Fundamental Limits of Anonymous Statistical Inference : Privacy-Preserving Crowdsourcing

Master Oral Exam

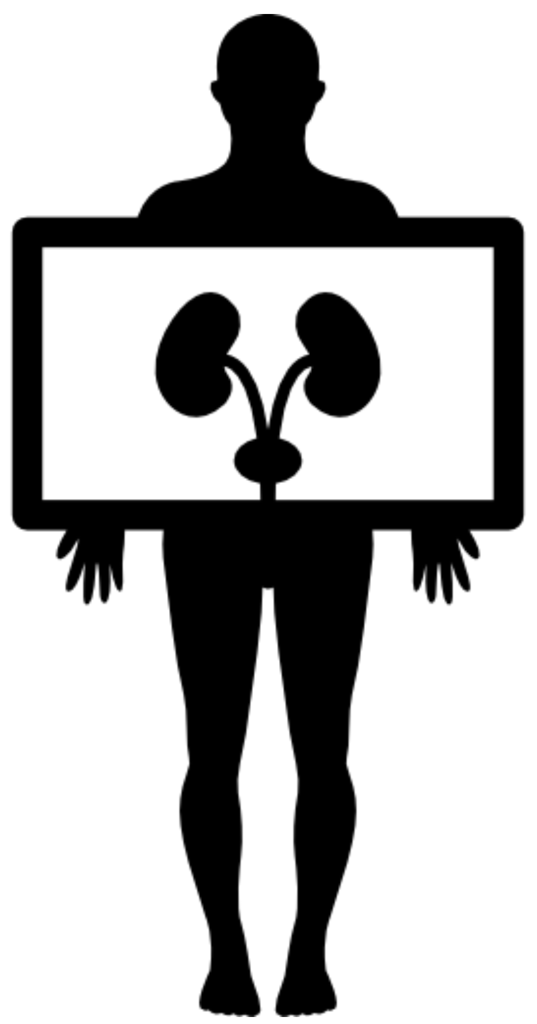
Wei-Ning Chen (wnchen@ntu.edu.tw)

Advisor : I-Hsiang Wang

National Taiwan University

Crowdsourcing Framework

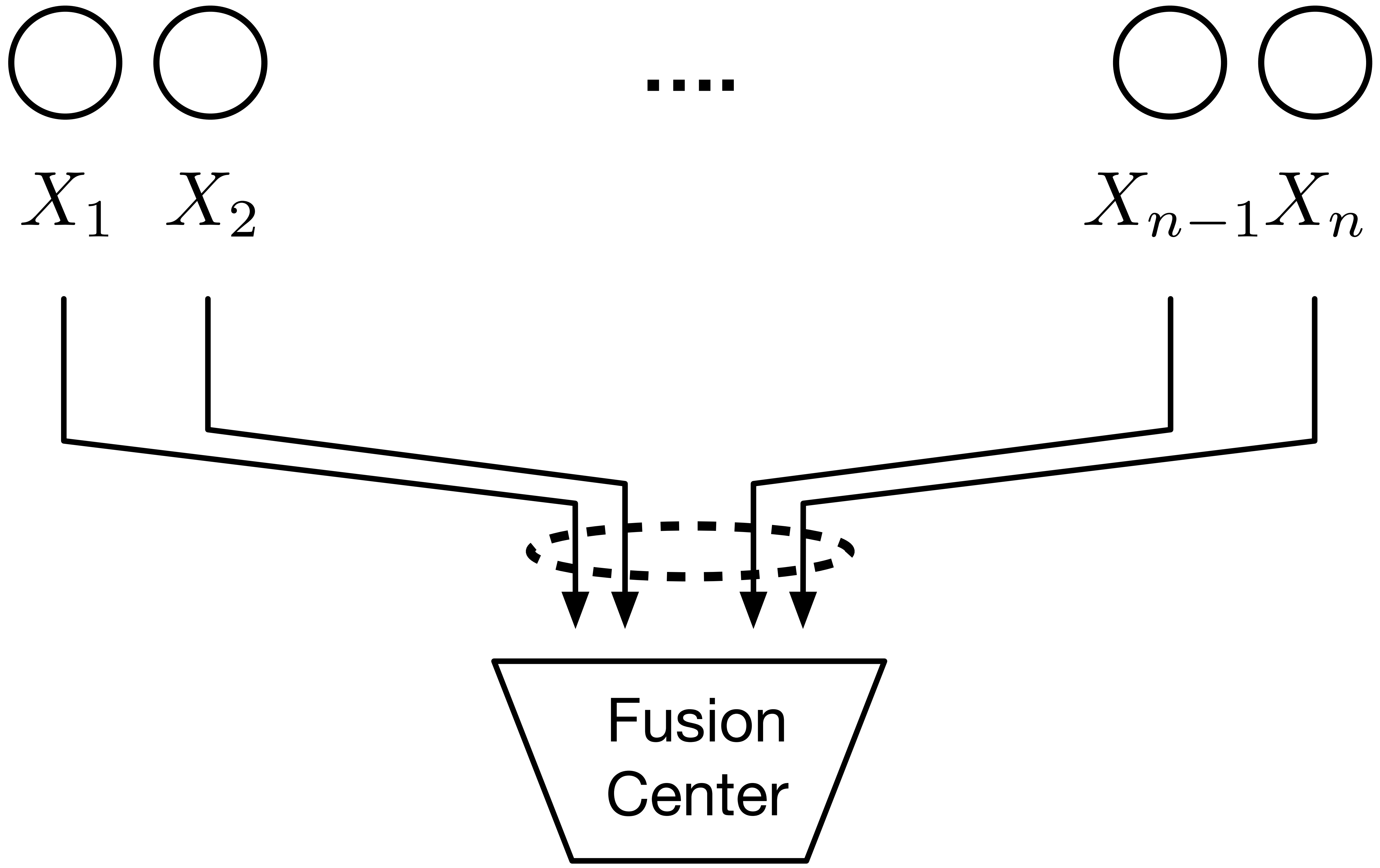
Tasks



$\mathcal{H}_0$: negative $\Rightarrow X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_0)$

$\mathcal{H}_1$: positive $\Rightarrow X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_1)$

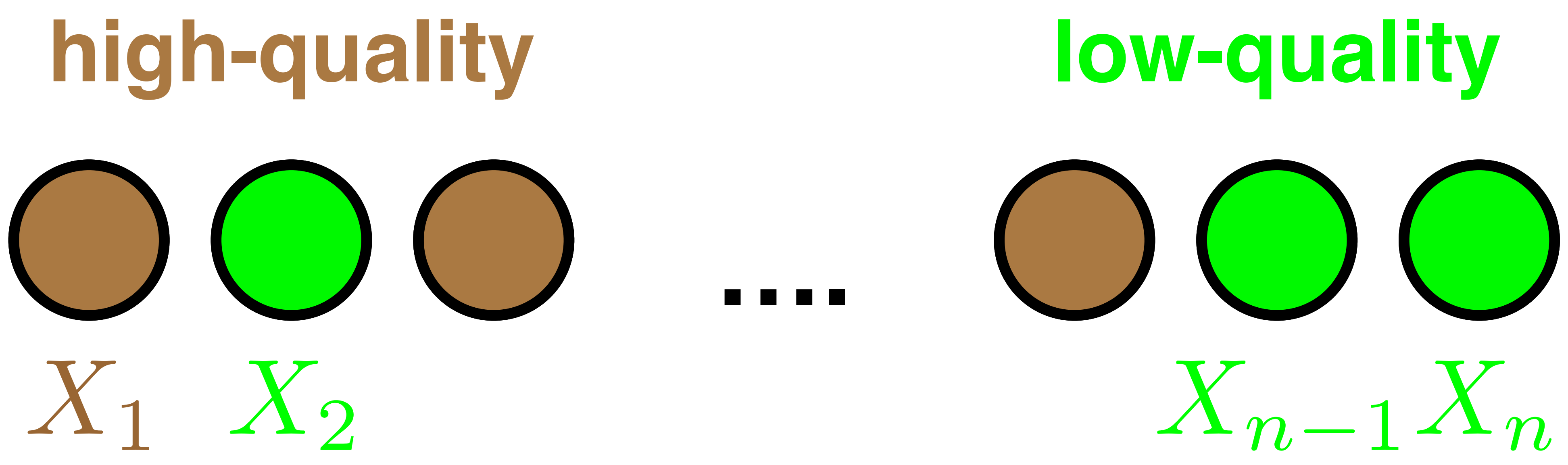
Workers



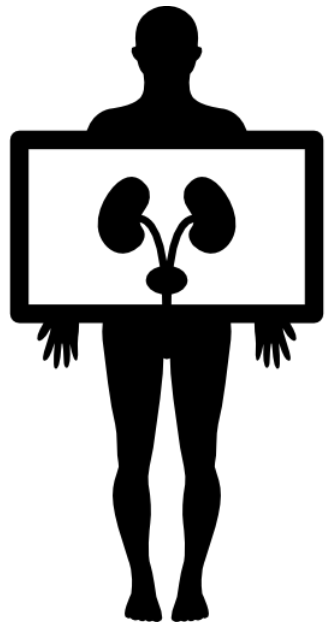
Goal : test the hypothesis

Heterogeneous Crowdsourcing

Workers



- Each worker has different ‘*ability/bias*’ [1]
  - *e.g. spammers or malicious workers*
  - can be grouped according to prior knowledge
- Answers no longer identically distributed



$$\mathcal{H}_0 : \text{negative} \Rightarrow X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_0)$$

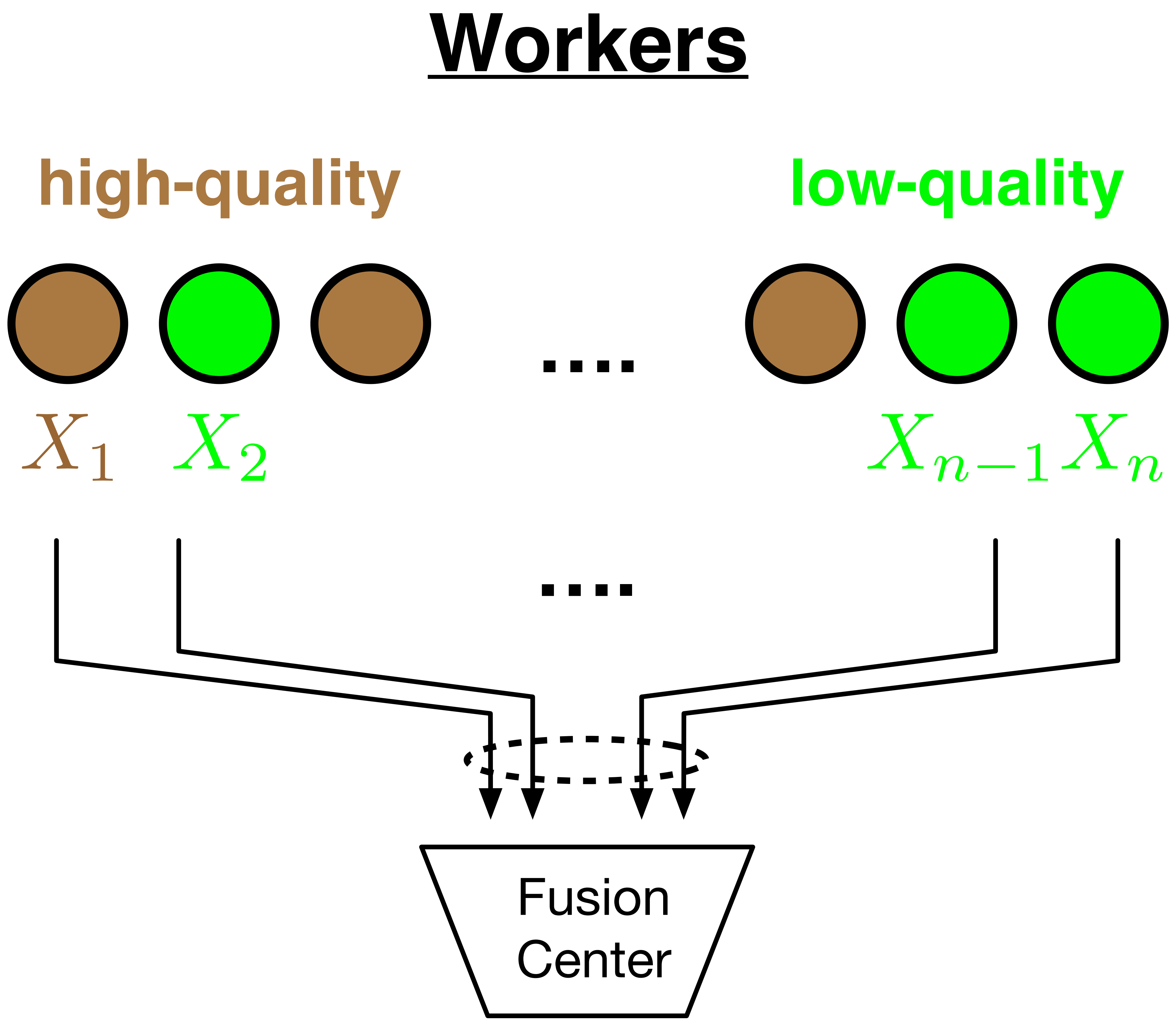
$$\mathcal{H}_1 : \text{positive} \Rightarrow X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_1)$$

[1] Panagiotis G. Ipeirotis, et. al *“Quality Management on Amazon Mechanical Turk,”* Proceedings of the ACM SIGKDD Workshop on Human Computation, 2010

Hardness : No Group Information

- Fusion center doesn't know the group each worker belongs to, due to
  - *Privacy*
  - *Identification cost*

- To address the anonymity issue, we propose
  - *<u>Using golden tasks to estimate the group info.</u>*
  - *<u>Testing the hypothesis anonymously</u>*



No group information available !

