Prospectus Presentation

Minimax Theory of Constrained Hypothesis Testing

Presenter: Yikun Li

Committee members: Prof. Matey Neykov (chair), Prof. Miklos Racz, Prof. Feng Ruan

Content

Introduction

- Problem Formulation
- Literature Review

Main Results

- Lower Bounds
- Upper Bounds
- Discussion

Experiments

- Numerical Simulation
- Discussion

Summary

- Summary
- Open Problems
- Plan for Future Works

Content

Problem Formulation

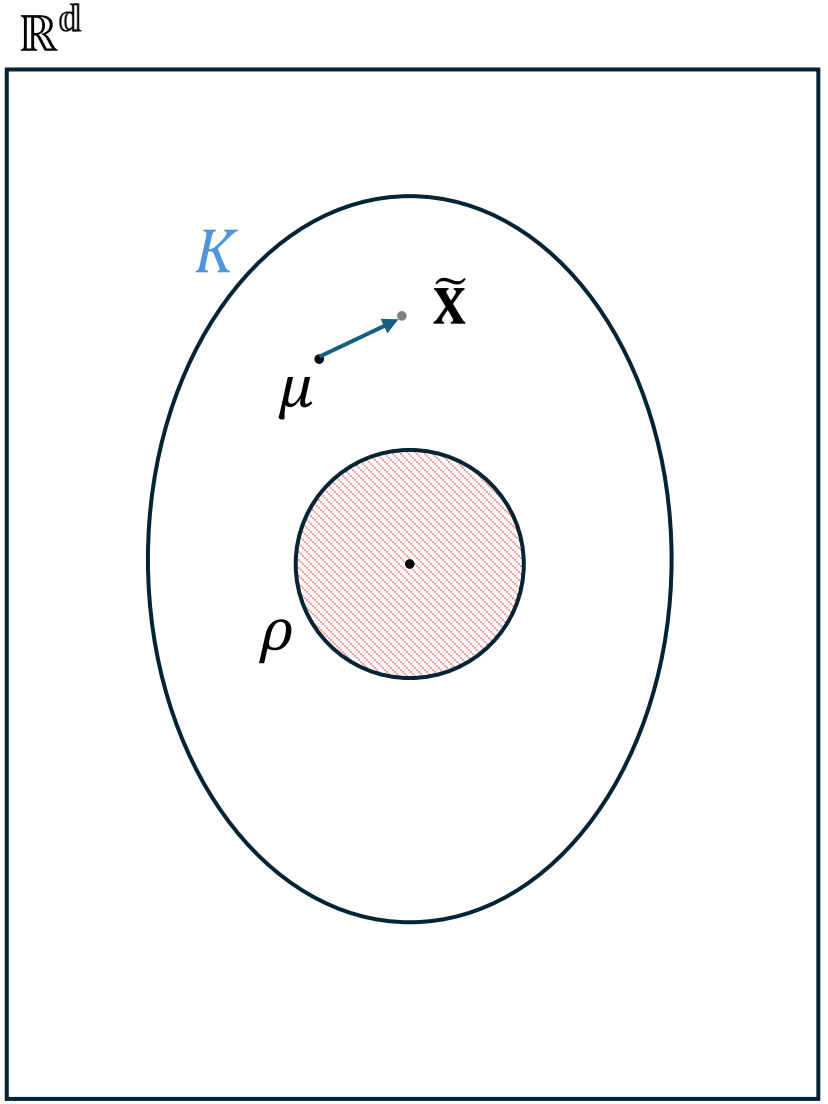
Problem studied: hypothesis testing

Assumptions:

- Gaussian sequence model (true distribution of the data);
- Known covariance matrix $\mathbf{I}_d$, mean vector $\mu \in K$; (prior constraints)

Authentic samples $\widetilde{\mathbf{X}} \coloneqq \{\widetilde{\mathrm{X}}_1, \ldots, \widetilde{\mathrm{X}}_\mathrm{N}\}$ i.i.d. from Gaussian $\mathcal{N}(\mu, \mathbf{I_d})$

$H_0: \mu = \mathbf{0}$;
$H_1: \|\mu\|_2 \geq \rho, \mu \in K$.



