Nearly-tight sample complexity bounds for learning mixtures of Gaussians

Hassan Ashtiani
(McMaster)

Shai Ben-David
(Waterloo)

Nick Harvey
(UBC)

Abbas Mehrabian
(McGill)

Yaniv Plan
(UBC)

Chris Liaw (UBC)

NeurIPS, December 2018

A fundamental statistical learning roblem

Given a sample drawn from some unknown probability distribution,

Learn as much as possible about that sample-generating distribution.

The first two questions to ask:

What prior knowledge does the learn have?

What does the learn wish to deduce about that distribution?

The most ambitious framework

- Assume the learner starts with no prior knowledge (namely, no assumptions about the unknown distributions).

- The learner aims to find what that distribution is (namely, come up with a reliable tight estimation of that distribution, say in the ***total variance*** sense).

Such an ambitious task is provably impossible

Proof idea:

Consider a discrete finite domain set. Denote its size by N.

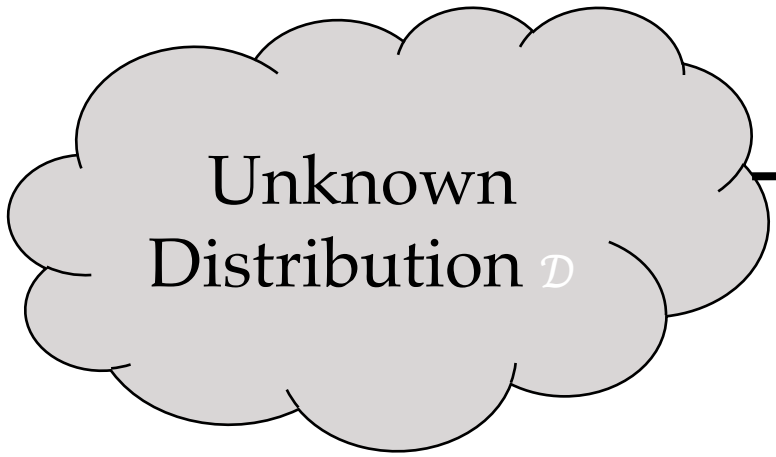
Any sample of size n< sqrt(N) containing no repetitions, is as likely to have been generated by a uniform distribution over X as it is to have been generated by a uniform distribution over a subset of size |X|/2 (that contains all members of that sample).

Both types of problem restriction are important

- Assuming the generating distribution belongs to (or is well approximated by a member of) a restricted family of distributions (e.g., mixtures of Gaussian distributions).

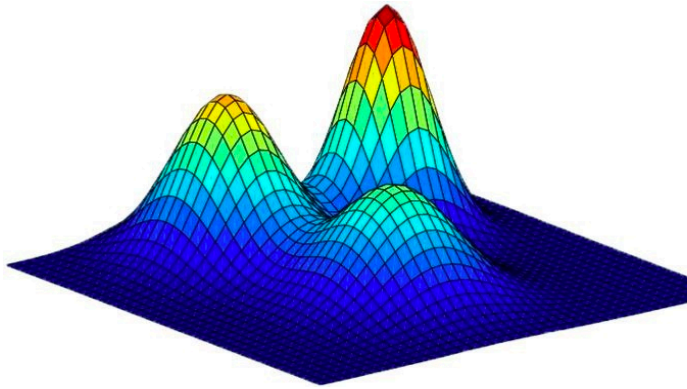
- Require the learner to discover only some limited information about the unknown distribution (being able to predict labels, cluster, find means or higher order moments, etc.)

Density estimation

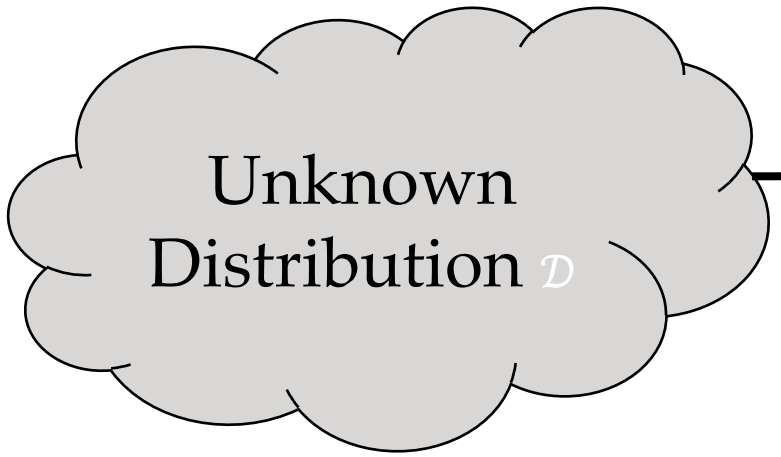


i.i.d. samples

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

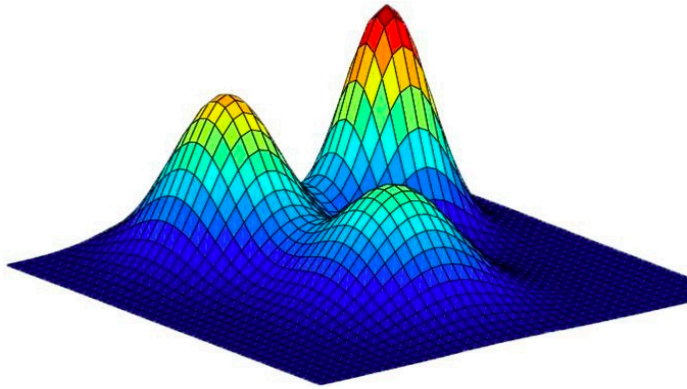


Density estimation



i.i.d. samples

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

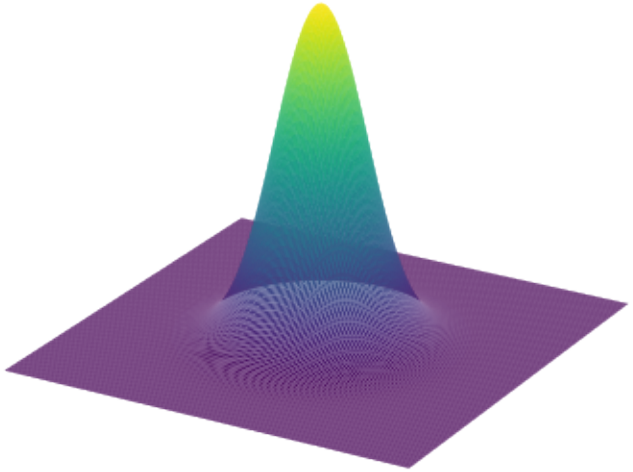


Fundamental & well-studied problem with many applications!

