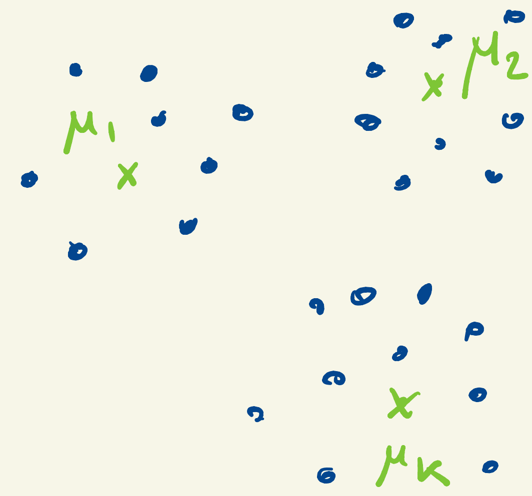
Information-computation gap in High-Dimensional clustering.

Bertrand Even[1], Christophe Giraud[1], Nicolas Verzelen[2]

[1] Labo de maths d'Orsay, université Paris Saclay

[2] MISTEA, INRAE Montpellier

- <ins>clustering</ins> = partitioning a set of points into K groups



- <ins>Model</ins>: we observe $X_1, \dots, X_m \in \mathbb{R}^d$

  $\exists\ G^*$ partition of $\{1, \dots, m\}$ in K groups:

  $$X_i \sim \mathcal{N}(\mu_k, \cancel{\sigma^2} I_d) \quad \forall i \in G^*_k$$

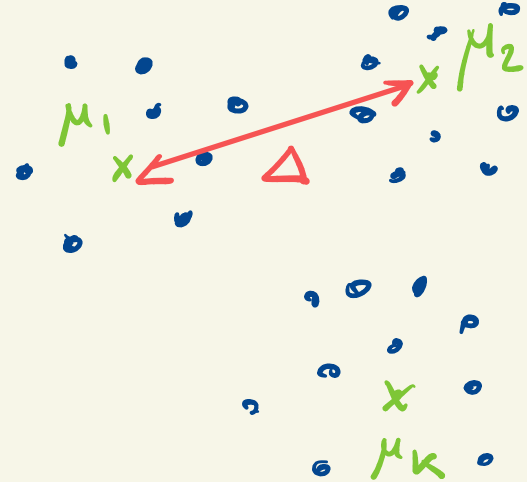
  w.l.o.g

- <ins>Goal</ins>: recover G^*

  $\nearrow$ exactly: $\hat{G} = G^*$

  $\searrow$ partially: $\mathrm{err}(\hat{G}) :=$ proportion of points well classified $\geq c > 0$.

- **Separation:** $\Delta^2 = \min_{k \neq l} \|\mu_k - \mu_l\|^2$



- **Assumption:** $|G_k^*| \simeq \frac{n}{K}$

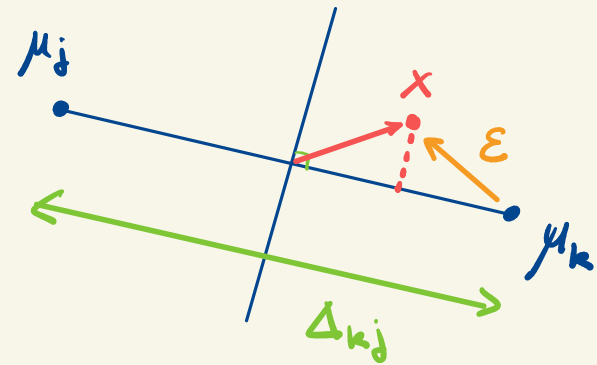
- **Our focus:**
  - → high dimension : $d \geq n$
  - → condition on Δ to recover exactly / partially G^*
    - → without computational constraints
    - → with computational constraints

- **Plan**

1. Detour on High-Dimensional classification
2. Information-Computation gap in HD clustering
3. Proving computational barriers