Problem Formulation

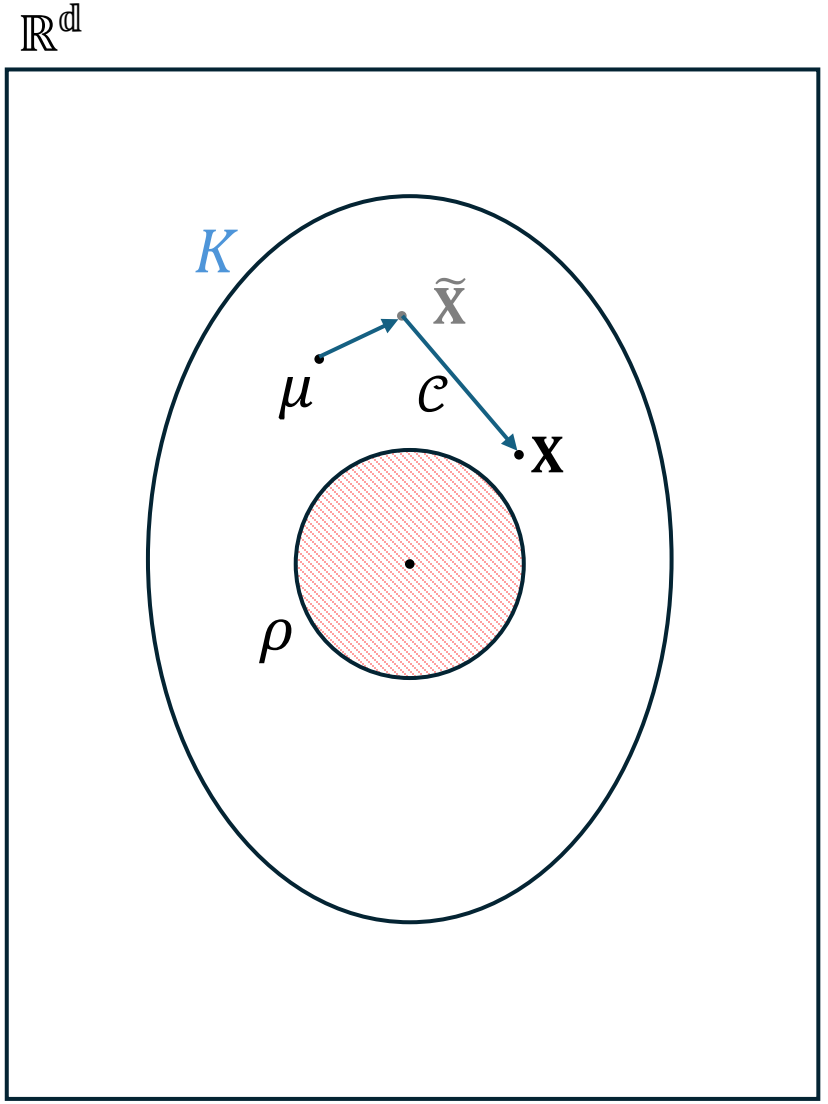
Problem studied: hypothesis testing

Additional Assumptions:

- Strong contamination adversary $\mathcal{C}$ (robustness)
- Powerful but limited contamination fraction ϵ

$\tilde{\mathbf{X}}$ (unobserved) $\longrightarrow$ $\mathbf{X} \coloneqq \mathcal{C}(\tilde{\mathbf{X}})$ (observed)

$H_0: \mu = \mathbf{0}$;
$H_1: \|\mu\|_2 \geq \rho, \mu \in K$.