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

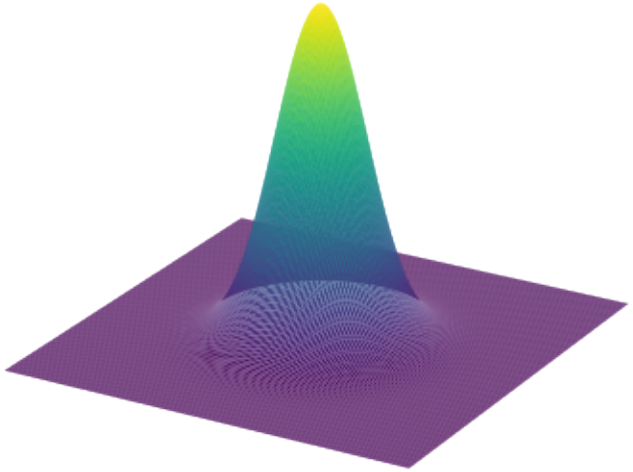
Q [D '16]: "For a distribution class $\mathcal{F}$, is there a complexity measure that characterizes the sample complexity of $\mathcal{F}$?"

Learning Gaussians

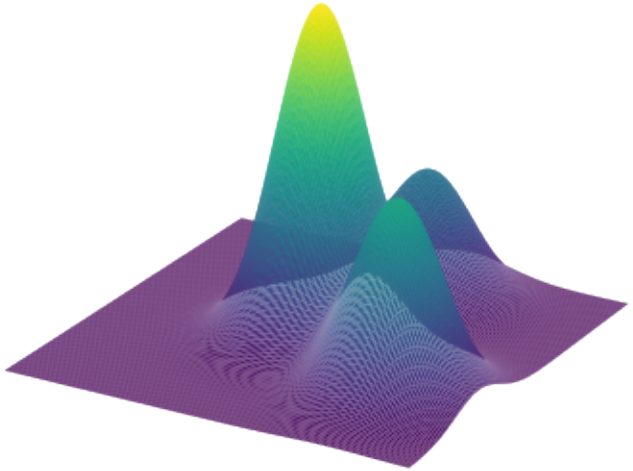


Single Gaussian in $\mathbb{R}^d$.
$O(d^2/\epsilon^2)$ samples are sufficient to recover Gaussian up to L_1-error ϵ.

Learning Gaussians



Single Gaussian in $\mathbb{R}^{d\uparrow}$.
$O(d^2/\varepsilon^2)$ samples are sufficient to recover Gaussian up to $L_{1\downarrow}$-error ϵ.



Mixture of k Gaussians in $\mathbf{R^d}$
Q: Are $O(kd^2/\varepsilon^2)$ samples sufficient?
Know that $O(kd^{2\uparrow}/\epsilon^{4\uparrow})$ are sufficient. [Ashtiani &BD '17]

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

Main Tool: Sample Compression Schemes

- *"Other things being equal, simpler explanations are generally better…"* [William of Ockham]
- One manifestation of this in learning theory is "sample compression".

A *sample compression scheme* is a way to replace every training sample with a small subsample that conveys all the relevant information contain in the full sample.

For example, any set of points on the line, labeled by membership in some interval, can be replaced by just the leftmost and rightmost positive labeled points.

Littlestone, Warmuth 1986 showed that the existence of such a scheme for a class of binary valued functions implies PAC learnability of that class.

Moran, Yehudayoff 2016 showed that the convers holds as well.

Main Contribution

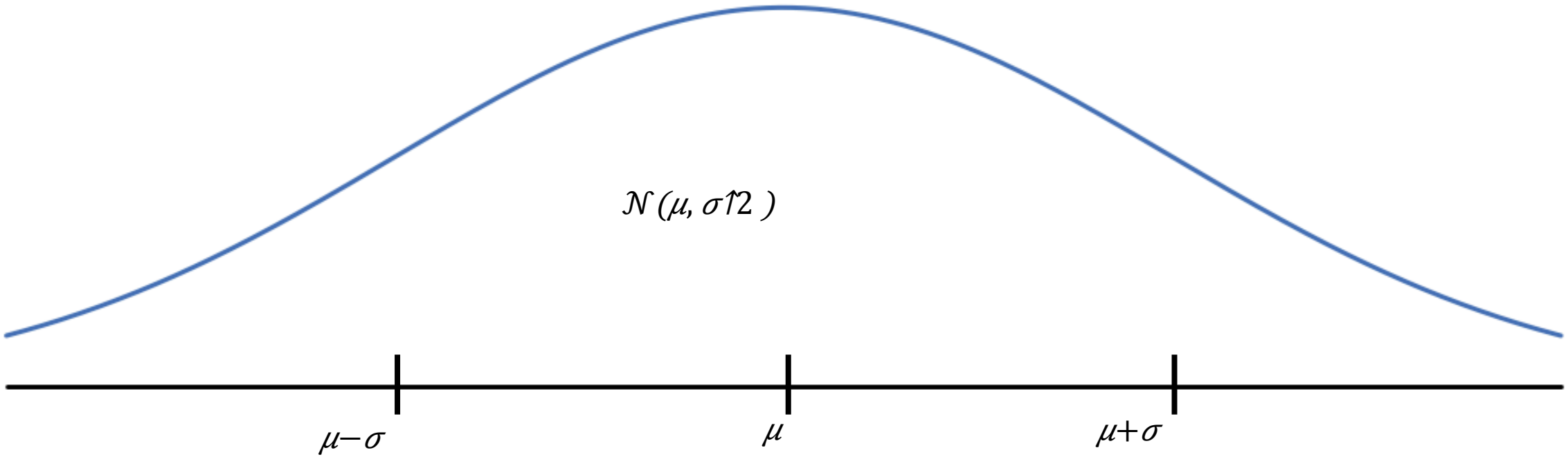
We introduce a **simple** & **sample-efficient** technique for density estimation via **compression schemes**.

