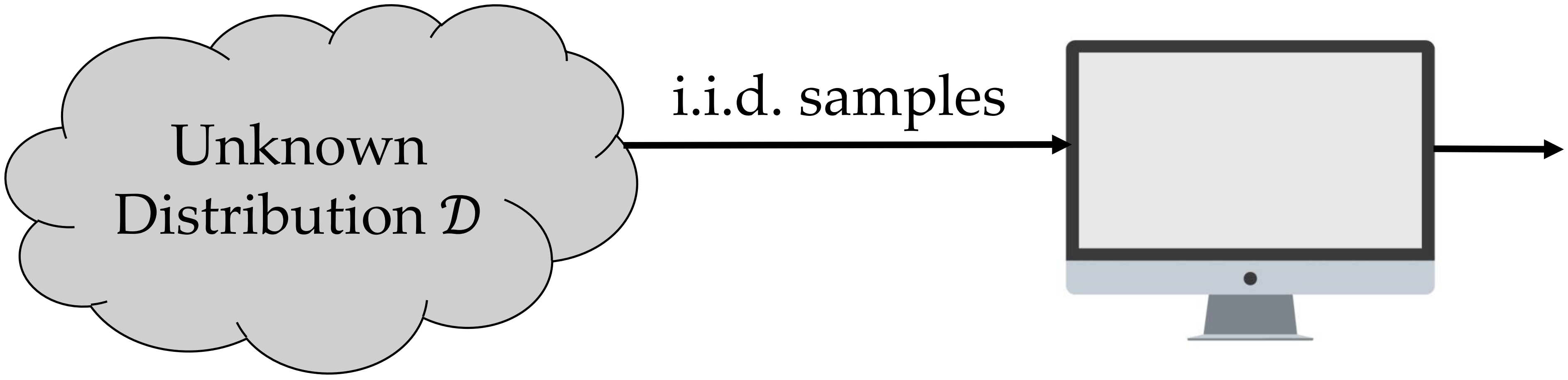
Unknown
Distribution $\mathcal{D}$

i.i.d. samples

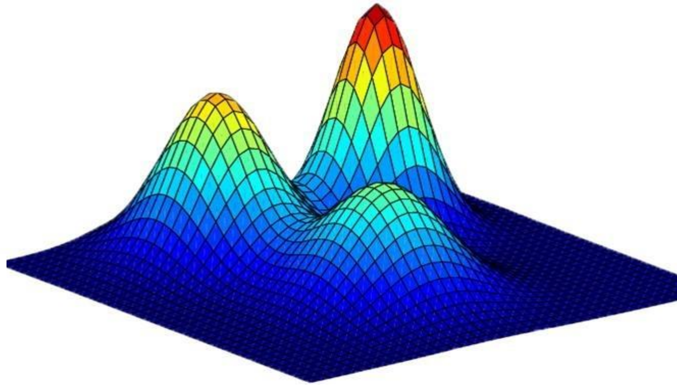
$\widehat{\mathcal{D}} \approx \mathcal{D}$
(in L_1-distance)



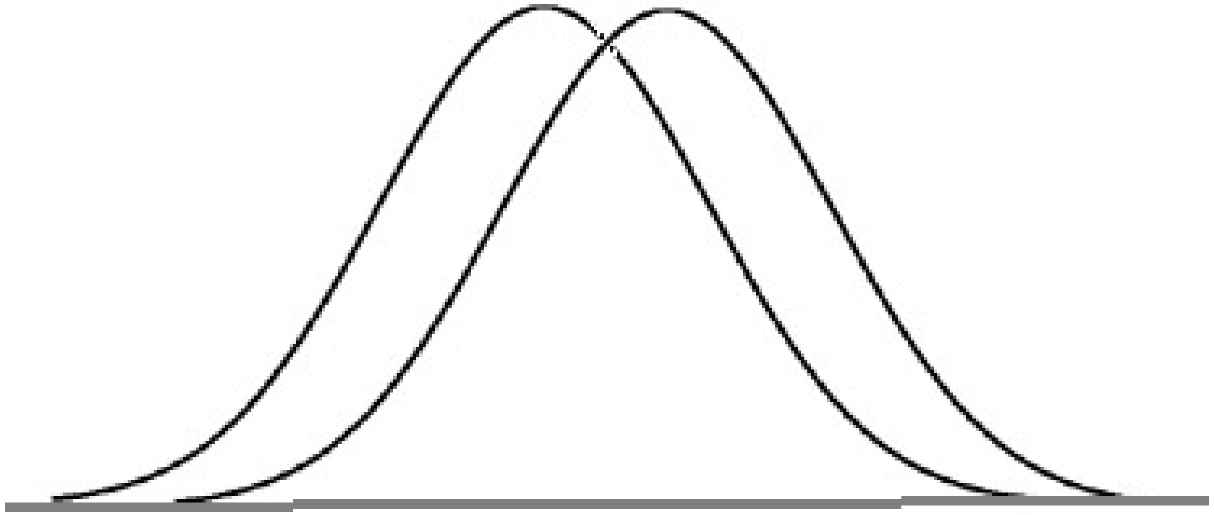
Density estimation



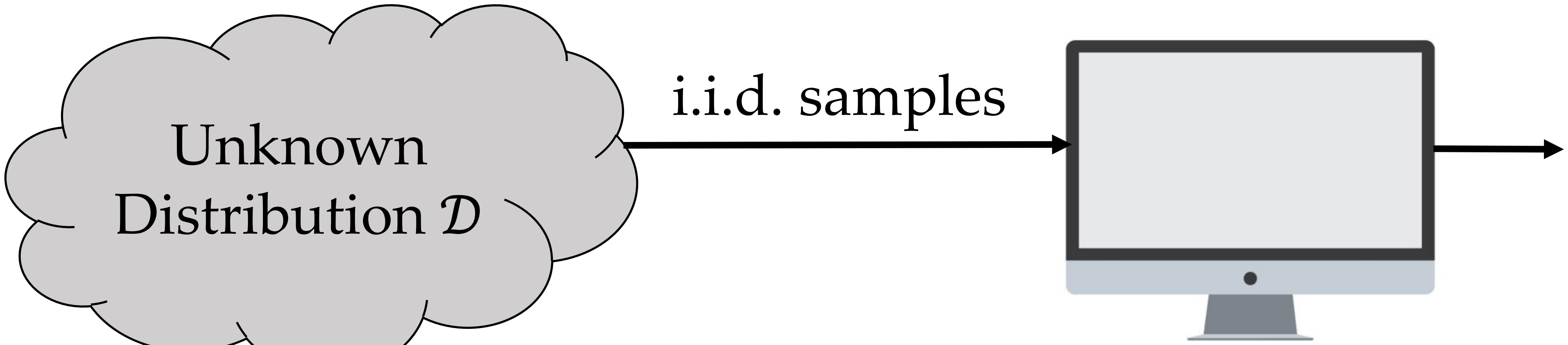
$\widehat{\mathcal{D}} \approx \mathcal{D}$
(in L_1-distance)



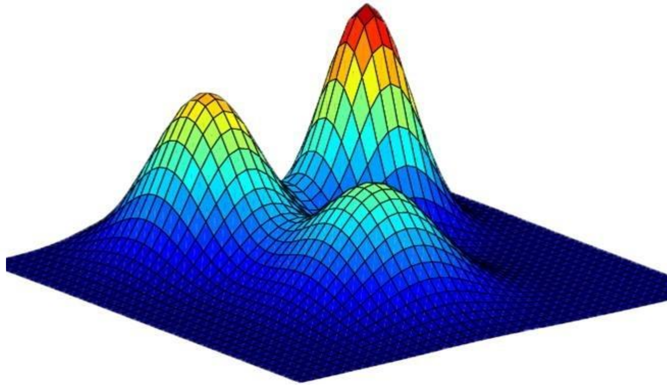
- ***L_1*-distance?**



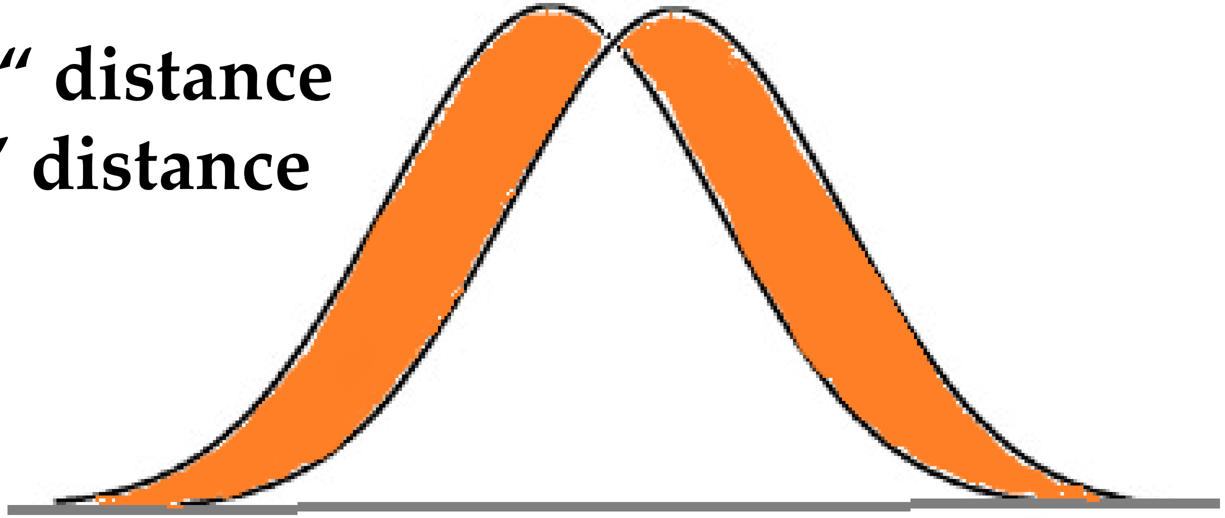
Density estimation



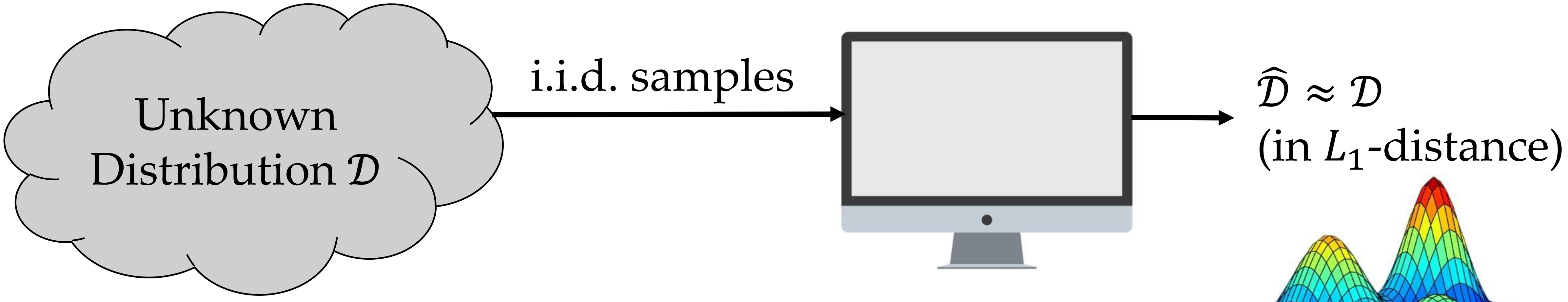
$\widehat{\mathcal{D}} \approx \mathcal{D}$
(in L_1-distance)

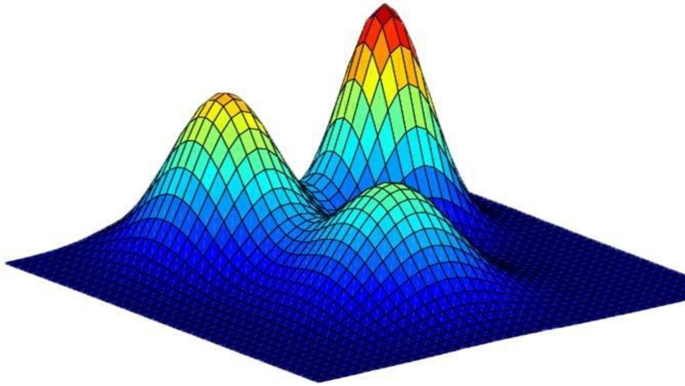


- ***L_1*-distance?**
- **“Total variation“ distance**
- **“The statistical” distance**



Density estimation



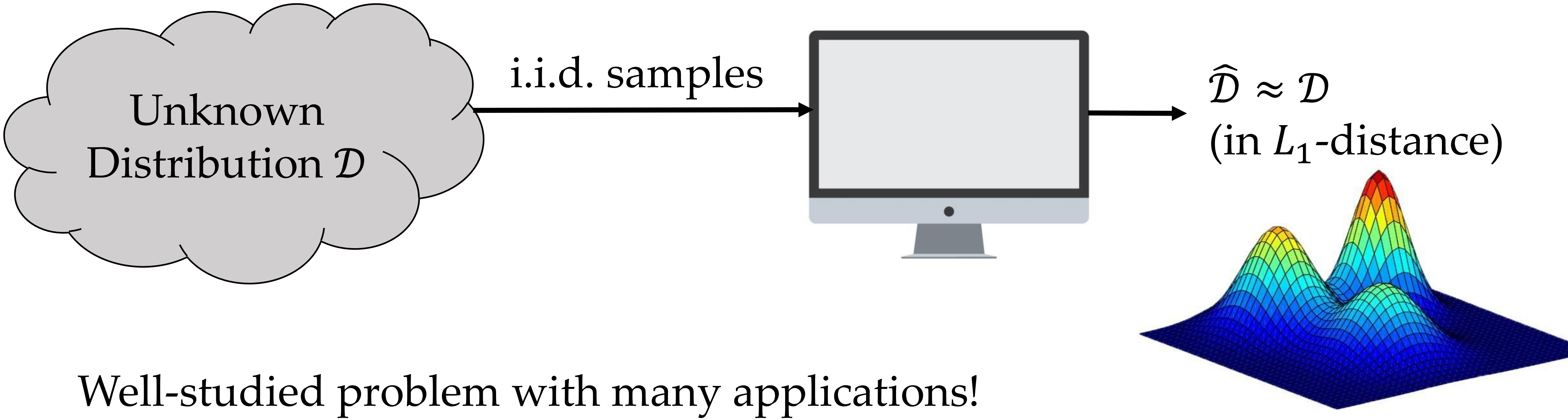


Well-studied problem with many applications!

[Feldman et al. '06; Suresh et al. '14; Ashtiani et al. '17; Diakonikolas et al. '14-'18, etc.]

Q [D '16]: "For a distribution class $\mathcal{F}$, is there a complexity measure that characterizes the **sample complexity** of $\mathcal{F}$?"

Density estimation



Well-studied problem with many applications!

[Feldman et al. ’06; Suresh et al. ’14; Ashtiani et al. ’17; Diakonikolas et al. ‘14-’18, etc.]

Q [D ‘16]: “For a distribution class $\mathcal{F}$, is there a complexity measure that characterizes the **sample complexity** of $\mathcal{F}$?”

“VC-dimension” of distribution learning?

The case of Gaussian Mixture Models

Studied for over a century!