Problem Formulation

Problem studied: hypothesis testing

- ρ overly small ⟶ large Type II errors
- ρ overly large ⟶ information not fully utilized

Question: **what is the exact rate of ρ ?**

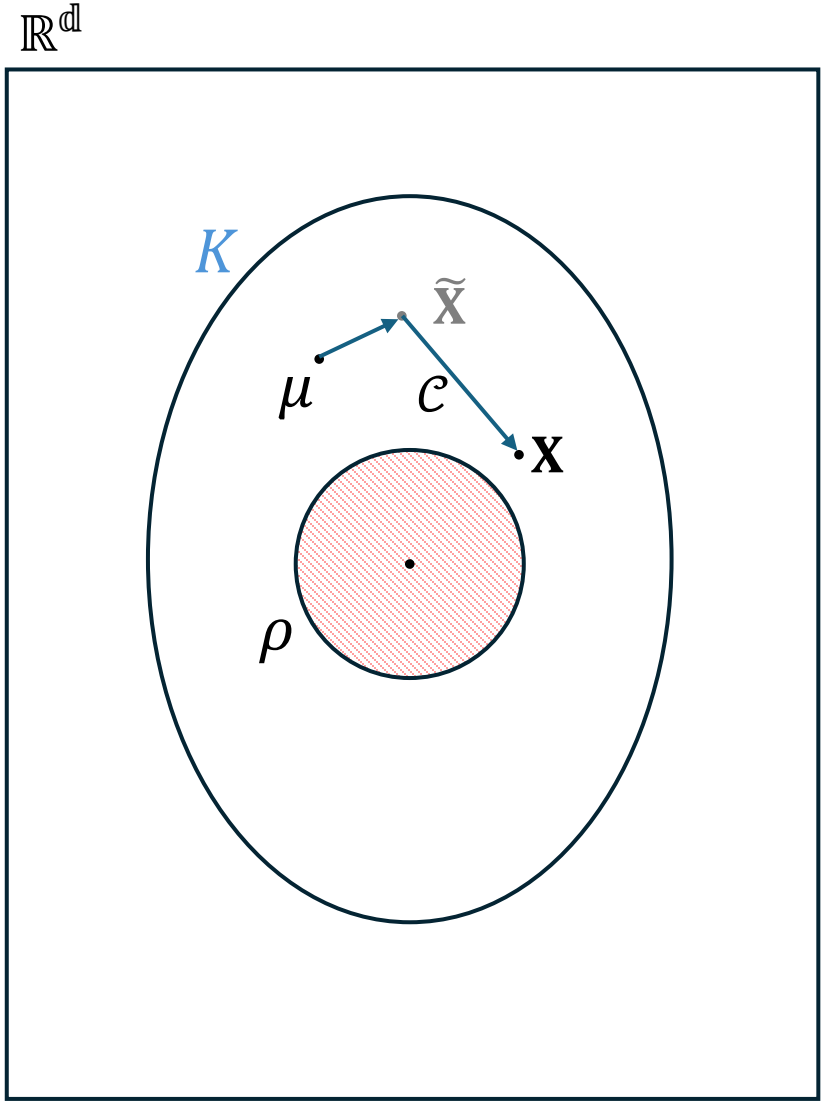
Acceptable tests: $A_s := \left\{ \phi : \sup_{\mathcal{C}} \mathbb{P}_0 \left(\phi(\mathcal{C}(\widetilde{\mathbf{X}})) = 1 \right) \leq \alpha \right\},$

Exact rate: $\rho_{\text{critical}} := \inf_{\rho} \left\{ \rho : \rho > 0, \sup_{\|\mu\|_2 \geq \rho, \mu \in K} \sup_{\mathcal{C}} \mathbb{P}_\mu \left(\phi(\mathcal{C}(\mathbf{X})) = 0 \right) \leq \alpha, \phi \in A_s \right\}$

optimize over the worst case

worst case of parameter

worst case of adversary



Literature Review

Information-theoretic aspect

1973, Le Cam. Le Cam's approach for minimax lower bound.