Our scheme allows compressing any large enough sample generated by a mixture of Gaussians to a subsample whose size is independent of the size of the full sample.

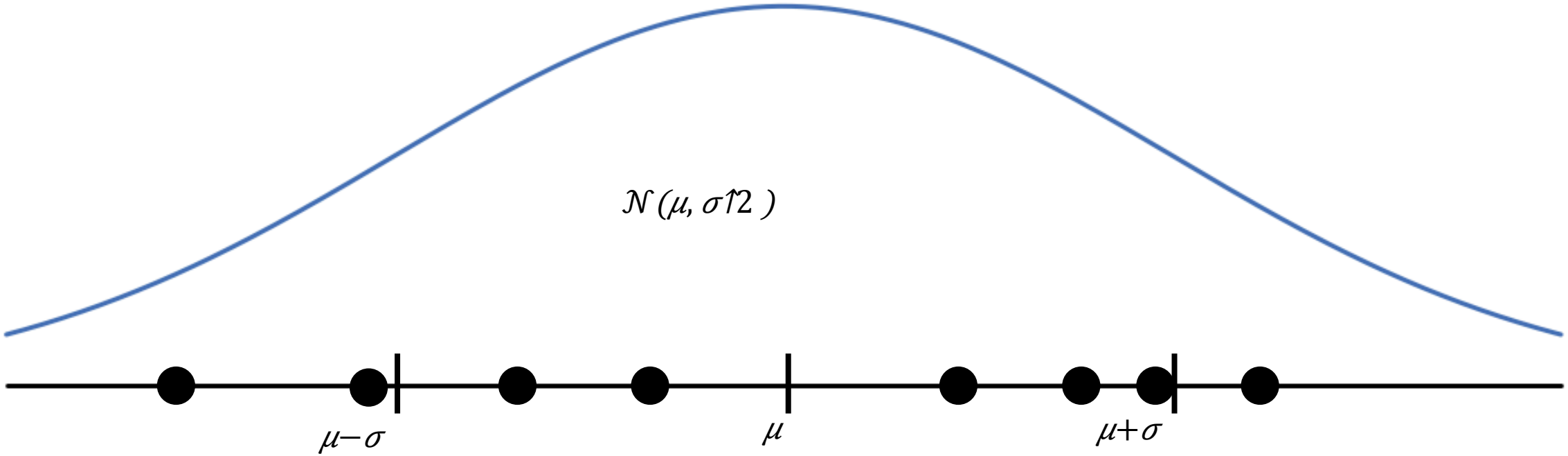
Our compression scheme yields upper bounds of $O(kd^2/\varepsilon^2)$ (up to logarithmic factors) of mixtures of k Gaussians in dimension d.

We also prove matching lower bounds on the sample complexity of this task.

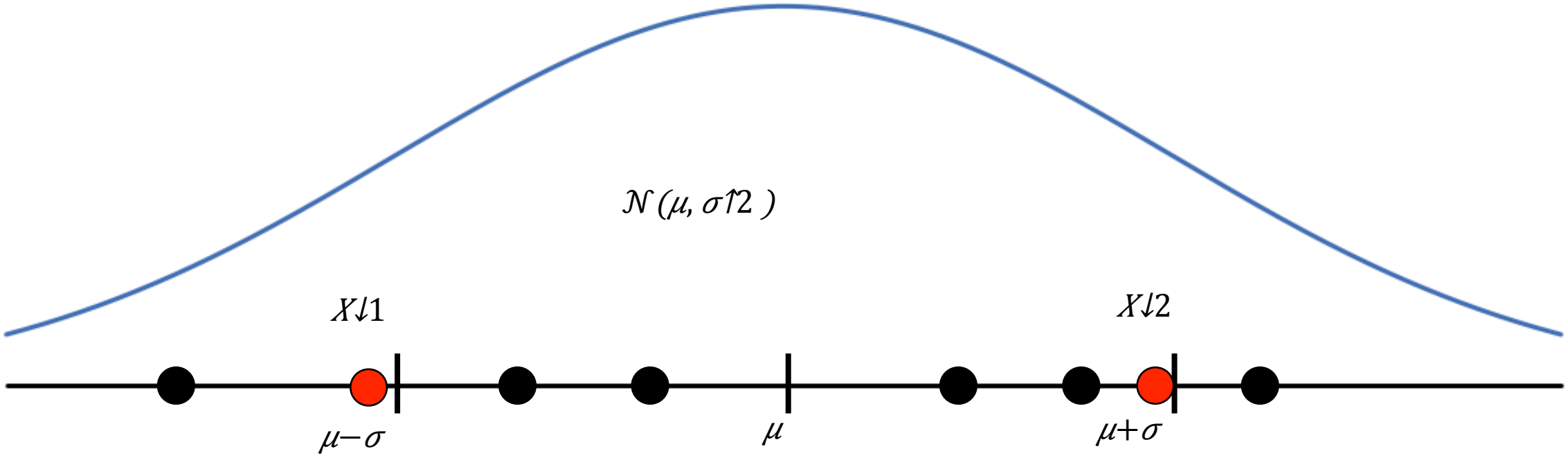
Compressing Gaussians in $\mathbb{R}$



Compressing Gaussians in $\mathbb{R}$



Compressing Gaussians in $\mathbb{R}$



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

Two samples are **sufficient** to **encode** $\mathcal{N}(\mu, \sigma^2)$.