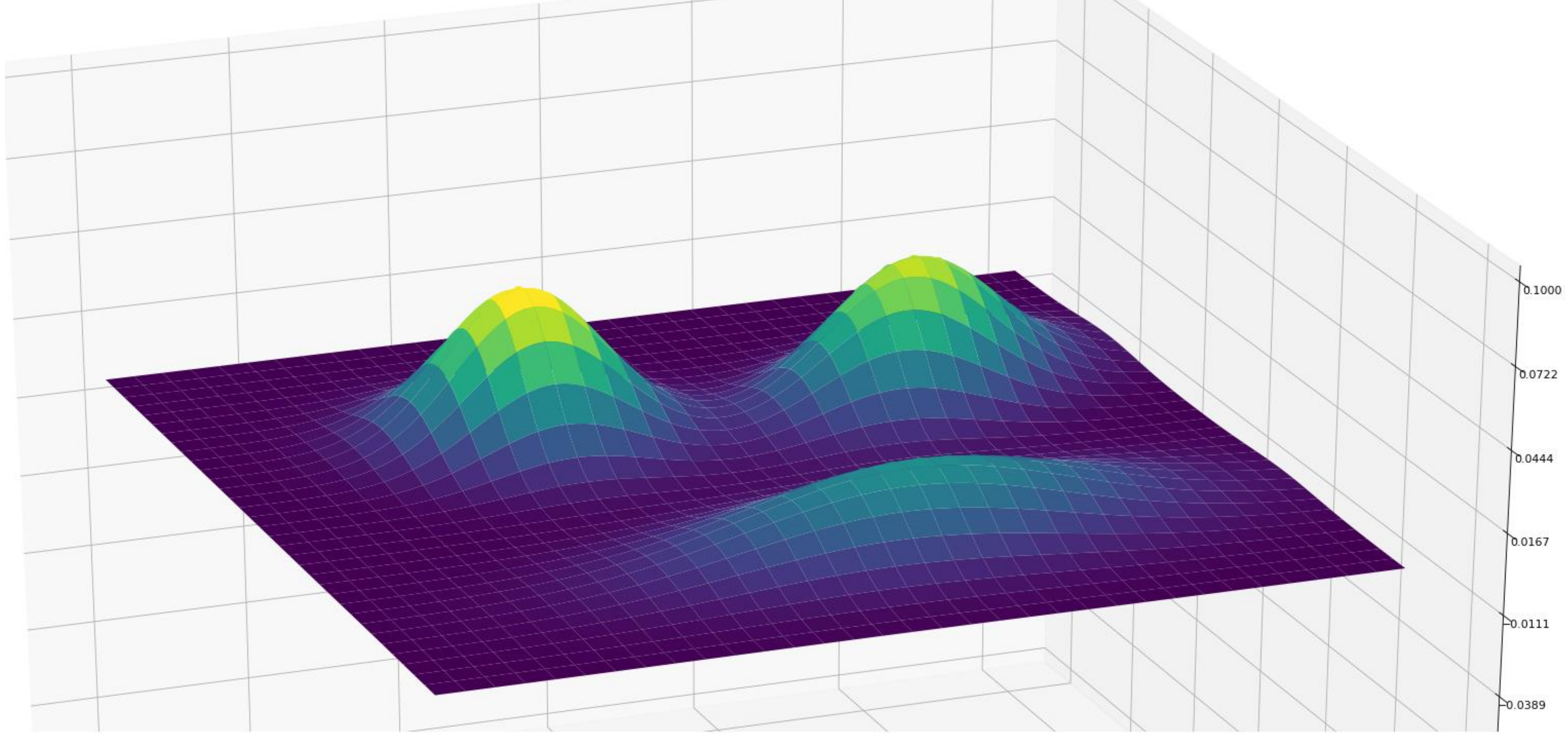
- Popular in practice
- One of the most basic universal density approximators
- Building blocks for more sophisticated density classes
- Natural way of extending Gaussians to multi-modal distributions

The case of Gaussian Mixture Models

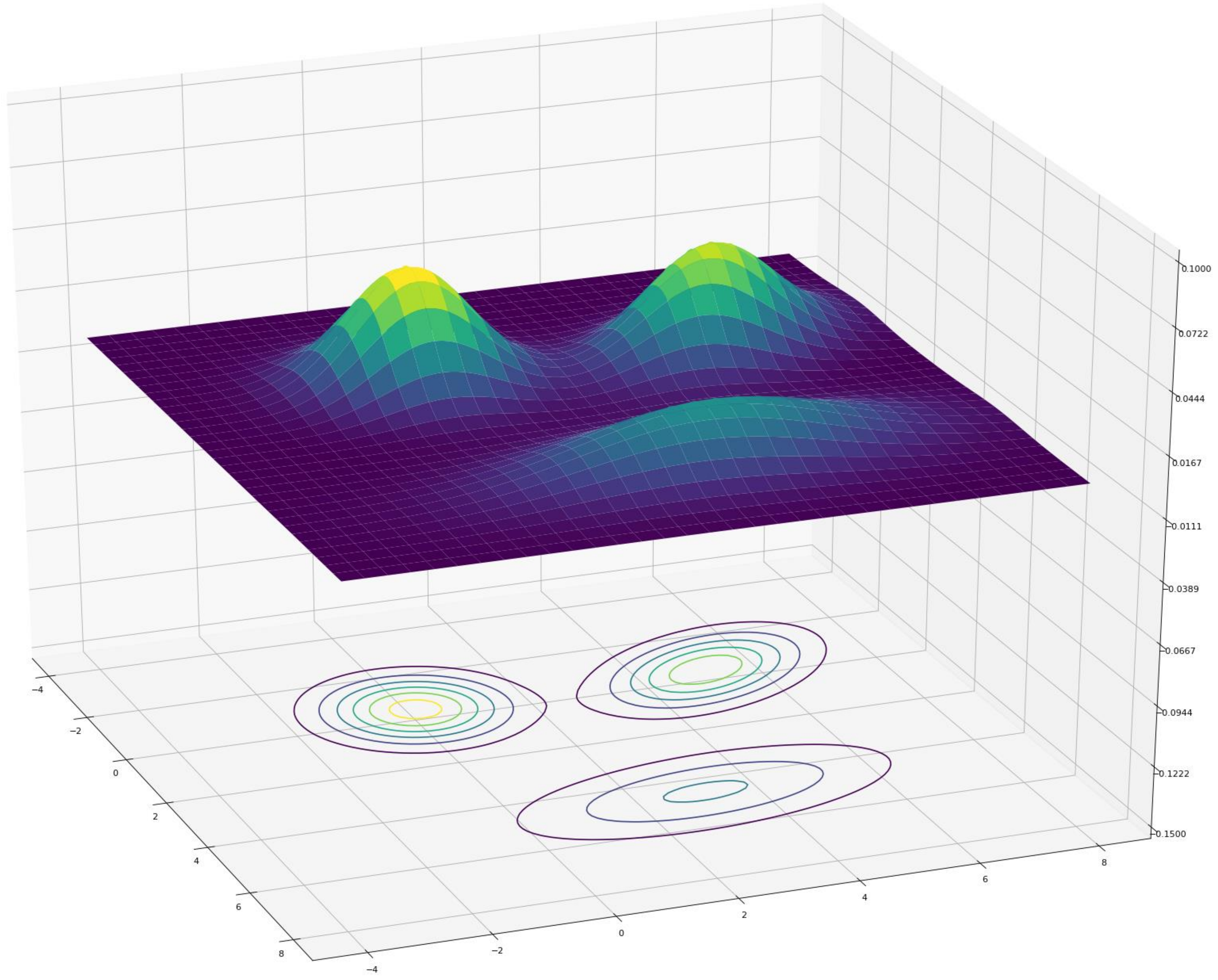
Studied for over a century!

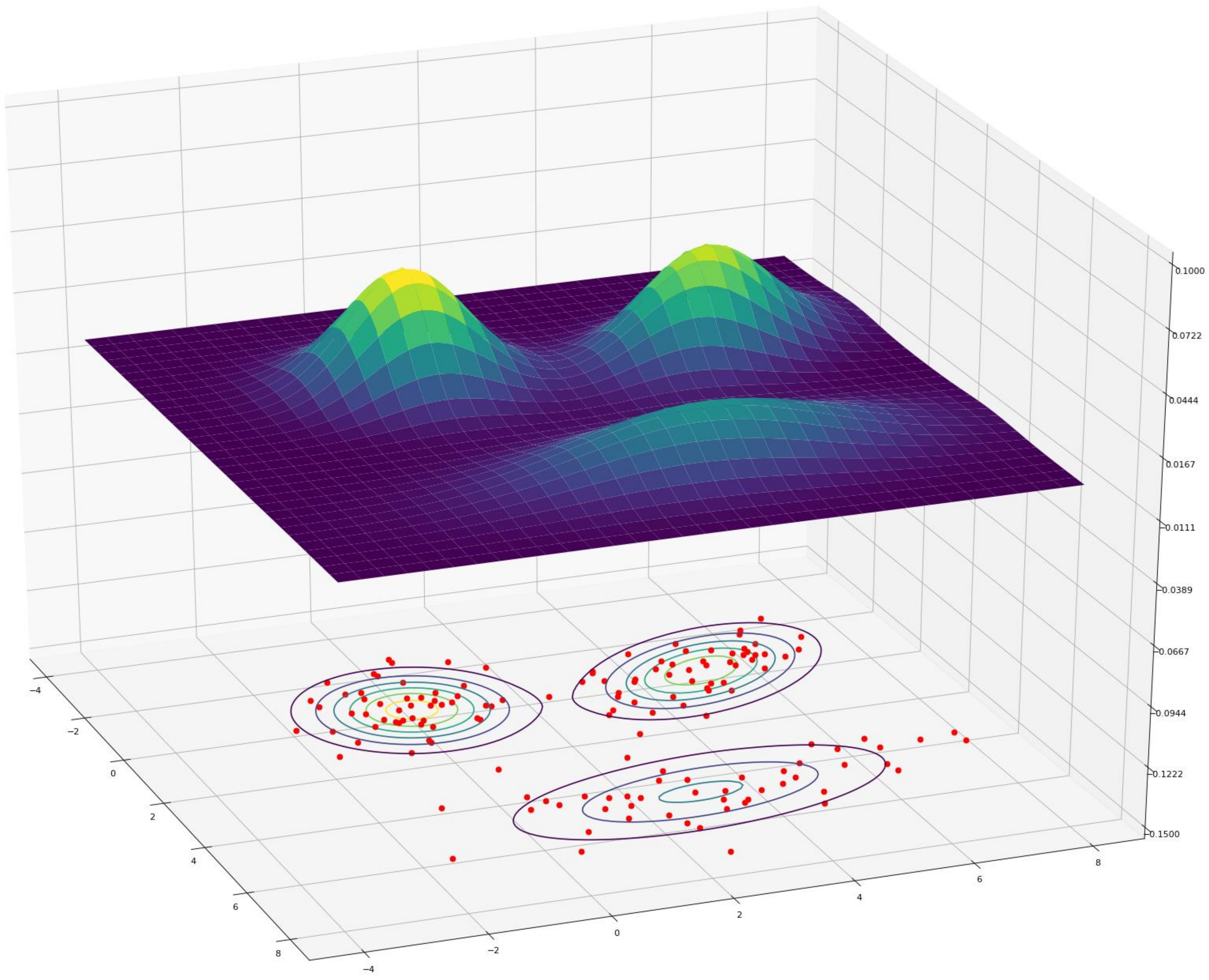
- Popular in practice
- One of the most basic universal density approximators
- Building blocks for more sophisticated density classes
- Natural way of extending Gaussians to multi-modal distributions

Surprisingly, not yet fully understood
- Sample complexity
- Computational complexity



- $f(x) = w_1 N(x|\mu_1, \Sigma_1) + w_2 N(x|\mu_2, \Sigma_2) + w_3 N(x|\mu_3, \Sigma_3)$

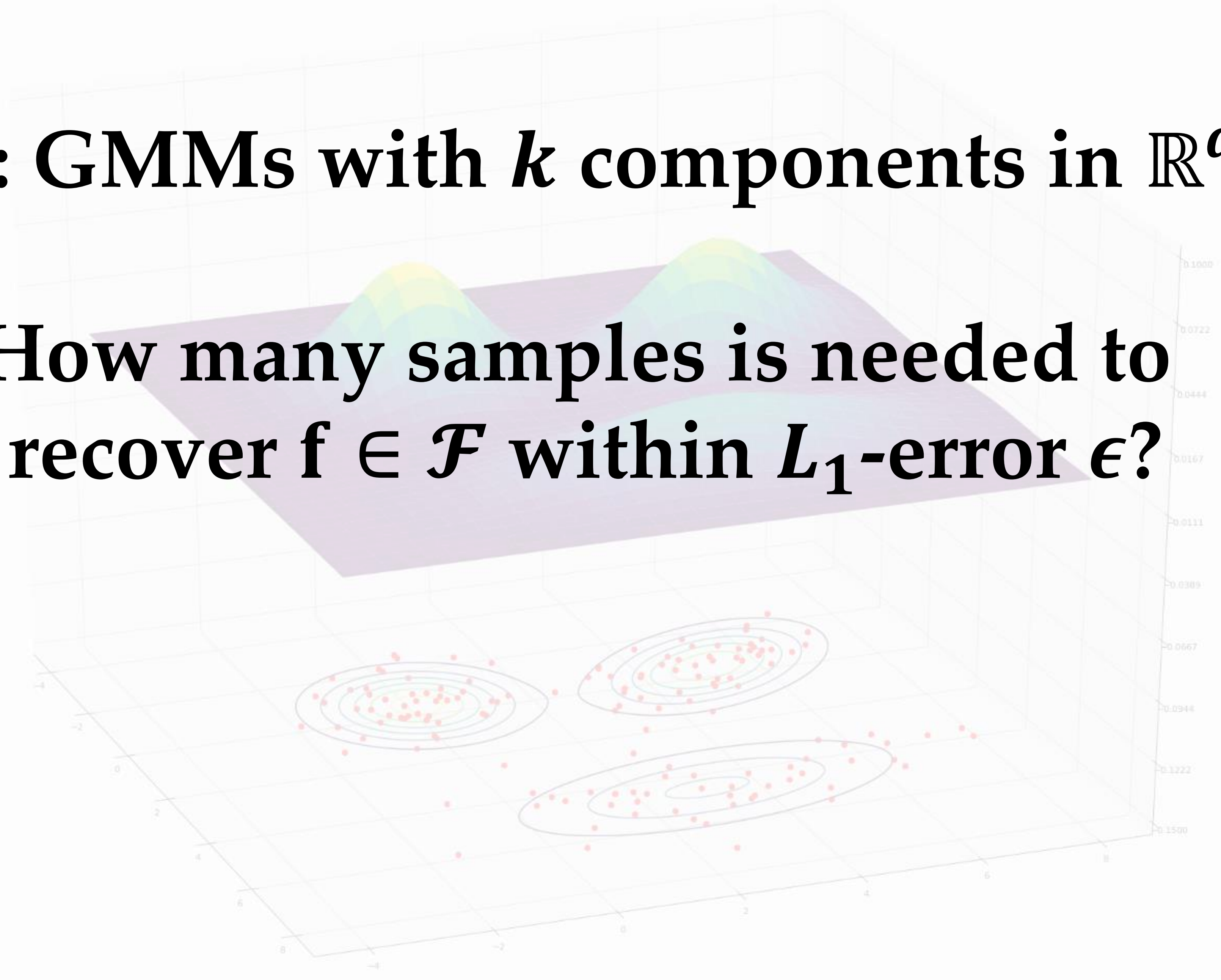

0.1000
0.0722
0.0444
0.0167
−0.0111
−0.0389
−0.0667
−0.0944
−0.1222
−0.1500
−4
−2
0
2
4
6
8
−4
−2
0
2
4
6
8



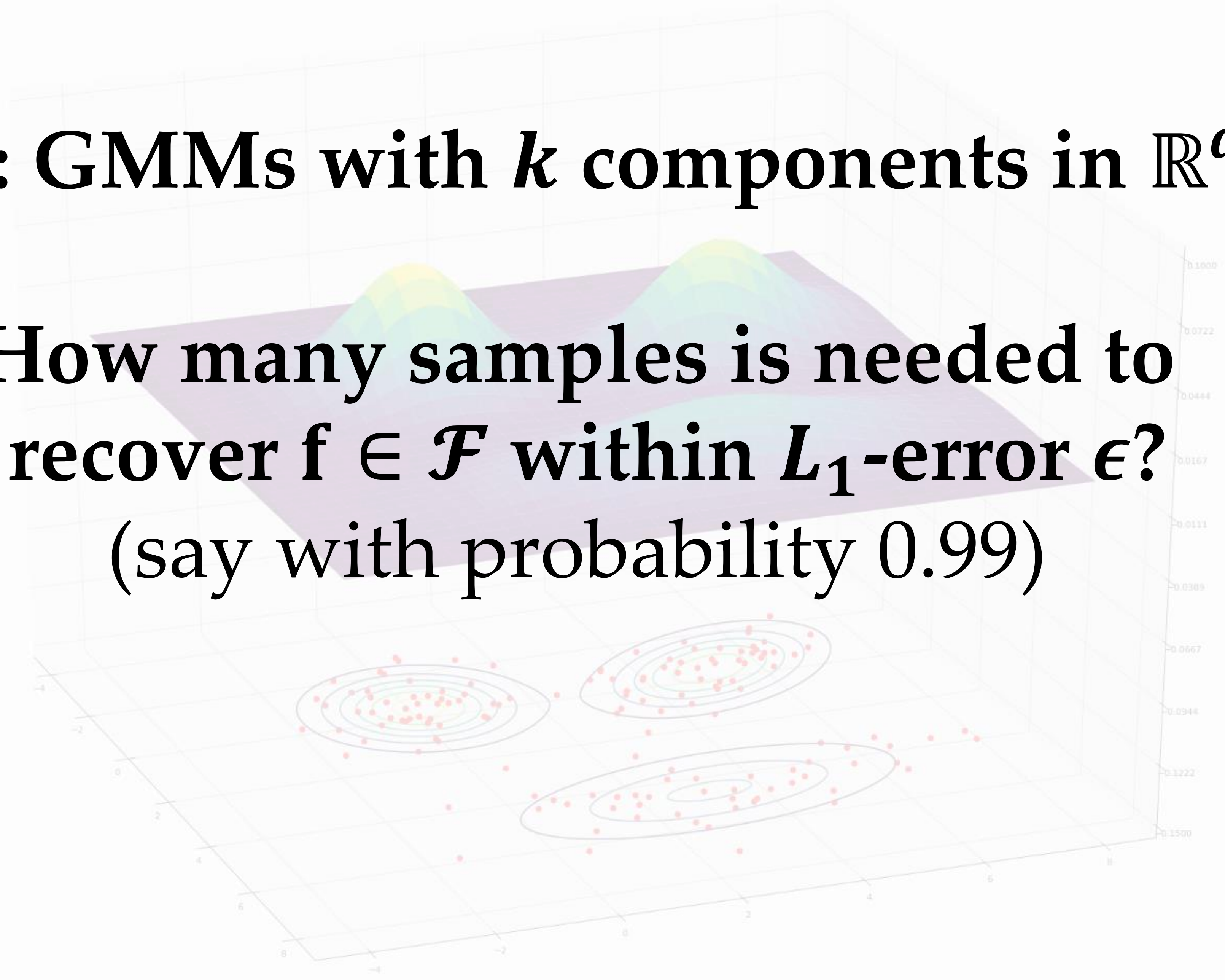
0.1000
0.0722
0.0444
0.0167
-0.0111
-0.0389
-0.0667
-0.0944
-0.1222
-0.1500
−4
−2
0
2
4
6
8
−4
−2
0
2
4
6
8

$\mathcal{F}$: GMMs with k components in $\mathbb{R}^d$

How many samples is needed to recover f $\in \mathcal{F}$ within L_1-error ϵ?



