Nearly-tight sample complexity bounds for learning mixtures of Gaussians


Hassan Ashtiani
(McMaster)


Shai Ben-David
(Waterloo)


Nick Harvey
(UBC)


Abbas Mehrabian
(McGill)

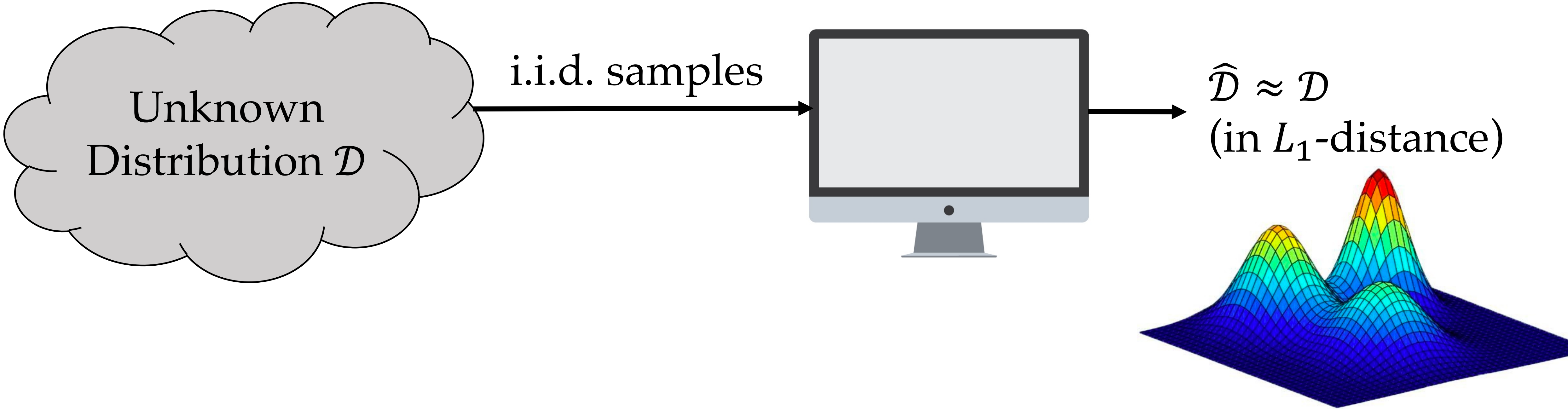

Yaniv Plan
(UBC)

Chris Liaw (UBC)

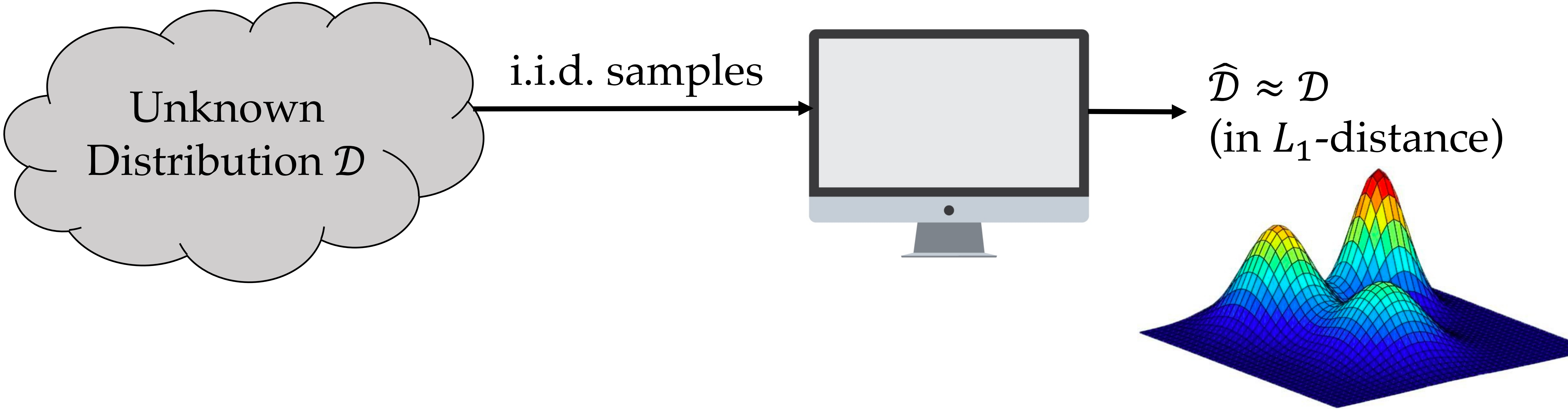
NeurIPS, December 2018

Poster #100

Density estimation



Density estimation

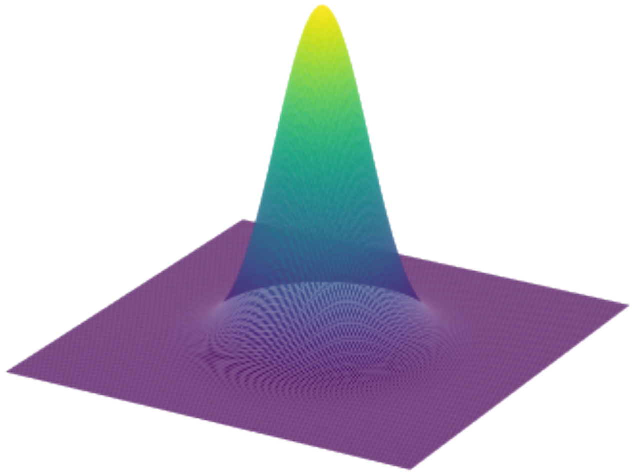


Fundamental & well-studied problem with many applications!