~1990, Ingster, Suslina, et al. Foundational methodology and rates established for paradigm testing problem.

1992, Donoho, et al. Minimax rate established for constrained estimation problem

2002, Baraud. Establish the minimax rate of testing under ellipsoid constraints

~2020, Canonne, Narayanan, et al. Minimax rate under robust settings

2024 Neykov, et al. Minimax rate for estimation under star-shaped constraints and robust settings

Information theoretic limits of robust sub-Gaussian mean estimation under star-shaped constraints

Akshay Prasadan and Matey Neykov

Department of Statistics & Data Science, Carnegie Mellon University
Department of Statistics and Data Science, Northwestern University

aprasada@andrew.cmu.edu, mneykov@northwestern.edu

Abstract

We obtain the minimax rate for a mean location model with a bounded star-shaped set $K \subseteq \mathbb{R}^n$ constraint on the mean, in an adversarially corrupted data setting with Gaussian noise. We assume an unknown fraction $\epsilon \leq 1/2 - \kappa$ for some fixed $\kappa \in (0, 1/2]$ of N observations are arbitrarily corrupted. We obtain a minimax risk up to proportionality constants under the squared ℓ_2 loss of $\max(\eta^{*2}, \sigma^2\epsilon^2) \wedge d^2$ with

$$\eta^* = \sup\left\{\eta > 0 : \frac{N\eta^2}{\sigma^2} \leq \log \mathcal{M}_K^{\text{loc}}(\eta, c)\right\},$$

where $\log \mathcal{M}_K^{\text{loc}}(\eta, c)$ denotes the local entropy of the set K, d is the diameter of K, σ^2 is the variance, and c is some sufficiently large absolute constant. A variant of our algorithm achieves the same rate for settings with known or symmetric sub-Gaussian noise, with a smaller breakdown point, still of constant order. We further study the case of unknown sub-Gaussian noise and show that the rate is slightly slower: $\max(\eta^{*2}, \sigma^2\epsilon^2 \log(1/\epsilon)) \wedge d^2$. We generalize our results to the case when K is star-shaped but unbounded.

Contents

Literature Review

Computational aspect

2016, Diakonikolas et al. First general computationally efficient framework under robust settings for estimation

2023, Canonne, Narayanan, et al. Efficient framework transferred to robust testing scenario without constraint

The Full Landscape of Robust Mean Testing: Sharp Separations between Oblivious and Adaptive Contamination

Clément Canonne* University of Sydney
Samuel B. Hopkins† MIT
Jerry Li‡ Microsoft Research
Allen Liu§ MIT
Shyam Narayanan¶ MIT

July 21, 2023

Abstract

We consider the question of Gaussian mean testing, a fundamental task in high-dimensional distribution testing and signal processing, subject to adversarial corruptions of the samples. We focus on the relative power of different adversaries, and show that, in contrast to the common wisdom in robust statistics, there exists a strict separation between adaptive adversaries (strong contamination) and oblivious ones (weak contamination) for this task. Specifically, we resolve both the information-theoretic and computational landscapes for robust mean testing. In the exponential-time setting, we establish the tight sample complexity of testing $\mathcal{N}(0, I)$ against $\mathcal{N}(\alpha v, I)$, where $\|v\|_2 = 1$, with an ε-fraction of adversarial corruptions, to be

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^3}{\alpha^4}, \min\left(\frac{d^{2/3}\varepsilon^{2/3}}{\alpha^{8/3}}, \frac{d\varepsilon}{\alpha^2}\right)\right)\right),$$

while the complexity against adaptive adversaries is

$$\tilde{\Theta}\left(\max\left(\frac{\sqrt{d}}{\alpha^2}, \frac{d\varepsilon^2}{\alpha^4}\right)\right),$$

which is strictly worse for a large range of vanishing ε, α. To the best of our knowledge, ours is the first separation in sample complexity between the strong and weak contamination models.