$\mathcal{F}$: GMMs with k components in $\mathbb{R}^d$



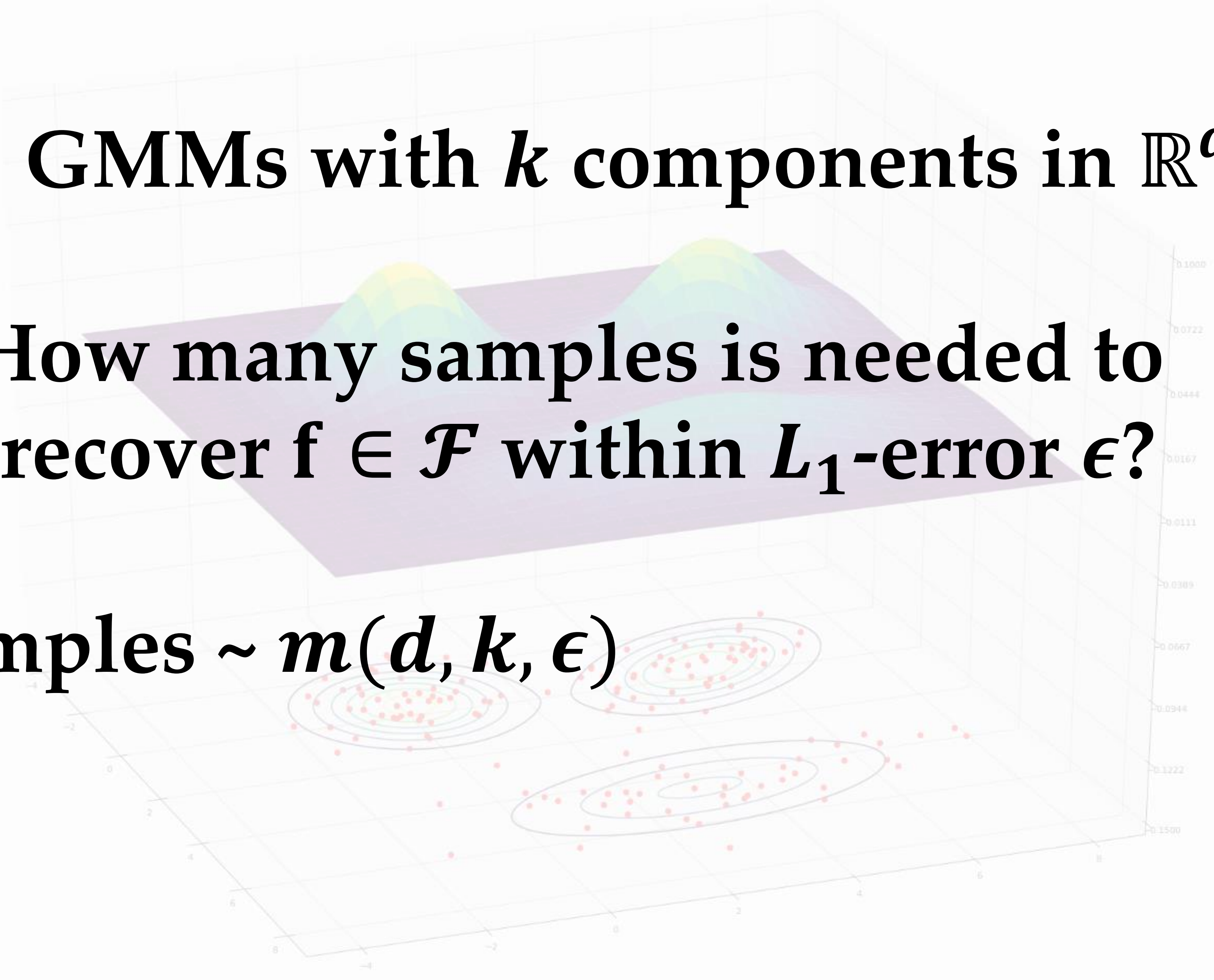
How many samples is needed to recover $\mathrm{f} \in \mathcal{F}$ within L_1-error ϵ?

(say with probability 0.99)

$\mathcal{F}$: **GMMs with k components in $\mathbb{R}^d$**

How many samples is needed to recover f $\in \mathcal{F}$ within L_1-error ϵ?

#samples ~ $m(d, k, \epsilon)$



$\mathcal{F}$: GMMs with k components in $\mathbb{R}^d$

How many samples is needed to recover f $\in \mathcal{F}$ within L_1-error ϵ?

#samples ~ $m(d, k, \epsilon)$

#samples ~ $m(d, k, \epsilon, f)$ (Worst-Case/Minimax)

No dependence on $||\mu||, \sigma_{max}, \sigma_{min}, \frac{\sigma_{max}}{\sigma_{min}}, \ldots$

Outline

We introduce **distribution compression schemes:**

A generic and simple technique for proving
sample complexity upper bounds
for density estimation

Outline

We introduce **distribution compression schemes:**

A generic and simple technique for proving sample complexity upper bounds for density estimation

For mixture of Gaussians with k components in $\mathbb{R}^d$:

- We show $\tilde{O}\left(\frac{kd^2}{\epsilon^2}\right)$ is sufficient
- We show $\tilde{\Omega}\left(\frac{kd^2}{\epsilon^2}\right)$ is necessary

*Note: $\tilde{O}$ and $\tilde{\Omega}$ hide polylog(kd/ϵ) factors.

Outline

We introduce **distribution compression schemes:**

A generic and simple technique for proving sample complexity upper bounds for density estimation

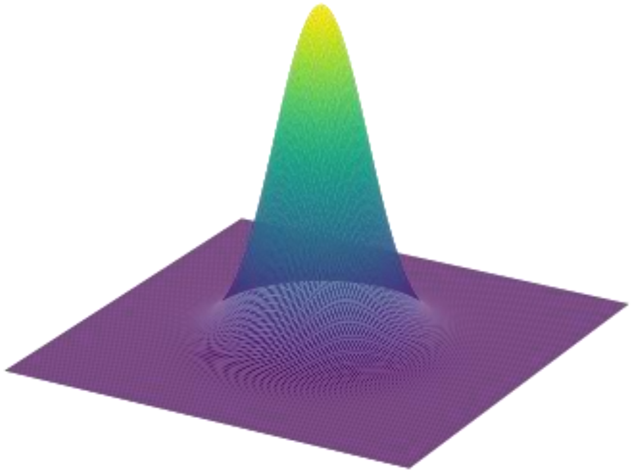
For mixture of Gaussians with k components in $\mathbb{R}^d$:

- We show $\tilde{O}\left(\frac{kd^2}{\epsilon^2}\right)$ is sufficient
- We show $\widetilde{\Omega}\left(\frac{kd^2}{\epsilon^2}\right)$ is necessary

Settles the sample complexity of GMMs
(within logarithmic factors)

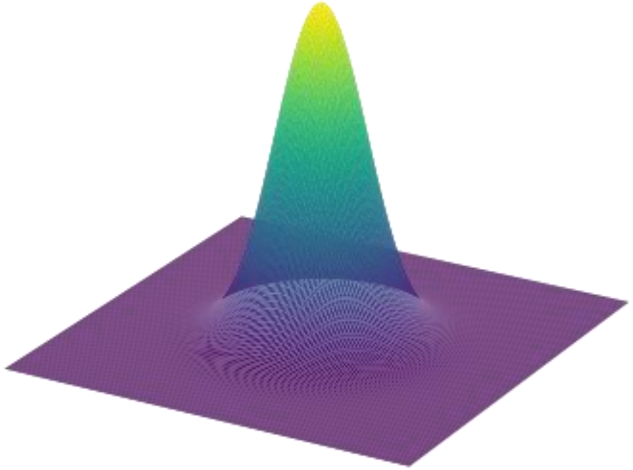
*Note: $\tilde{O}$ and $\widetilde{\Omega}$ hide polylog(kd/ϵ) factors.

Learning Gaussians



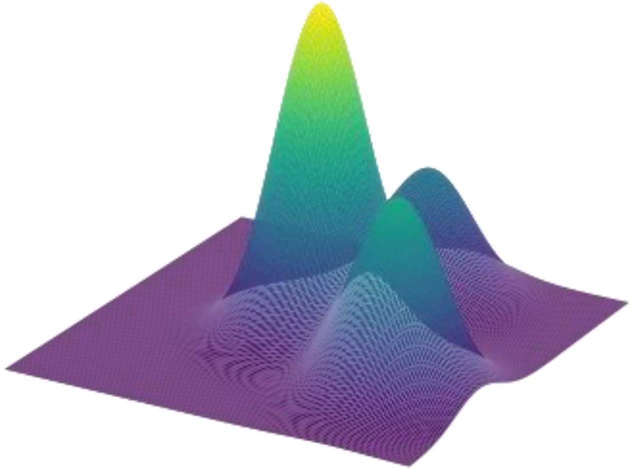
Single Gaussian in $\mathbb{R}^{\boldsymbol{d}}$.
$O\left(\frac{d^2}{\epsilon^2}\right) = O\left(\frac{\#params}{\epsilon^2}\right)$ samples are sufficient.

Learning Gaussians



Single Gaussian in $\mathbb{R}^d$.

$O\left(\frac{d^2}{\epsilon^2}\right) = O\left(\frac{\#params}{\epsilon^2}\right)$ samples are sufficient.



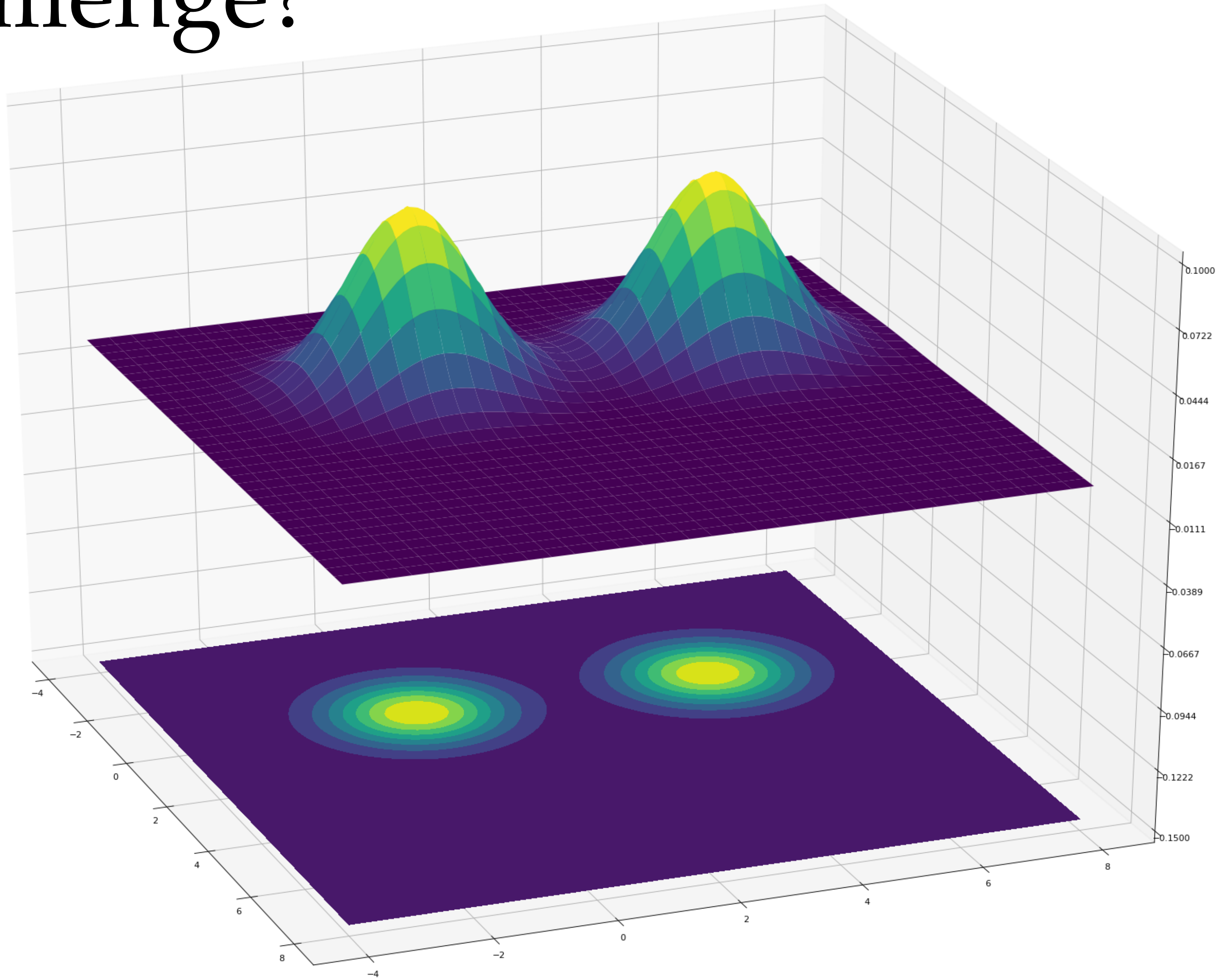
Mixture of k Gaussians in $\mathbb{R}^d$.

Q: Are $O\left(\frac{kd^2}{\epsilon^2}\right) = O\left(\frac{\#params}{\epsilon^2}\right)$ samples sufficient? **(Open problem)**

Note: We aim to recover density, *not* parameters of the mixture.

Where is the challenge?

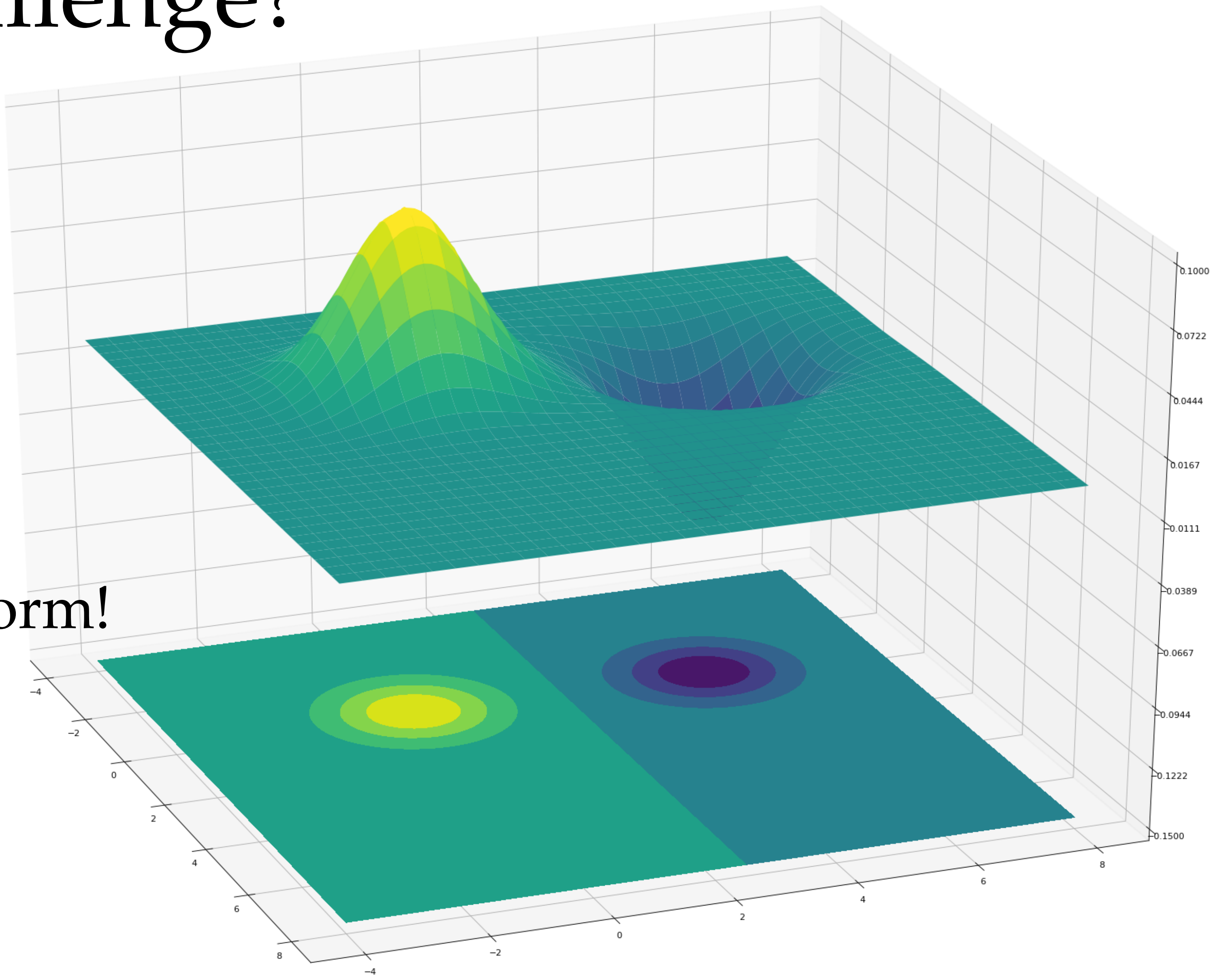
- For a moment look at this as a binary classification problem.