Organization

Part I : Group Recovery with Golden Tasks

- Mathematical Formulation and Previous Works
- Main Results : Converse, Achievability, and Impossibility Results
- Sketch of Proofs

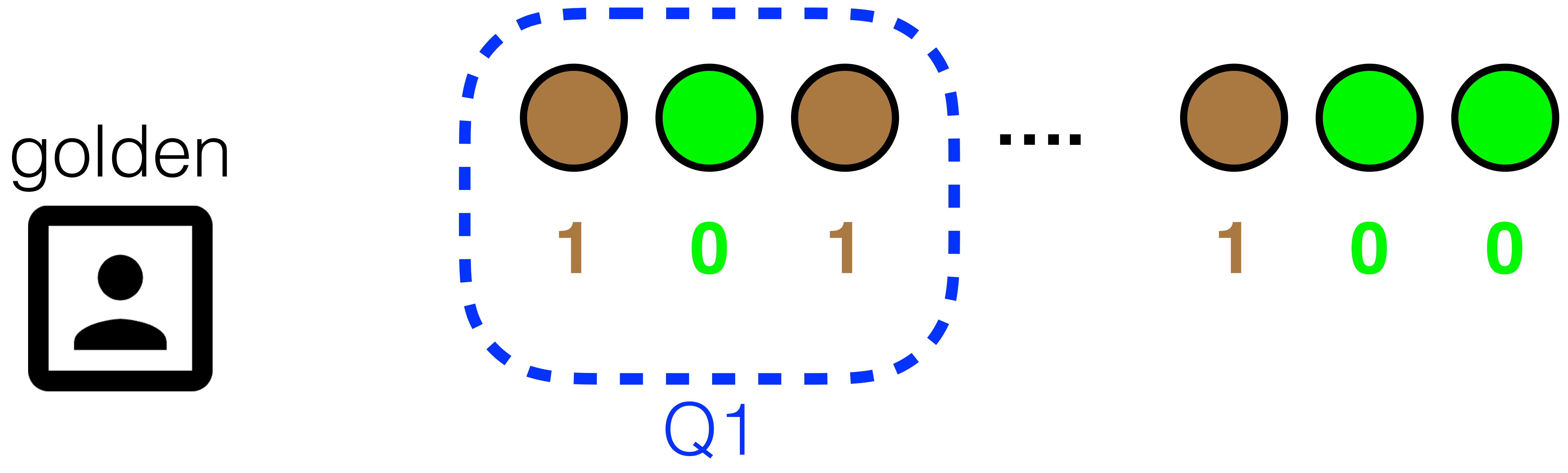
Part II : Anonymous Hypothesis Testing

- Formulation
- Main Results : Optimal Decision Rule and Asymptotic Behavior
- Sketch of Proofs
- Extensions

Part III : Conclusion and Future Directions

Part I : Group Recovery with Golden Tasks

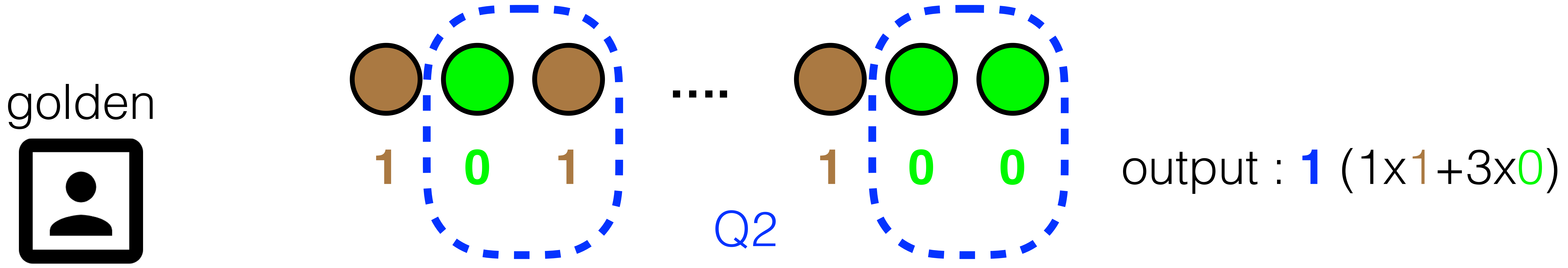
Golden Questions for Group Recovery



output : **2** (2x1+1x0)

- Assumptions on golden questions
  - Answers are (almost) *deterministic*
  - Workers from different groups (green/brown) respond different answers (**0**/**1**)
- Allowed to query the golden questions to a *subset of workers*
- Collect the *aggregation* of answers

Golden Questions for Group Recovery



- Assumptions on golden questions
  - Answers are (almost) *deterministic*
  - Workers from different groups (green/brown) respond different answers (**0**/**1**)
- Allowed to query the golden questions to a *subset of workers*
- Collect the *aggregation* of answers

Golden Questions for Group Recovery

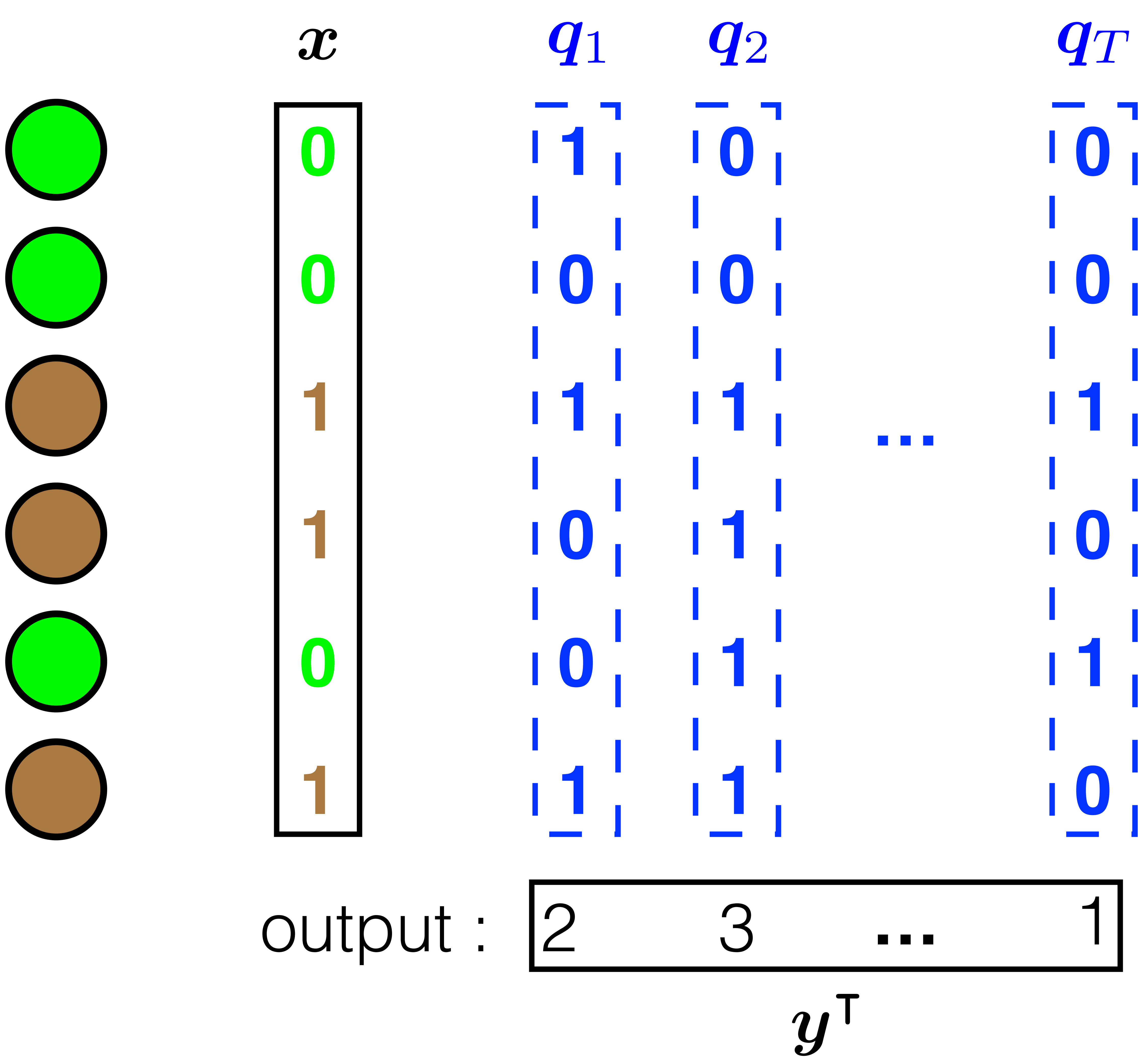
How many queries required to recover the group info. ?

- Workers from different groups respond different answers

- Allowed to query the golden questions to a *subset of workers*
- Collect the *aggregation* of answers

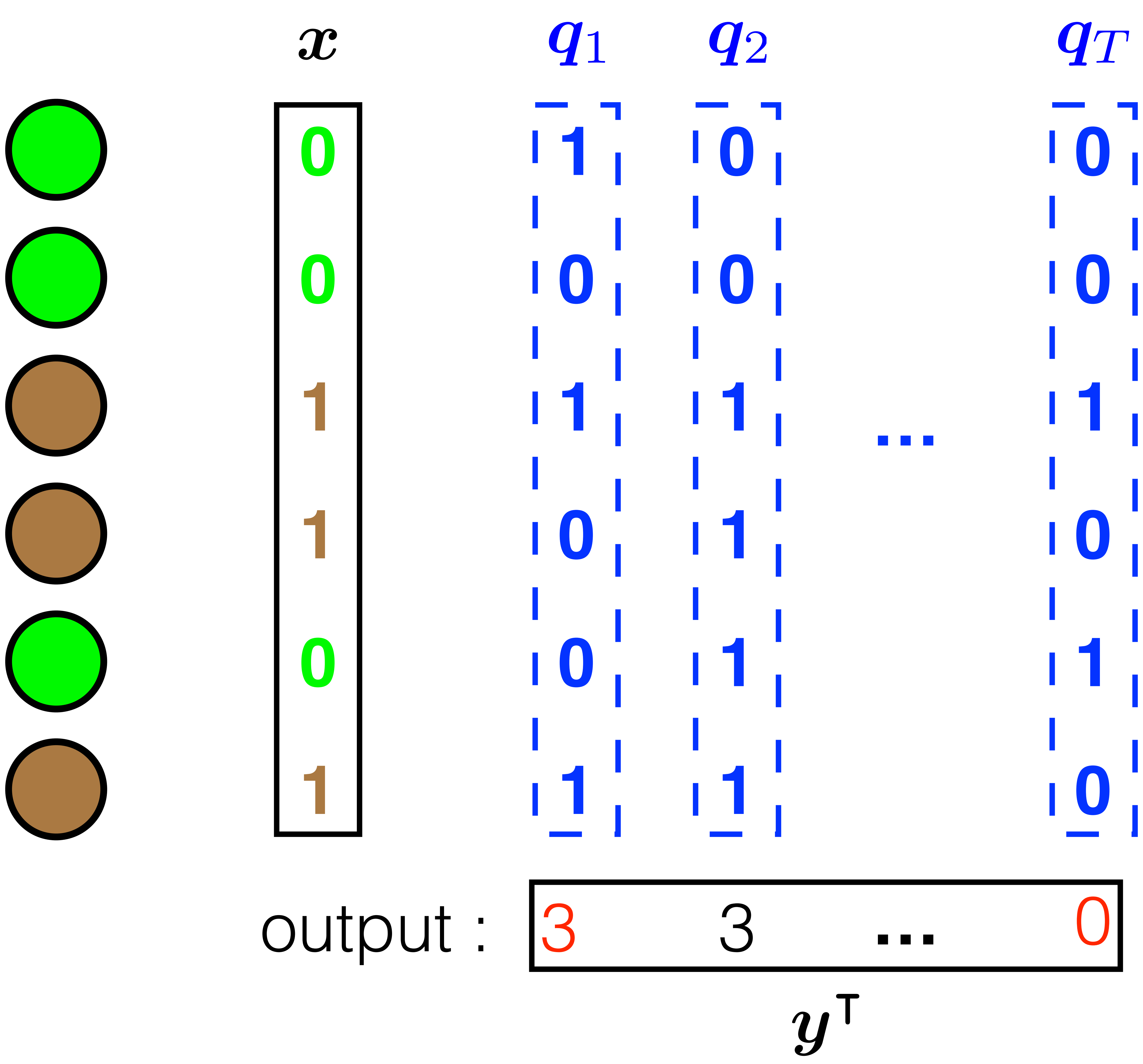
Golden Tasks for Group Recovery

Equivalent Linear Inverse Problem



noiseless : $$y = \begin{bmatrix} q_1^\intercal \\ q_2^\intercal \\ \vdots \\ q_T^\intercal \end{bmatrix} x \triangleq \mathbf{Q}x$$

Golden Tasks for Group Recovery



Equivalent Linear Inverse Problem

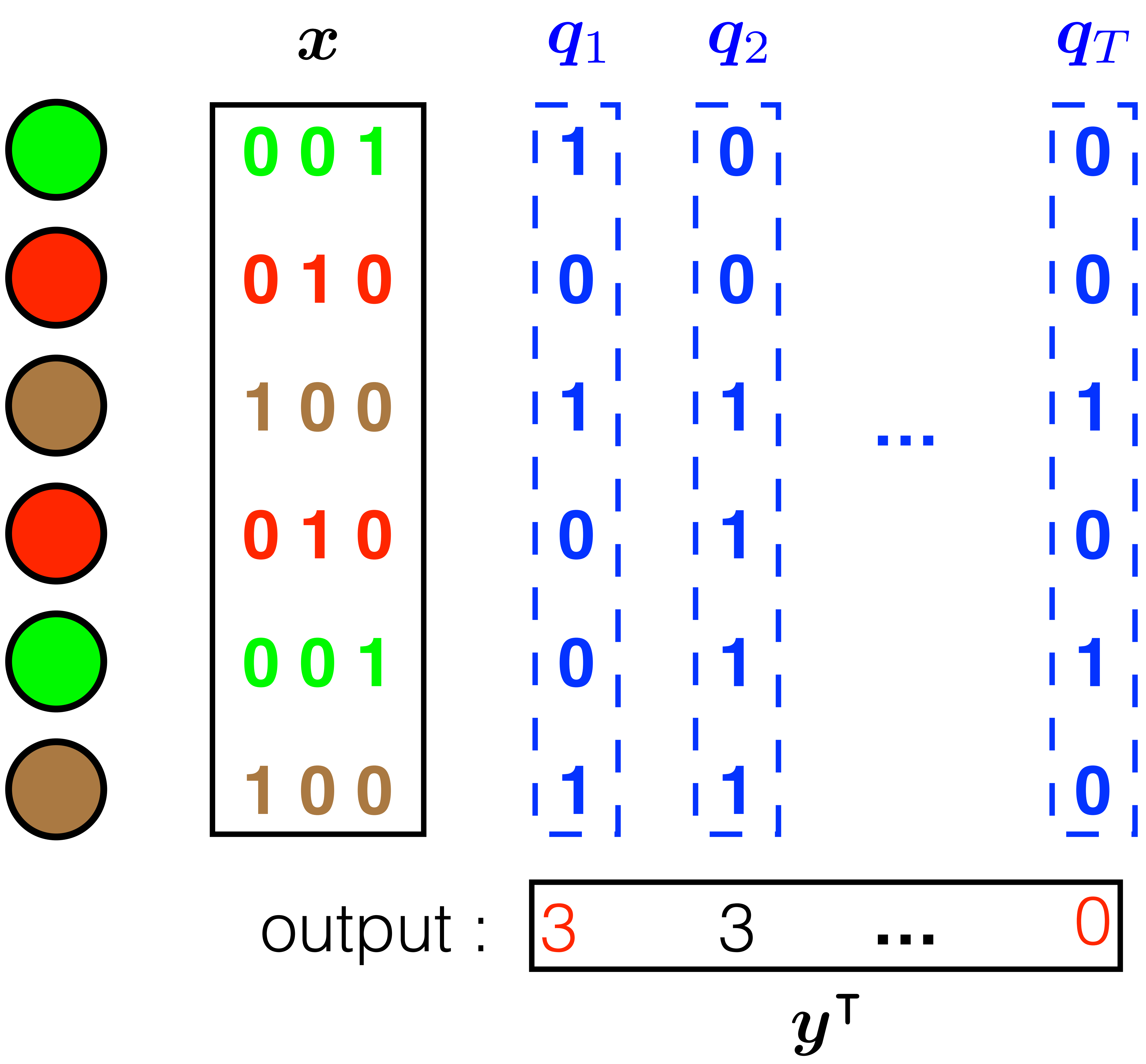
noiseless : $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \\ \vdots \\ \boldsymbol{q}_T^\top \end{bmatrix} \boldsymbol{x} \triangleq \mathbf{Q}\boldsymbol{x}$

noisy : $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{q}_1^\top \\ \boldsymbol{q}_2^\top \\ \vdots \\ \boldsymbol{q}_T^\top \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_T \end{bmatrix} \triangleq \mathbf{Q}\boldsymbol{x} + \boldsymbol{\Delta}$

assumption : $|\Delta_i| \leq \delta_n \; (\Leftrightarrow \|\boldsymbol{\Delta}\|_\infty \leq \delta_n)$

Golden Tasks for Group Recovery

recover column by column !



