Low-degree Lower bounds for latent models: example of clustering

Bertrand Even

Joint work with Christophe Giraud and Nicolas Verzelen

Motivation: computation information gaps

In some high dimensional statistical problems, **computation-information gap** are conjectured;

$$\boxed{\inf_{any\ \hat{f}} err(\hat{f}) << \inf_{\hat{f}\ poly-time} err(\hat{f})}$$

Question: How can we quantify the optimal performance over estimators that are computable in **polynomial time**? How can we give evidence of the existence of a **Computational gap**?

Low-Degree framework

(Tests: Hopkins ’18,
Estimation: Schramm/Wein ’22)

- The notion of NP-hardness is worst case: not suitable for statistical problems where we seek to have **average-case hardness.**

- We need to consider a **model of computation:** one of them is the model of low-degree polynomials. We analyse the best performance achieved by a multivariate polynomial of **low-degree D.**

- Low degree $D = log(n)^{1+\eta}$ is used as a **proxy** for algorithms computable in polynomial time. Can approximate most of classical statistical methods: spectral methods, **AMP**…

$\boxed{\text{Stat. Physics predictions}}$

Goal: Given a statistical problem, what is the optimal performance of $\log(n)^{1+\eta}$-degree polynomial?

Contribution: Strategy for proving low degree lower bounds in some latent variable models with gaussian noise.

Builds on Schramm/Wein 22 which proves a generic formula for Gaussian Additive models. In addition, we leverage the presence of latent variables in the models we consider;

- Clustering
- Sparse Clustering
- Bi-clustering

We can characterize almost completely the different information-computational landscapes

Example: Computation-Information Gap in Clustering Gaussian mixtures

Isotropic Gaussian Mixture Model

Observation: $Y_1, \ldots, Y_n \in \mathbb{R}^d$. Feature Matrix $Y \in \mathbb{R}^{n \times d}$.

Model:

- Hidden partition $G^* = G_1^*, \ldots, G_K^*$ into K groups.
- Hidden vectors $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$.
- $Y_i \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}\left(\mu_k, I_p\right)$, if $i \in G_k^*$.

Goal: Estimate G^*

- **Separation**: $\Delta^2 := \dfrac{1}{2} \min_{k \neq l} \|\mu_k - \mu_l\|^2$.
- **Assumption**: the partition G^* satisfies $|G_k^*| \asymp \dfrac{n}{K}$ for all $k \in [1, K]$.
- **Questions**: What are the conditions on Δ^2, d, n and K to recover exactly/partially the partition G^*