Where is the challenge?

- For a moment look at this as a binary classification problem.

- The decision boundary has a simple quadratic form!

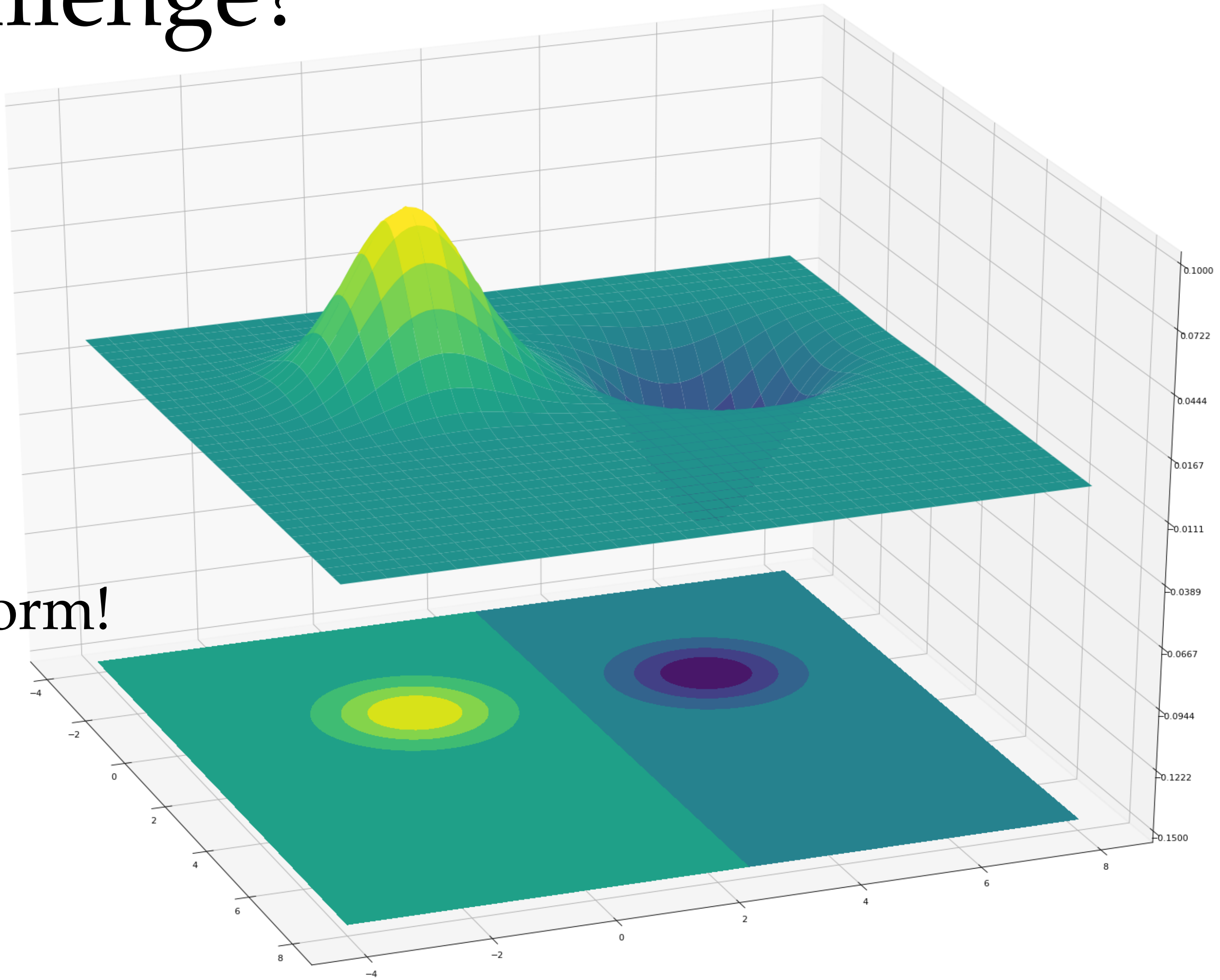


Where is the challenge?

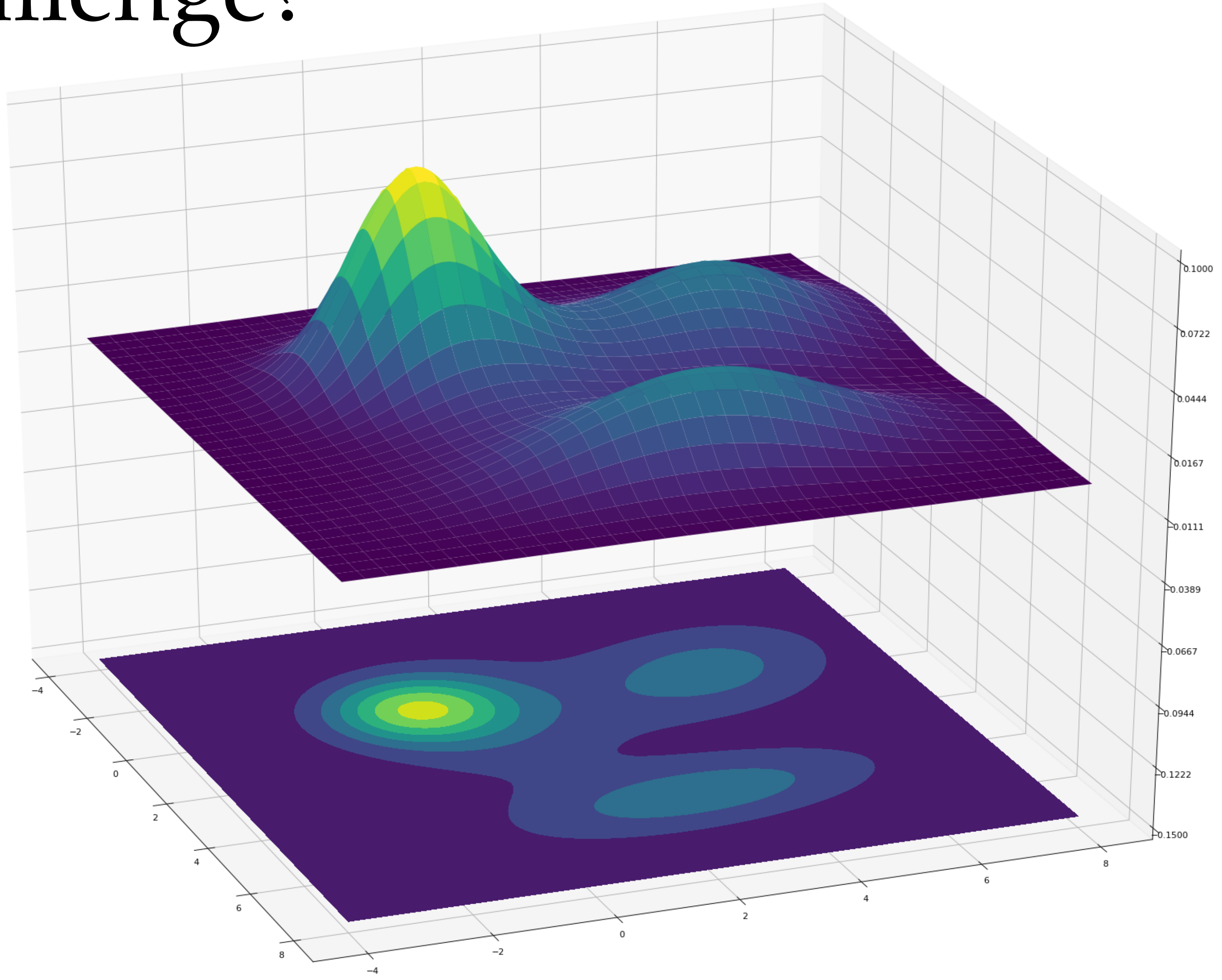
- For a moment look at this as a binary classification problem.

- The decision boundary has a simple quadratic form!

- VC-dim = $O(d^2)$

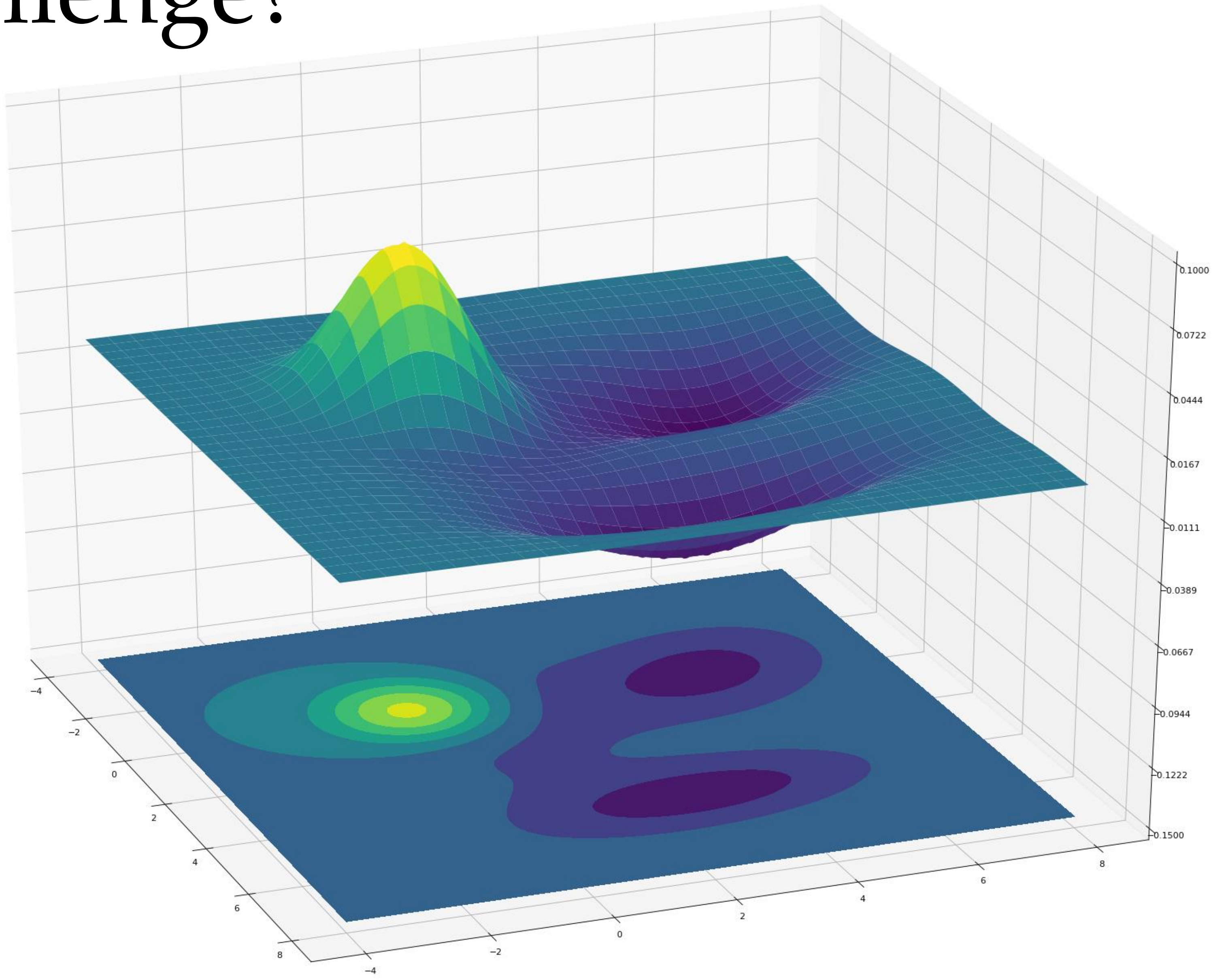


Where is the challenge?



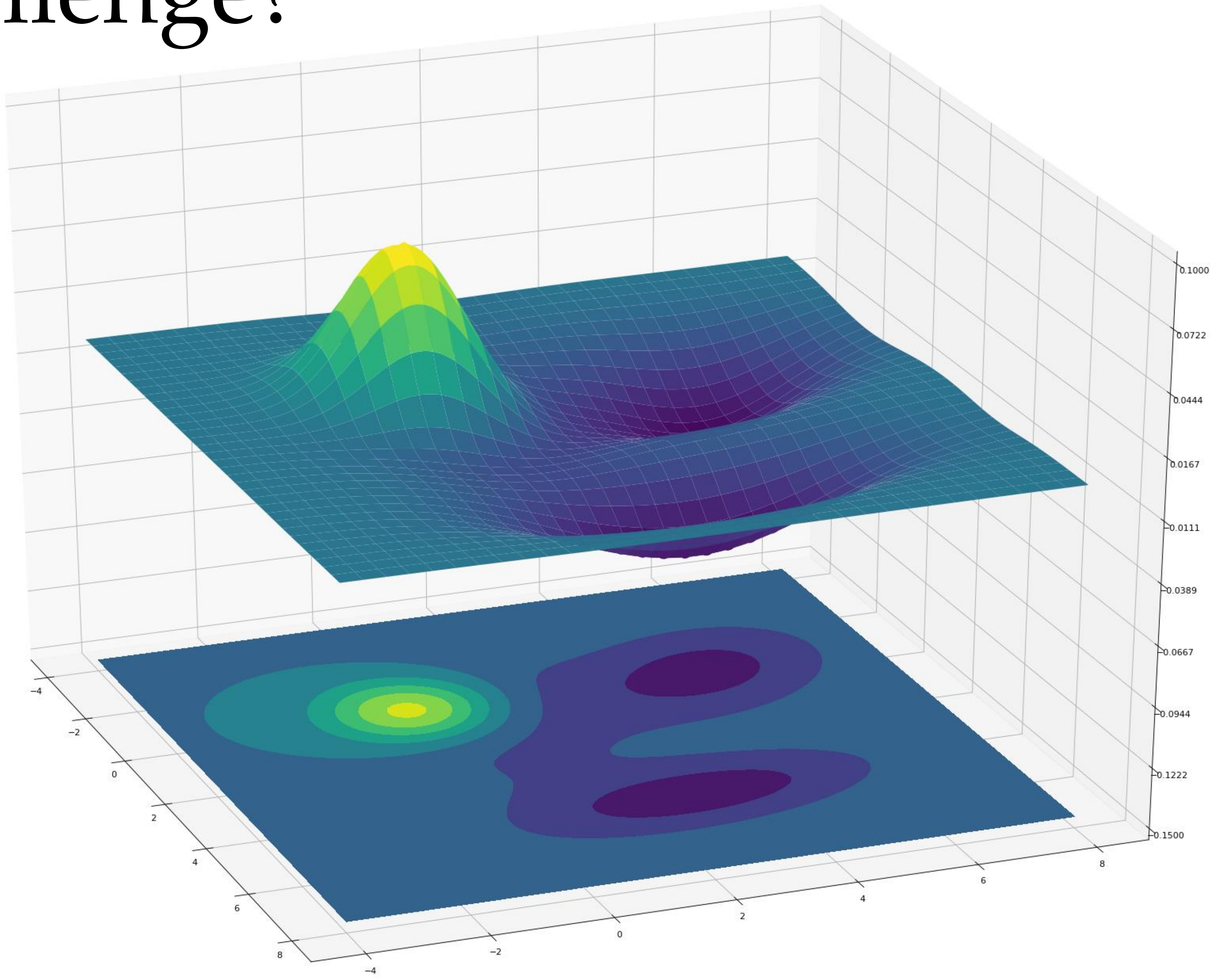
Where is the challenge?

- The decision boundary becomes very complex when the number of components is higher



Where is the challenge?

- The decision boundary becomes very complex when the number of components is higher
- VC-dimension?