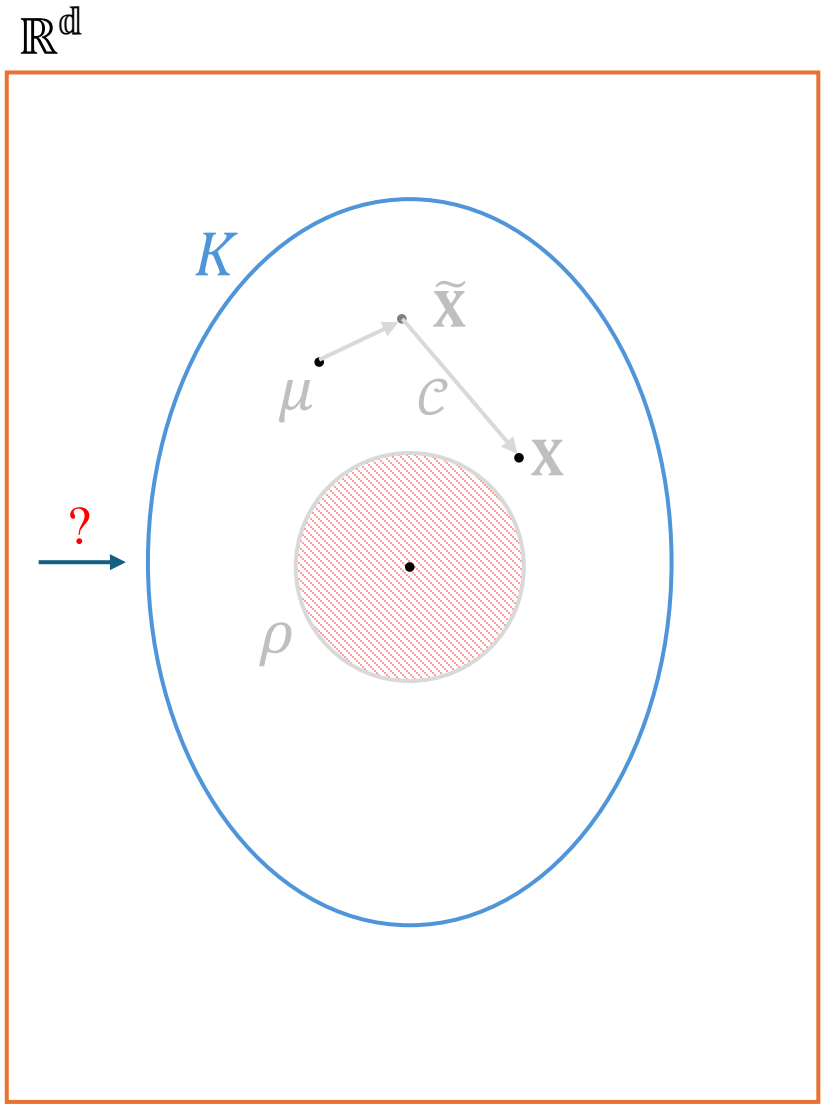
In the polynomial-time setting, we close a gap in the literature by providing a polynomial-time algorithm against adaptive adversaries achieving the above sample complexity $\tilde{\Theta}(\max(\sqrt{d}/\alpha^2, d\varepsilon^2/\alpha^4))$, and a low-degree lower bound (which complements an existing reduction from planted clique) suggesting that all efficient algorithms require this many samples, even in the oblivious-adversary setting.

*clement.canonne@sydney.edu.au. Supported by an ARC DECRA (DE230101329) and an unrestricted gift from Google Research.
†samhop@mit.edu. Supported by NSF Award No. 2238080 and MLA@CSAIL
‡jerrl@microsoft.com.
§cliu568@mit.edu Supported by an NSF Graduate Research Fellowship and a Fannie and John Hertz Foundation Fellowship.
¶shyamsn@mit.edu. Supported by an NSF Graduate Fellowship and a Google Fellowship.