① High-dimensional classification

a/ with $\mu_1, \dots, \mu_K$ known



- consider 2 means μ_j, μ_k and $x \in \mathbb{R}^d$

$$S_{kj}(x) = \left\langle x - \frac{\mu_j + \mu_k}{2}, \mu_k - \mu_j \right\rangle$$

if $x = \mu_k + \underset{\mathcal{N}(0, I_d)}{\varepsilon}$

$$= \left\langle \frac{\mu_k - \mu_j}{2} + \varepsilon, \mu_k - \mu_j \right\rangle = \frac{1}{2} \Delta_{kj}^2 + \Delta_{kj} \mathcal{N}(0,1)$$

so

$$\mathbb{P}_{x \sim \mathcal{N}(\mu_k, I_d)} \left[\exists j : S_{kj}(x) < 0 \right] = \sum_j \mathbb{P}\left[\mathcal{N}(0,1) \leq -\frac{1}{2} \Delta_{kj} \right]$$

$$\leq K e^{-\Delta^2/8}$$

- Setting $\hat{k}(x) = \underset{k'}{\operatorname{argmax}} \ \min_{j : j \neq k'} S_{k'j}(x)$

  $\mathbb{P}[\text{1 point misclassified}] \leq K e^{-\Delta^2/8}$

  $\mathbb{P}[\text{at least 1 out of } n \text{ points misclassified}] \leq nK e^{-\Delta^2/8}$

- So if $\mu_1, \dots, \mu_K$ known, we need

$$\Delta^2 \gtrsim \begin{cases} \log(K) & \text{for partial recovery} \\ \log(n) & \text{for exact recovery.} \end{cases}$$

b/ with $\mu_1, \dots, \mu_K$ unknown:

we rely on estimators $\hat\mu_1, \dots, \hat\mu_K$ computed with sample size $m = \frac{n}{K}$:

$$\hat\mu_k = \mu_k + \frac{1}{\sqrt{m}} \xi_k \leftarrow \mathcal{N}(0, I_d)$$

$$\rightsquigarrow \hat S_{kj}(x) = \left\langle x - \frac{\hat\mu_k + \hat\mu_j}{2}, \hat\mu_k - \hat\mu_j \right\rangle$$

$$\hat S_{kj}(\mu_k + \varepsilon) = \left\langle \frac{\mu_k - \mu_j}{2} + \varepsilon - \frac{\xi_k + \xi_j}{\sqrt{m}}, \mu_k - \mu_j + \frac{\xi_k - \xi_j}{\sqrt{m}} \right\rangle$$

$$= \frac{1}{2}\Delta_{kj}^2 + \left(1 + O\left(\tfrac{1}{\sqrt{m}}\right)\right)\left[\Delta_{kj}\,\mathcal{N}(0,1) + \underbrace{\frac{\langle \varepsilon, \xi_k - \xi_j \rangle}{\sqrt{m}}}_{\simeq \sqrt{\frac{d}{m}}\,\mathcal{N}'(0,1)}\right]$$

$$\geq 0$$

$\uparrow$

if $\mathcal{N}(0,1) \geq -\Delta_{jk}$ and $\mathcal{N}'(0,1) \geq -\sqrt{\frac{dK}{n}}\,\Delta_{kj}^2$

So $\quad \mathbb{P}[\text{1 point misclassified}] \lesssim K \exp\left(-c\, \Delta^2 \wedge \frac{m\Delta^4}{Kd}\right)$

$\mathbb{P}[\text{at least 1 out of } m \text{ points misclass.}] \lesssim mK \exp\left(-c\, \Delta^2 \wedge \frac{m\Delta^4}{Kd}\right)$

So, with estimated means we need

$$\Delta^2 \overset{(*)}{\gtrsim} \log(\square) \vee \underbrace{\sqrt{\frac{dK}{m} \log(\square)}}_{\substack{\text{curse of dimensionality} \\ \text{for } d \geq \frac{m}{K} \log(\square)}}$$

with $\square = \begin{matrix} K \\ m \end{matrix}$ for $\begin{matrix} \text{partial} \\ \text{exact} \end{matrix}$ recovery

- Is $(*)$ the minimal separation for clustering?