Equivalent Linear Inverse Problem

noiseless : $$y = \begin{bmatrix} q_1^\intercal \\ q_2^\intercal \\ \vdots \\ q_T^\intercal \end{bmatrix} x \triangleq \mathbf{Q}x$$

noisy : $$y = \begin{bmatrix} q_1^\intercal \\ q_2^\intercal \\ \vdots \\ q_T^\intercal \end{bmatrix} x + \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \vdots \\ \Delta_T \end{bmatrix} \triangleq \mathbf{Q}x + \mathbf{\Delta}$$

assumption : $|\Delta_i| \leq \delta_n \; (\Leftrightarrow \|\mathbf{\Delta}\|_\infty \leq \delta_n)$

Query Complexity

- Recovery criterion
  - Lossless recovery : $\hat{\boldsymbol{x}} = \boldsymbol{x}$
  - Lossy recovery with distortion : $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_1 \leq k_n$

- Query complexity $T^*(k_n, \delta_n)$: minimum # of queries required to recover

- Also known as *pooled data decoding*, *histogram query*, *coin weighing*, etc.
  - [2] specified the query complexity for noiseless query, lossless recovery : $T^* = \Theta\left(\frac{n}{\log n}\right)$
  - [3] studied the query complexity for k-sparse data : $T^* = \Theta\left(\frac{k}{\log k} \log\left(\frac{n}{k}\right)\right)$
  - [4,5] studied random noise, and proposed AMP decoding
  - Independently, [6,7] also suggested similar results, and studied erasure errors

[2] I.-H. Wang, et. al "*Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms*," ISIT, 2016
[3] I.-H. Wang, et. al "*Extracting Sparse Data via Histogram Queries,*" Allerton, 2016
[4] Ahmed El Alaoui, et. al "*Decoding from Pooled Data: Phase Transitions of Message Passing*," ISIT, 2017
[5] J. Scarlett, et. al "*Phase Transitions in the Pooled Data Problem,*" NIPS, 2017
[6] Nader H. Bshouty, et. al *"On the Coin Weighing Problem with the Presence of Noise"*
[7] Nader H. Bshouty, *"Optimal Algorithms for the Coin Weighing Problem with a Spring Scale,"* COLT, 2009

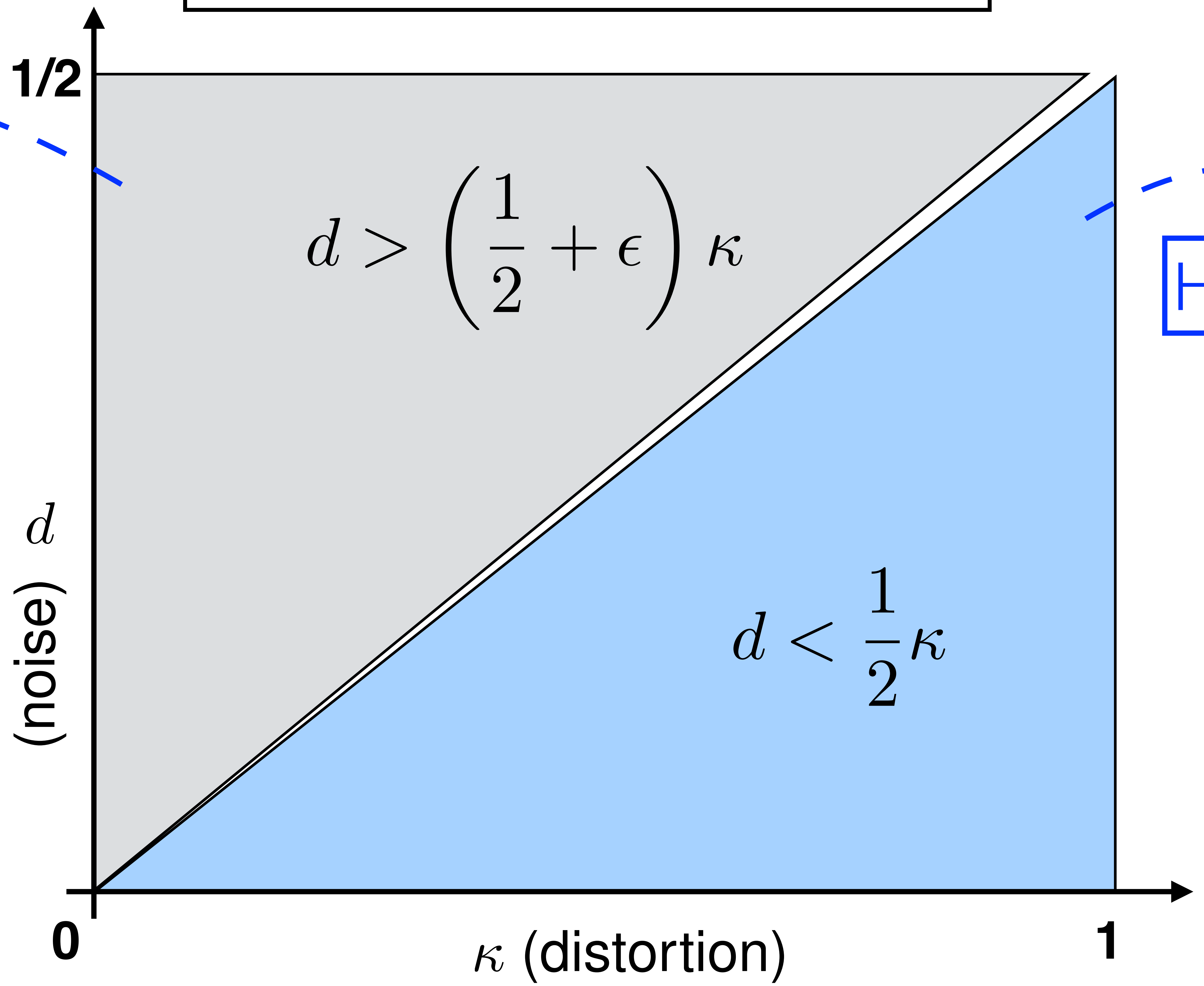
Main Results

$$\delta_n = \Theta\left(n^d\right), \; k_n = \Theta\left(n^\kappa\right)$$

Low SNR regime

$$T^* = \Omega\left(\exp\left(n^\epsilon\right)\right)$$

non-polynomial !



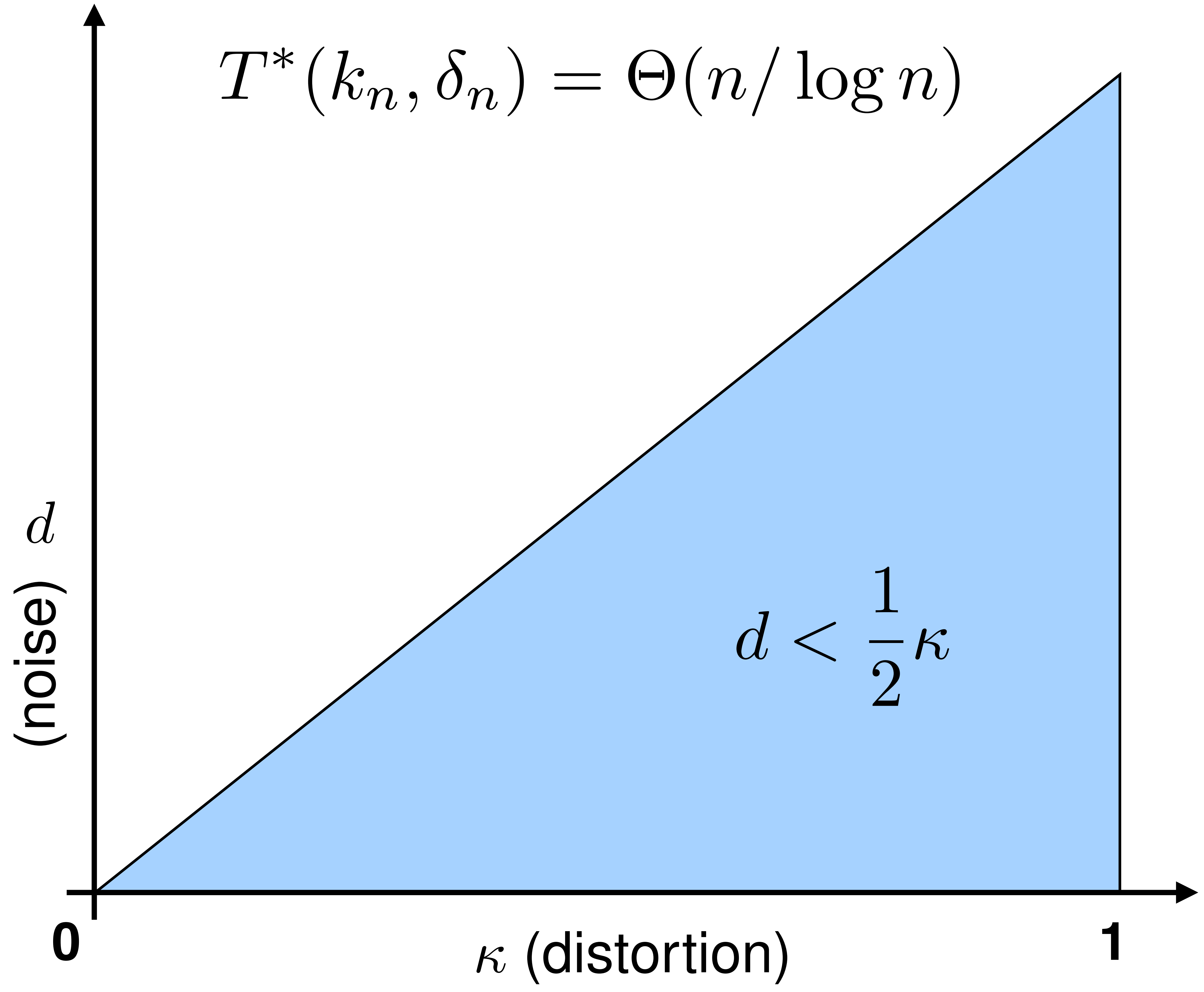
High SNR regime

$$T^* = \Theta\left(\frac{n}{\log n}\right)$$

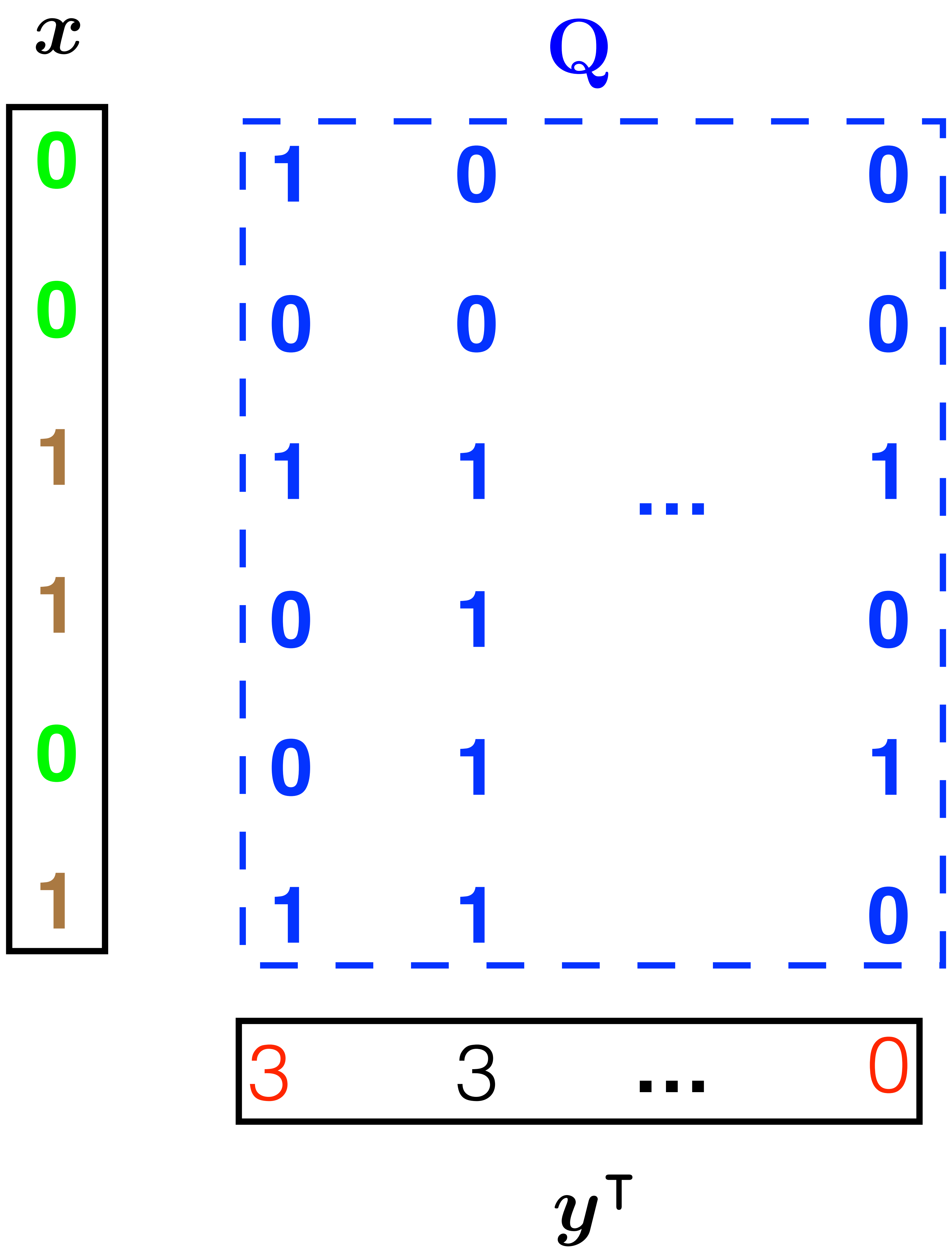
sub-linear !