Literature Review

Overall Motivation

- Under certain geometric constraints, how does the minimax lower bound changed?
- Does the matching upper bound still exist?
- Is the algorithm computationally efficient?



Content

Introduction

- Problem Formulation
- Literature Review

Main Results

- Lower Bounds
- Upper Bounds
- Discussion

Experiments

- Numerical Simulation
- Discussion

Summary

- Summary
- Open Problems
- Plan for Future Works

Main Results

Preliminaries (Quadratically convex orthosymmetric (QCO) set)

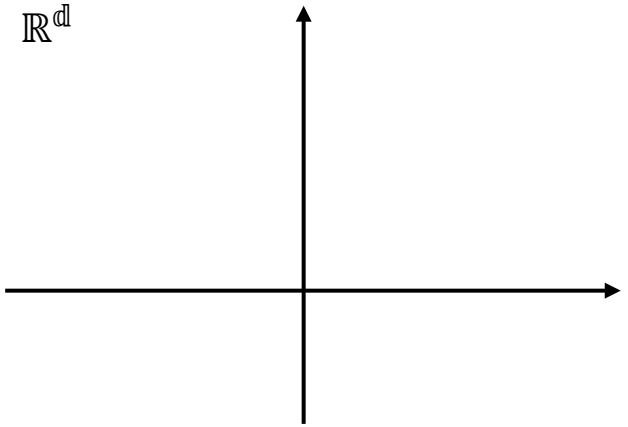
Definition 2.1 (Quadratically convex orthosymmetric (QCO) set). *Given a set $K \subset \mathbb{R}^d$, we say K is a quadratically convex orthosymmetric (QCO) set if it satisfies the following conditions:*

(1), K is convex;

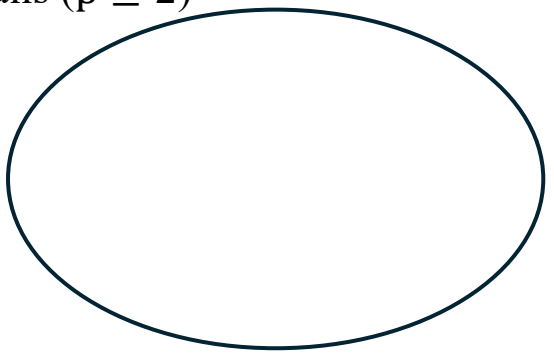
(2), K is quadratically convex, which means that K^2 is also convex, where K^2 is defined as

$$K^2 := \left\{ \left(\theta_1^2, \ldots, \theta_d^2\right)^\top \mid (\theta_1, \ldots, \theta_d)^\top \in K \right\};$$

(3), K is orthosymmetric, which means that if $\theta = (\theta_1, \ldots, \theta_d)^\top \in K$, then $\theta_\eta := (\eta_1\theta_1, \ldots, \eta_d\theta_d)^\top \in K$, where $\eta_i \in \{-1, 1\}, 1 \le i \le d$.



ℓ_p balls (p ≥ 2)