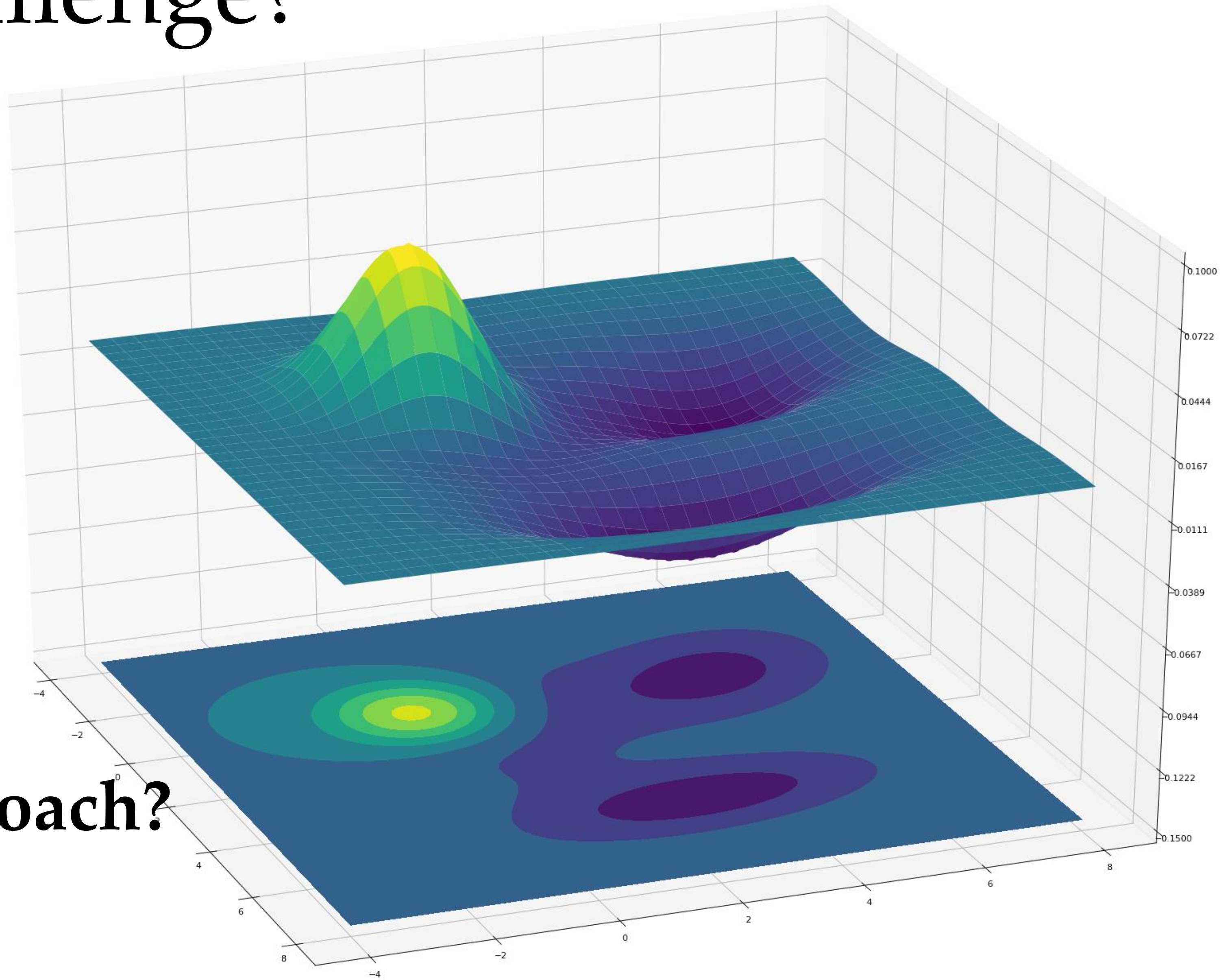


Where is the challenge?

- The decision boundary becomes very complex when the number of components is higher

- VC-dimension?

A more intuitive approach?

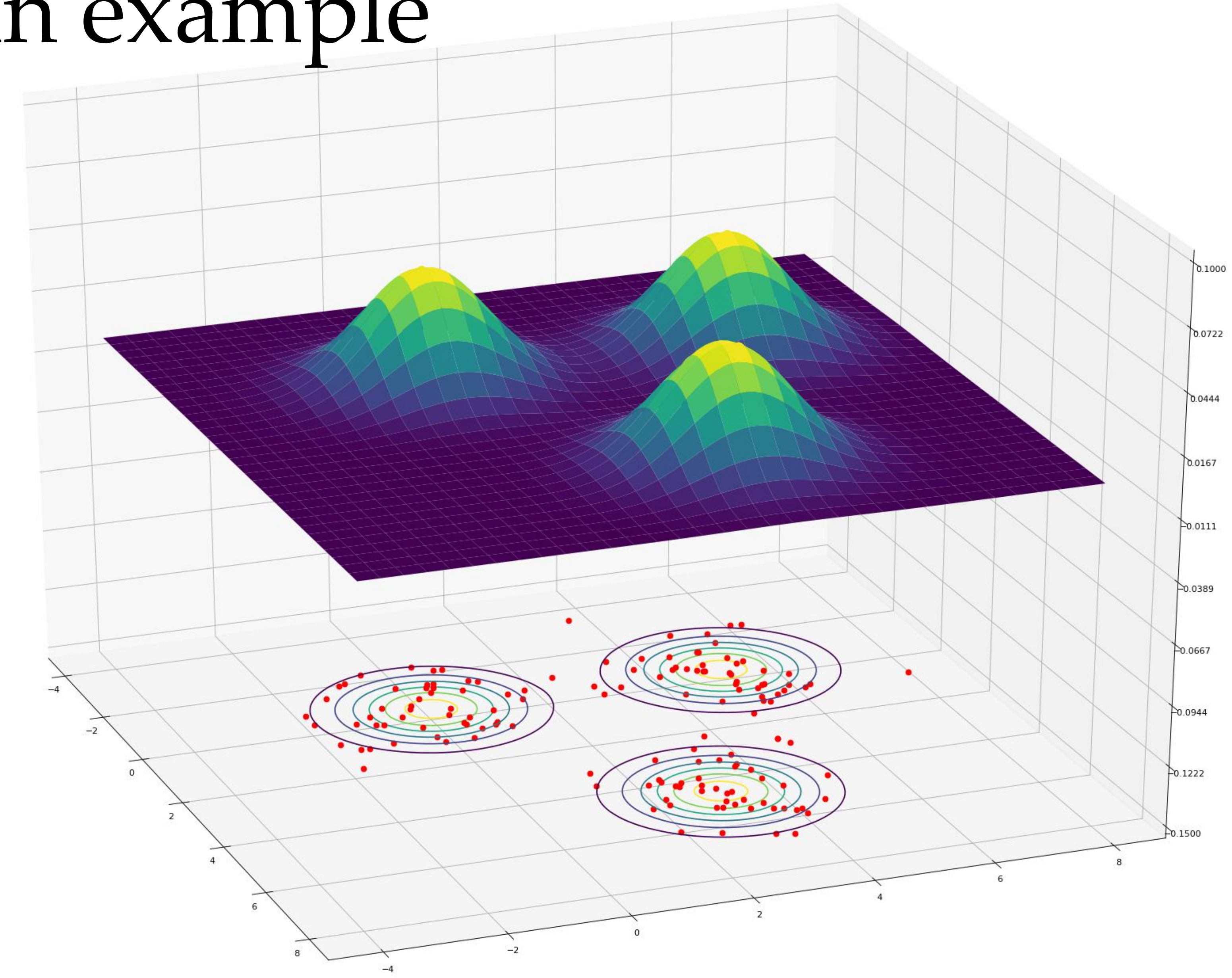


Compression: an example

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = I$$
$$w_1 = w_2 = w_3 = 1/3$$