Regime I : High SNR

High SNR regime



Regime I : Achievability



- Random sampling
  - $(\mathbf{Q})_{i,j} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2)$
- Probability of failure

$$P_f(\boldsymbol{x}; k_n, \delta_n) \triangleq P\{\exists \text{ another consistent } \tilde{\boldsymbol{x}}\}$$

- If # queries is $\Omega(n/\log n)$, then

$$P_f(\boldsymbol{x}; k_n, \delta_n) \to 0$$

  - Apply Chernoff's bound for the failure event

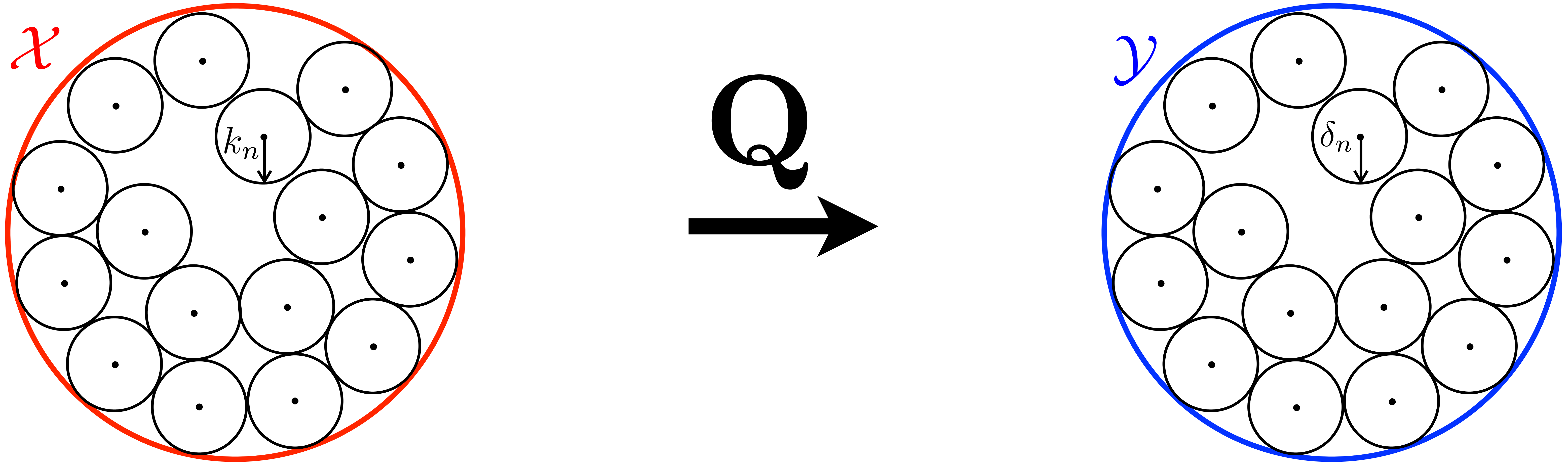
Regime I : Converse

- Necessary condition :

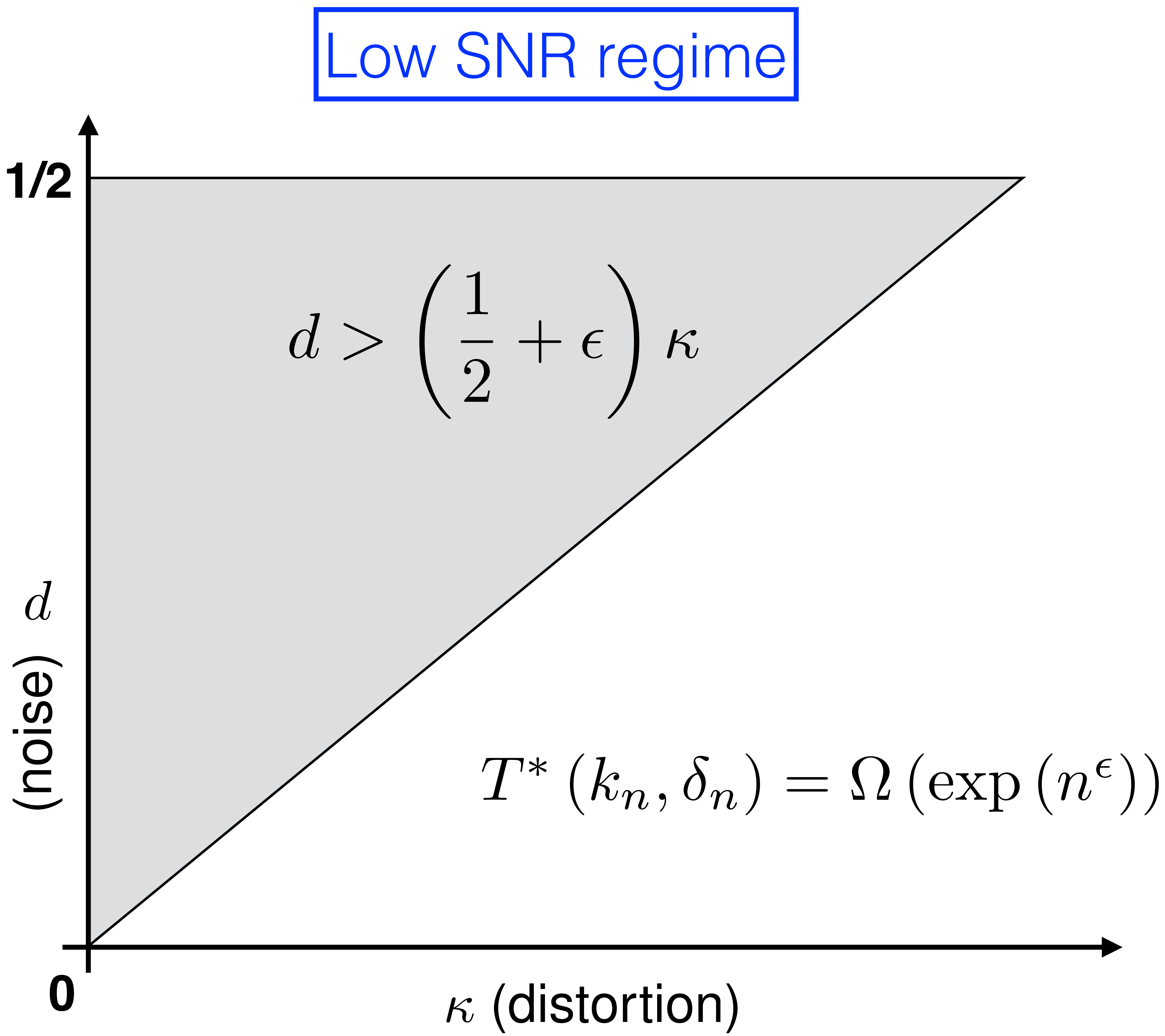
$$\forall \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X},\ \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 > k_n \implies \|\mathbf{Q}\boldsymbol{x} - \mathbf{Q}\tilde{\boldsymbol{x}}\|_\infty > 2\delta_n$$

- Packing inequality

$$2\delta_n\text{-packing number on } \mathcal{Y} \geq \tfrac{1}{2}k_n\text{-packing number on } \mathcal{X}$$



Regime II : Low SNR



Regime II : Impossibility of Polynomial Queries

Idea : without sufficient queries, $\exists$ more than one $\boldsymbol{x}$ consistent with the response $\boldsymbol{y}$

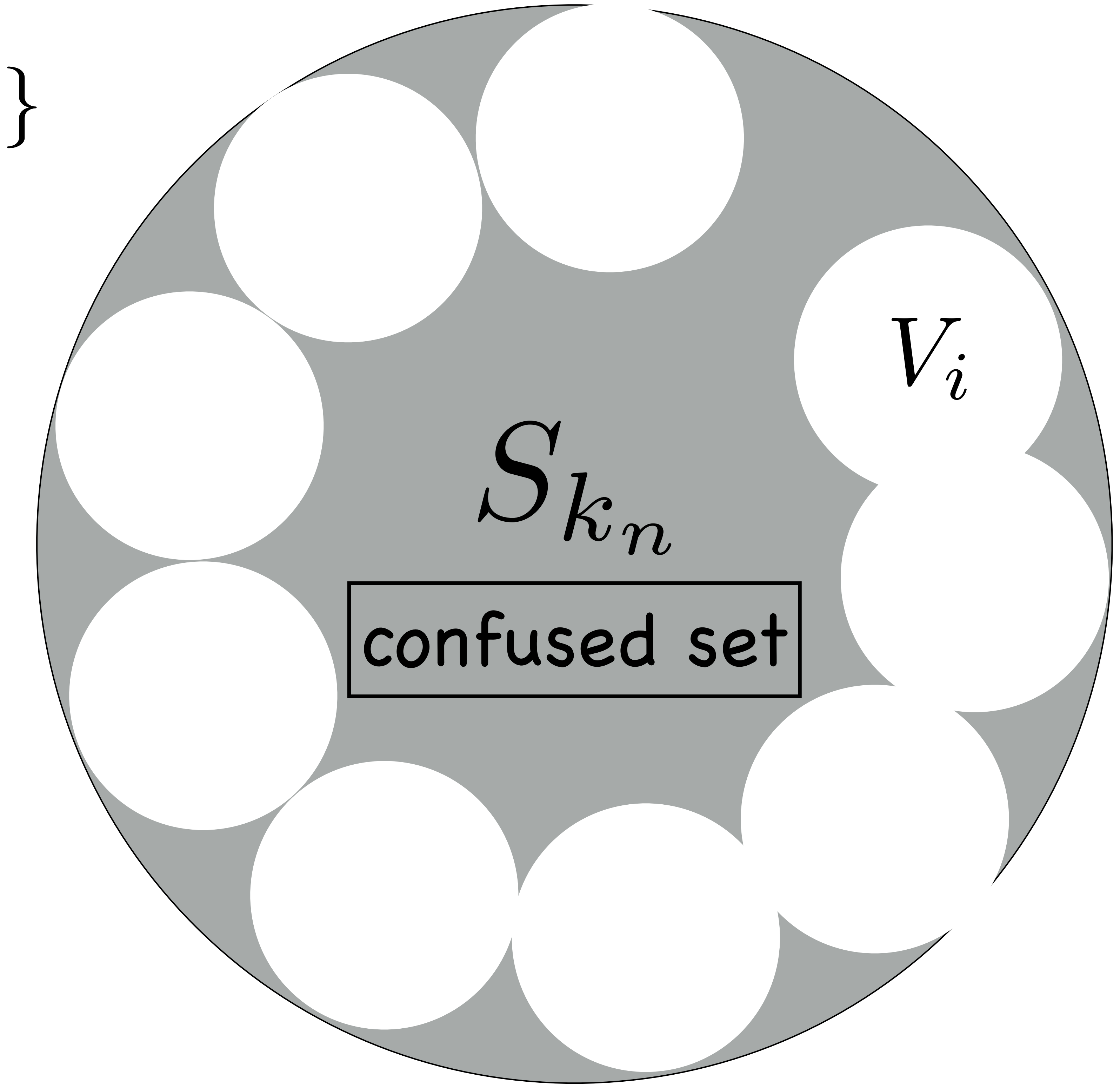
1. *Initial : consider all possible pairs*

$$S_{k_n} \triangleq \{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mid \boldsymbol{x}, \tilde{\boldsymbol{x}} \in \{0,1\}^n,\ \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_1 = k_n,\ \|\boldsymbol{x}\|_1 = \|\tilde{\boldsymbol{x}}\|_1\}$$

2. *After each query : remove inconsistent pairs*

$$V_i \triangleq \{(\boldsymbol{x}, \tilde{\boldsymbol{x}}) \in S_{k_n} \mid |\boldsymbol{q}_i^\intercal (\boldsymbol{x} - \tilde{\boldsymbol{x}})| > \delta_n\}$$