hyperrectangles

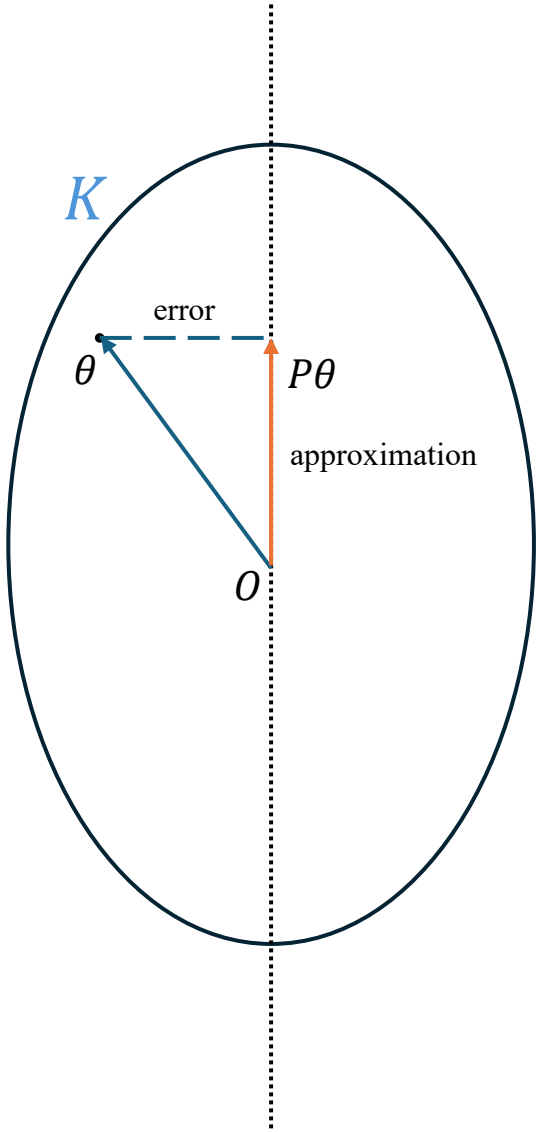
Main Results

Preliminaries (Kolmogorov widths)

Definition 2.2 (Kolmogorov k-width, also known as Kolmogorov N-width)**.** *Let $\mathcal{X}$ be a Banach space equipped with the norm $\|\cdot\|$, and $K \subset \mathcal{X}$ is a subset. The Kolmogorov k-width is defined as*

$$D_k(K) = \inf_{V_k} \sup_{\theta \in K} \min_{\tilde{\theta} \in V_k} \left\| \theta - \tilde{\theta} \right\|_2, \tag{2.1}$$

where $V_k \subset \mathbb{R}^d$ is any k-dimensional subspace of $\mathbb{R}^d$.



- For ℓ_2 norm approximation, $\tilde{\theta}$ can be expressed as $\tilde{\theta} = P\theta$, where P is the projection operator onto V_k

Motivation: the minimax risk of k-dimensional linear approximations

Main Results

Preliminaries (Optimal dimensions and projections)

First optimal dimensions: $k_1^\star(K,\sigma,N) := \max\left\{ j \mid 0 \le j \le d, D_{j-1}(K) > \frac{j^{\frac{1}{4}}}{\sqrt{N}}\sigma \right\}$

Second optimal dimensions: $k_2^\star(K,\sigma,N,\epsilon) := \max\left\{ j \mid 0 \le j \le d, D_{j-1}(K) > \frac{j^{\frac{1}{4}}\sqrt{\epsilon}}{N^{\frac{1}{4}}}\sigma \right\}$

The corresponding projections are first/second optimal projections $P_1^\star, P_2^\star$

Motivation: interaction between the geometry and the data properties

Main Results

Preliminaries (Notations)

- d: full dimension
- N: sample size
- μ: mean vector
- K: geometric constraints of μ
- α: fixed Type I and Type II errors tolerance

- $\widetilde{\mathbf{X}} \coloneqq \{\widetilde{\mathrm{X}}_1, \ldots, \widetilde{\mathrm{X}}_\mathrm{N}\}$: original, unobserved samples from $\mathcal{N}(\mu, \mathbf{I_d})$
- $\mathbf{X} \coloneqq \{X_1, \ldots, X_N\}$: contaminated samples from $\mathcal{C}$ and $\widetilde{\mathbf{X}}$
- C: index set of the corrupted samples