$\boxed{\text{Without any computational constraints?}}$

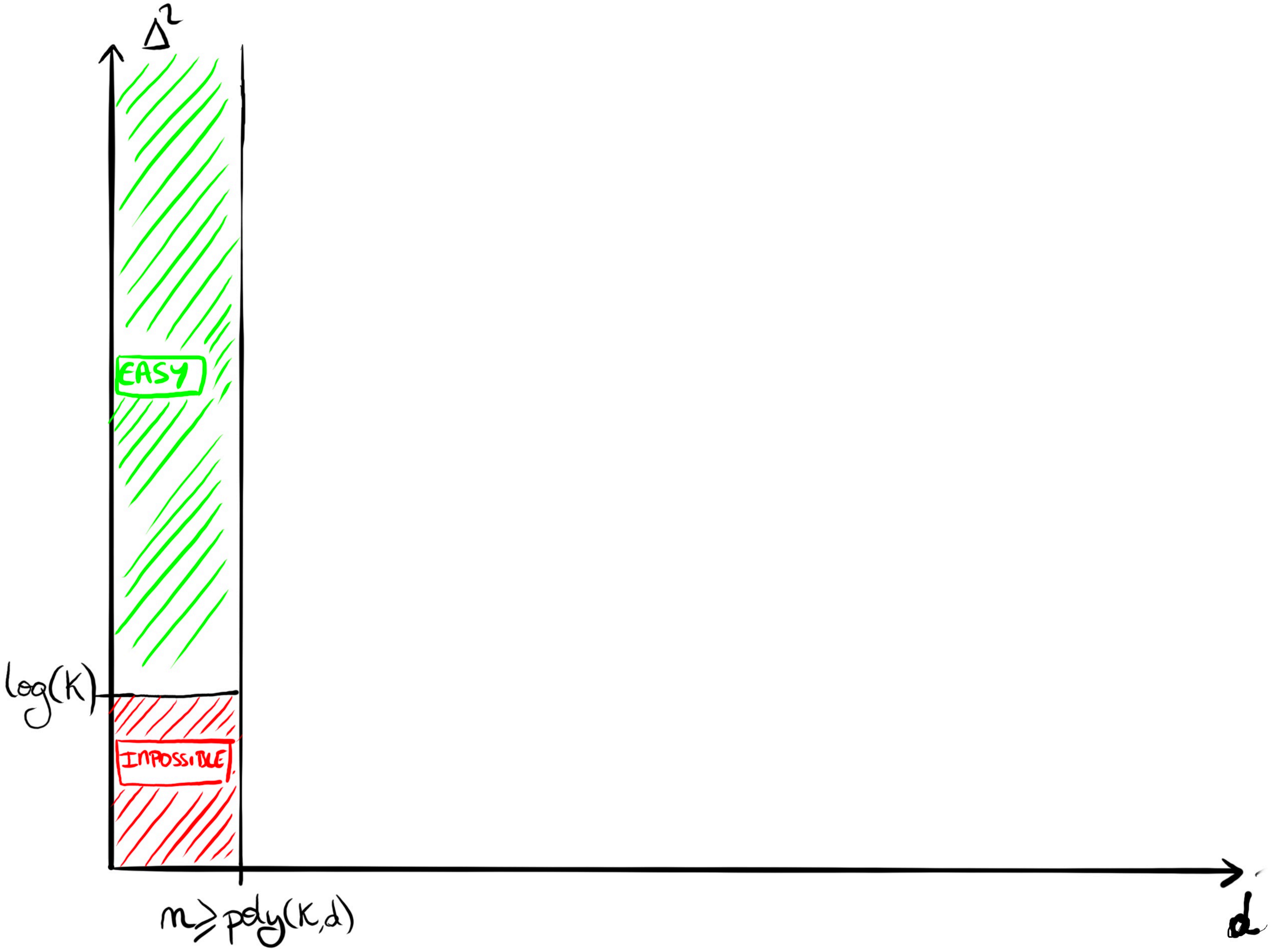
$\boxed{\text{With algorithms computable in polynomial time?}}$

Low-dimensional regime $n \geq poly(d, K)$

- Frontier for **partial recovery** is $log(K)$

- Liu/Li; when $\Delta^2 \geq \log(K)^{1+c}$, algorithm **computable in polynomial times** achieves partial recovery