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

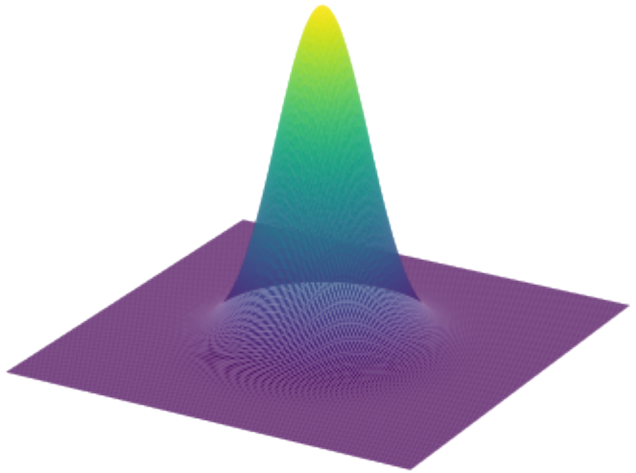
Q [D ‘16]: “For a distribution class $\mathcal{F}$, is there a complexity measure that characterizes the sample complexity of $\mathcal{F}$?”

Learning Gaussians

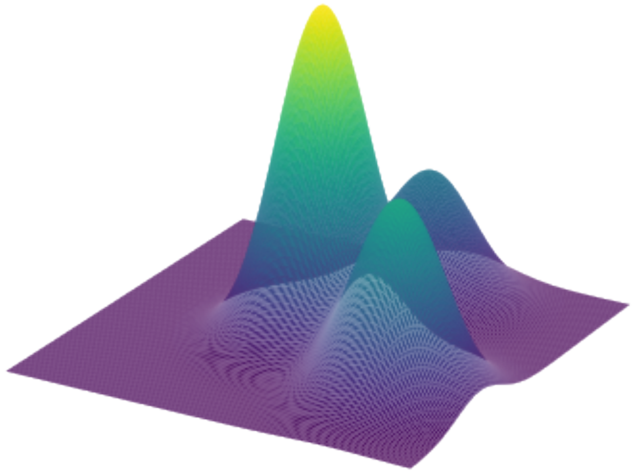


Single Gaussian in $\mathbb{R}^d$.
$O\left(\frac{d^2}{\epsilon^2}\right)$ samples are sufficient to recover Gaussian up to L_1-error ϵ.

Learning Gaussians



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



Mixture of k Gaussians in $\mathbb{R}^d$.
Q: Are $O\left(\frac{kd^2}{\epsilon^2}\right)$ samples sufficient?
Know that $\tilde{O}\left(\frac{kd^2}{\epsilon^4}\right)$ are sufficient. [Ashtiani et al. '17]

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

Main Contribution

- *"Other things being equal, simpler explanations are generally better…"*
  [William of Ockham]
- One manifestation of this in learning theory is "sample compression".
  [e.g. Littlestone, Warmuth '86; Moran, Yehudayoff '16]

Main Contribution

- *“Other things being equal, simpler explanations are generally better…”*
  [William of Ockham]
- One manifestation of this in learning theory is “sample compression”.
  [e.g. Littlestone, Warmuth ‘86; Moran, Yehudayoff ‘16]

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

Main Contribution

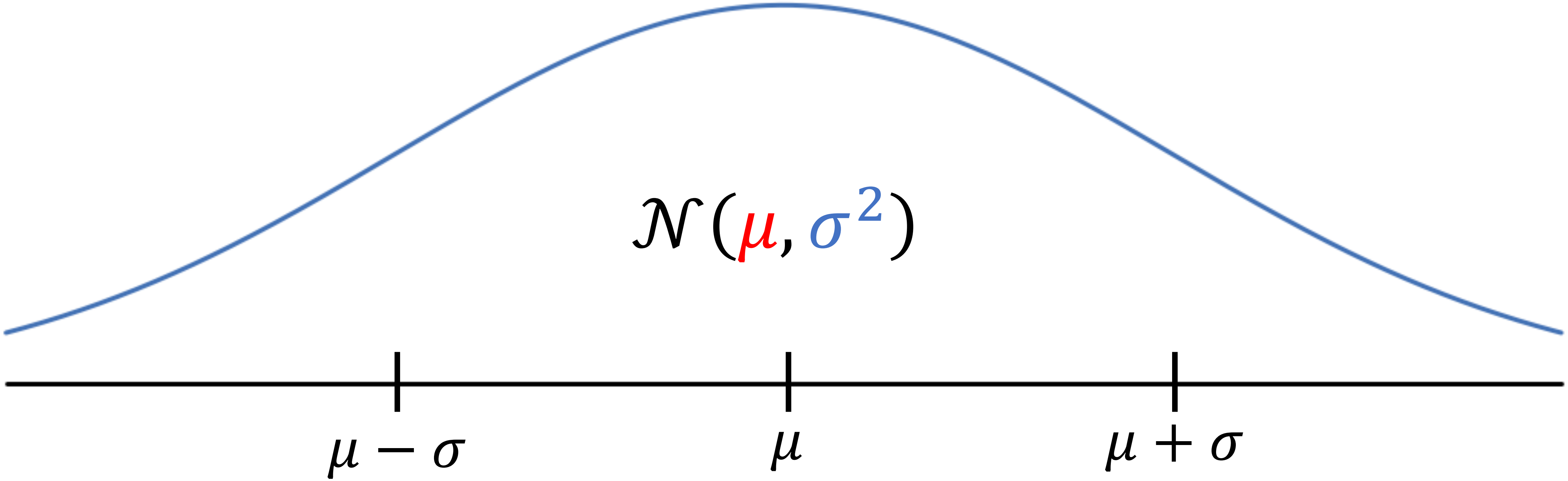
- *"Other things being equal, simpler explanations are generally better…"* [William of Ockham]
- One manifestation of this in learning theory is "sample compression". [e.g. Littlestone, Warmuth '86; Moran, Yehudayoff '16]