② Information-Computation gap in HD clustering

a) Without computational constraints

Theorem : EGV '24

Partial / Exact recovery minimax impossible if

$$\Delta^2 \lesssim \log(\square) \vee \sqrt{\frac{dK}{n} \log(\square)}$$

and possible with exact Kmeans if

$$\Delta^2 \overset{(*)}{\gtrsim} \log(\square) \vee \sqrt{\frac{dK}{n} \log(\square)}$$

Remarks :

→ $\Delta^2 \gtrsim \log(K)$ already known for d small (Kwon and Caramis COLT 2020)

→ for $K=2$, tight rate for exact recovery in Ndaoud (AOS 2022)

b) <u>with computational constraints</u>

- Is clustering possible in polynomial time when $(*)$ holds?
  - → in low dimension: (kind of) yes [Liu and Li 2022]
  - → in high-dimension?

Theorem (informal) EGV '24

For $d \geq n$, under computational constraints (specified later)

$$\Delta^2 \overset{\log^\beta}{\gtrsim} \sqrt{\frac{dK^2}{n}} \wedge \sqrt{d}$$

is required and enough for partial recovery

⇝ information-computation gap

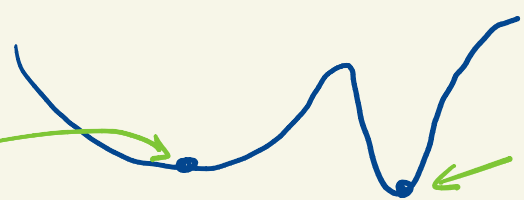
Remarks

→ when $\Delta^2 \overset{\log}{\gtrsim} \sqrt{d}$: recovery possible with hierarchical clustering
with single linkage

→ when $\Delta^2 \overset{\log}{\gtrsim} \sqrt{\frac{dK^2}{n}}$: recovery possible with SDP relaxation of Kmeans
(G.V. '19)

→ computational gap conjectured in Lesieur et al. (2016)
based on the computation of fixed points of State Evolution of AMP
$\Leftrightarrow$ local minima of Bethe free energy
$\Leftrightarrow$ replica theory predicts that multiple minima $\Leftrightarrow$ gap

local minima achieved by "local" search with non informative initialization



global optima hard to achieve

③ Proving computational barriers

→ **Worst case complexity**: proving that a problem is NP-hard (e.g. minimizing Kmeans exactly)

↳ not our case, as we consider some random instances with separation

→ **Reduction**: to a problem that we believe to be hard (e.g. planted clique)

→ **Computation model**: prove that some classes of algorithms fail.

Ex:
- SQ algorithms
- SoS algorithms
- local algorithms (MCMC) (landscape analysis)
- low degree polynomials

  ↳ our choice here

Rmk: there are connections between these notions and also with tools from statistical physics (replica symmetry and cavity method).

Recipe 1: from clustering to estimation

combinatorial $\rightsquigarrow$ continuous

- **Partnership matrix**:

  - $\Pi^G_{ij} := \mathbb{1}_{i \overset{G}{\sim} j} \in \{0,1\}^{n\times n}$

  - $\Pi^* := \Pi^{G^*}$

- **Estimation error**

$$R(\hat\Pi) := \frac{1}{n(n-1)} \sum_{i\neq j} (\hat\Pi_{ij} - \Pi_{ij})^2$$

- **Relation to clustering**

$$R(\Pi^{\hat G}) \leq 2\, \underbrace{\mathrm{err}(\hat G)}_{\text{proportion of misclustered points}}$$

So

$$\inf_{\hat\Pi \text{ poly-time}} R(\hat\Pi) \leq 2 \inf_{\hat G \text{ poly-time}} \mathrm{err}(\hat G)$$

Recipe 2: introducing a generative model

$\rightarrow k_1, \dots, k_n \overset{iid}{\sim} \mathcal{U}[\{1, \dots, K\}]$

and $G^*_k = \{i : k_i = k\}$

$\rightarrow \mu_1, \dots, \mu_K \overset{iid}{\sim} \mathcal{U}\left[\left\{-\frac{\Delta}{\sqrt d}, +\frac{\Delta}{\sqrt d}\right\}^d\right]$

$\Rightarrow \|\mu_j - \mu_k\|^2 \simeq \Delta^2$

We can investigate $\mathbb{E}\left[R(\hat\Pi)\right]$

expectation relative to prior + data generation

Low degree polynomials : (Schramm & Wein 2022)

We restrict to $\hat{M}$ s.t.

$$\hat{M}_{ij} = f_{ij}(X) \text{ with } f_{ij} \in \mathbb{R}_D[x]$$

$$D = O(\log(n))$$

⤳ approximate spectral

AMP

etc...