- **No computational gap** in the low dimensional regime

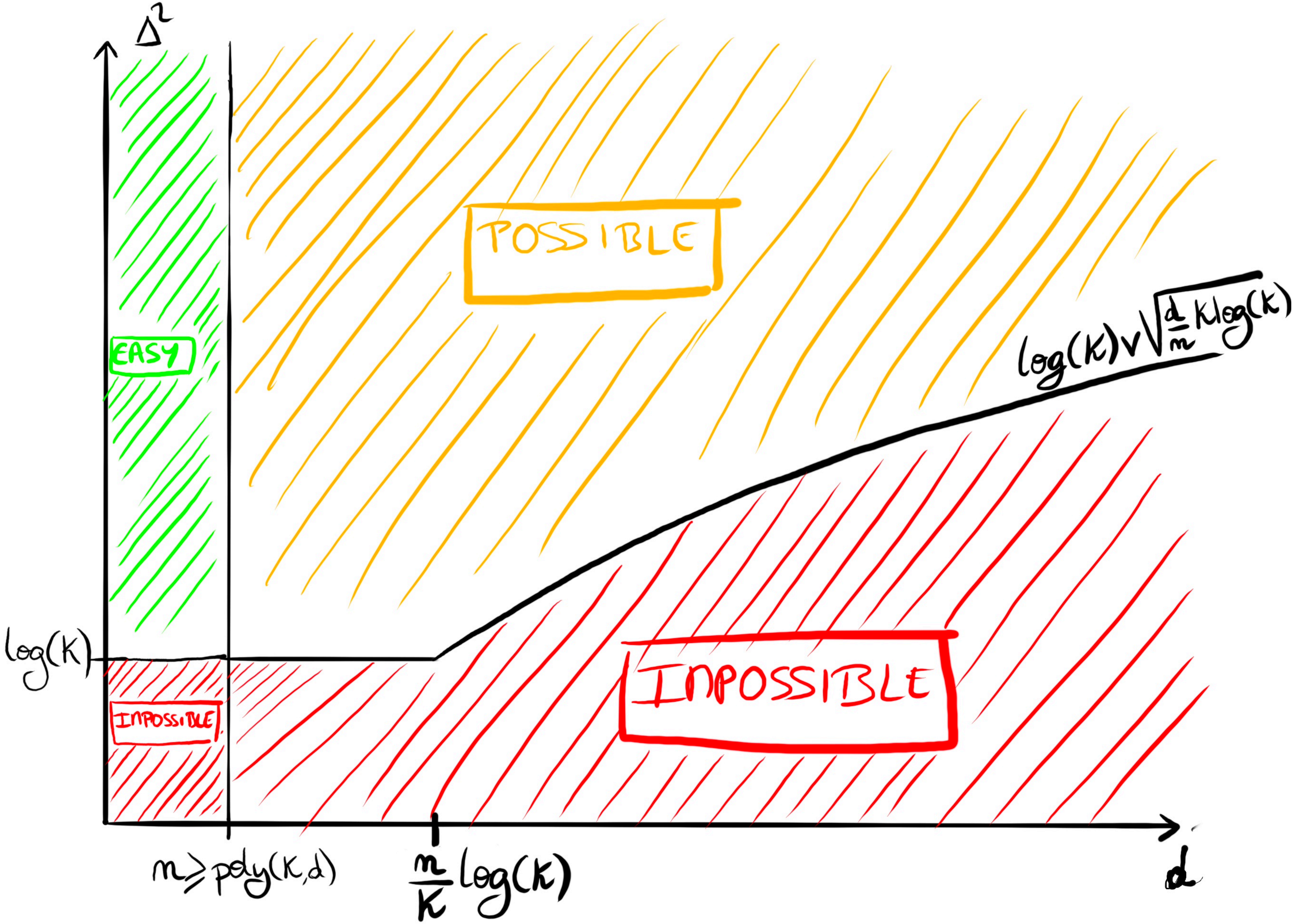

Δ^2
EASY
$\log(K)$
IMPOSSIBLE
$n \geq poly(K,d)$
d

Information threshold

Frontier for partial recovery: $\Delta^2 \asymp \log(K) + \sqrt{\frac{d}{n} K \log(K)}$.

And, the exact K-means estimator is optimal both for partial and perfect recovery.

Optimal estimator is **not computable in polynomial time.**



Δ^2
POSSIBLE
EASY
$\log(K) \vee \sqrt{\frac{d}{n} K \log(K)}$
$\log(K)$
IMPOSSIBLE
IMPOSSIBLE
$n \geq \mathrm{poly}(K,d)$
$\frac{n}{K}\log(K)$
d

Low degree Lower bound

Equivalent problem: estimate with LD polynomials $x = 1_{1 \overset{G^*}{\sim} 2}$