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

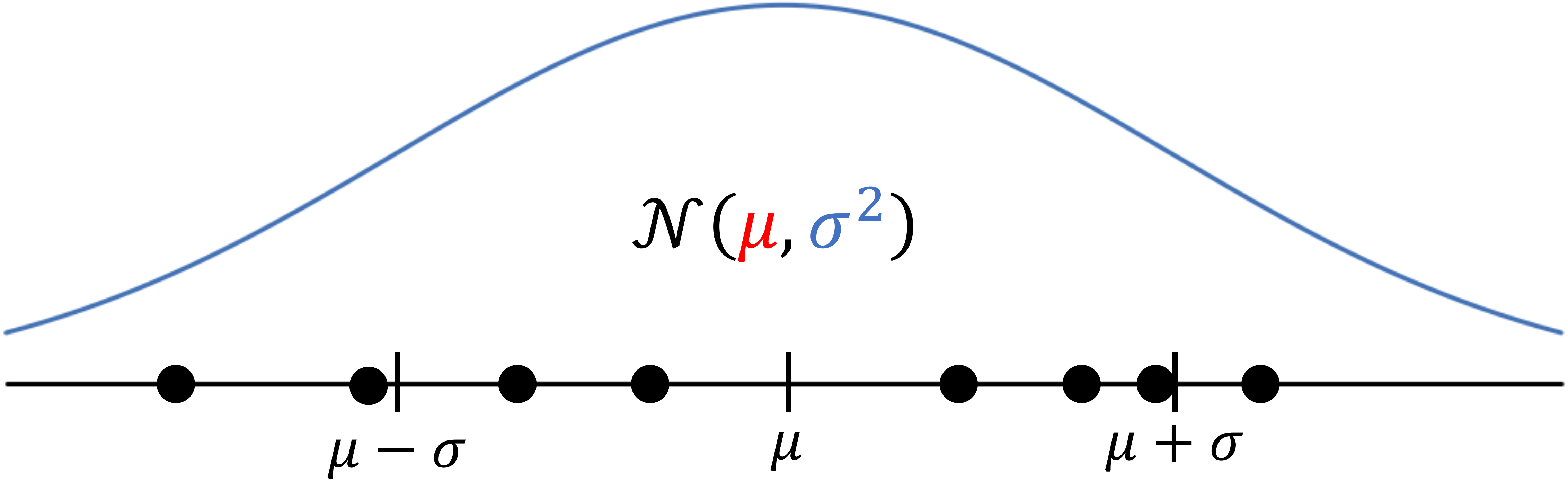
- Application: $\tilde{O}(kd^2/\epsilon^2)$ samples suffice to learn mixtures of k Gaussians in $\mathbb{R}^d$.
- We also show nearly-matching lower bound of $\tilde{\Omega}(kd^2/\epsilon^2)$.

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

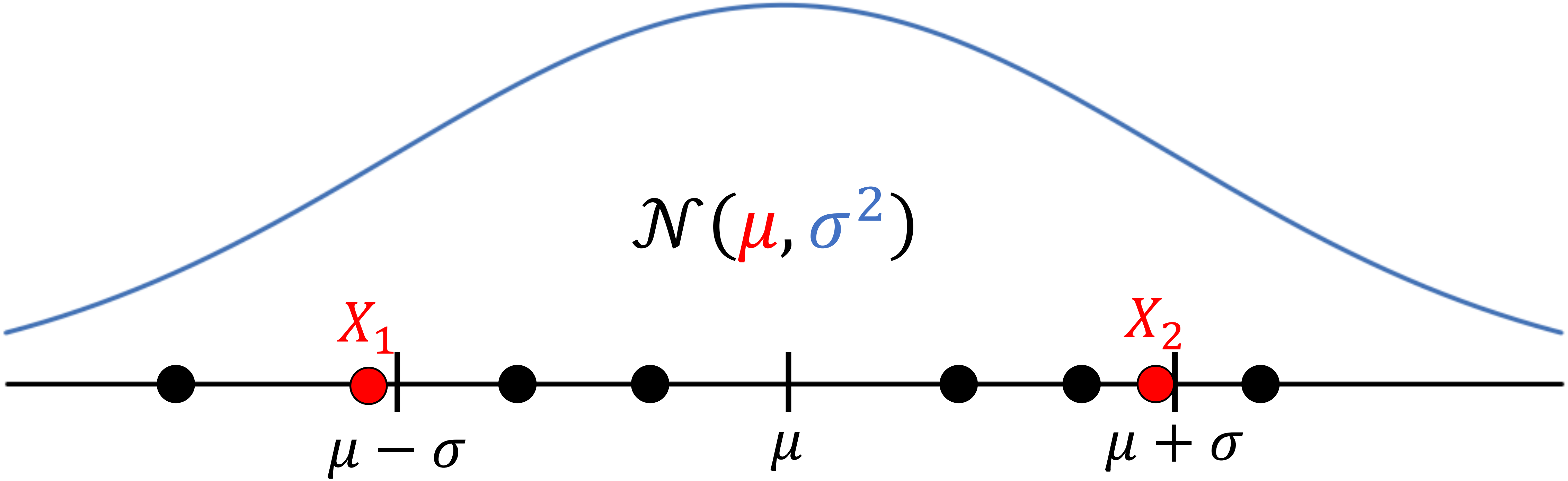
Compressing Gaussians in $\mathbb{R}$



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)