Compression Framework

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

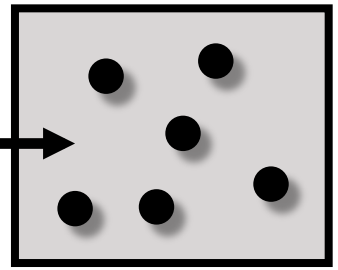
Knows $\mathcal{D}$, $\mathcal{F}$

Knows $\mathcal{F}$

Compression Framework

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

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

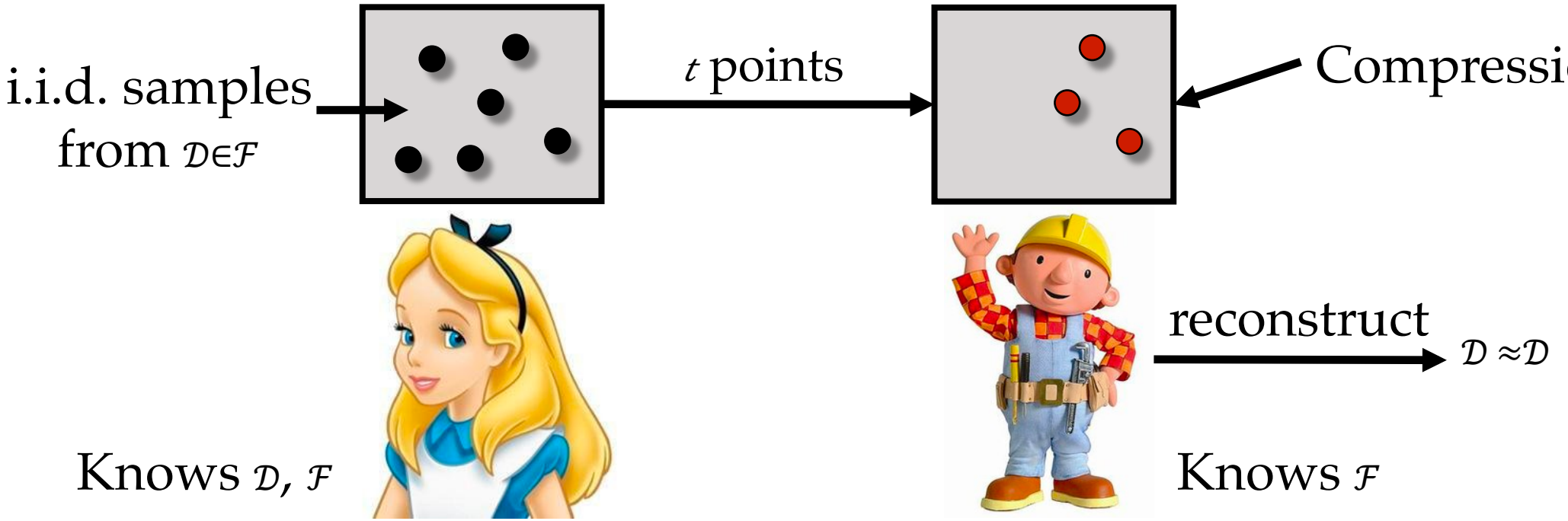


Knows $\mathcal{D}$, $\mathcal{F}$

Knows $\mathcal{F}$

Compression Framework

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



If Alice sends t points and Bob approximates $\mathcal{D}$ then we say $\mathcal{F}$ has compression of size t.

Compression Theorem

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size t then sample complexity to learn $\mathcal{F}$ (up to L_1-error ϵ) is $O(t/\epsilon^2)$.

$O(\cdot)$ hides polylog factors

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

Compression Theorem

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size t then sample complexity to learn $\mathcal{F}$ (up to L_1-error ϵ) is $O(t/\epsilon^2)$.

$O(\cdot)$ hides polylog factors

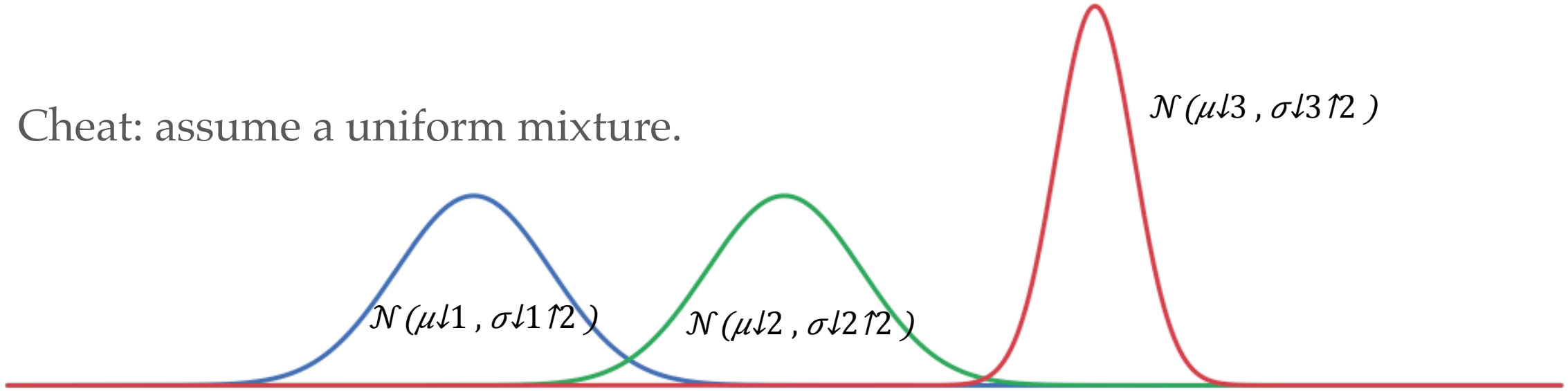
Small compression schemes imply **sample-efficient** algorithms.

Proof idea.

- Compression is used to find small set of "representative" distributions.
- Now, we can learn with respect to a finite class.