3. *Until : no more confused pair*



at least $\dfrac{|S_{k_n}|}{\max_i |V_i|}$ queries required

Regime II : Impossibility of Polynomial Queries

- Lower bound on query complexity

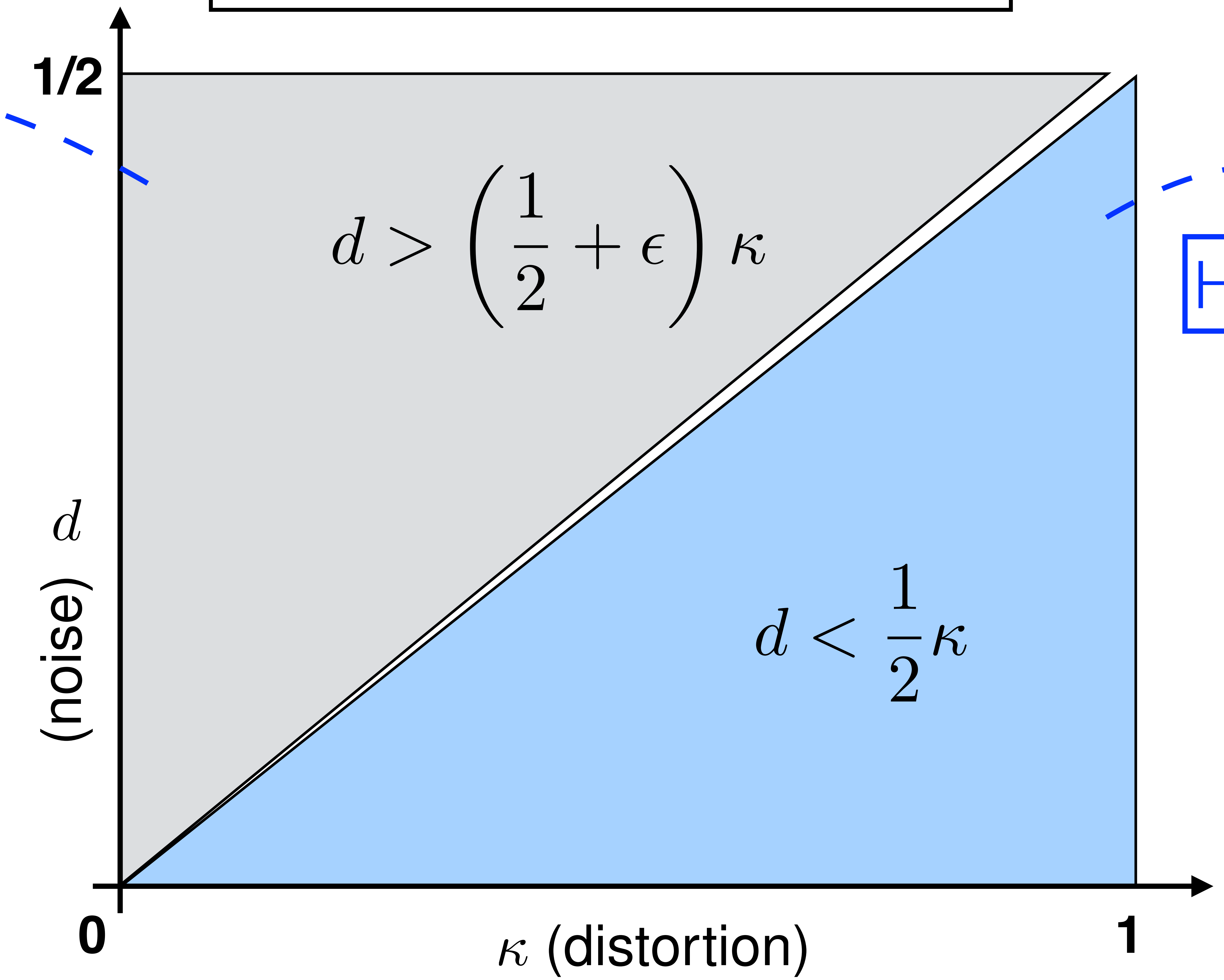
$$T^*(k_n, \delta_n) \geq \frac{|S_{k_n}|}{\max_{i \in \{1,2,\ldots,T\}} |V_i|}$$

$$\geq C \exp\left(\frac{\delta_n^2}{k_n}\right) = C \exp\left(n^{2d-\kappa}\right)$$

solving the optimization over V, and apply Chernoff ineq

Summary

$$\delta_n = \Theta\left(n^d\right),\ k_n = \Theta\left(n^\kappa\right)$$



Low SNR regime

$$T^* = \Omega\left(\exp\left(n^\epsilon\right)\right)$$

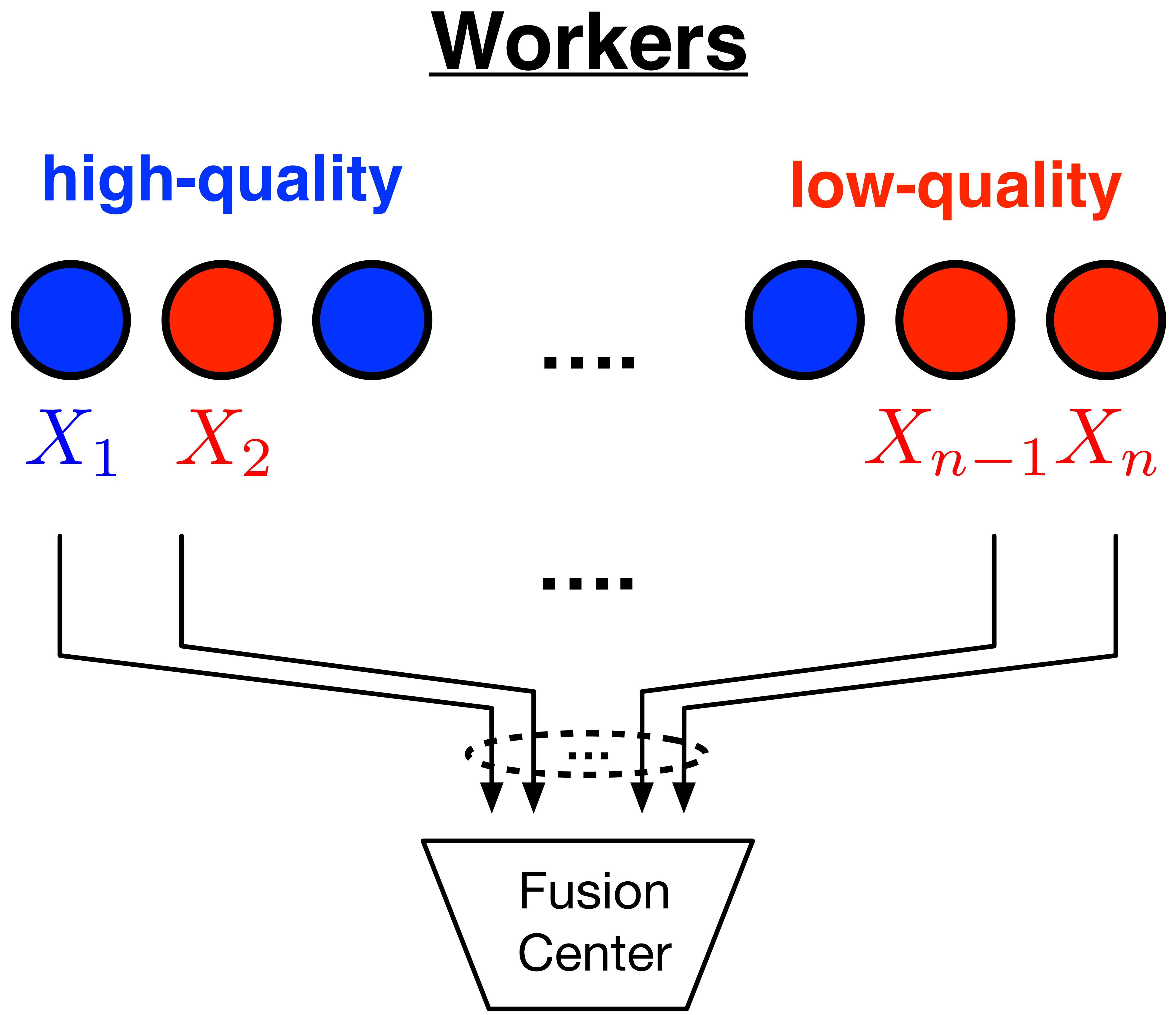
High SNR regime

$$T^* = \Theta\left(\frac{n}{\log n}\right)$$

Part II : Anonymous Hypothesis Testing

Test Hypothesis Anonymously

- Sometimes we don't need the group info.
  - *e.g. the homogeneous setting*
- Goal: deign a good decision rule for *all possible scenario*
- Quantify *price of anonymity*



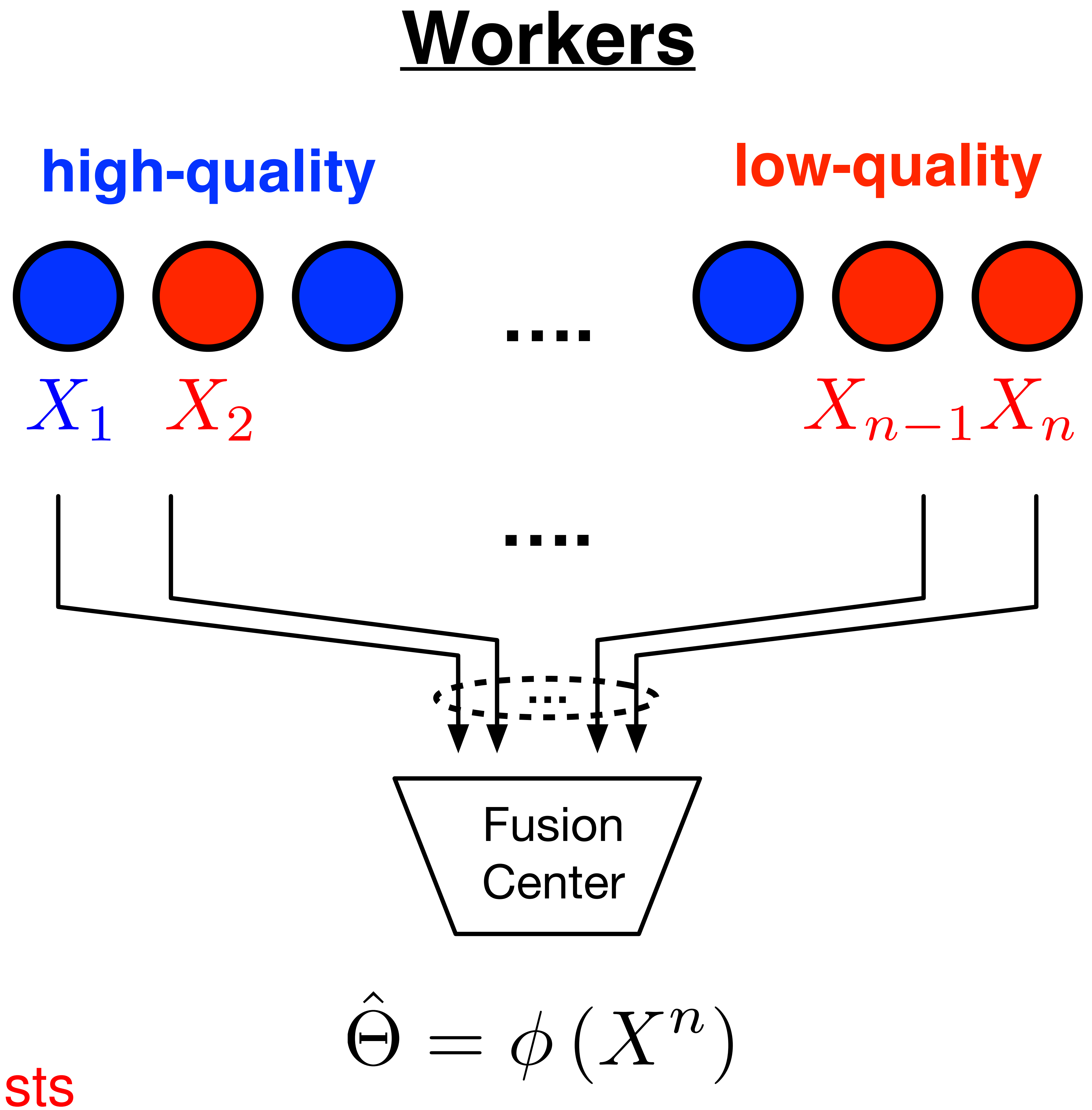
No group information available !

Heterogeneous Distributed Detection

- Heterogeneity: K group of workers
  - Workers in group $\mathcal{I}_k$ follows distribution $P_{\theta;k}$

$$X_i \overset{\text{i.i.d.}}{\sim} P_{\theta;k}, \text{ for } i \in \mathcal{I}_k$$

  - The k-th group has $n\alpha_k$ workers, $\sum_{k=1}^{K} \alpha_k = 1$
- Neyman-Pearson setting: $\theta \in \{0, 1\}$
  - Minimize Type-II error prob. while keeping type-I error prob. small ($\leq \epsilon$)
  - Minimum Type-II error probability: $\beta^{(n)}(\epsilon, \alpha_1, ..., \alpha_K)$
  - Error exponent: $E(\epsilon, \boldsymbol{\alpha}) \triangleq \lim_{n\to\infty} \left\{-\frac{1}{n}\log_2 \beta^{(n)}(\epsilon, \boldsymbol{\alpha})\right\}$, if it exists



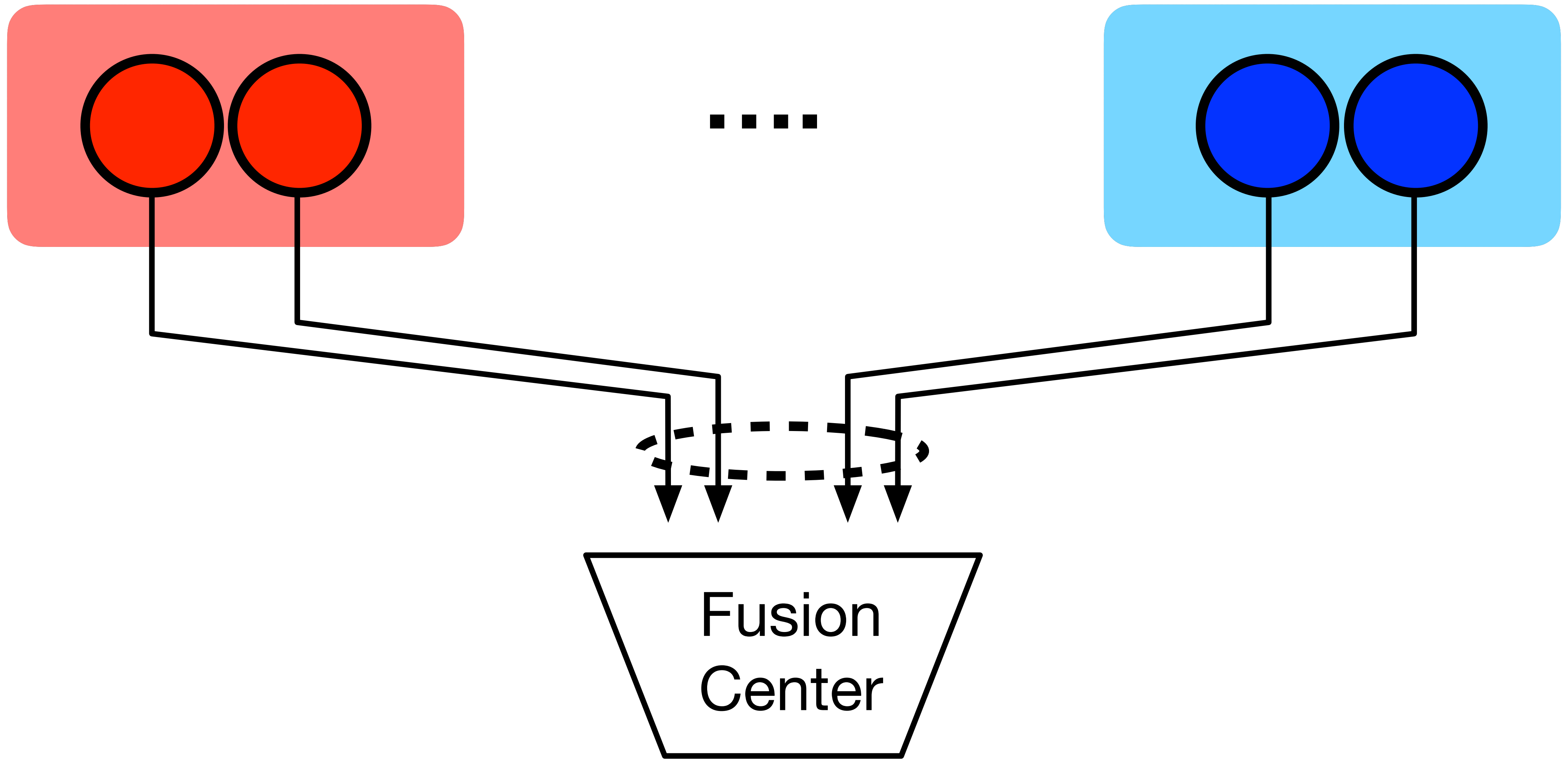
Effect of Heterogeneity without Anonymity