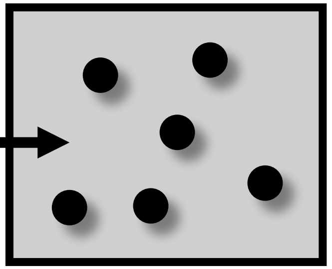
Knows $\mathcal{D}$, $\mathcal{F}$ 

 Knows $\mathcal{F}$

Compression Framework

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

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

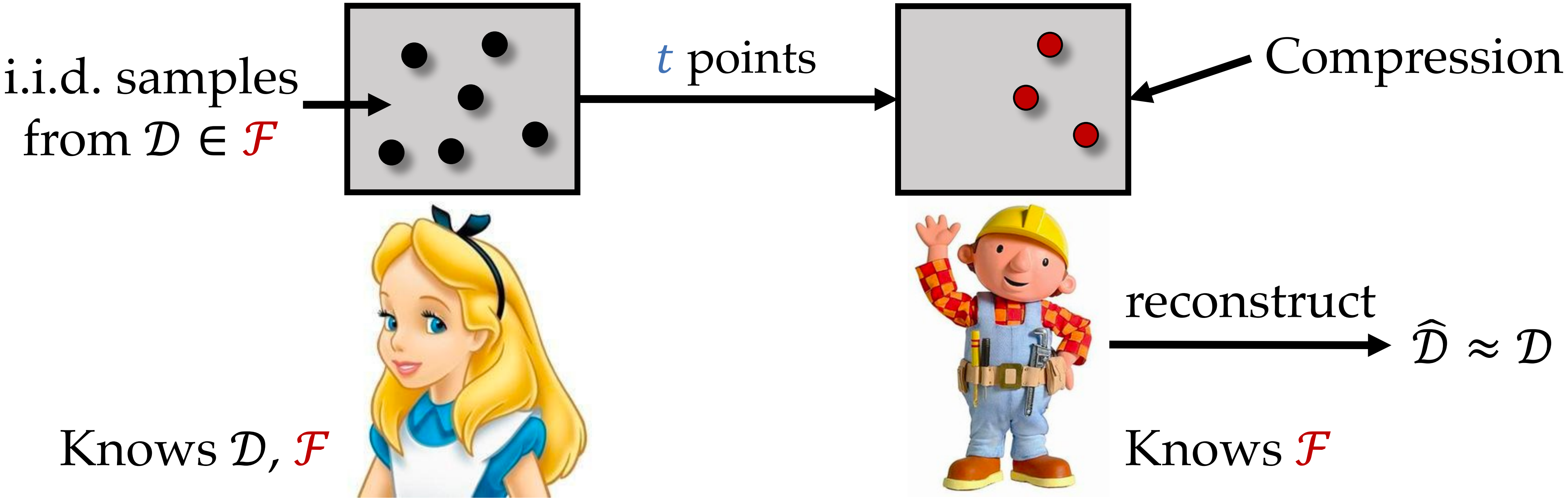


Knows $\mathcal{D}$, $\mathcal{F}$

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 $\boldsymbol{t}$ then sample complexity to learn $\mathcal{F}$ (up to L_1-error ϵ) is

$$\tilde{\boldsymbol{O}}\left(\frac{\boldsymbol{t}}{\boldsymbol{\epsilon}^2}\right).$$

$\tilde{O}(\cdot)$ 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

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

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

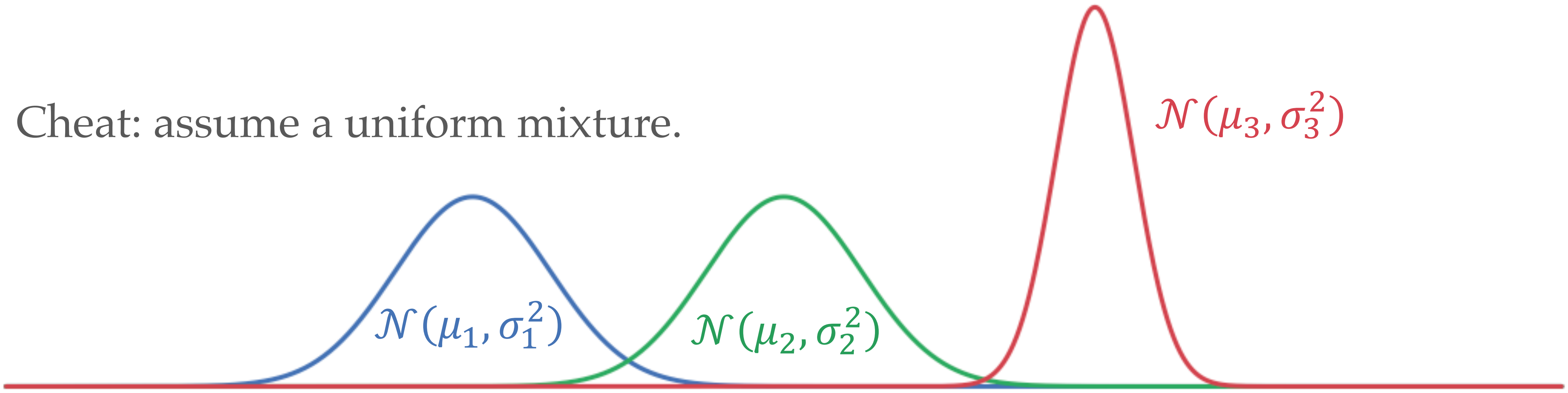
Small compression schemes imply **sample-efficient** algorithms.

Proof idea.