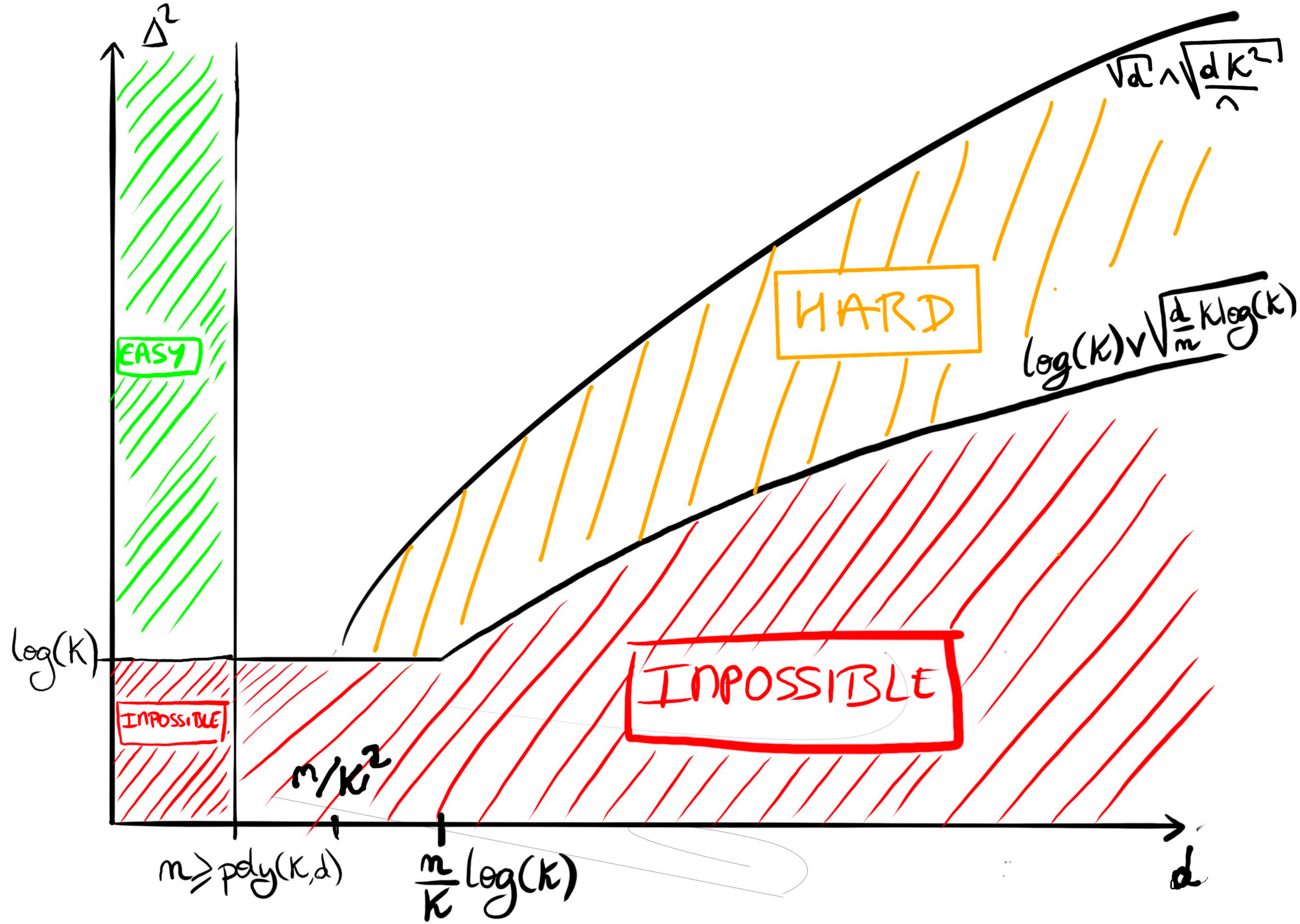
Hardness of clustering when

$$\Delta^2 \leq_{\log} \sqrt{d} \wedge \sqrt{\frac{dK^2}{n}} + 1$$

Deviation of the norm of a Gaussian

BBP threshold: eigenvalue/vectors of YY^T are informative.

Computational gap when K large and $d \gtrsim n/K^2$



Δ^2
$\sqrt{d} \wedge \sqrt{\frac{dK^2}{n}}$
HARD
$\log(K) \vee \sqrt{\frac{d}{n} K \log(K)}$
EASY
$\log(K)$
IMPOSSIBLE
IMPOSSIBLE
n/K^2
$n \geq \mathrm{poly}(K,d)$
$\frac{n}{K}\log(K)$
d

Is it possible to recover $G*$ in **polynomial time** above this threshold?

- If $\Delta^2 \geq_{\log} 1 + \sqrt{d}$, then **perfect recovery** is possible in **polynomial time** using a **Hierarchical clustering procedure.** (Optimal when $n \leq K^2$)

- In very high dimension $d \geq n$ and above the **BBP** threshold $\Delta^2 \gtrsim \sqrt{\dfrac{dK^2}{n}}$,