Example: Two Group (K=2)

Group 1:

$X_i \overset{\text{i.i.d.}}{\sim} P_\theta$

$n(1-\alpha)$ workers



Group 2:

$X_i \overset{\text{i.i.d.}}{\sim} Q_\theta$

$n\alpha$ workers

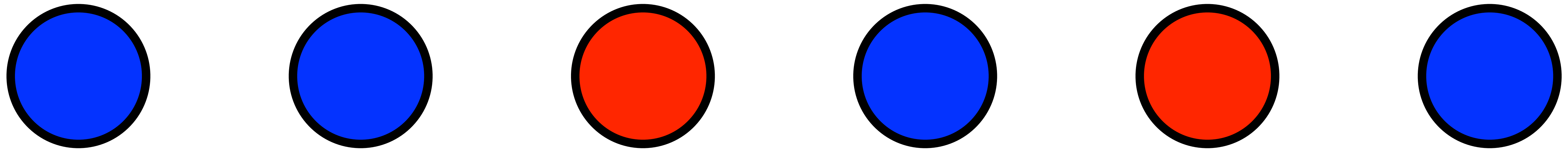
When FC is informed of the group that each worker belongs to:

$$\Rightarrow E_{\text{informed}}(\epsilon, \alpha) = (1-\alpha)D\left(P_0 \| P_1\right) + \alpha D\left(Q_0 \| Q_1\right)$$

weighted combination of ‘resolvability’ of different groups!

Composite Hypothesis Testing

- Not sure about which group each worker belongs to?
  $\Rightarrow$ design algo. with performance guarantee **for all possible scenarios**



worker ID i	1	2	3	4	5	6
group assignment $\sigma(i)$	2	2	1	2	1	2

- Formally speaking:

$$\begin{cases} \mathcal{H}_0 : X^n \sim \mathbb{P}_{0;\sigma} \triangleq \prod_{i=1}^n P_{0;\sigma(i)}, \text{ for some } \sigma \\ \mathcal{H}_1 : X^n \sim \mathbb{P}_{1;\sigma} \triangleq \prod_{i=1}^n P_{1;\sigma(i)}, \text{ for some } \sigma \end{cases}$$

$$\sigma : [n] \to [K], \; s.t. \; |\{i : \sigma(i) = k\}| = n\alpha_k$$

$$\boldsymbol{P}_\theta \triangleq \begin{bmatrix} P_{\theta;1} \\ P_{\theta;2} \\ \vdots \\ P_{\theta;K} \end{bmatrix}$$

group distributions

Composite Hypothesis Testing

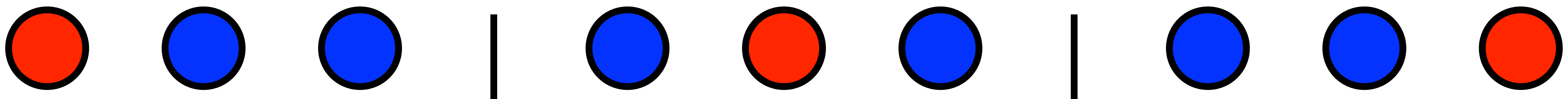
$$\begin{cases} \mathcal{H}_0 : X^n \sim \mathbb{P}_{0;\sigma} \triangleq \prod_{i=1}^n P_{0;\sigma(i)}, & \text{for some } \sigma \\ \mathcal{H}_1 : X^n \sim \mathbb{P}_{1;\sigma} \triangleq \prod_{i=1}^n P_{1;\sigma(i)}, & \text{for some } \sigma \end{cases}$$

$\sigma : [n] \to [k]$, s.t. $|\{i|\sigma(i) = k\}| = n\alpha_k$

- Example: $K = 2$, $\boldsymbol{\alpha} = (\frac{1}{3}, \frac{2}{3})$ (red : blue$= 1 : 2$)

Red: $X_i \overset{\text{i.i.d.}}{\sim} P_{\theta;1}$

Blue: $X_i \overset{\text{i.i.d.}}{\sim} P_{\theta;2}$



σ	red blue blue	blue red blue	blue blue red
possible dist. under $\mathcal{H}_0$	$P_{0;1}P_{0;2}P_{0;2}$	$P_{0;2}P_{0;1}P_{0;2}$	$P_{0;2}P_{0;2}P_{0;1}$
possible dist. under $\mathcal{H}_1$	$P_{1;1}P_{1;2}P_{1;2}$	$P_{1;2}P_{1;1}P_{1;2}$	$P_{1;2}P_{1;2}P_{1;1}$

Minimax Neyman-Pearson Formulation

- Probability of errors:

$$\mathsf{P_F}^{(n)}(\phi) \triangleq \max_{\sigma} \mathbb{P}_{0;\sigma}\{\phi(X^n)=1\} \text{ (the worst case Type-I error probability)}$$

$$\mathsf{P_M}^{(n)}(\phi) \triangleq \max_{\sigma} \mathbb{P}_{1;\sigma}\{\phi(X^n)=0\} \text{ (the worst case Type-II error probability)}$$

- Neyman-Pearson Regime :

$$\beta^{(n)}(\epsilon, \boldsymbol{\alpha}) \triangleq \min_{\phi} \mathsf{P_M}^{(n)}(\phi)$$

$$\text{s.t. } \mathsf{P_F}^{(n)}(\phi) < \epsilon$$

Huber[1973], Kuznetsov[1982], Veeravalli [1994], etc.

- Type-II error exponent:

$$E(\epsilon, \boldsymbol{\alpha}) \triangleq \lim_{n\to\infty} \left\{ -\frac{1}{n} \log_2 \beta^{(n)}(\epsilon, \boldsymbol{\alpha}) \right\}$$

Main Contribution : Optimal Test

- An intuitive test : first *estimate the group assignment* σ, then do LRT

  $\Rightarrow$ Generalized likelihood ratio test

$$\phi(x^n) = \begin{cases} 1, & \text{if } \ell(x^n) < \tau \\ \gamma, & \text{if } \ell(x^n) = \tau \\ 0, & \text{if } \ell(x^n) > \tau \end{cases}$$

is this optimal ?

$$\ell_{\text{GLRT}}(x^n) \triangleq \frac{\sup_\sigma \mathbb{P}_{0;\sigma}(x^n)}{\sup_\sigma \mathbb{P}_{1;\sigma}(x^n)}$$

- Optimal Decision Rule :

$$\ell(x^n) \triangleq \frac{\sum_\sigma \mathbb{P}_{0;\sigma}(x^n)}{\sum_\sigma \mathbb{P}_{1;\sigma}(x^n)}$$

mixture likelihood ratio test

likelihood ratio between uniform mixture under $\mathcal{H}_0$ to $\mathcal{H}_1$

Main Contribution : Type-II Error Exponent

- A generalized 'divergence' :

$$D_{\boldsymbol{\alpha}}\left(\boldsymbol{P};\boldsymbol{Q}\right) \triangleq \min_{\boldsymbol{U}\in(\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^{K} \alpha_k D\left(U_k \,\|\, Q_k\right)$$

$$\text{s.t.} \quad \boldsymbol{\alpha}^\intercal \boldsymbol{U} = \boldsymbol{\alpha}^\intercal \boldsymbol{P}$$

  - ► Plays a similar role as KL divergence in simple hypothesis testing

recall

$$\boldsymbol{P} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix}$$

- Type-II error exponent :

$$E(\epsilon, \boldsymbol{\alpha}) = D_{\boldsymbol{\alpha}}\left(\boldsymbol{P};\boldsymbol{Q}\right)$$

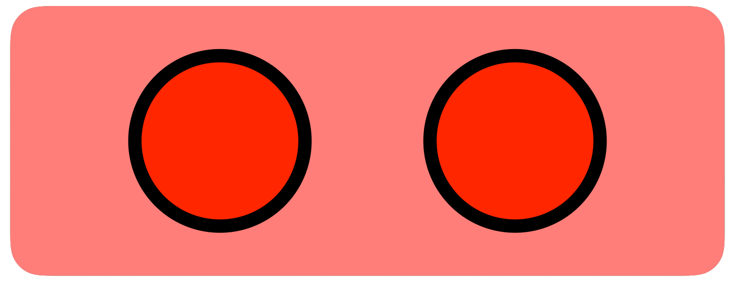
  - ► Independent of ϵ, convex in $\boldsymbol{\alpha}$

- Compared to informed case :

$$E_{\text{informed}}(\epsilon, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \alpha_k D\left(P_{0;k} \,\|\, P_{1;k}\right)$$

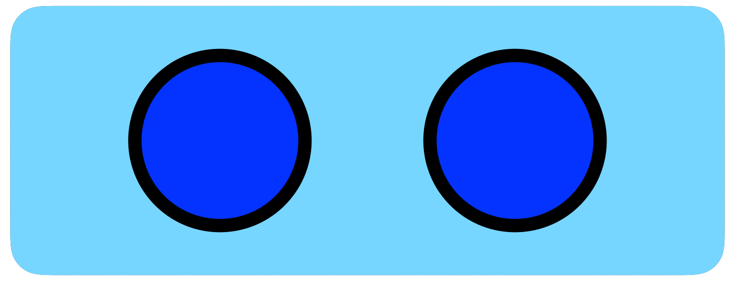
Main Contribution : Type-II Error Exponent

Example (K=2)



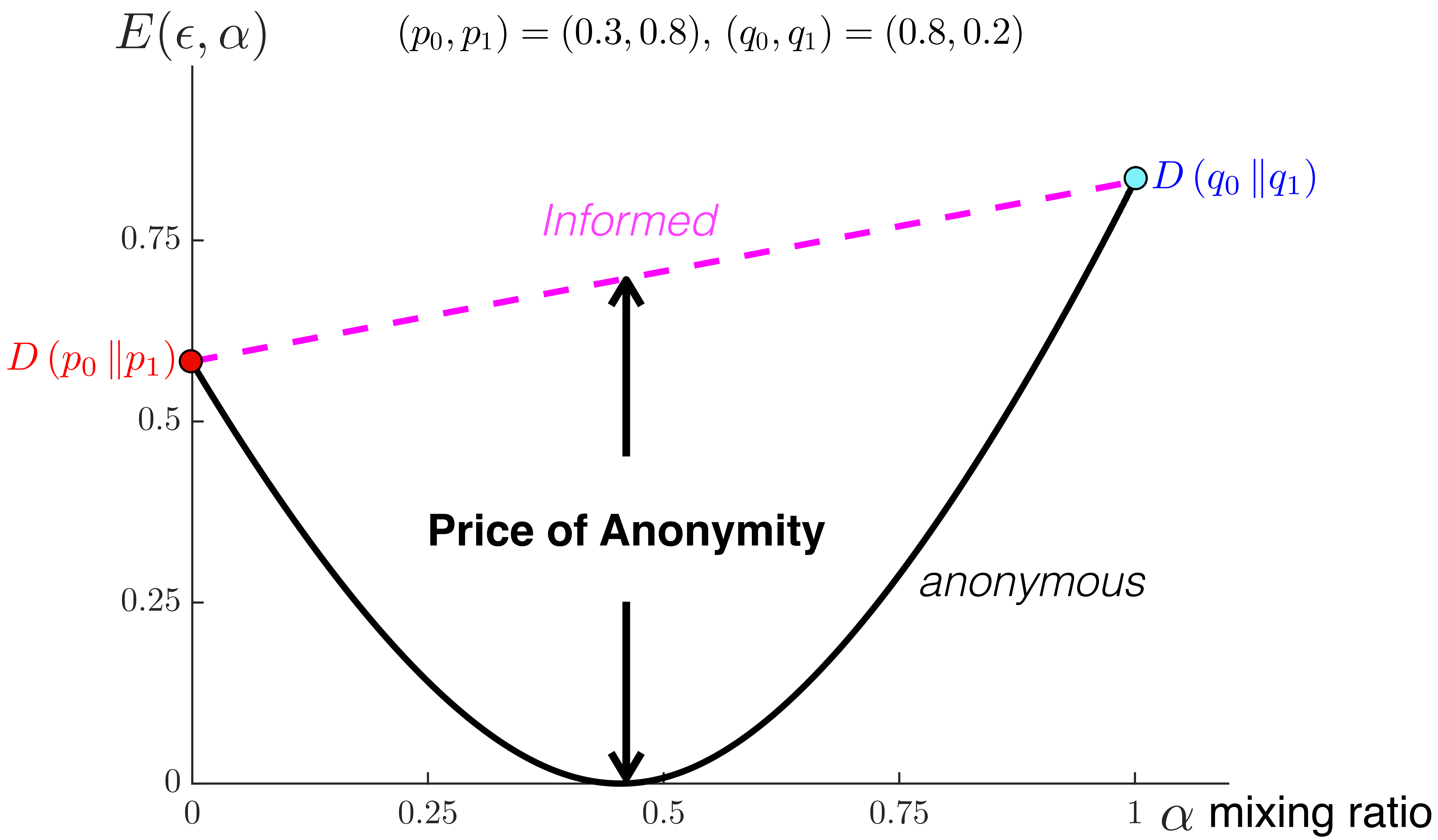
$n(1-\alpha)$ workers

$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p_\theta)$



$n\alpha$ workers

$X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(q_\theta)$



Sketch of Proof : Optimal Test

- Idea :

  1) *'Symmetric test'* (tests depend only on the empirical distribution of x^n) is the best

  2) Among all symmetric tests, the *mixture likelihood ratio test (MLRT)* is optimal

Sketch of Proof : Optimal Test

- Idea :

  1) *'Symmetric test'* (tests depend only on the empirical distribution of x^n) is the best