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

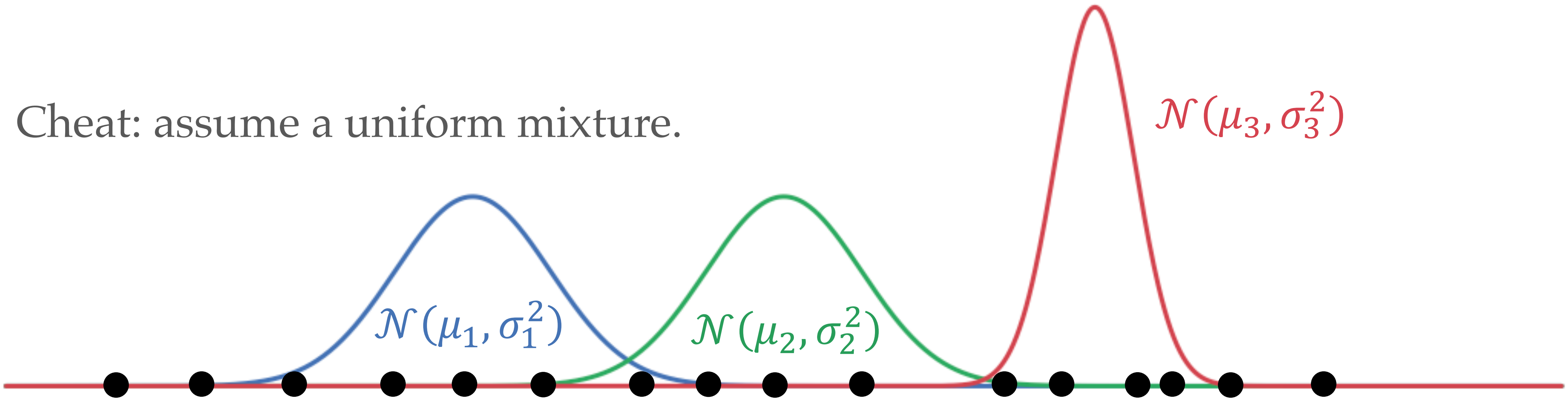
Compression Of Mixtures

Cheat: assume a uniform mixture.



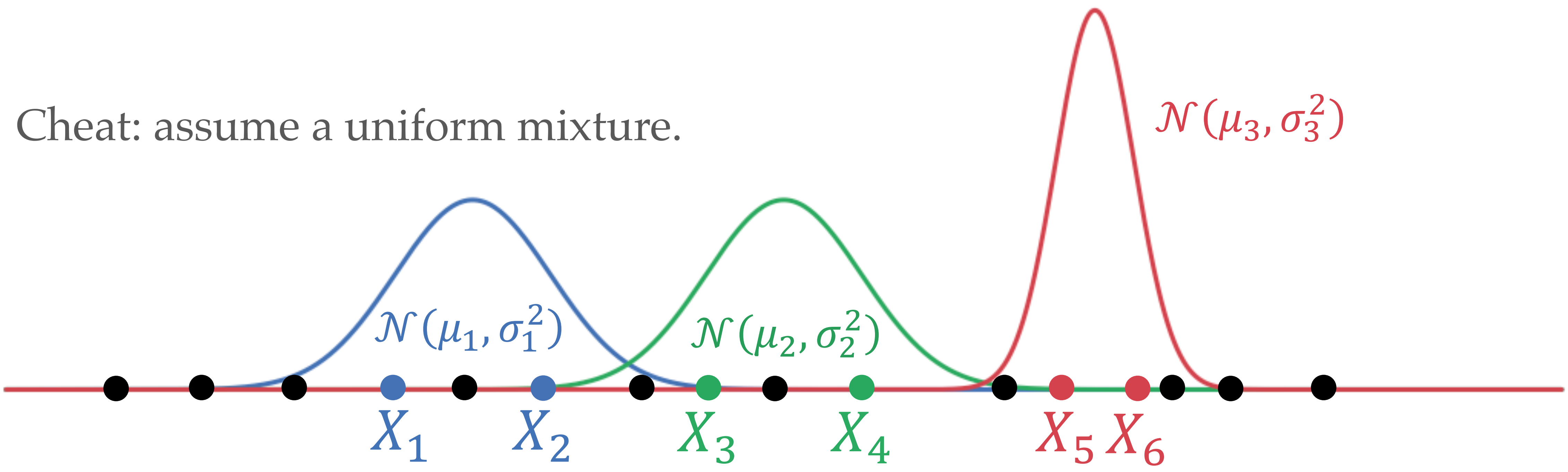
Compression Of Mixtures

Cheat: assume a uniform mixture.



Compression Of Mixtures

Cheat: assume a uniform mixture.



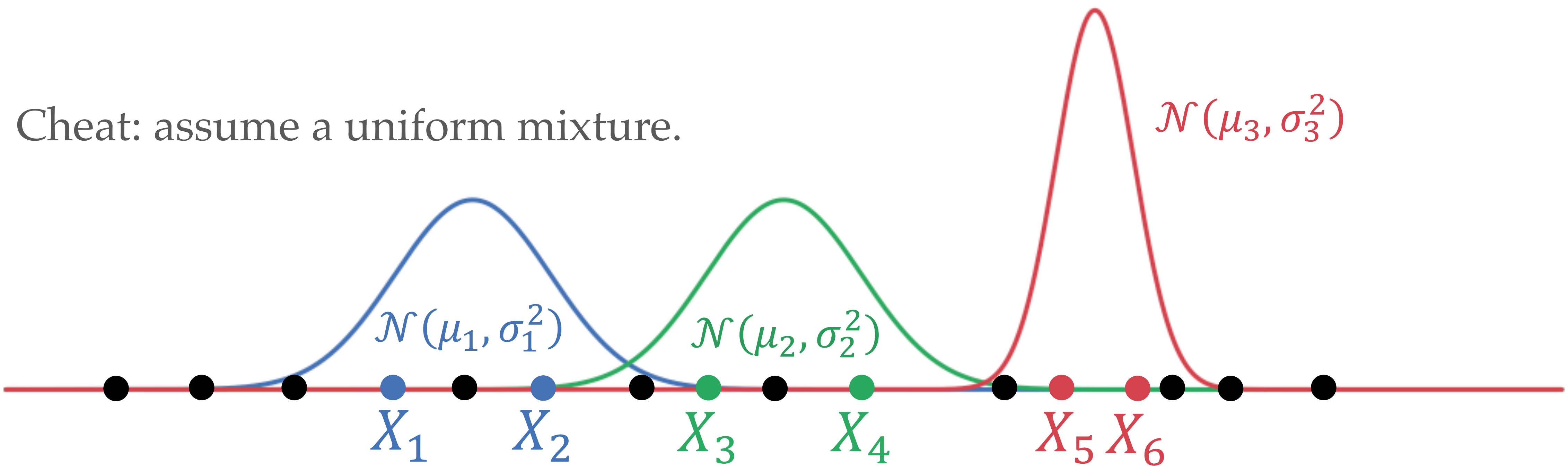
$X_1 \approx \mu_1 - \sigma_1$
$X_2 \approx \mu_1 + \sigma_1$