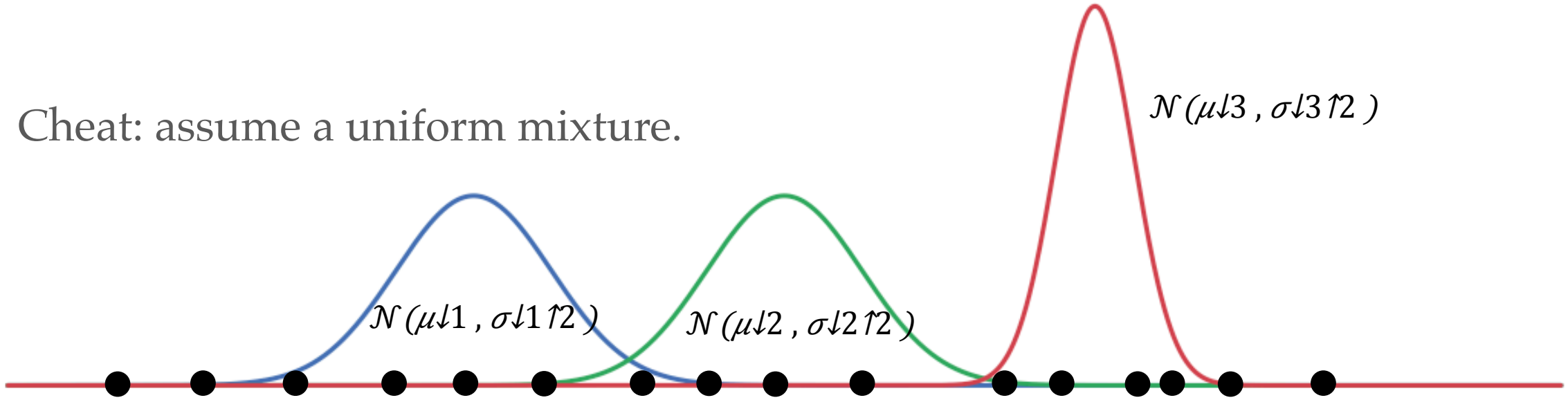
Compression Of Mixtures

Cheat: assume a uniform mixture.



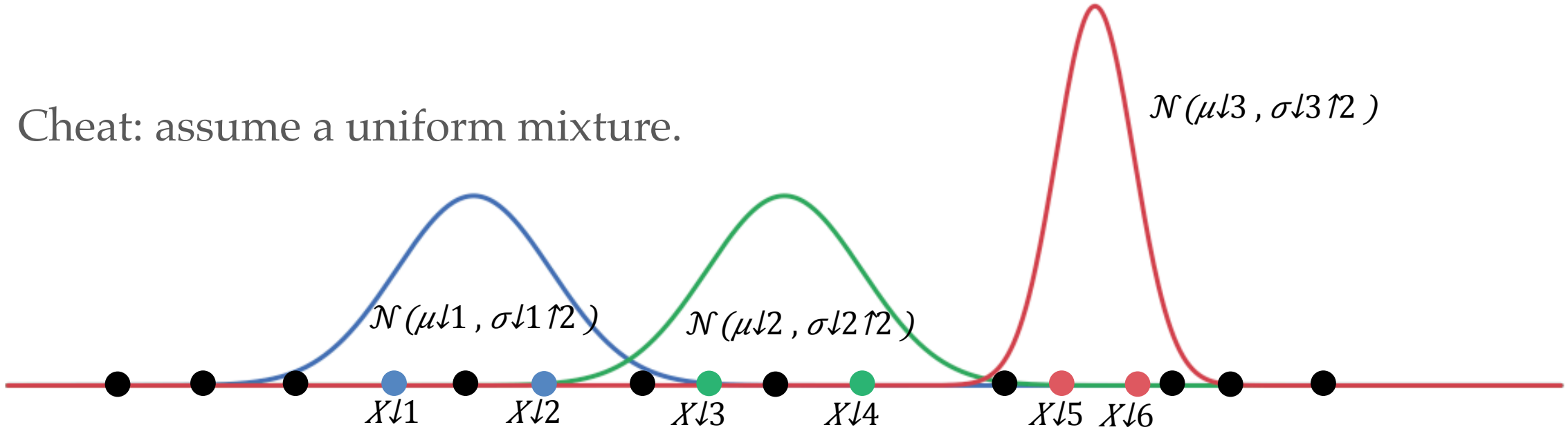
Compression Of Mixtures

Cheat: assume a uniform mixture.



Compression Of Mixtures

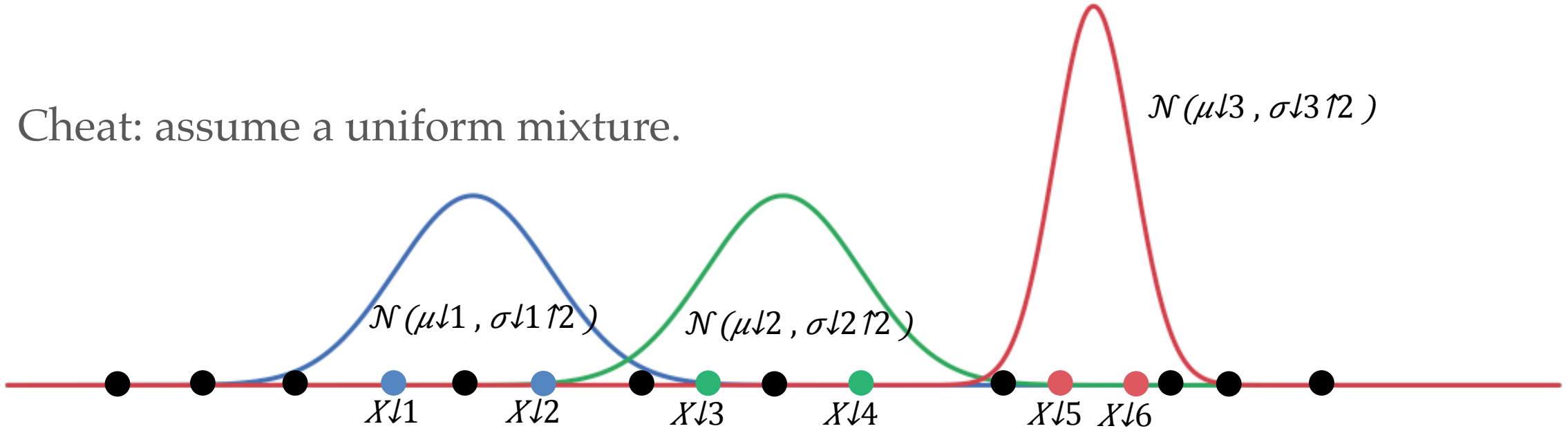
Cheat: assume a uniform mixture.



$$X_1 \approx \mu_1 - \sigma_1 \qquad X_3 \approx \mu_2 - \sigma_2 \qquad X_5 \approx \mu_3 - \sigma_3$$
$$X_2 \approx \mu_1 + \sigma_1 \qquad X_4 \approx \mu_2 + \sigma_2 \qquad X_6 \approx \mu_3 + \sigma_3$$

Compression Of Mixtures

Cheat: assume a uniform mixture.