  2) Among all symmetric tests, the *mixture likelihood ratio test (MLRT)* is optimal

step 1

$$\psi(\cdot) \longrightarrow \boxed{\text{symmetrization}} \longrightarrow \phi(\cdot)$$

$$\phi(x^n) = \frac{1}{n!} \sum_{\tau:\ \text{all permutations}} \psi\left(\tau(x^n)\right)$$

step 2

ϕ is better : $\mathsf{P}_\mathsf{F}(\phi) \le \mathsf{P}_\mathsf{F}(\psi)$, and $\mathsf{P}_\mathsf{M}(\phi) \le \mathsf{P}_\mathsf{M}(\psi)$

proof

$$\begin{aligned}
\mathsf{P}_\mathsf{F}(\phi) &= \max_\sigma \mathsf{E}_{\mathbb{P}_{0;\sigma}} \left[\frac{1}{n!} \sum_\tau \psi \circ \tau(X^n) \right] \\
&= \max_\sigma \frac{1}{n!} \sum_\tau \mathsf{E}_{\mathbb{P}_{0;\sigma}} \left[\psi \circ \tau(X^n) \right] \\
&\le \frac{1}{n!} \sum_\tau \max_\sigma \mathsf{E}_{\mathbb{P}_{0;\sigma}} \left[\psi \circ \tau(X^n) \right] \\
&= \mathsf{P}_\mathsf{F}(\psi)
\end{aligned}$$

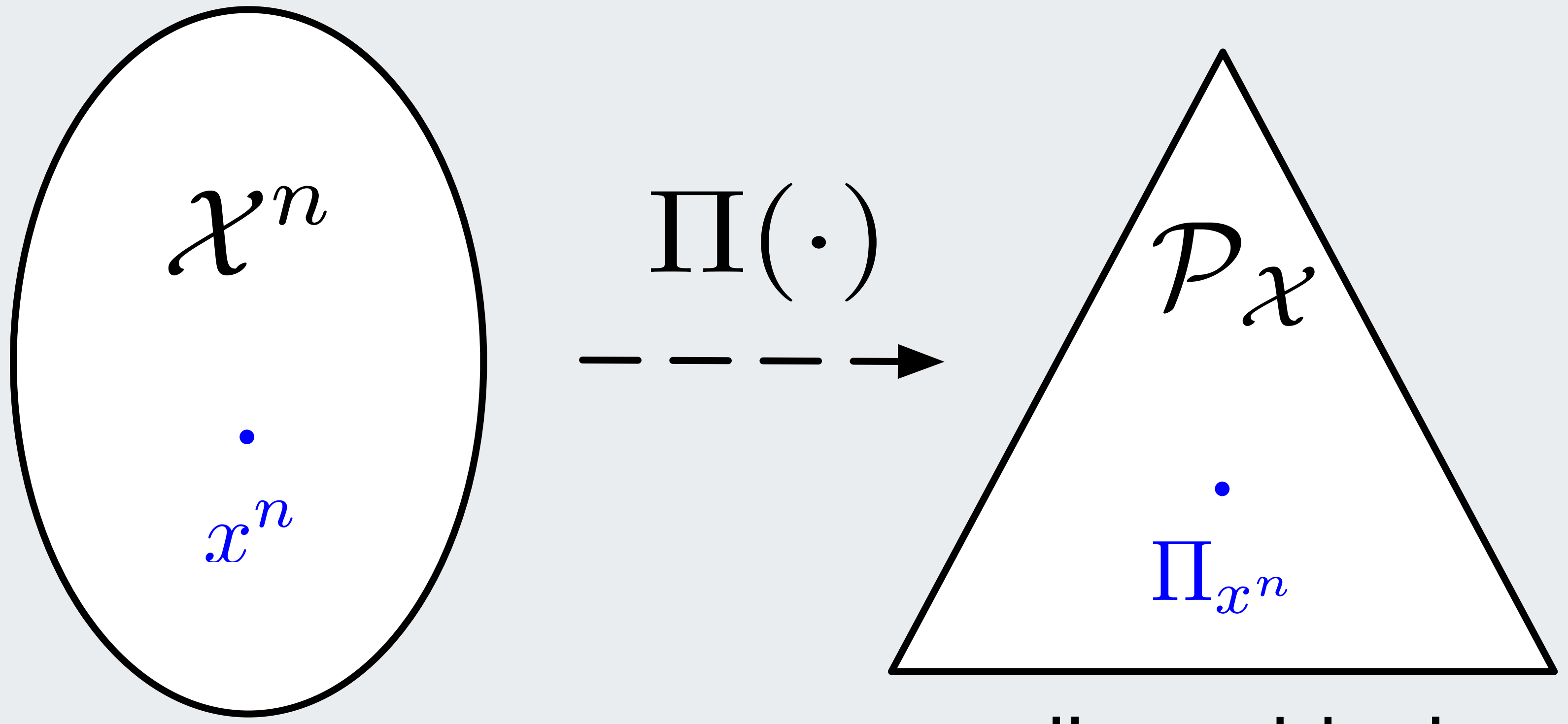
the empirical distribution contains sufficient information !

Sketch of Proof : Optimal Test

- Idea :

  1) *'Symmetric test'* (tests depend only on the empirical distribution of x^n) is the best

  2) Among all symmetric tests, the *mixture likelihood ratio test (MLRT)* is optimal



all empirical distributions

observation :

independent of σ !

$$\mathbb{P}_{\theta;\sigma}\left(T(\Pi_{x^n})\right) \triangleq \tilde{\mathbb{P}}_\theta(\Pi_{x^n})$$

collection of x^n with all possible orderings

Equivalent *simple* hypothesis testing on $\mathcal{P}_{\mathcal{X}}$

$$\begin{cases} \mathcal{H}_0 : \mathbb{P}_{0;\sigma}, \text{ for some } \sigma \\ \mathcal{H}_1 : \mathbb{P}_{1;\sigma}, \text{ for some } \sigma \end{cases} \Rightarrow \begin{cases} \tilde{\mathcal{H}}_0 : \tilde{\mathbb{P}}_0 \\ \tilde{\mathcal{H}}_1 : \tilde{\mathbb{P}}_1 \end{cases}$$

Neyman-Pearson lemma:

$$\ell(x^n) = \frac{\tilde{\mathbb{P}}_0(\Pi_{x^n})}{\tilde{\mathbb{P}}_1(\Pi_{x^n})} = \frac{\sum_\sigma \mathbb{P}_{0;\sigma}(x^n)}{\sum_\sigma \mathbb{P}_{1;\sigma}(x^n)}$$

Asymptotic Regime : Sanov's Theorem

i.i.d simple hypothesis testing

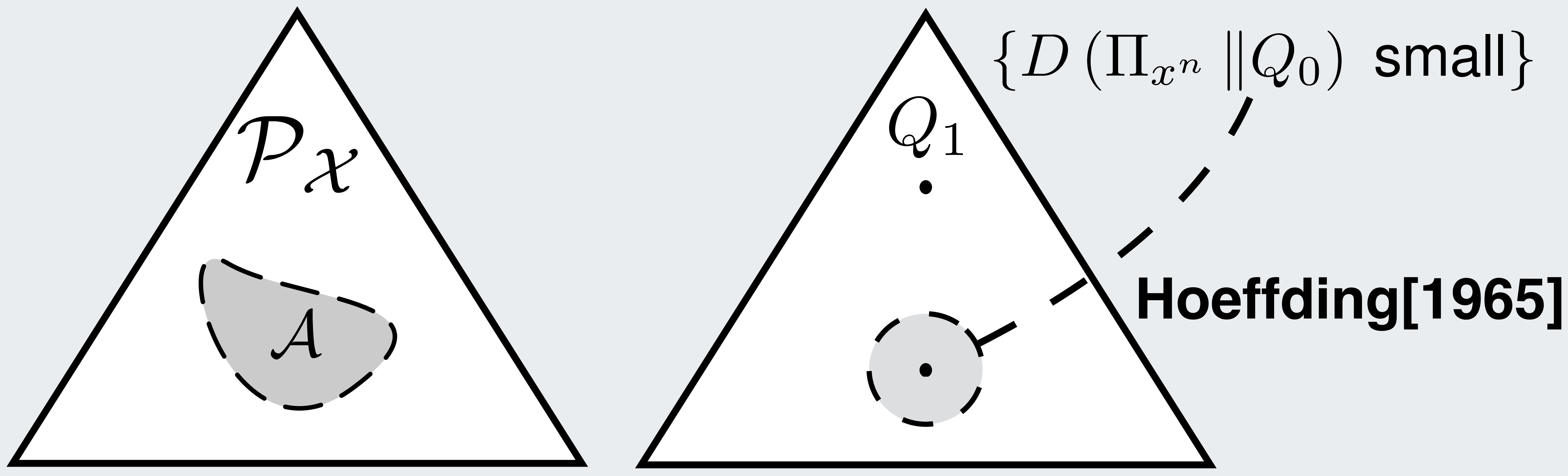
$$\mathcal{H}_\theta : X^n \sim (Q_\theta)^{\otimes n}$$

heterogeneous anonymous testing

$$\mathcal{H}_\theta : X^n \sim \mathbb{P}_{\theta;\sigma} \text{ for some } \sigma$$

Sanov's Theorem

$$Q_\theta^{\otimes n}(x^n : \Pi_{x^n} \in \mathcal{A}) \approx 2^{-n\left(\min_{U \in \mathcal{A}} D(U \| Q_\theta)\right)}$$



$\Longrightarrow$ type-II error exponent : $D(Q_0 \| Q_1)$

Find exponents of large deviation events:

For any σ, we have

$$\mathbb{P}_{\theta;\sigma}(\Pi_{x^n} \in \mathcal{A}) \approx 2^{-n\left(\min_{\boldsymbol{\alpha}^\intercal \boldsymbol{U} \in \mathcal{A}} D_{\boldsymbol{\alpha}}(\boldsymbol{U}; \boldsymbol{P}_\theta)\right)}$$

with the rate function being

$$D_{\boldsymbol{\alpha}}(\boldsymbol{P}; \boldsymbol{Q}) \triangleq \min_{\boldsymbol{V} \in (\mathcal{P}_\mathcal{X})^K} \sum_{k=1}^{K} \alpha_k D(V_k \| P_{\theta;k})$$

$$\text{s.t. } \boldsymbol{\alpha}^\intercal \boldsymbol{V} = \boldsymbol{\alpha}^\intercal \boldsymbol{U}$$

$\Longrightarrow$ type-II error exponent : $D_{\boldsymbol{\alpha}}(\boldsymbol{P}; \boldsymbol{Q})$

Key Step : non-i.i.d. Sanov's Theorem

Theorem :

For any σ, $\mathbb{P}_{\theta;\sigma}\left(\Pi_{x^n} \in \mathcal{A}\right) \approx 2^{-n\left(\min\limits_{\boldsymbol{\alpha}^\intercal \boldsymbol{U} \in \mathcal{A}} D_{\boldsymbol{\alpha}}(\boldsymbol{U};\boldsymbol{P}_\theta)\right)}$, with the rate function being

$$D_{\boldsymbol{\alpha}}\left(\boldsymbol{P};\boldsymbol{Q}\right) \triangleq \min_{\boldsymbol{V} \in (\mathcal{P}_{\mathcal{X}})^K} \sum_{k=1}^{K} \alpha_k D\left(V_k \,\|\, P_{\theta;k}\right)$$

s.t. $\boldsymbol{\alpha}^\intercal \boldsymbol{V} = \boldsymbol{\alpha}^\intercal \boldsymbol{U}$

	$n\alpha_1$	$n\alpha_2$
$\mathbb{P}_{\theta;\sigma}$	i.i.d. $P_{\theta;1}$	i.i.d. $P_{\theta;2}$
x^n	sub-type U_1	sub-type U_2

Recall : $Q^{\otimes n}\left(\Pi_{x^n}\right) \approx 2^{-nD(\Pi_{x^n} \| Q)}$

$\mathbb{P}_{\theta;\sigma}(\Pi_{x^n}) \approx 2^{-n(\alpha_1 D(U_1 \| P_\theta) + \alpha_2 D(U_2 \| Q_\theta))}$

- minimize over all sub-types : $\{U_1, U_2 : \alpha_1 U_1 + \alpha_2 U_2 = V\}$
- minimize over all types : $V \in \mathcal{A}$

Key Step : non-i.i.d. Sanov's Theorem

Theorem :

For any σ, $\mathbb{P}_{\theta;\sigma}\left(\Pi_{x^n} \in \mathcal{A}\right) \approx 2^{-n\left(\min\limits_{V \in \mathcal{A}} d(V, \boldsymbol{P}_\theta)\right)}$, with the rate function being

$$d(V, \boldsymbol{P}_\theta) \triangleq \min_{\substack{\boldsymbol{U} \in (\mathcal{P}_\mathcal{X})^n \\ \boldsymbol{\alpha}^\intercal \boldsymbol{U} = V}} \sum_{k=1}^{K} \alpha_k D\left(U_k \| P_{\theta;k}\right)$$