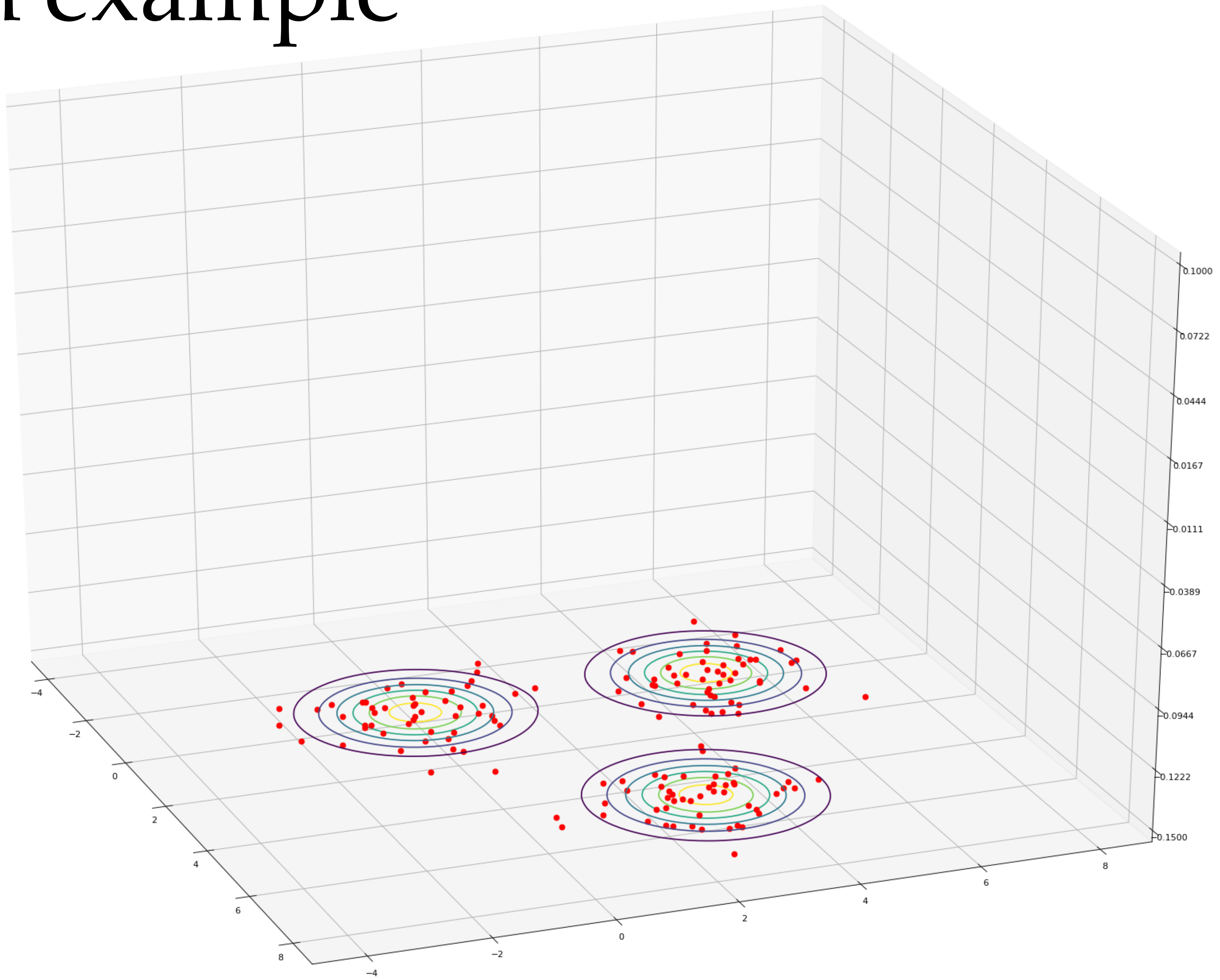
but

μ_1, μ_2, μ_3 are unknown



Compression: an example

Given S where $|S| > 1/\epsilon$

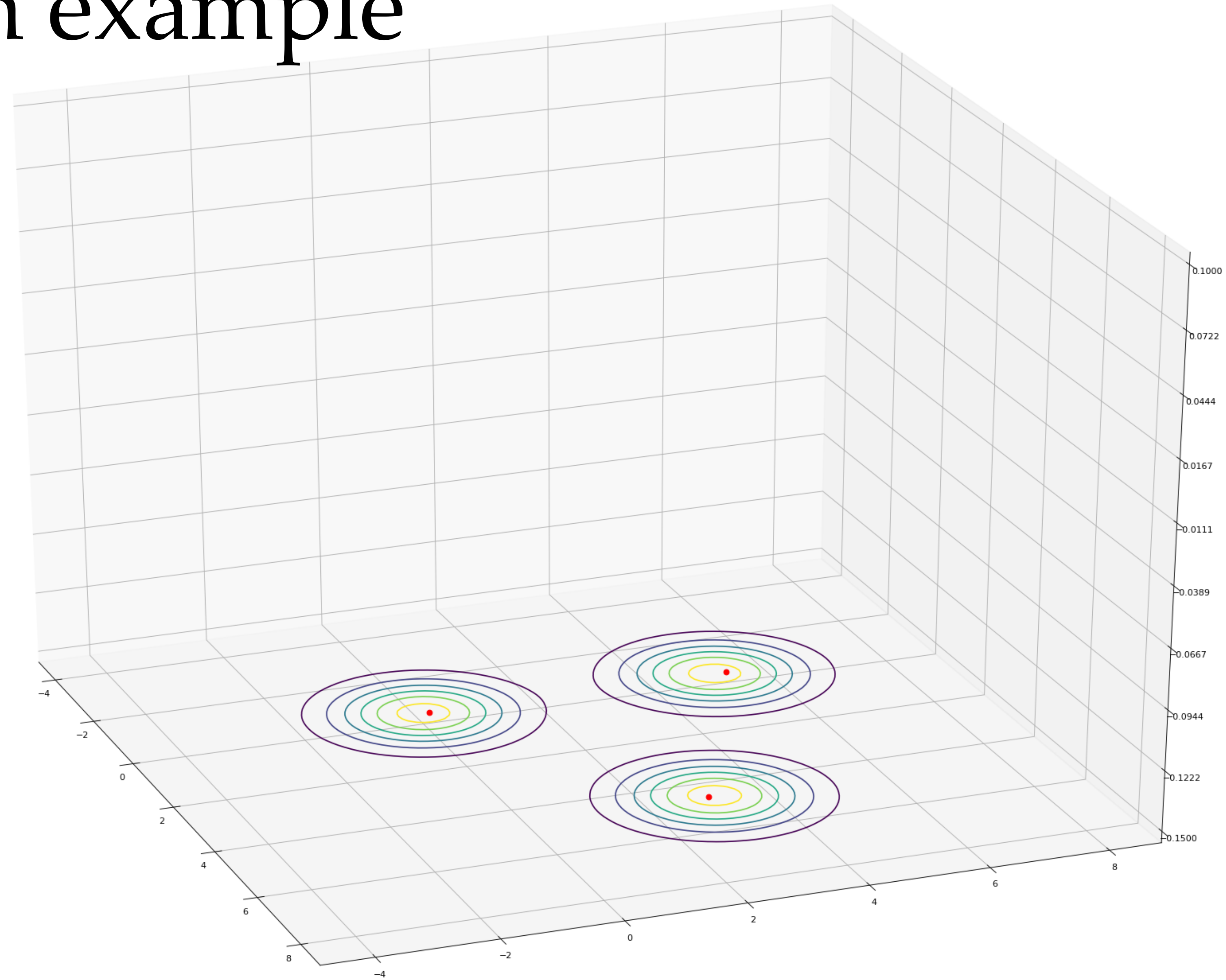


Compression: an example

Given S where $|S| > 1/\epsilon$

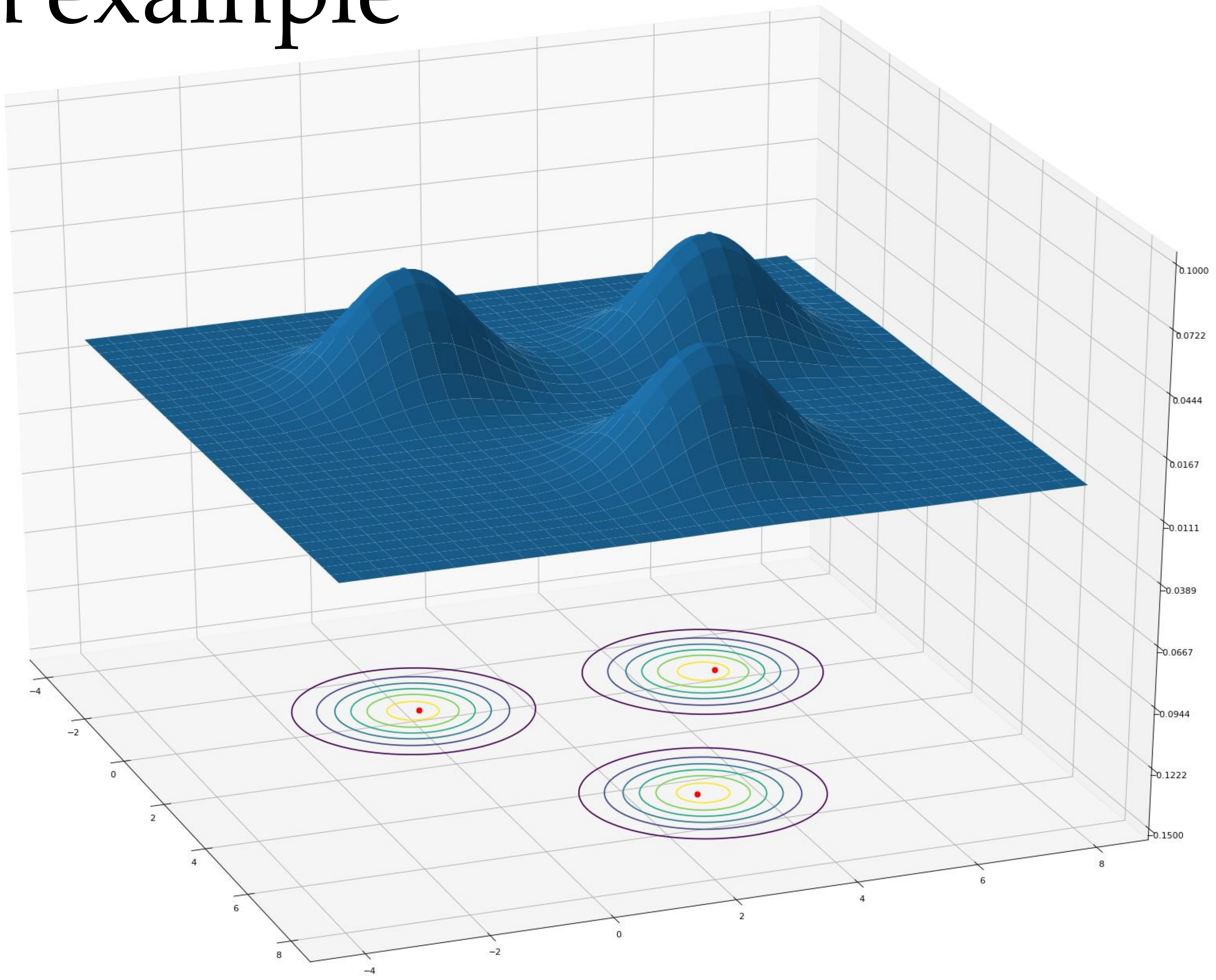
w.h.p. there exists

$Z = \{x_1, x_2, x_3\} \subset S$



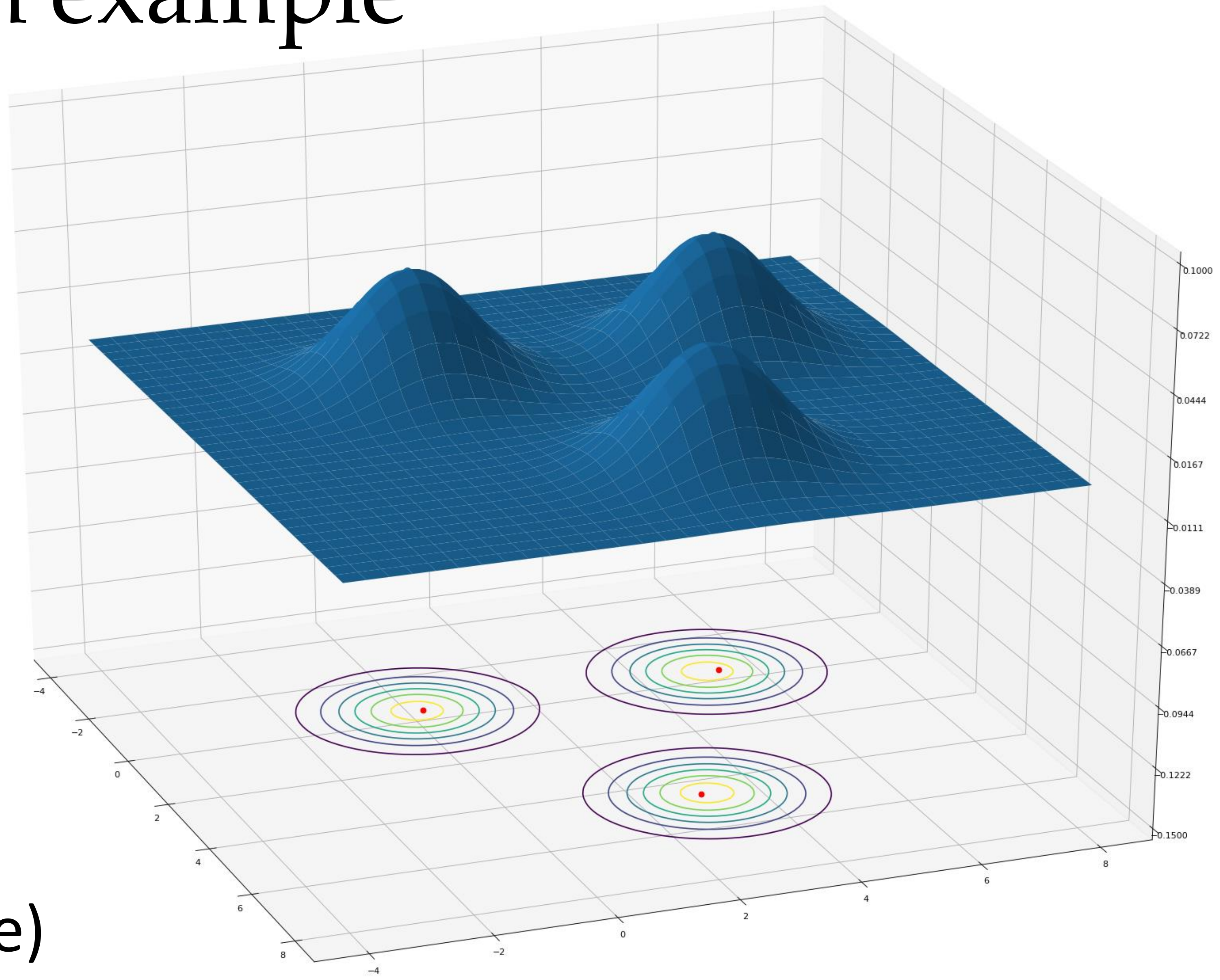
Compression: an example

Given S where $|S| > 1/\epsilon$
w.h.p. there exists
$Z = \{x_1, x_2, x_3\} \subset S$
based on which
the true distribution
can be reconstructed
up to error ϵ



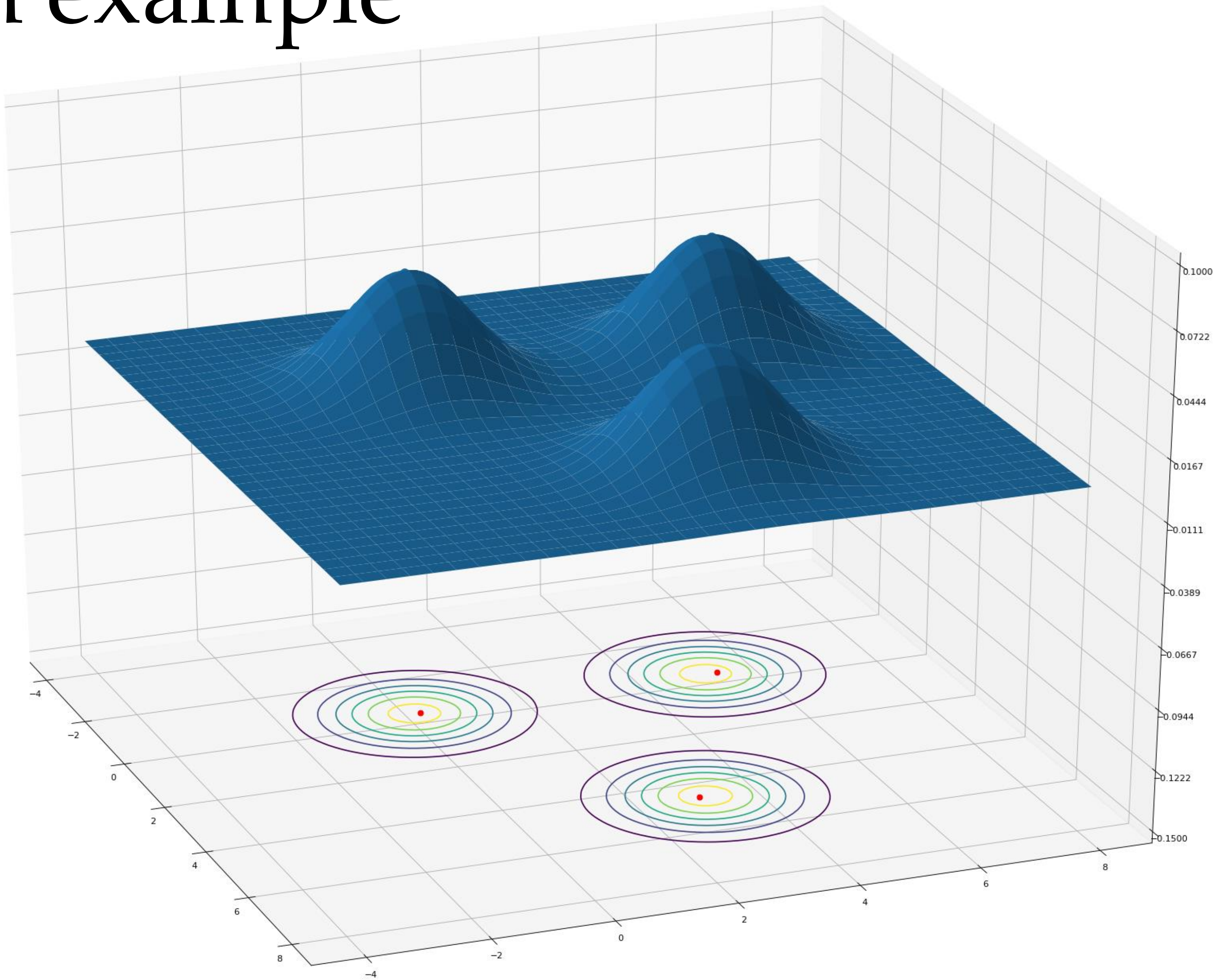
Compression: an example

Given S where $|\boldsymbol{S}| > \mathbf{1}/\boldsymbol{\epsilon}$
w.h.p. there exists
$\boldsymbol{Z} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\} \subset S$
based on which
the true distribution
can be reconstructed
up to error ϵ
(The decoder is fixed
before seeing the sample)



Compression: an example

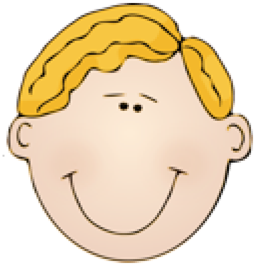
This class of distributions admits $\left(3, \frac{1}{\epsilon}\right)$-compression



Compression Framework

$\mathcal{F}$: a class of distributions (e.g. Gaussians)

Knows $\mathcal{D}$, $\mathcal{F}$

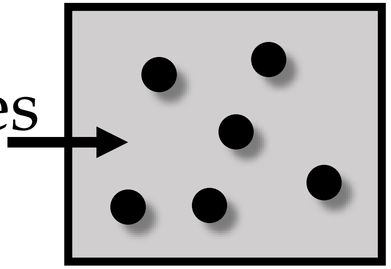


Knows $\mathcal{F}$

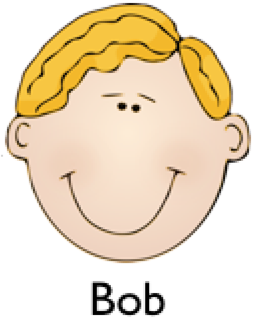
Compression Framework

$\mathcal{F}$: a class of distributions (e.g. Gaussians)

m i.i.d. samples from $\mathcal{D} \in \mathcal{F}$



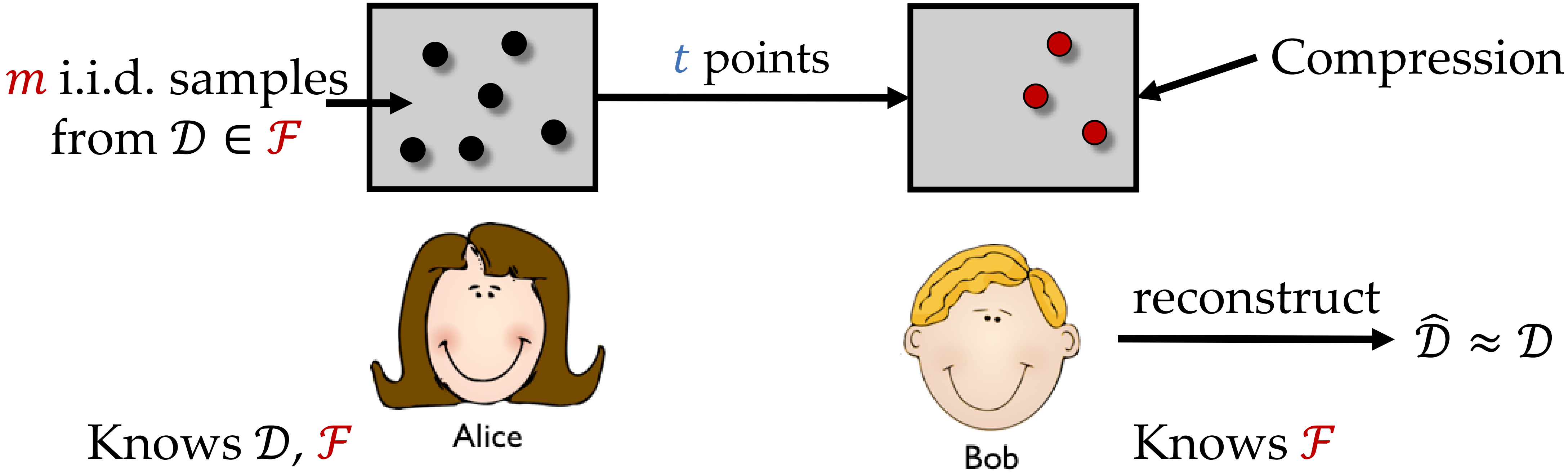
Knows $\mathcal{D}$, $\mathcal{F}$



Knows $\mathcal{F}$

Compression Framework

$\mathcal{F}$: a class of distributions (e.g. Gaussians)



If Alice sends t points from m points and Bob approximates $\mathcal{D}$ then we say $\mathcal{F}$ admits (t, m)-compression.

Distribution Compression Schemes

Theorem **[ABHLMP '18]** If $\mathcal{F}$ has a compression scheme of size (t, m) then sample complexity of learning $\mathcal{F}$ is

$$\tilde{O}\left(\frac{t}{\epsilon^2} + m\right)$$

$\tilde{O}(\cdot)$ hides polylog factors

Small compression schemes imply **sample-efficient** algorithms.

Distribution Compression Schemes

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size (t, m) then sample complexity of learning $\mathcal{F}$ is

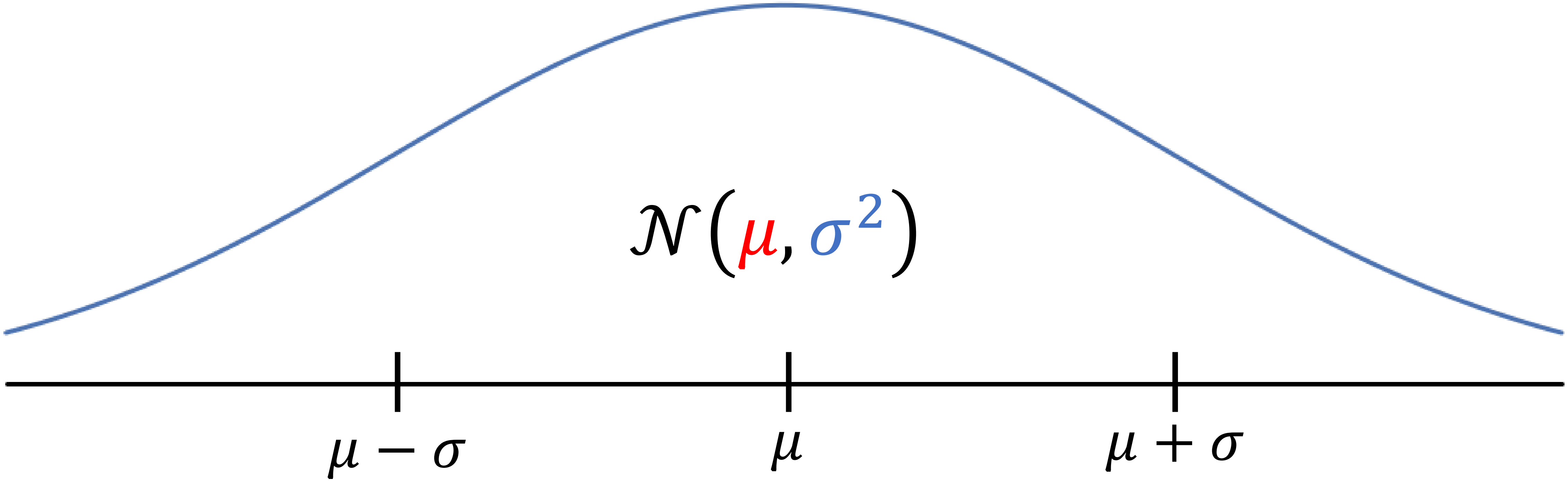
$$\tilde{O}\left(\frac{t}{\epsilon^2} + m\right)$$

$\tilde{O}(\cdot)$ hides polylog factors

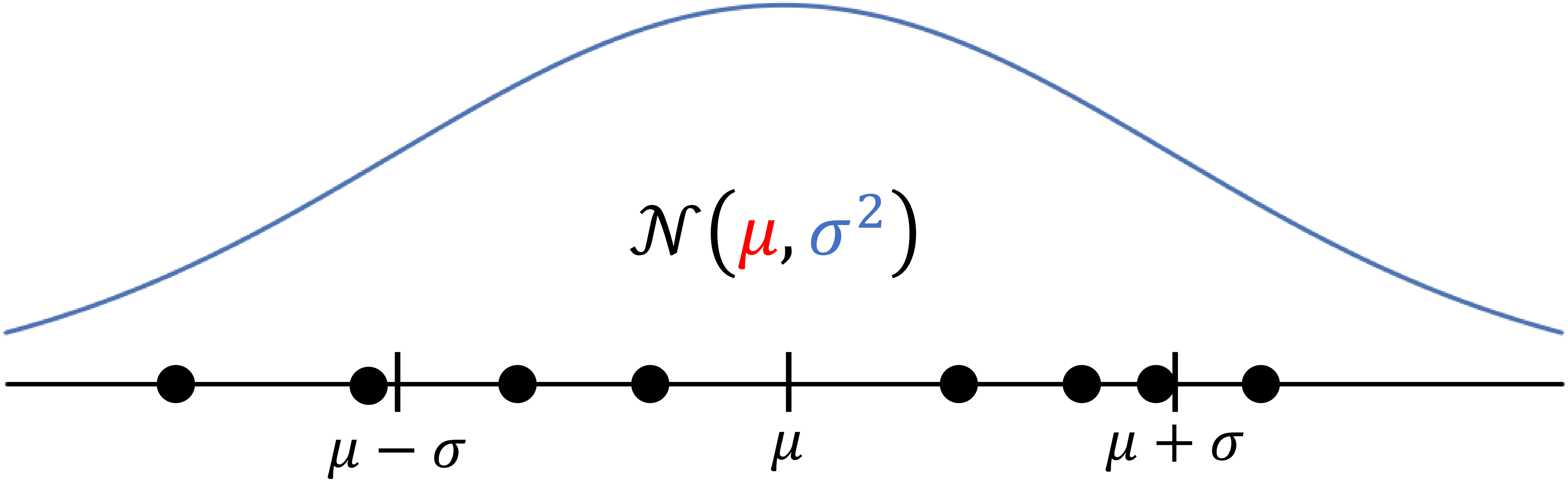
Small compression schemes imply **sample-efficient** algorithms.