The goal: lower bound

$$\mathrm{MMSE}_D := \inf_{f_{ij} \in \mathbb{R}_D[x]} \mathbb{E}[R(f(X))]$$

for $D \simeq \log(n)$.

Theorem EGV'24

$$\boxed{\text{if } \Delta^2 \underset{\log^\beta}{\leq} \sqrt{\frac{dK^2}{n}} \wedge \sqrt{d} \text{ then } \quad \mathrm{MMSE}_{O(\log(n))} = \frac{1}{K} - \frac{1+o(1)}{K^2}}$$

Remark: $\tilde{M}_{ij} = \frac{1}{K}$ for $i \neq j$

fulfills

$$\mathbb{E}[R(\tilde{M})] = \frac{1}{K} - \frac{1}{K^2}$$

Sketch of proof:

1. focusing on a single entry
2. relating MMSE to cumulants (Schramm & Wein 22)
3. bounding cumulants (technical)

① Since the MMSE_D optimisation problem is separable

$$\mathrm{MMSE}_D = \inf_{g \in \mathbb{R}_D[x]} \mathbb{E}\Big[\big(g(x) - \underbrace{M^*_{12}}_{=:m}\big)^2\Big]$$

② Relating MMSE to cumulants

(Schramm & Wein '22)

$$\mathrm{MMSE}_D = \|m - P_D m\|^2_{L^2} = \|m\|^2_{L^2} - \|P_D m\|^2_{L^2}$$

$$= \underbrace{\|m\|^2_{L^2}}_{=\frac{1}{K}} - \underbrace{\left[\sup_{g \in \mathbb{R}_D(x)} \frac{\langle m, g(x)\rangle_{L^2}}{\|g(x)\|_{L^2}}\right]^2}_{=:\mathrm{Corr}^2_D}$$

Below $\underset{m \times d}{X} = \underset{\mathbb{E}[X]}{Z} + E$ $\quad E_{ij} \overset{iid}{\sim} \mathcal{N}(0,1)$

Lemma (SW'22, translated in our setting)

$$\mathrm{Corr}^2_D \leq \sum_{\substack{\alpha \in \mathbb{N}^{m \times d} \\ |\alpha| \leq D}} \frac{\kappa_\alpha^2}{\alpha!} \qquad (**)$$

where $\cdot\ \kappa_\alpha = \mathrm{cumulant}(m, \dots, \underbrace{z_{ij}, \dots, z_{ij}}_{\alpha_{ij} \text{ times}}, \dots)$

$\cdot\ \alpha! = \prod_{ij} \alpha_{ij}!$

Proof:

- Inequality from Jensen

$$\mathbb{E}\left[g(x)^2\right] \geq \mathbb{E}_E\left[\mathbb{E}_Z[g(Z+E)]^2\right]$$

- expansion on Hermite polynomials
- Linear algebra
- recognize recursion of cumulants

$\square$

How can we exploit (**) ?

(i) exploit the property

$$X \perp\!\!\!\perp Y \Rightarrow \text{cumulant}(X, Y) = 0$$

to detect the $K_\alpha = 0$ and

to prune $\sum_\alpha \frac{K_\alpha^2}{\alpha!}$

(ii) relate cumulants to moments

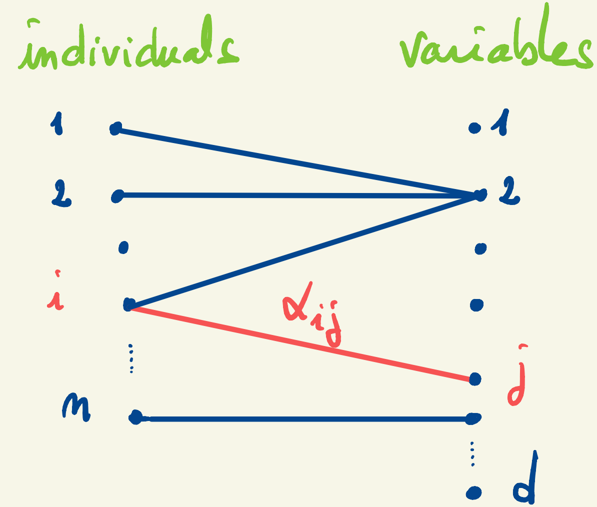
$$K_\alpha \overset{(***)}{=} \mathbb{E}[m Z^\alpha] - \sum_{\beta \lneq \alpha} K_\beta \binom{\alpha}{\beta} \mathbb{E}[Z^{\alpha-\beta}]$$