Also holds for Polish $\mathcal{X}$

x^n: sub-type U_1 | sub-type U_2

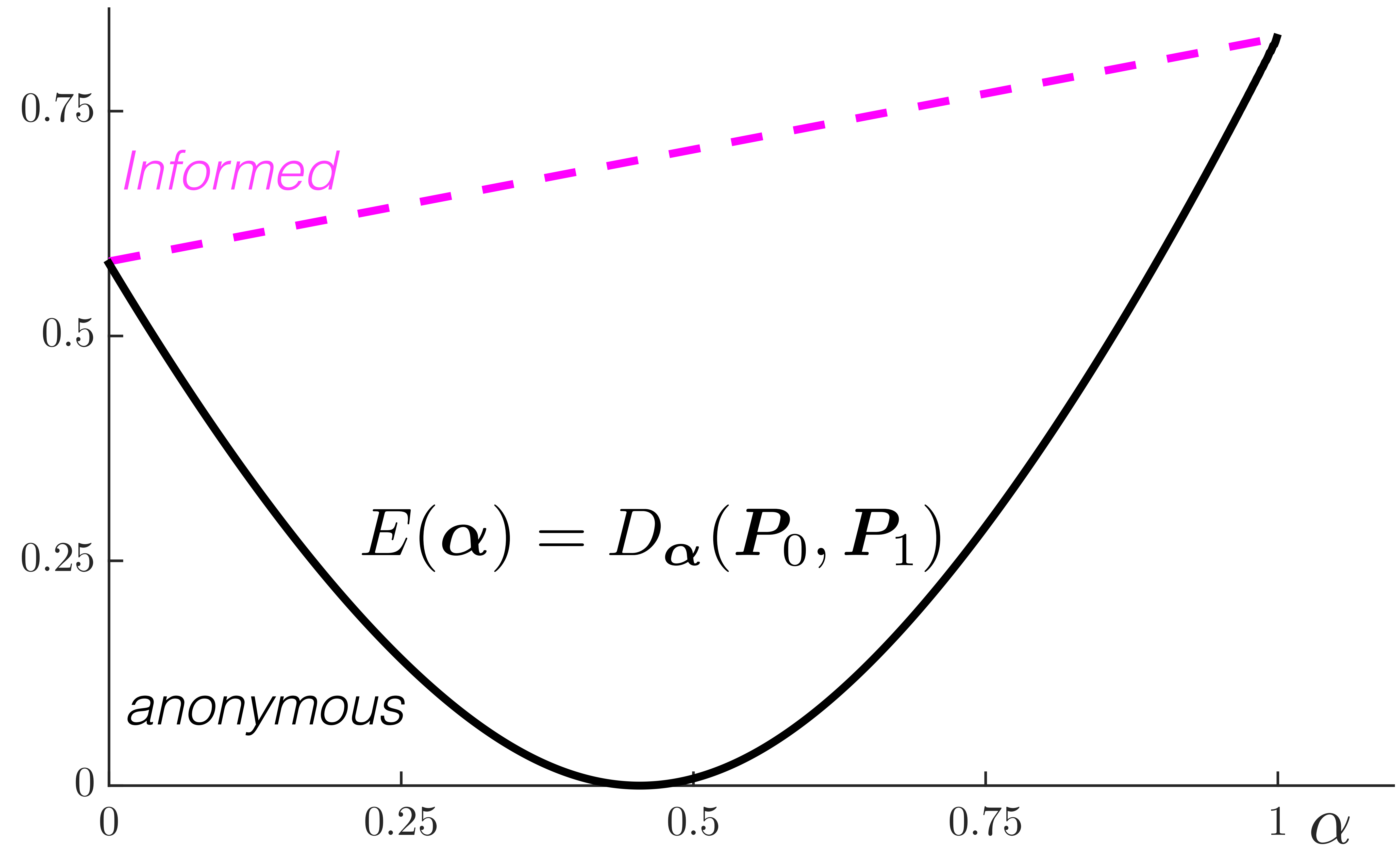
$$\mathbb{P}_{\theta;\sigma}(\Pi_{x^n}) \approx 2^{-n(\alpha_1 D(U_1 \| P_\theta) + \alpha_2 D(U_2 \| Q_\theta))}$$

- minimize over all sub-types : $\{U_1, U_2 : \alpha_1 U_1 + \alpha_2 U_2 = V\}$
- minimize over all types : $V \in \mathcal{A}$

Summary

- ***Optimal decision rule*** : mixture likelihood ratio test (MLRT) ~~GLRT~~

- ***Asymptotic*** :



Generalized divergence

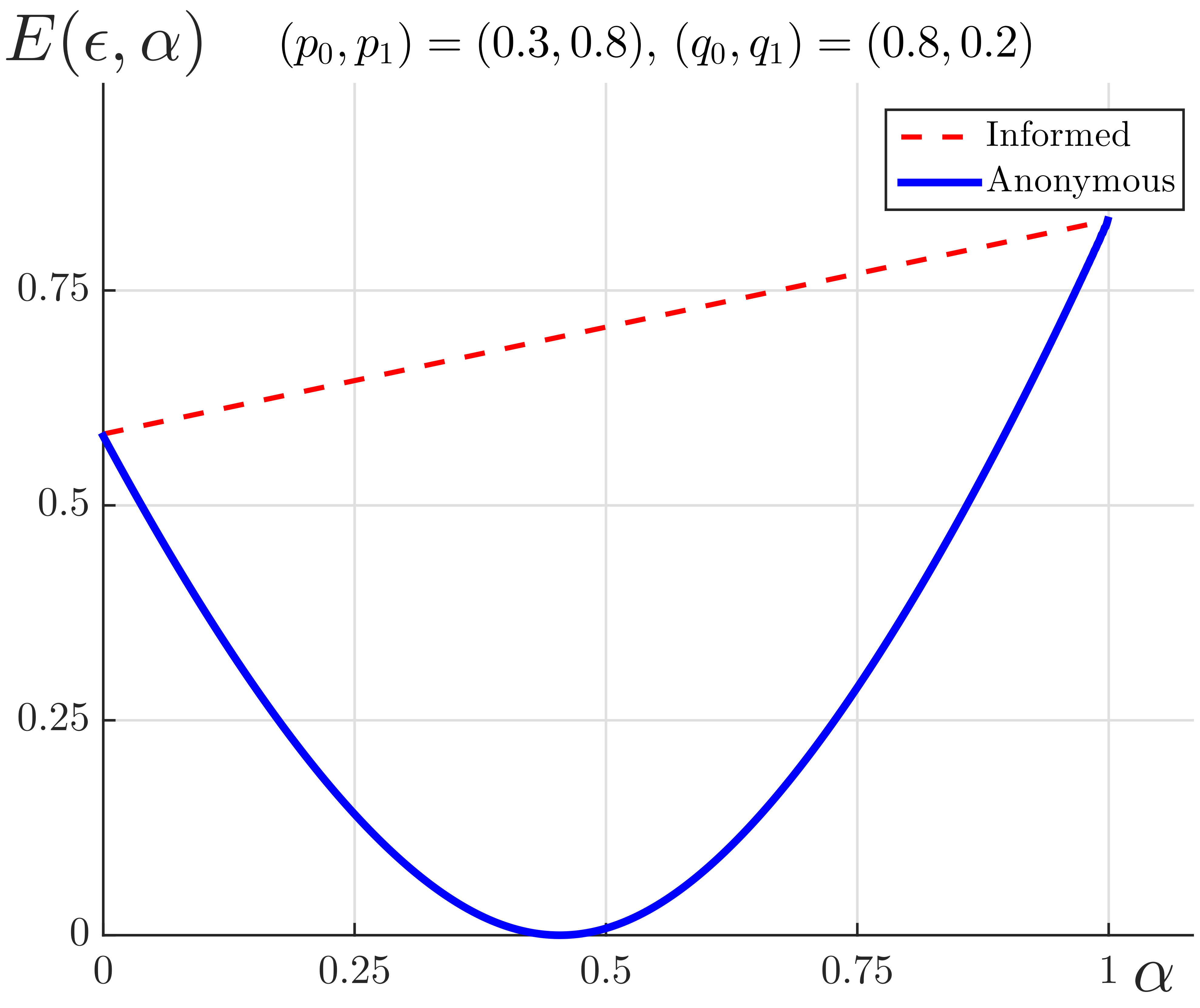
$$D_{\boldsymbol{\alpha}}(\boldsymbol{U}, \boldsymbol{P}_\theta) \triangleq \min_{\boldsymbol{V} \in (\mathcal{P}_\mathcal{X})^K} \sum_{k=1}^{K} \alpha_k D\left(V_k \,\|\, P_{\theta;k}\right)$$

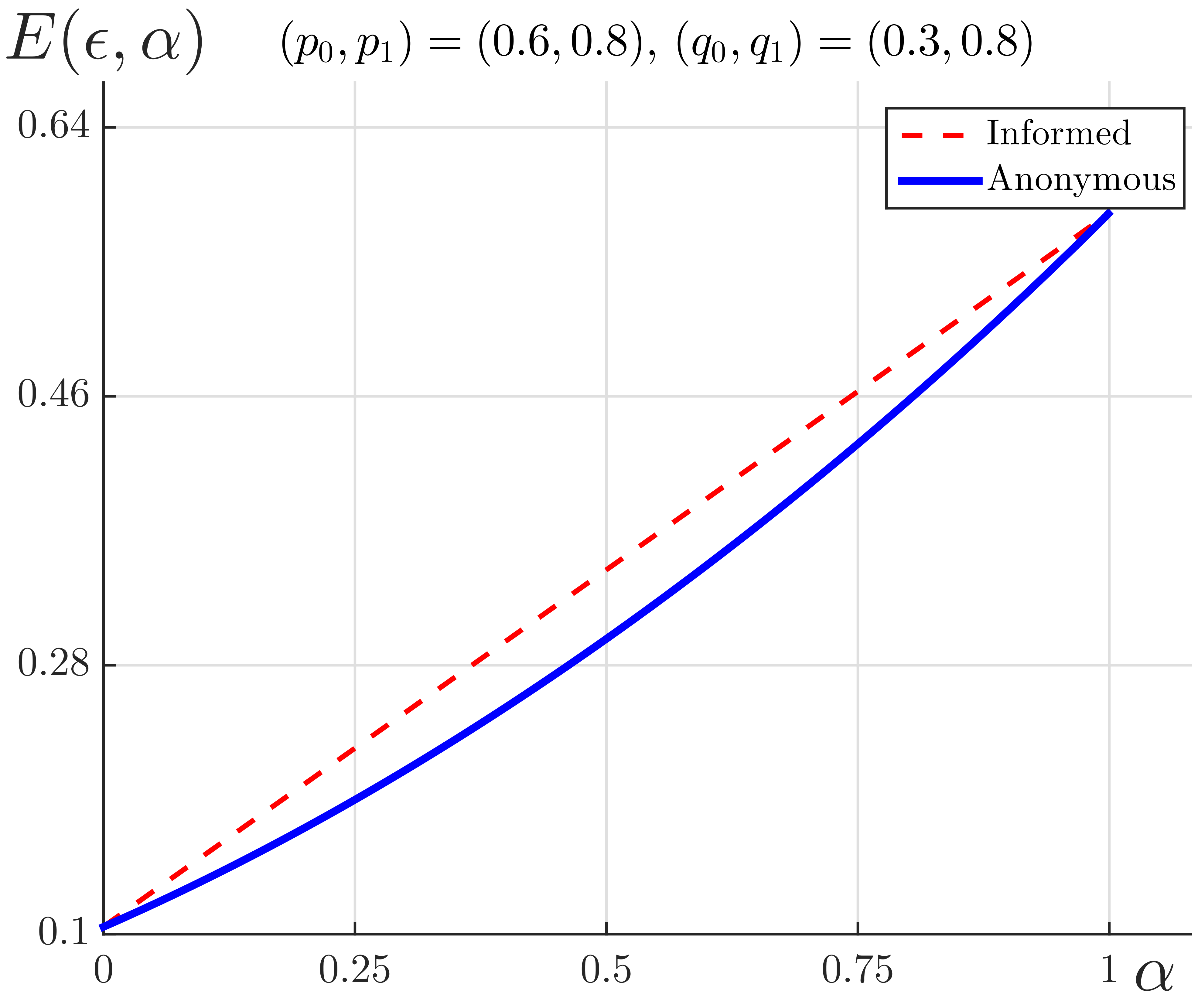
s.t. $\boldsymbol{\alpha}^\intercal \boldsymbol{V} = \boldsymbol{\alpha}^\intercal \boldsymbol{U}$

extended to Chernoff regime by solving information projection !

Part III : Conclusion and Future Directions

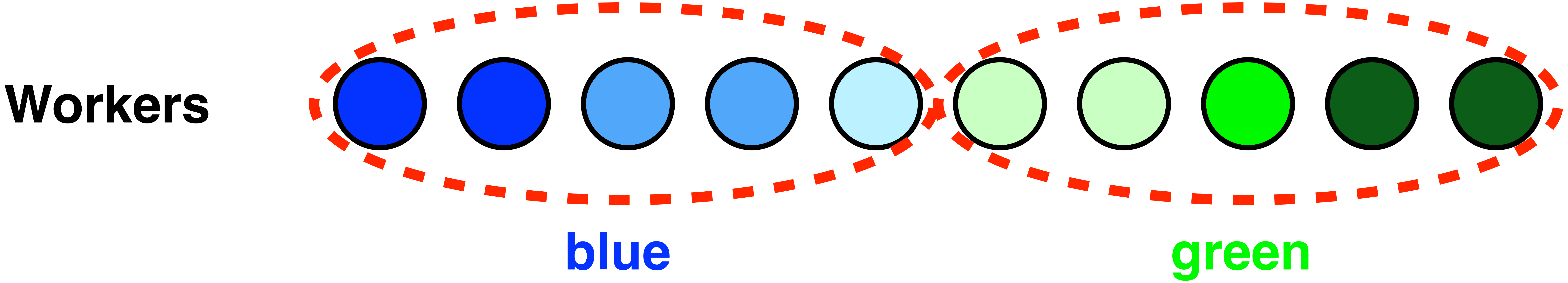
Is group-recovery necessary ?





Estimate '*price of anonymity*', and decide whether or not to recover group info.

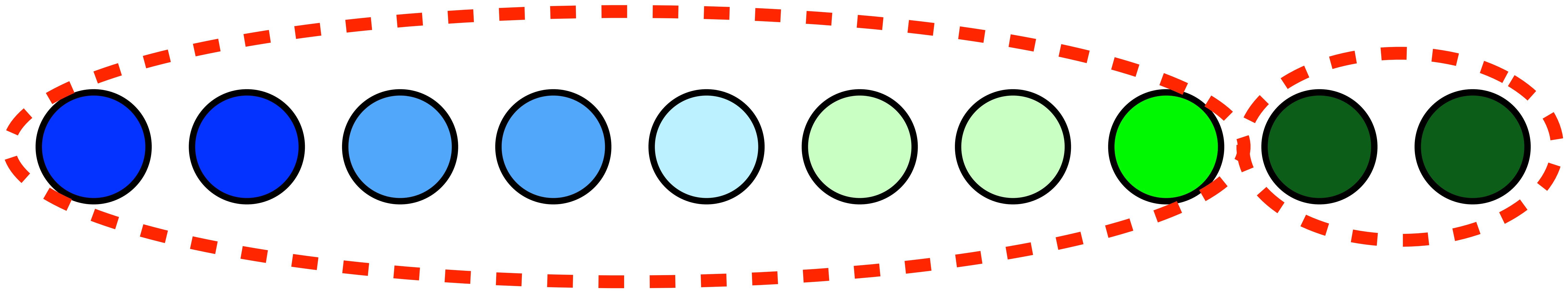
Future Work : Partially Recovery



- Partially recover the group
  - Exact recovery is too expensive
  - *Clustering* different groups
- Difficulties
  - How to *optimally* cluster groups
  - How to evaluate the optimal type-II exponent
  - How to trade off

Future Work : Partially Recovery

Workers



- Partially recover the group
  - Exact recovery is too expensive
  - *Clustering* different groups
- Difficulties
  - How to *optimally* cluster groups
  - How to evaluate the optimal type-II exponent
  - How to trade off

NP-hard !

approximation / bounds

Thanks for your attention !