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

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

Compression 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

$$\tilde{O}\left(\frac{kt}{\epsilon^2}\right).$$

$\tilde{O}(\cdot)$ 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

$$\tilde{O}\left(\frac{kt}{\epsilon^2}\right).$$

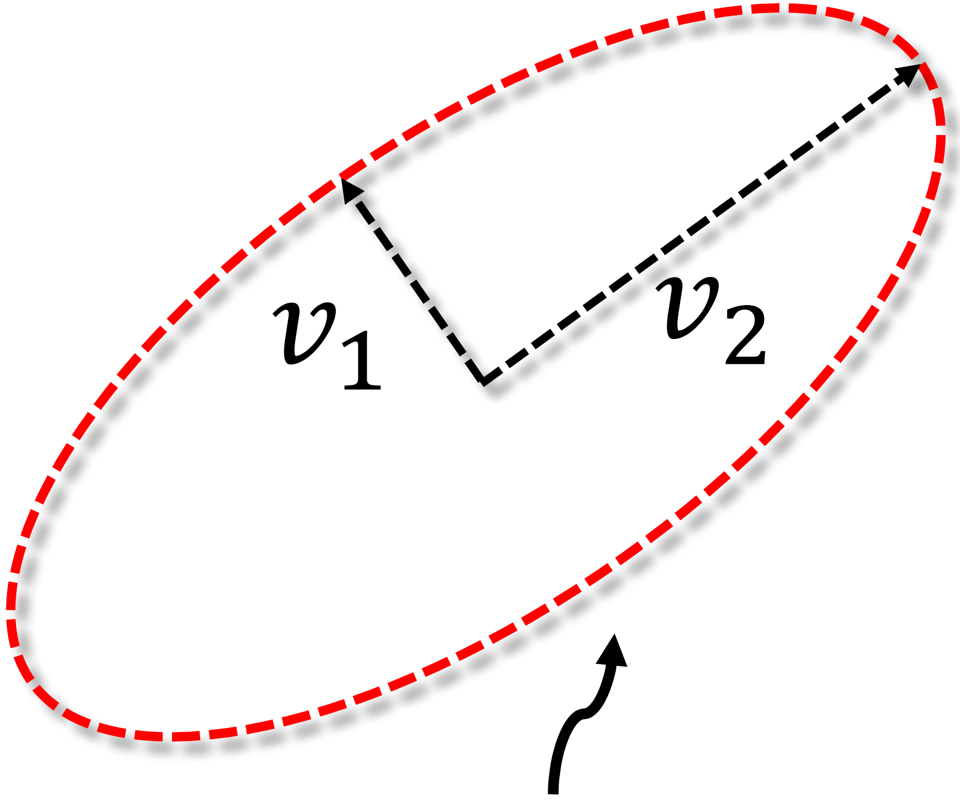
$\tilde{O}(\cdot)$ hides polylog factors

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

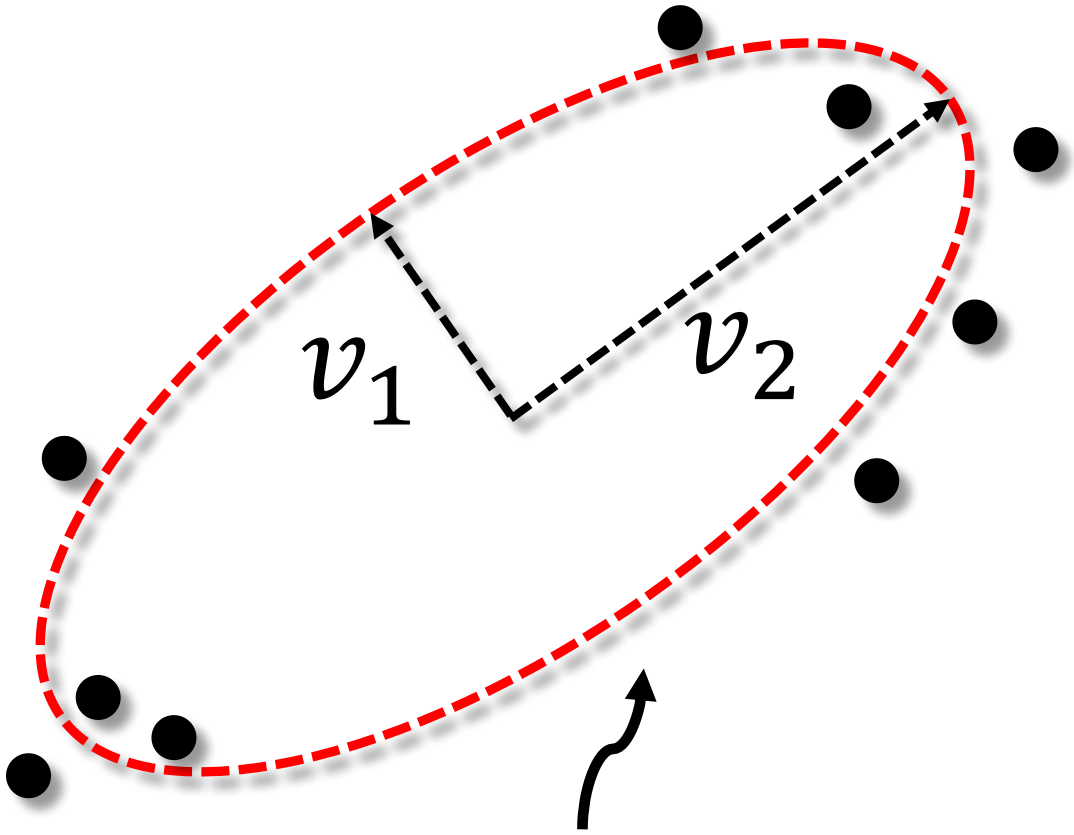
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.