(iii) upper-bound the moments

$$\mathbb{E}[m Z^\alpha] \quad \text{and} \quad \mathbb{E}[Z^{\alpha-\beta}]$$

(iv) bound K_α by induction from (***)

(i) represent $\alpha \in \mathbb{N}^{m \times d}$ as weighted bipartite graph G_α



<u>Lemma</u>: If $K_\alpha \neq 0$ then

- G_α^+ connex
- individuals 1 and 2 $\in G_\alpha^+$
- each variable $j \in G_\alpha^+$ connected to at least 2 individuals.

(iii) Bounds on the moments

Define: $c_\alpha :=$ # connected components of G_α^+

$l_\alpha =$ # nodes of G_α^+

Then $$\mathbb{E}[m Z^\alpha] \leqslant \left(\frac{\Delta}{\sqrt{d}}\right)^{|\alpha|_1} \left(\frac{|\alpha|_1^{|\alpha|_1}}{K^{l_\alpha - \frac{1}{2}|\alpha|_1 - c_\alpha}} \wedge \frac{1}{K} \right)$$

Idea: reminder: $k_1, \dots, k_m \overset{iid}{\sim} \mathcal{U}\{1, \dots, K\}$

$\mu_1, \dots, \mu_K \overset{iid}{\sim} \mathcal{U}\left\{-\frac{\Delta}{\sqrt{d}}, \frac{\Delta}{\sqrt{d}}\right\}^d$

and $$Z^\alpha = \prod_{i,j} \mu_{k_i j}^{\alpha_{ij}} = \prod_{k,j} \mu_{kj}^{\sum_{i \in G_k} \alpha_{ij}}$$

So $$\mathbb{E}[Z^\alpha \mid k_1 \dots k_m] = \left(\frac{\Delta}{\sqrt{d}}\right)^{|\alpha|_1} \times \begin{cases} 1 & \text{if } \sum_{i \in G_k} \alpha_{ij} \text{ even } \forall k, j \\ 0 & \text{otherwise} \end{cases}$$

Hence $\mathbb{E}[Z^\alpha] = \left(\frac{\Delta}{\sqrt{d}}\right)^{|\alpha|_1} \underbrace{\mathbb{P}_G\left[\sum_{i \in G_k} \alpha_{ij} \text{ even } \forall k, j\right]}_{\text{where delicate combinatorics kicks in } \dots}$ $\square$

(iv) Bound κ_α by induction

- $\kappa_0 = \mathbb{E}[m] = \frac{1}{K}$
- induction: from (***)

$$\kappa_\alpha \leq \left(\frac{\Delta}{\sqrt{d}}\right)^{|\alpha|_1} (1 + |\alpha|_1)^{|\alpha|_1} \left[\frac{|\alpha|_1^{|\alpha|_1}}{K^{l_\alpha - \frac{1}{2}|\alpha|_1 - 1}} \wedge \frac{1}{K}\right]$$

(v) Conclusion

$$\mathrm{Corr}_D^2 \leq \sum_{|\alpha|_1 \leq D} \frac{\kappa_\alpha^2}{\alpha!} \leq \frac{1 + o(1)}{K^2}$$

for $\Delta^2 \overset{\log \beta}{\leq} \sqrt{\frac{dK^2}{n}} \wedge \sqrt{d}$ and $D = O(\log(n))$ $\square$

Take Home Message:

- Low degree polynomials are handy for proving the existence of computational barriers, at the price of spurious log factors.