If $\mathcal{F}$ has a compression of size $\boldsymbol{t}$ then
$\boldsymbol{k}$ mixtures of $\mathcal{F}$ have a compression of size $\approx \boldsymbol{kt}$.

Compression Theorem for Mixtures

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size t then sample complexity to learn k mixtures of $\mathcal{F}$ (up to L_1-error ϵ) is $O(kt/\epsilon^2)$.

$O(\cdot)$ hides polylog factors

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

Compression Theorem for Mixtures

Theorem [ABHLMP '18] If $\mathcal{F}$ has a compression scheme of size t then sample complexity to learn k mixtures of $\mathcal{F}$ (up to L_1-error ϵ) is $\tilde{O}(kt/\epsilon^2)$.

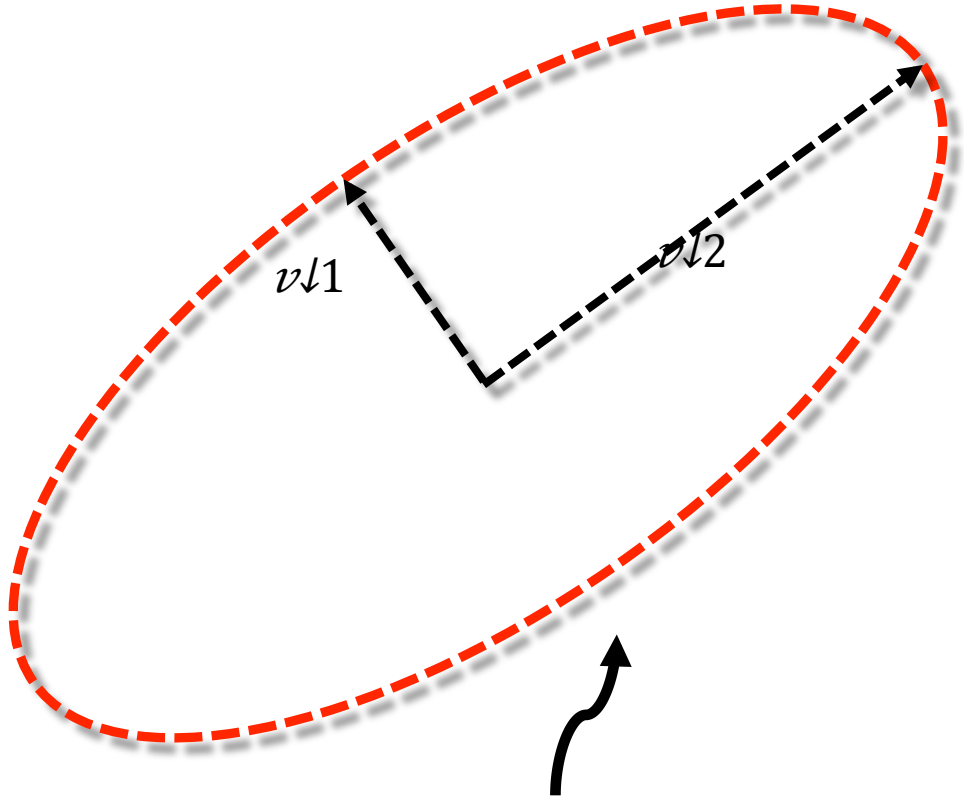
$\tilde{O}(\cdot)$ hides polylog factors

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

Q: Does an analogous statement hold for other notions of complexity (e.g. VC-dimension)?

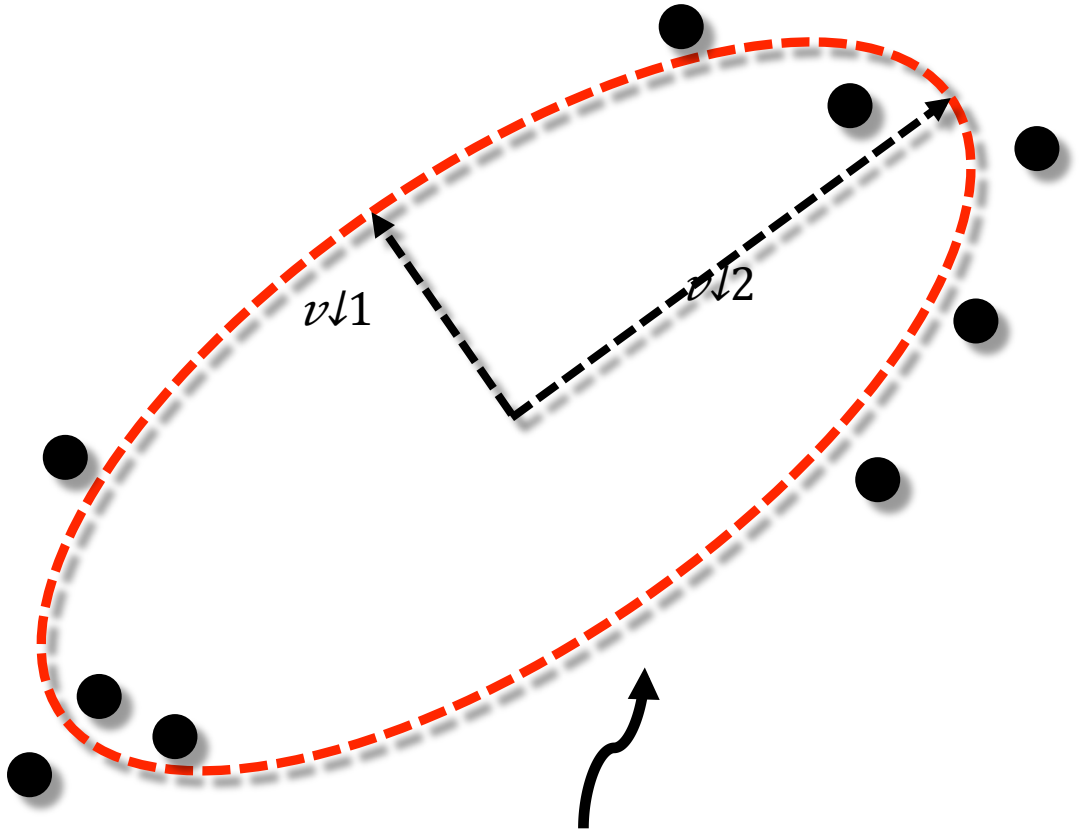
Application: Learning Mixtures of Gaussians