Ellipsoid defined by μ, Σ.

Application: Learning Mixtures of Gaussians

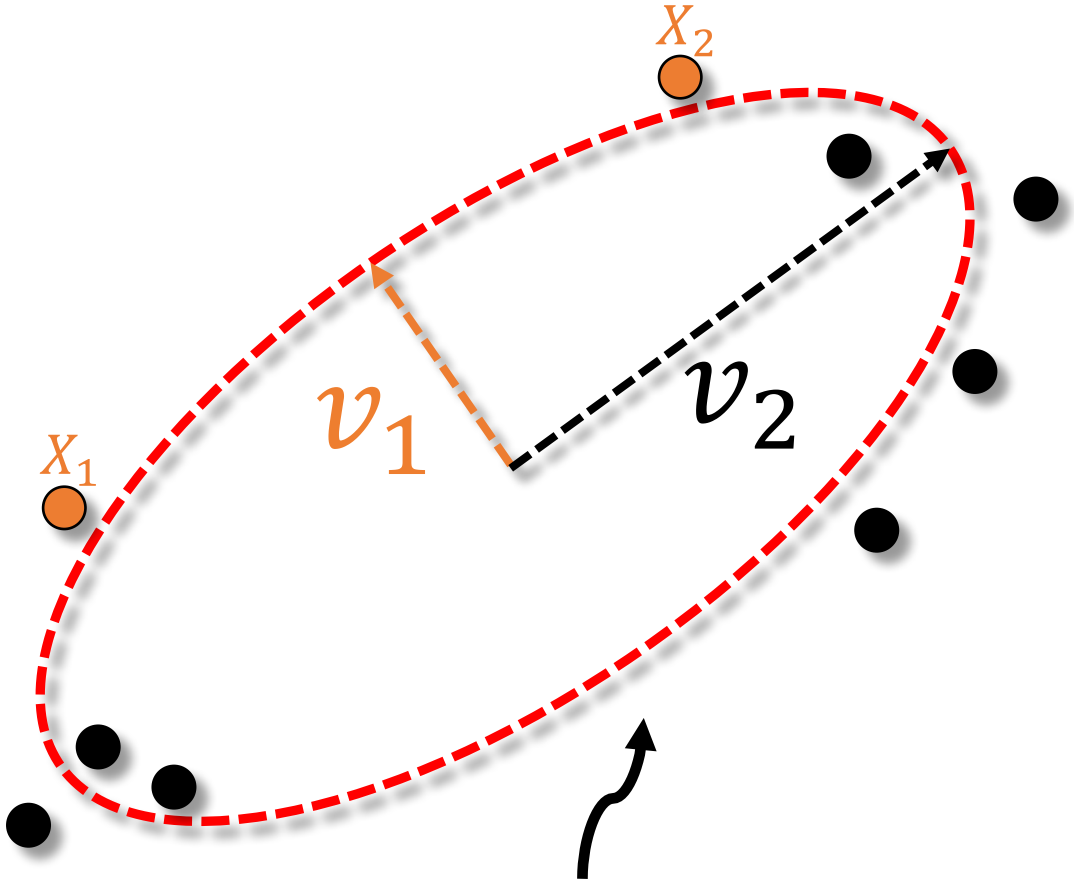
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.