- Minimal information separation:

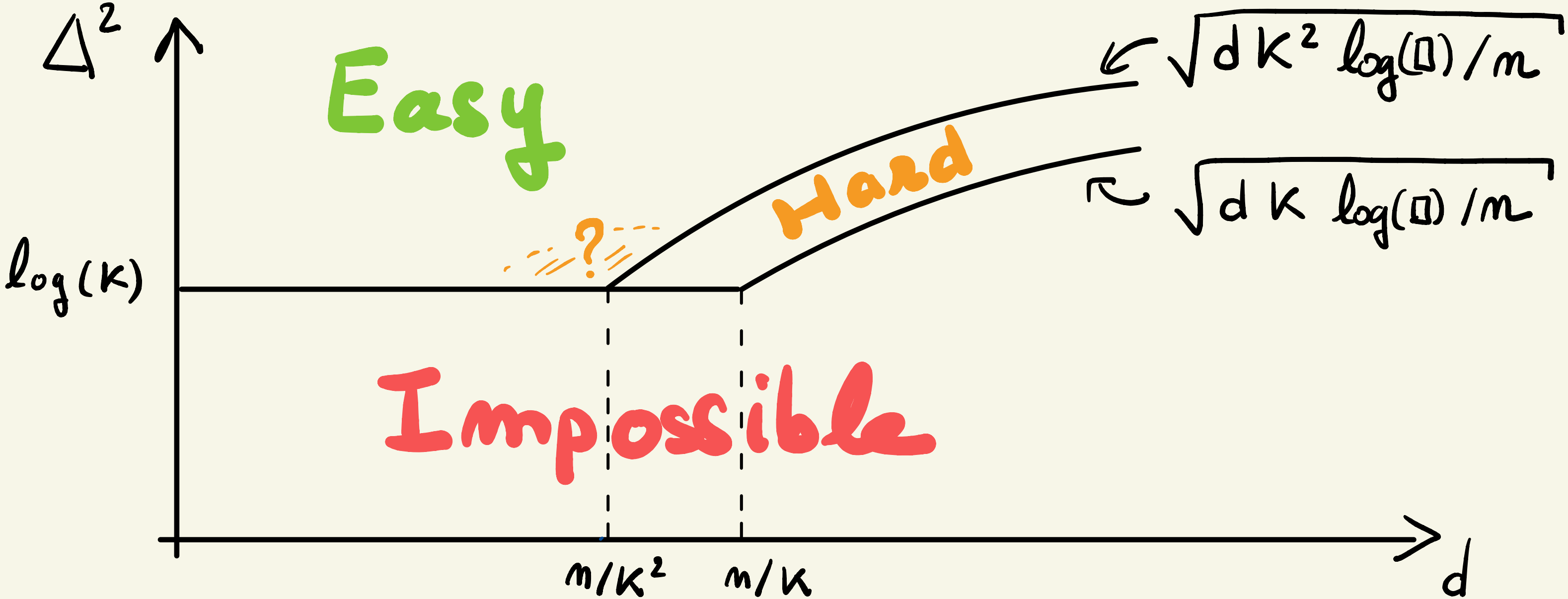
$$\Delta_I^2 \asymp \log(\square) \vee \sqrt{\frac{dK\log(\square)}{n}}$$

- Minimal computational separation: (conjectured)

$$\Delta_C^2 \asymp \log(\square) \vee \sqrt{\left(\frac{K^2}{n}\wedge 1\right) d\log(\square)}$$

proved for $d \geq n$ or d small

in progress for $\frac{n}{K^2} \leq d \leq n$



- What is special with $\Delta^2 \overset{\log}{\gtrsim} \sqrt{\frac{dK^2}{n}}$?

  $\rightsquigarrow$ related to BBP transition for "isotropic" $\mu_1 \dots \mu_K$:

  $n\Delta^4 \gtrsim dK^2$ is where K largest eigenvalues of the Gram matrix escape of the bulk.

- Remarkable feature: the Information-Computation gap disappear in an active setting.

active setting: we can sample each point multiple time.

with a total budget of T observations, minimal separation is

$$\Delta_*^2 \asymp \frac{n}{T}\left[\log(n) \vee \sqrt{\frac{dk}{n}\log(n)}\right]$$

and no computationnal barrier

Victor Thuot, Alexandra Carpentier, C.G., Nicolas Verzelen 2024.

why?

$\rightsquigarrow$ we can collect localized information