There is a classic analogue in supervised learning [Littlestone and Warmuth, 1986]

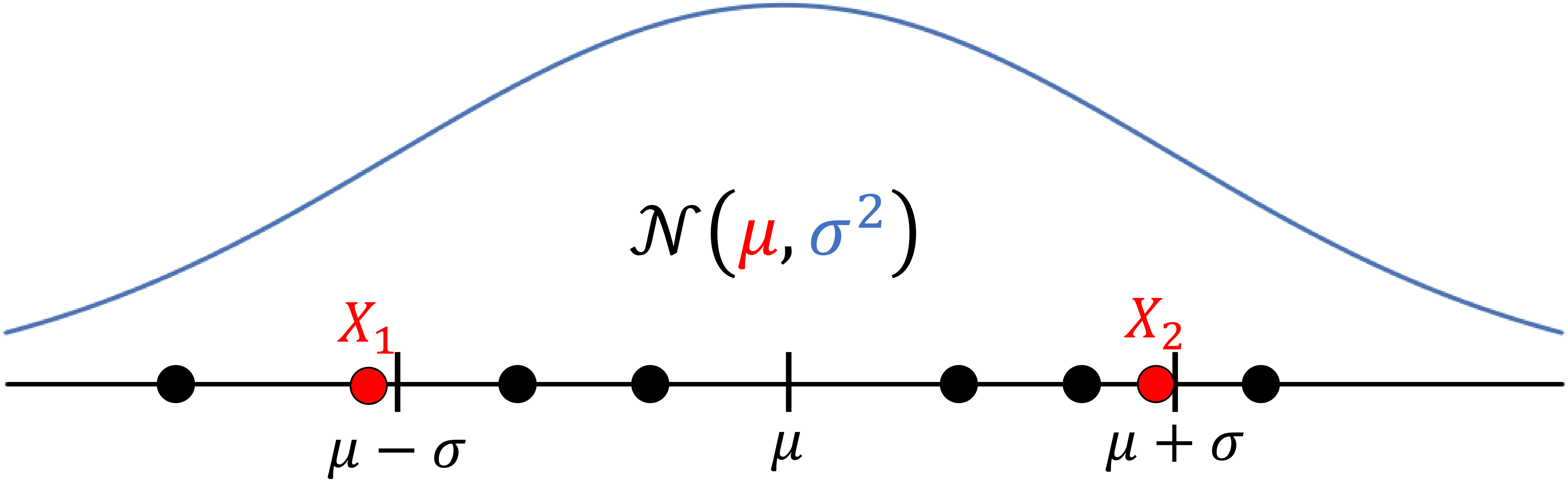
Compressing Gaussians in $\mathbb{R}$



Compressing Gaussians in $\mathbb{R}$



Compressing Gaussians in $\mathbb{R}$

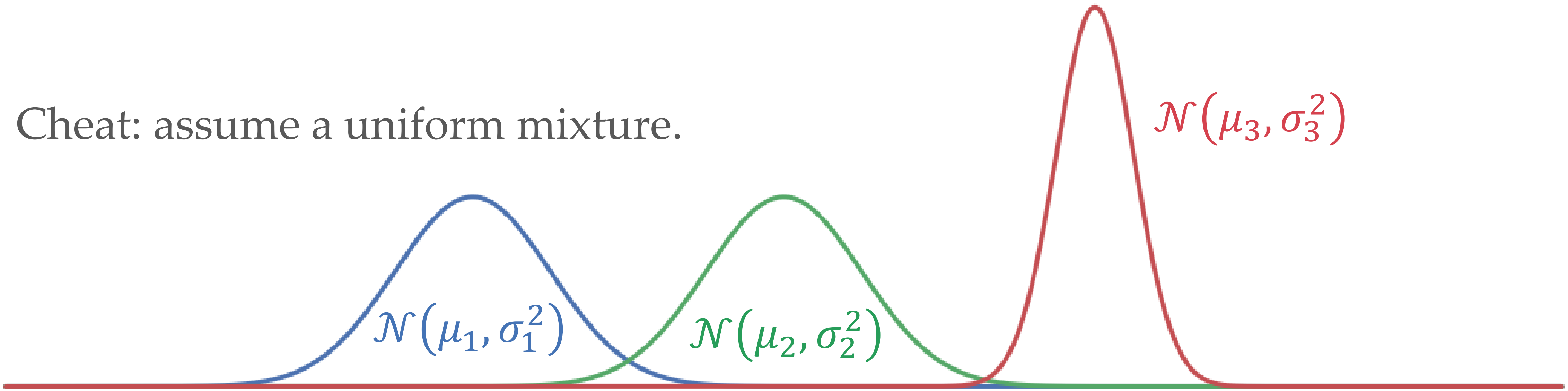


$$\frac{X_2 + X_1}{2} \approx \mu \qquad \frac{X_2 - X_1}{2} \approx \sigma$$

Admits $(2, \frac{1}{\epsilon})$-compression!

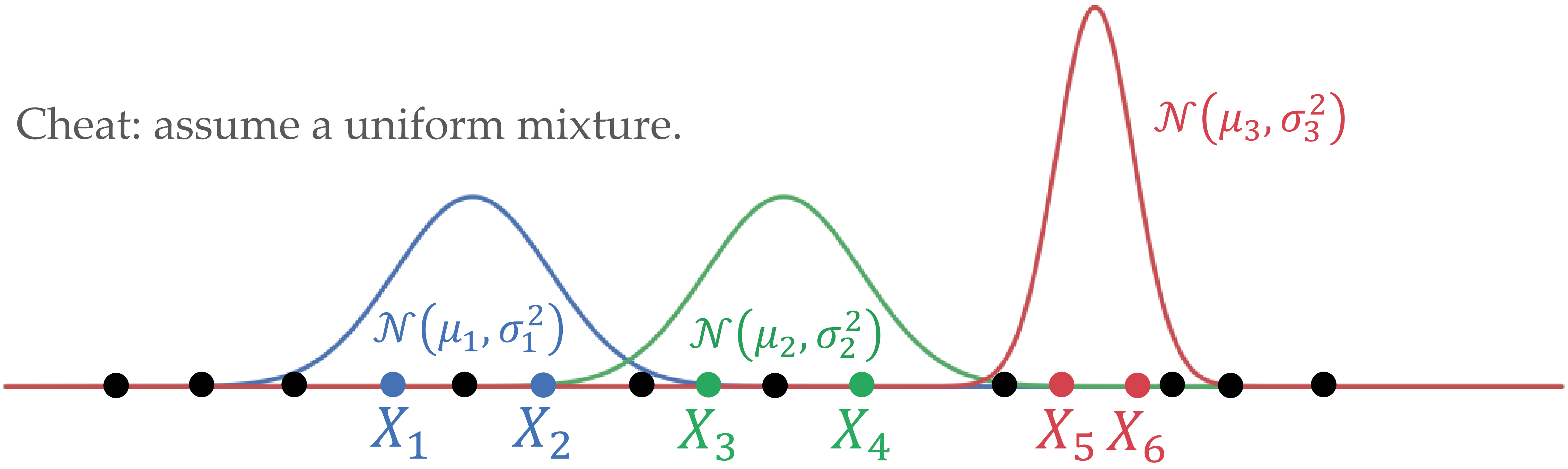
Compression of Mixtures

Cheat: assume a uniform mixture.



Compression of Mixtures

Cheat: assume a uniform mixture.



$X_1 \approx \mu_1 - \sigma_1$
$X_2 \approx \mu_1 + \sigma_1$

$X_3 \approx \mu_2 - \sigma_2$
$X_4 \approx \mu_2 + \sigma_2$

$X_5 \approx \mu_3 - \sigma_3$
$X_6 \approx \mu_3 + \sigma_3$

Compression Theorem for Mixtures

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size (t, m) then k mixtures of $\mathcal{F}$ is admits (kt, km) compression.

Distribution compression schemes extend to mixture classes automatically!

Compression Theorem for Mixtures

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size (t, m) then k mixtures of $\mathcal{F}$ is admits (kt, km) compression.

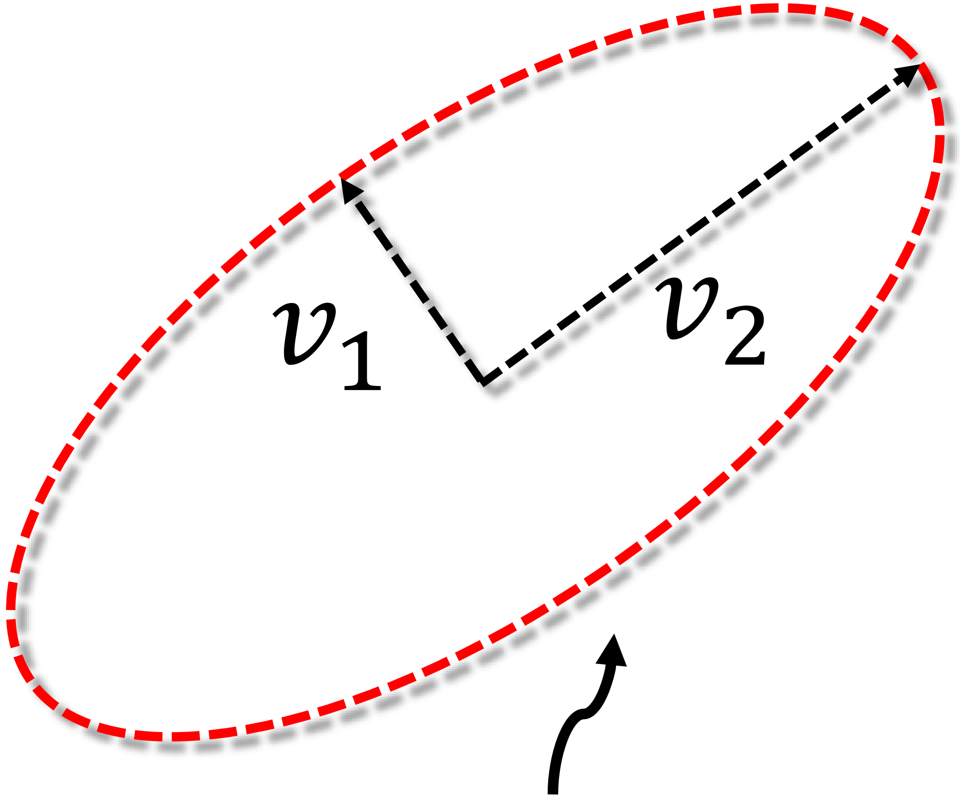
Distribution compression schemes extend to mixture classes automatically!

So for the case of GMMs in $\mathbb{R}^d$ it is enough to come up with a good compression scheme for a single Gaussian!

Learning Mixtures of Gaussians

Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.

Is $\tilde{O}\left(d^2, \frac{1}{\epsilon}\right)$ compression is possible?

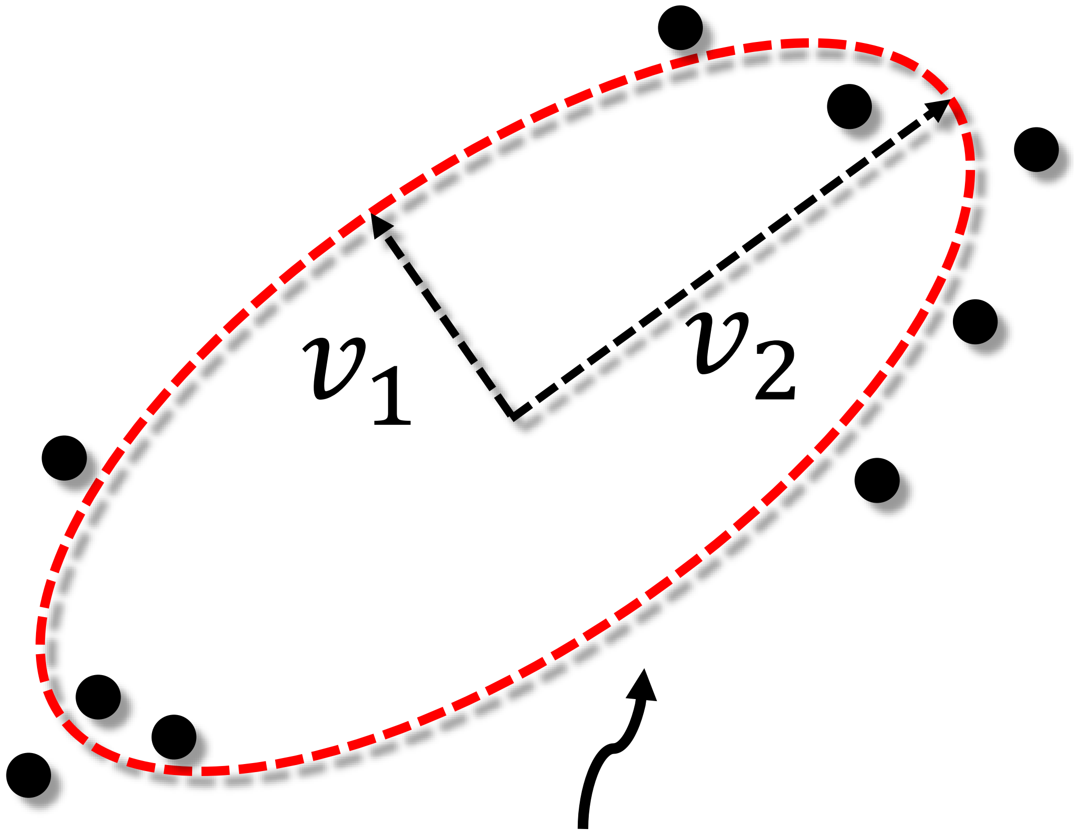


Ellipsoid defined by μ, Σ.