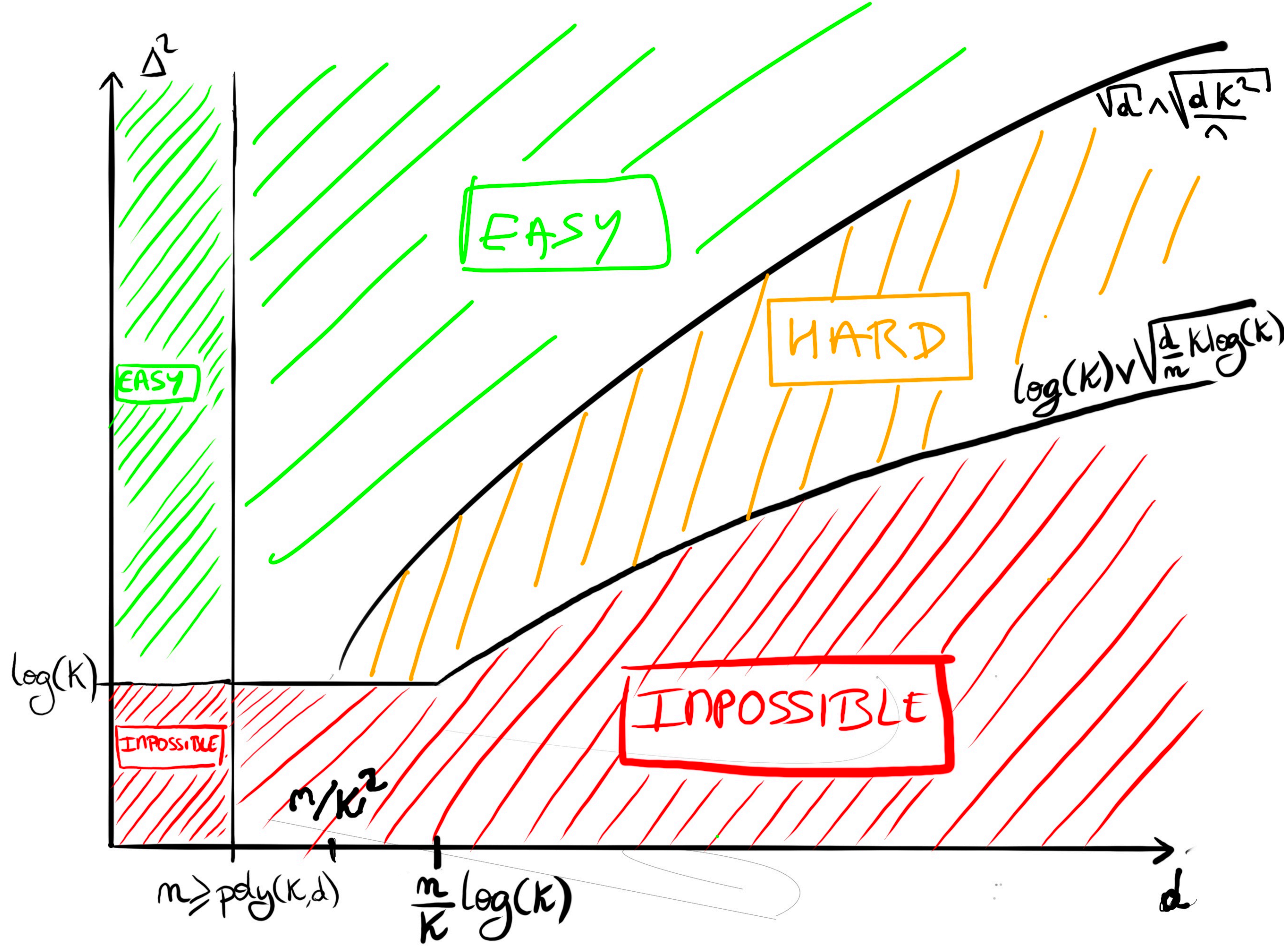
  partial recovery is possible in **polynomial time** using a **Convex Relaxation** of the exact K-means estimator. (Giraud Verzelen 2018)

What about when $n \geq K^2$ and $d \leq n$?

Spectral Projection Method

- Compute the K leading eigenvectors of YY^T.
- Project the dataset on those eigenvectors.
- Proceed with a low-dimensional clustering procedure (Liu/LI or Hierarchical clustering).

Excepts when $n \in [K^2, K^c]$ and $p \leq \dfrac{n}{K}$, we are able to recover exactly $G*$ with high probability when $\Delta^2 \geq_{\log} 1 + \sqrt{\dfrac{pK^2}{n}}$.



Δ²
EASY
EASY
HARD
√d ∧ √(dK²/n)
log(K) ∨ √((d/n) K log(K))
log(K)
IMPOSSIBLE
IMPOSSIBLE
n/K²
n ≥ poly(K,d)
(n/K) log(K)
d

Conclusion

- Almost full characterization of the regimes of clustering a Gaussian Mixture Model

- Exhibit the existence of **Computational gap** in **high dimension**

- Analysis of **LD lower-bounds** can also be done with additional sparsity assumption or for more complex structures such as biclustering