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

Application: Learning Mixtures of Gaussians

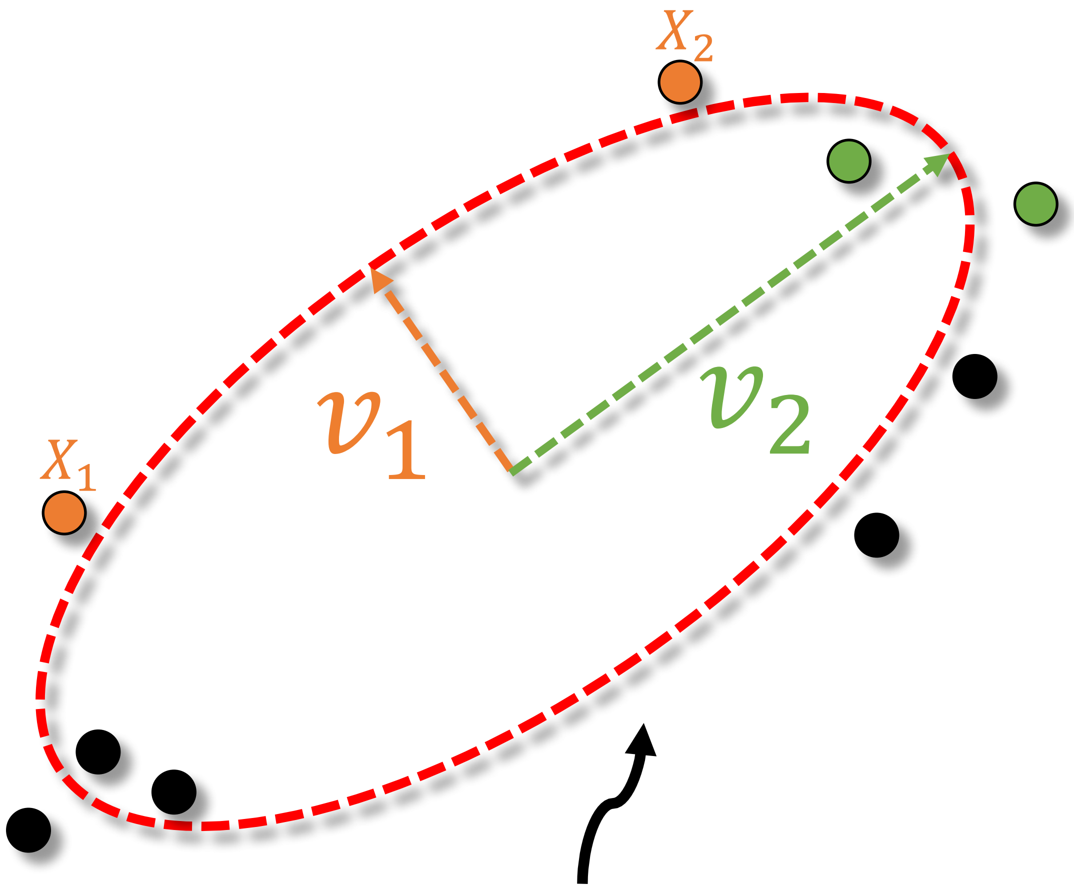
Encoding center and axes of ellipsoid is sufficient to recover $\mathcal{N}(\mu, \Sigma)$.



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 $\mathcal{N}(\mu, \Sigma)$.

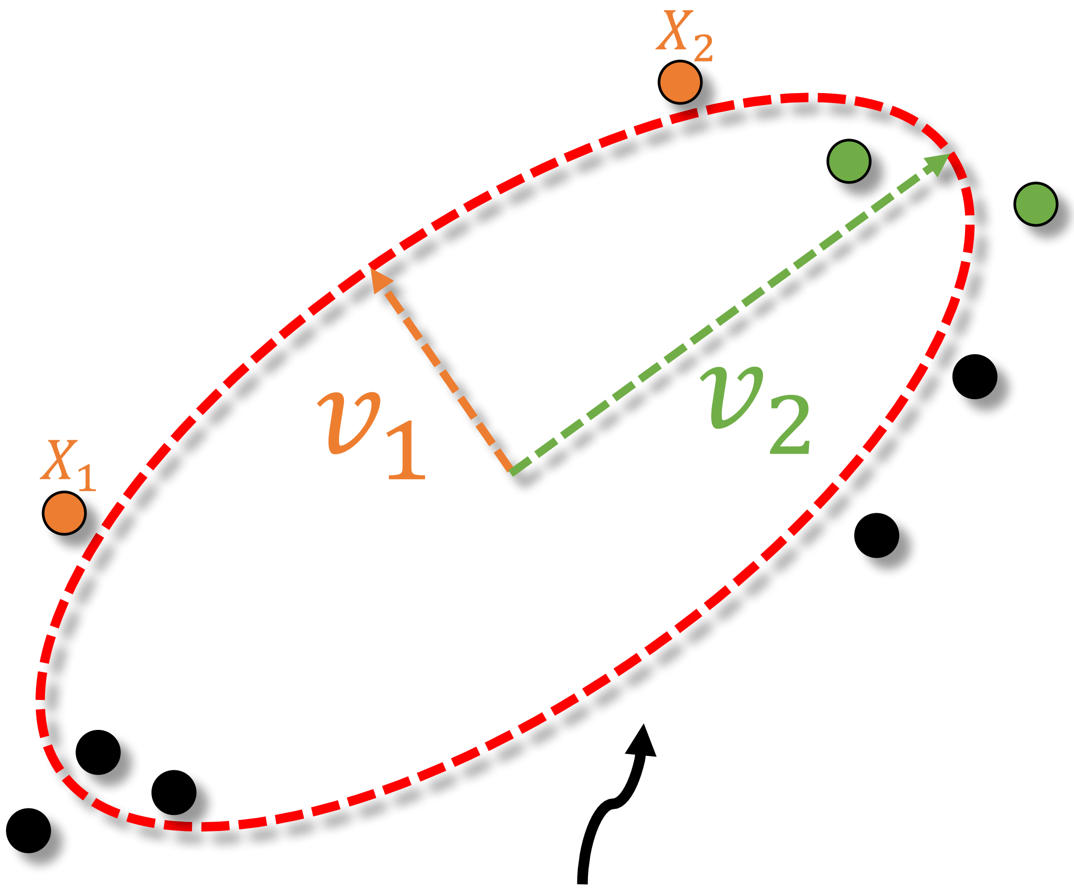


Ellipsoid defined by μ, Σ.
Points drawn from $\mathcal{N}(\mu, \Sigma)$.

Application: Learning Mixtures of Gaussians

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

In general, $\tilde{O}(d^2)$ 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

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

$\tilde{O}(\cdot)$ 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]
- This is nearly-tight! We show $\tilde{\Omega}(kd^2/\epsilon^2)$ samples are necessary.
  - Improves on previous bound of $\tilde{\Omega}(kd/\epsilon^2)$ [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!

See us at poster #100!