Learning Mixtures of Gaussians

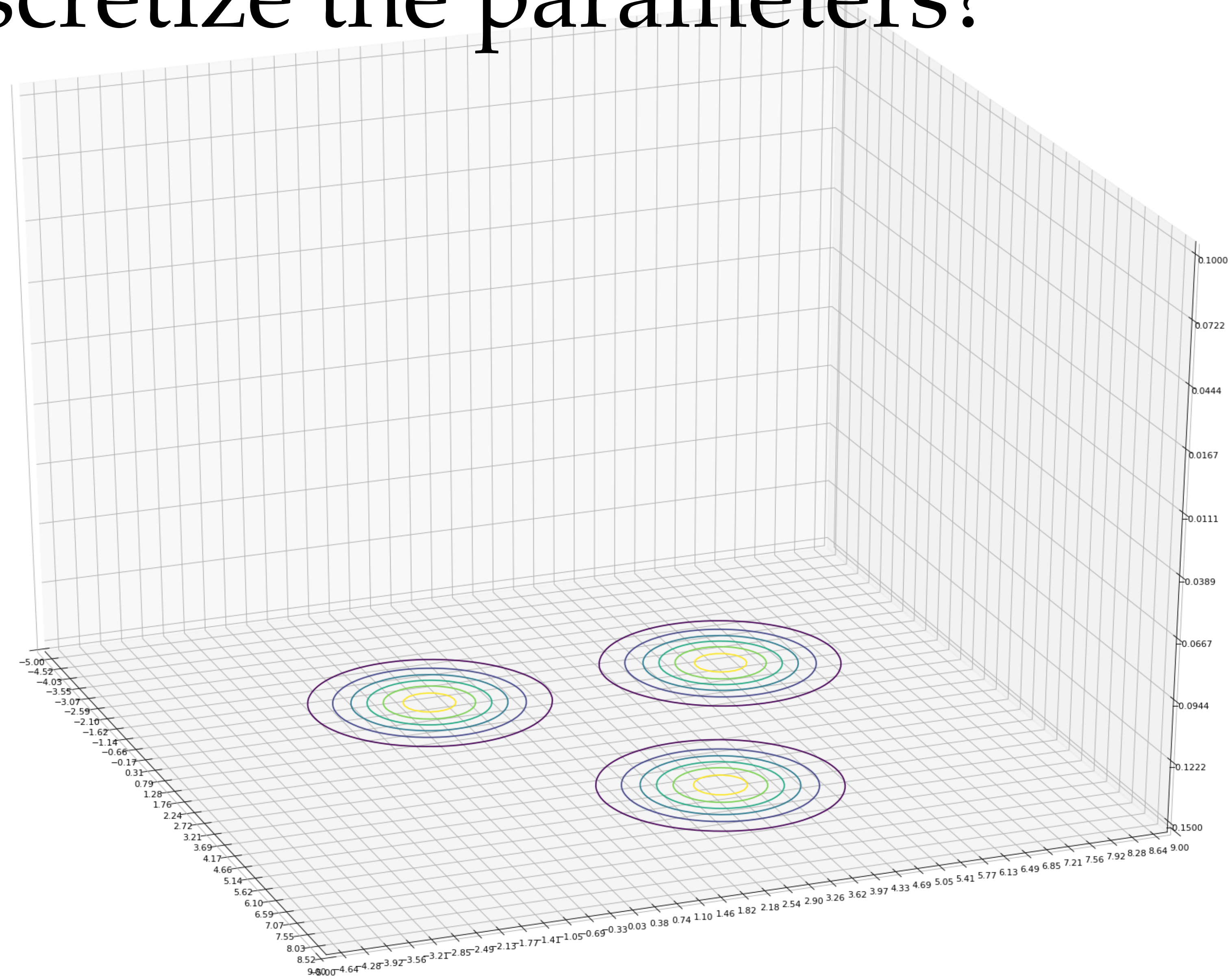
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.

Is $\tilde{O}\left(d^2, \frac{1}{\epsilon}\right)$ compression is possible?



Ellipsoid defined by μ, Σ.
Points drawn from $\mathcal{N}(\mu, \Sigma)$.

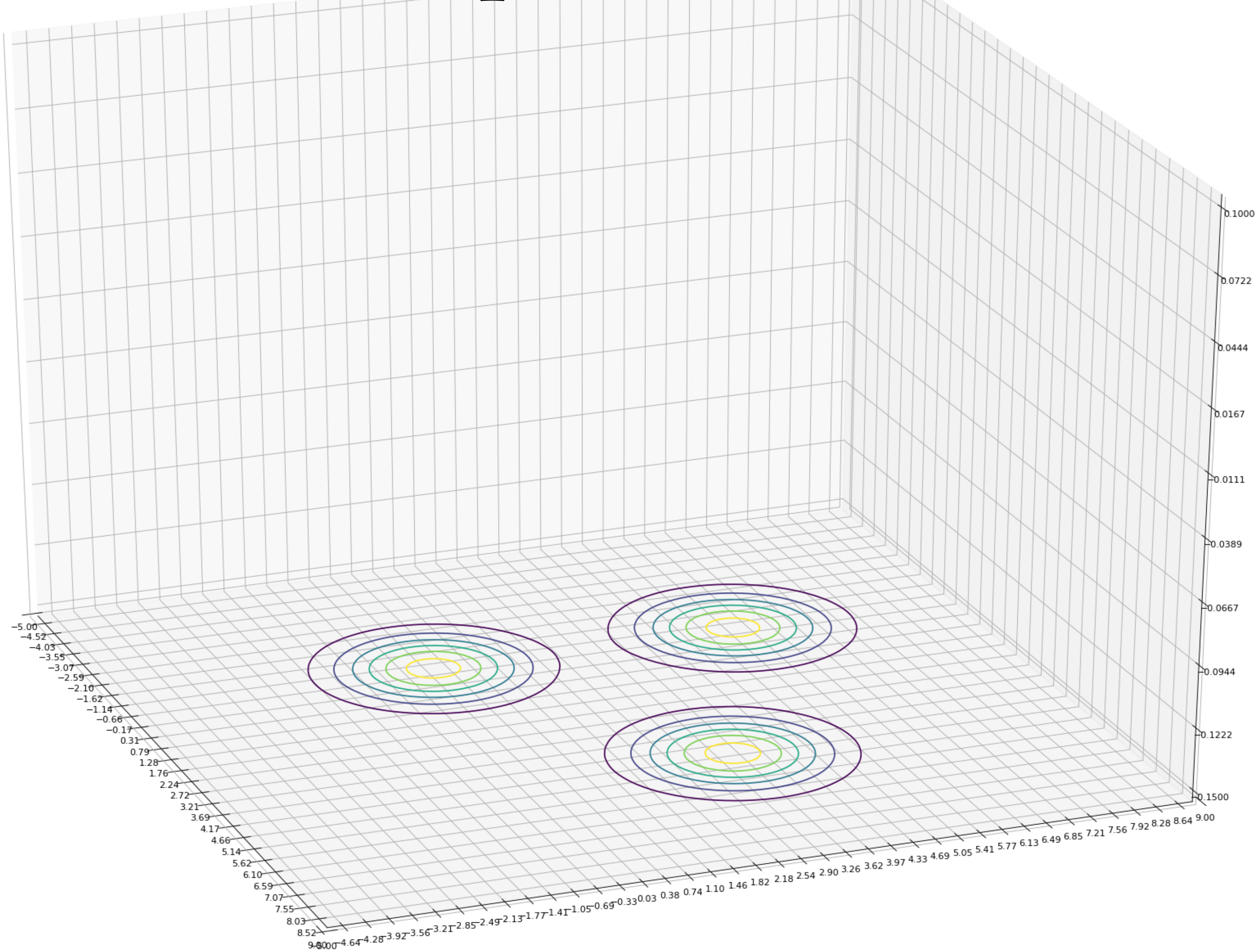
Why not just discretize the parameters?



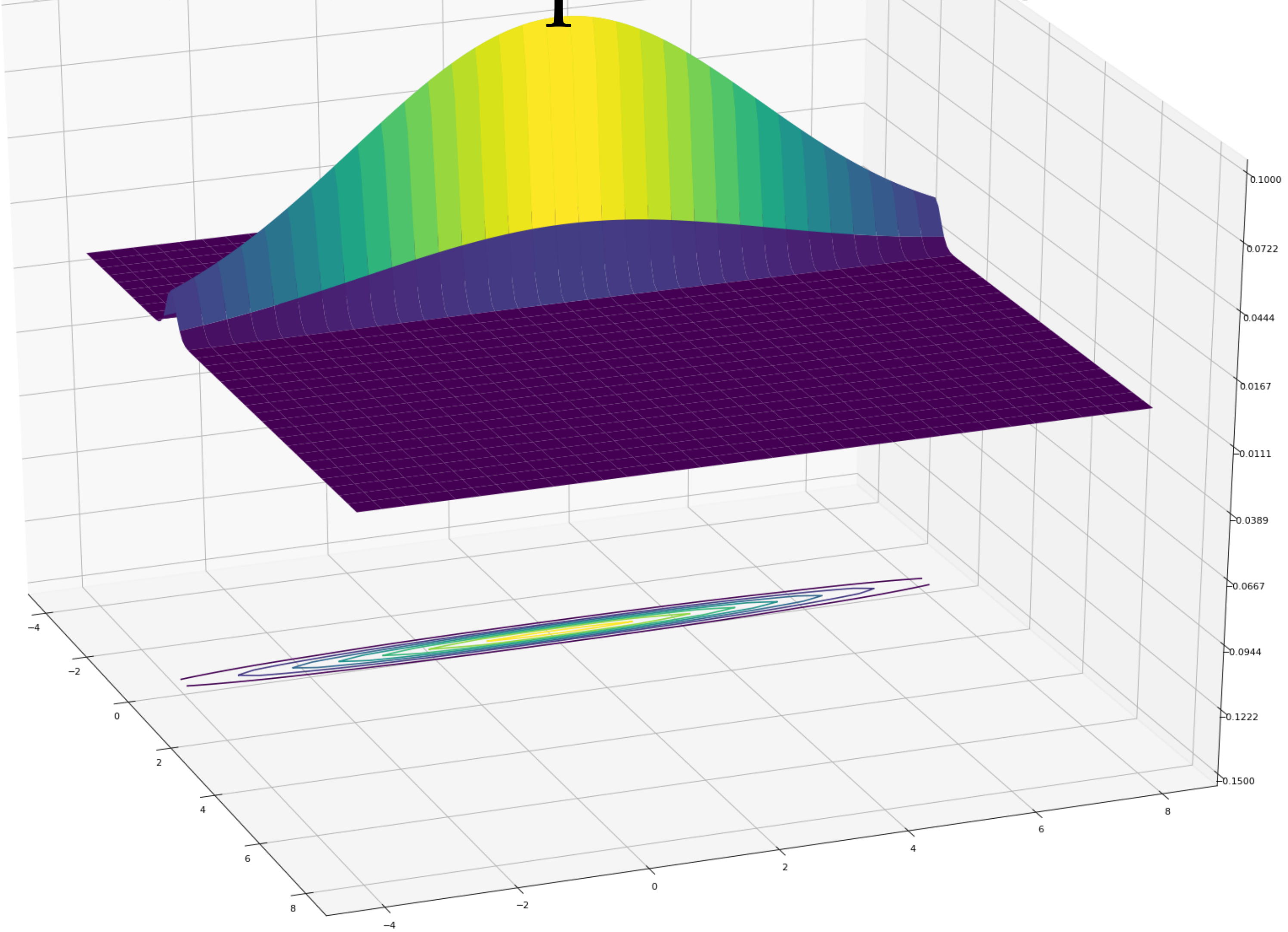
Why not just discretize the parameters?

Discretization does not work because…

- μ is unbounded
- Σ is unbounded
- And…



Why not just discretize the parameters?



$\frac{\sigma_{max}}{\sigma_{min}}$ can be large

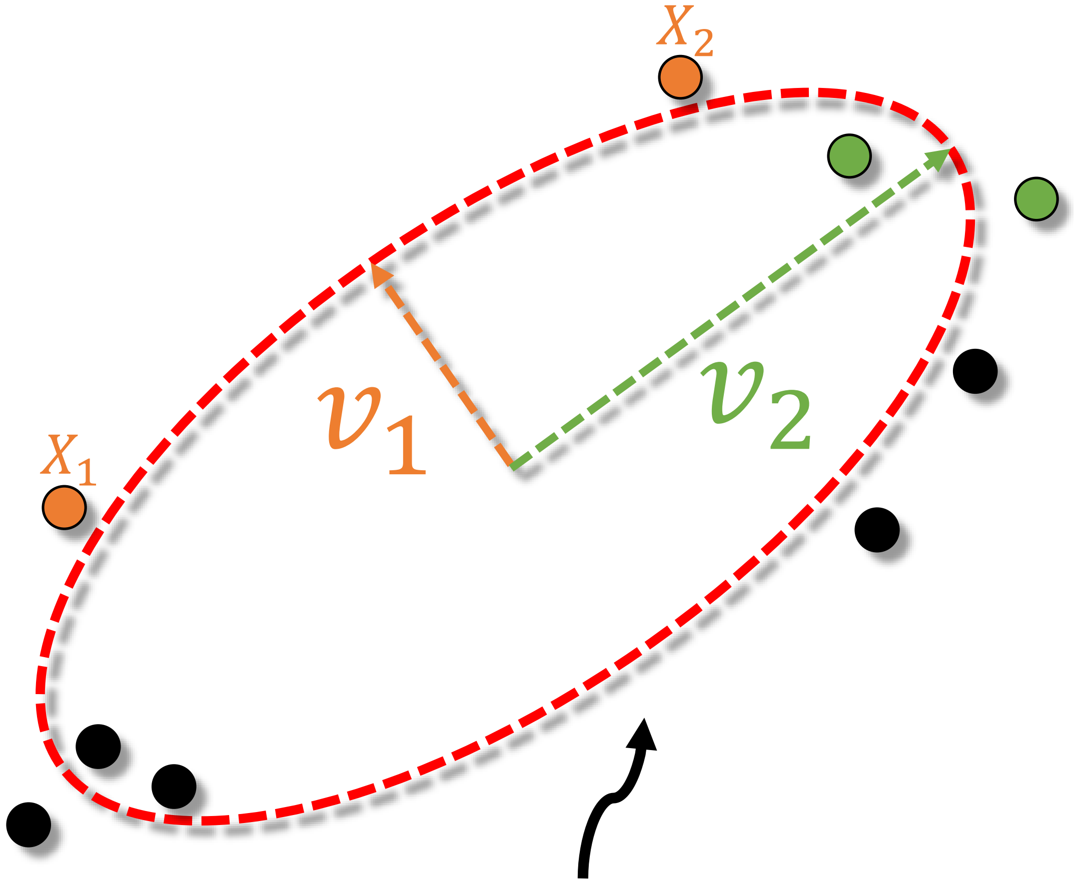
Not exactly a parameter estimation problem!

Learning Mixtures of Gaussians

Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.

Is $\tilde{O}\left(d^2, \frac{1}{\epsilon}\right)$ compression is possible?

The technical challenge is encoding the ***d* eigen-vectors “accurately” using only *d*^{2}** points.



Ellipsoid defined by μ, Σ.
Points drawn from $\mathcal{N}(\mu, \Sigma)$.

Application: Learning Mixtures of Gaussians

Theorem [ABHLMP '18] Sample complexity for learning mixtures of k Gaussians in $\mathbb{R}^d$ up to L_1-error ϵ is

$$\tilde{\mathbf{O}}\left(\frac{\boldsymbol{kd^2}}{\boldsymbol{\epsilon^2}}\right)$$

$\tilde{\mathrm{O}}(\cdot)$ hides polylog factors

- Improves upon:
  - $O(k^4 d^4/\epsilon^2)$ via a VC-dimension argument
  - $\tilde{O}(kd^2/\epsilon^4)$ [Ashtiani, Ben-David, Mehrabian '17]

Application: Learning Mixtures of Gaussians

Theorem [ABHLMP ’18] Sample complexity for learning mixtures of k Gaussians in $\mathbb{R}^d$ up to L_1-error ϵ is

$$\tilde{\mathbf{O}}\left(\frac{\boldsymbol{kd^2}}{\boldsymbol{\epsilon^2}}\right)$$

$\tilde{O}(\cdot)$ hides polylog factors

- Improves upon:
  - $O(k^4 d^4/\epsilon^2)$ via a VC-dimension argument
  - $\tilde{O}(kd^2/\epsilon^4)$ [Ashtiani, Ben-David, Mehrabian ‘17]
- We show this is nearly-tight!
  - $\tilde{\Omega}(kd^2/\epsilon^2)$ samples are necessary!
  - **Along the way we had to prove $\tilde{\Omega}(d^2/\epsilon^2)$ lower bound for Gaussians!**

Summary

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes
- Application
  - Almost-tight bounds for GMMs

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes
- Application
  - Almost-tight bounds for GMMs
- **Q**: What if the target is just "almost a GMM"?
  - Compression can be extended to the agnostic/robust setting!

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes
- Application
  - Almost-tight bounds for GMMs
- **Q**: What if the target is just "almost a GMM"?
  - Compression can be extended to the agnostic/robust setting!
- **Q**: Does compression size characterize sample complexity?
  - Still an open problem…
  - It is (almost) the case for supervised learning [Moran and Yehudayoff, 2016].

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes
- Application
  - Almost-tight bounds for GMMs
- **Q**: What if the target is just “almost a GMM”?
  - Compression can be extended to the agnostic/robust setting!
- **Q**: Does compression size characterize sample complexity?
  - Still an open problem…
  - It is (almost) the case for supervised learning [Moran and Yehudayoff, 2016].
- **Q**: Polynomial time algorithm for learning GMMs?

Summary

- Introduced compression schemes for density estimation
  - Simple and generic
    Naturally extends to mixture classes
- Application
  - Almost-tight bounds for GMMs
- **Q**: What if the target is just "almost a GMM"?
  - Compression can be extended to the agnostic/robust setting!
- **Q**: Does compression size characterize sample complexity?
  - Still an open problem…
  - It is (almost) the case for supervised learning [Moran and Yehudayoff, 2016].
- **Q**: Polynomial time algorithm for learning GMMs?

Thanks for listening!