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



Application: Learning Mixtures of Gaussians

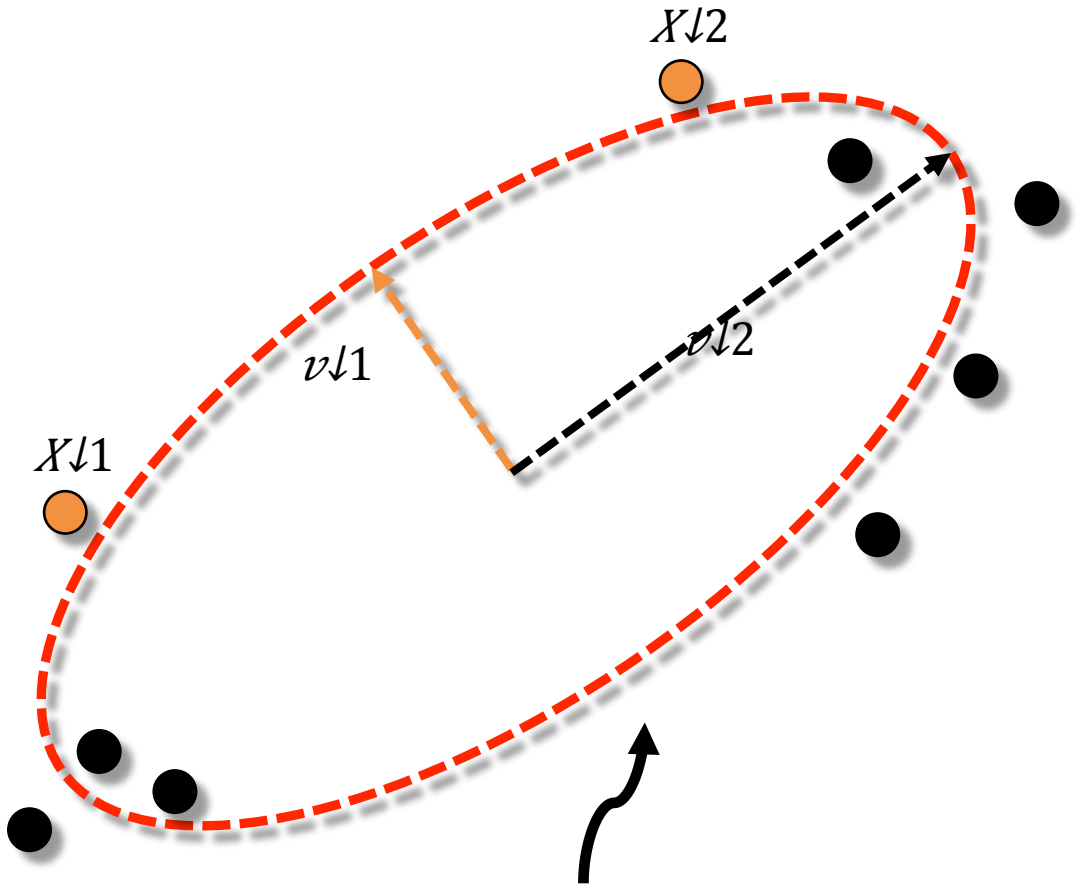
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.



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

Application: Learning Mixtures of Gaussians

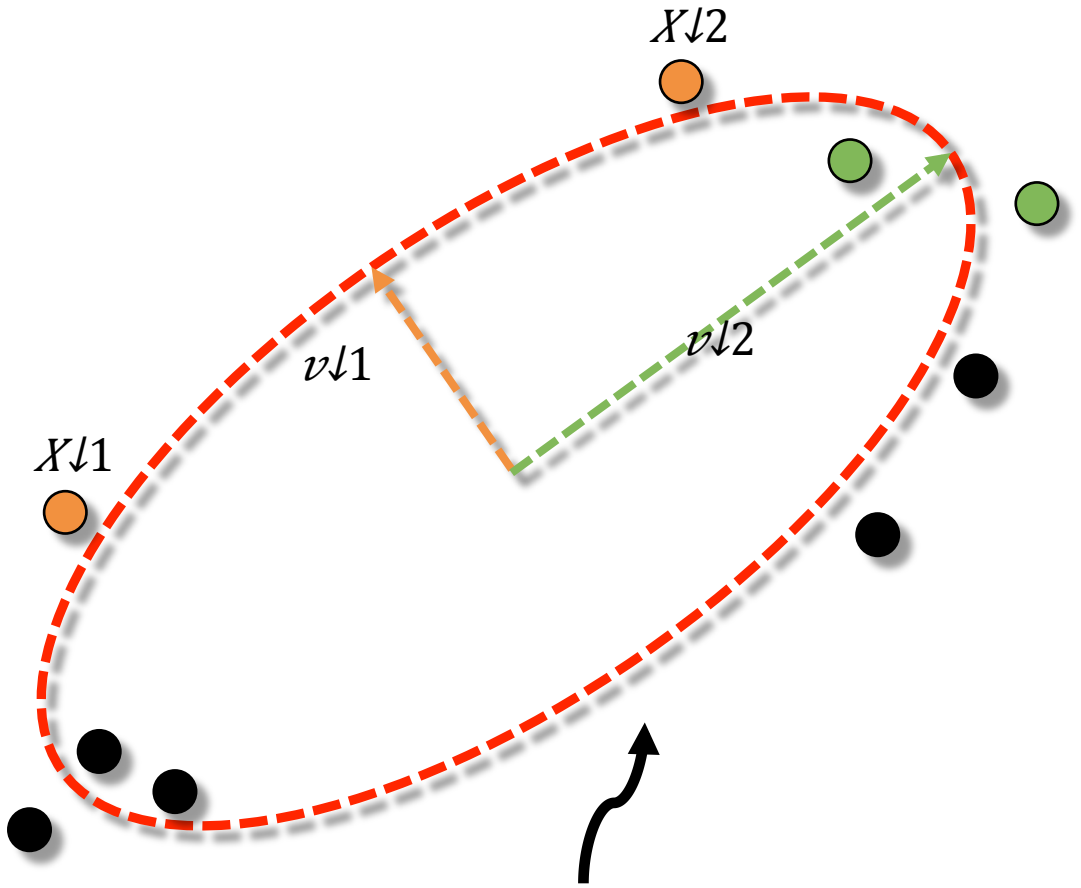
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.



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

Application: Learning Mixtures of Gaussians

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

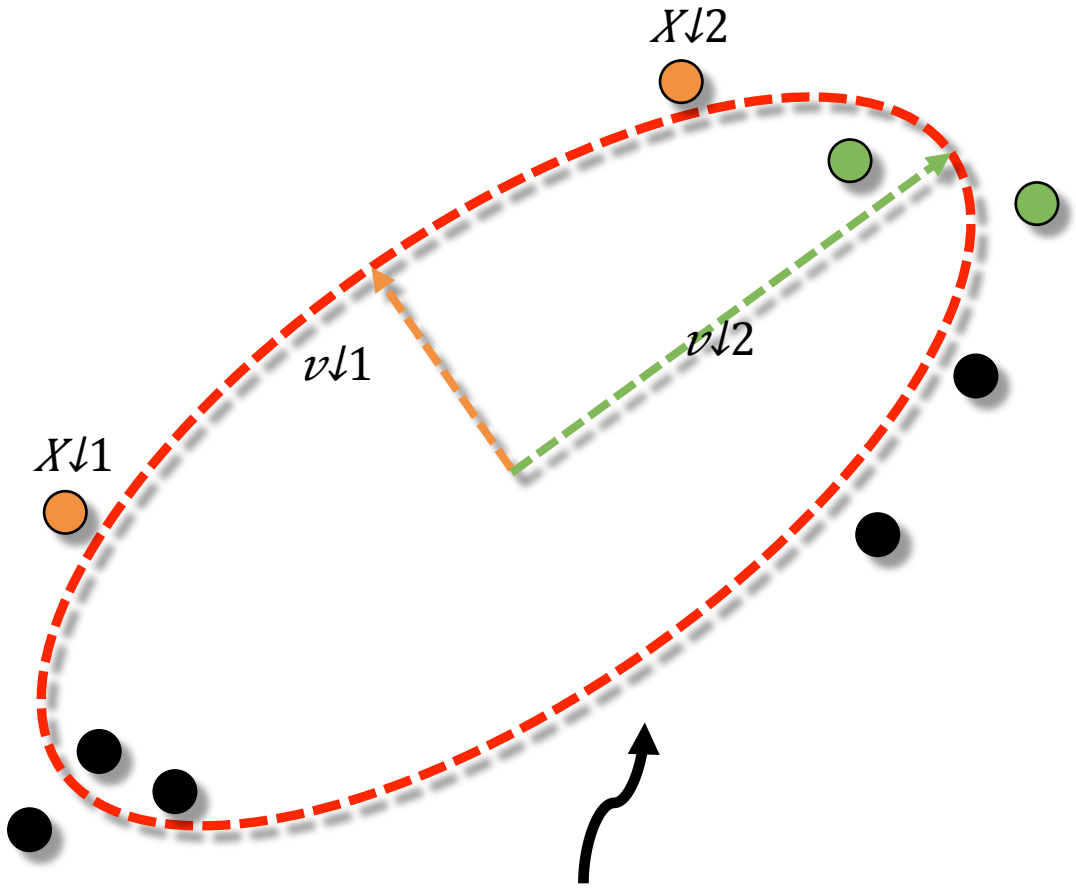


Ellipsoid defined by μ, Σ.
Points drawn from $\mathcal{N}(\mu, \Sigma)$.

Application: Learning Mixtures of Gaussians

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

In general, $O(d^2)$ compression is possible for Gaussians in $\mathbb{R}^d$.



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
$\tilde{\Theta}(kd^2/\epsilon^2)$

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

- Improves upon:
  - $O(k^4 d^4/\epsilon^2)$ via a VC-dimension argument
  - $O(kd^2/\epsilon^4)$ [Ashtiani, Ben-David, Mehrabian '17]
- This is nearly-tight! We show $\Omega(kd^2/\epsilon^2)$ samples are necessary.
  - Improves on previous bound of $\Omega(kd/\epsilon^2)$ [Suresh et al. NeurIPS '14]
- Compression ideas can be extended to agnostic learning as well.

Summary

- We introduced a compression framework for density estimation.
  - **Application:** improved upper bounds for learning mixtures of Gaussians.
  - **Q**: Other applications of compression?
  - **Q**: Can we get a more computationally-efficient algorithm?
- We also show a nearly-matching lower bound for learning mixtures of Gaussians.

Summary

- We introduced a compression framework for density estimation.
  - **Application:** improved upper bounds for learning mixtures of Gaussians.
  - **Q**: Other applications of compression?
  - **Q**: Can we get a more computationally-efficient algorithm?
- We also show a nearly-matching lower bound for learning mixtures of Gaussians.

Thank you!