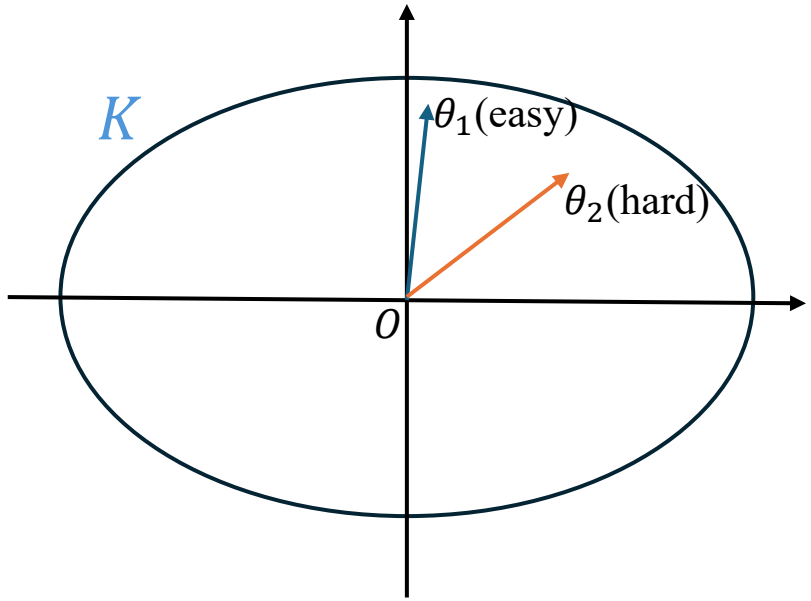
- $D_k(K)$: Kolmogorov k-width of K
- $k_1^\star, k_2^\star$: first/second optimal dimensions
- $P_1^\star, P_2^\star$: first/second optimal projections

Main Results (First Lower Bounds)

Important lemma. Let $K \subset \mathcal{X}$ be a QCO set and c be an arbitrary positive constant. Suppose that $D_{k-1}(\mathrm{K}) > c\sigma$ for some $k \geq 1$. Then there exists a vector $\theta \in K$ such that $\|\theta\|_2 = c\sigma$ while $\|\theta\|_\infty \leq \frac{c}{\sqrt{k}}\sigma$.

Motivation:

- The output vector θ is "smeared" with non-trivial total energy
- Construction of *least favor distribution*



Main Results (First Lower Bounds)

First lower bound. If $\rho \leq c(\alpha)\frac{(k_1^\star)^{1/4}}{\sqrt{N}}\sigma$, where $c(\alpha)$ is a constant that only depends on α, or $k_1^\star = 0$, then

$$inf_{\psi:\mathbb{P}_{\mathbb{0}}(\psi=1)\leq\alpha} sup_{\theta\in K,\|\theta\|\geq\rho} \mathbb{P}_\theta(\psi = 0) \geq \alpha$$

High-Level Proof:

- Small χ^2 divergence between $\mathbb{P}_{\mathbb{0}}^{\otimes N}$ and $\mathbb{E}_{\mu\sim\gamma}\mathbb{P}_\mu^{\otimes N}$ leads to impossibility of the test via the standard Le Cam's fuzzy hypothesis method

  γ is any prior on the mean μ

- Control the χ^2 divergence via the output vector θ

Main Results (First Lower Bounds)

First lower bound. If $\rho \le c(\alpha)\frac{(k_1^\star)^{1/4}}{\sqrt{N}}\sigma$, where $c(\alpha)$ is a constant that only depends on α, or $k_1^\star = 0$, then

$$inf_{\psi:\mathbb{P}_0(\psi=1)\le\alpha} sup_{\theta\in K,\|\theta\|\ge\rho} \mathbb{P}_\theta(\psi = 0) \ge \alpha$$

Remark:

- No adversary appears here?
- Classic rate $\asymp \frac{d^{1/4}}{\sqrt{N}}\sigma$ (Ingster)
- One intrinsic dimensional parameter is $\boldsymbol{k_1^\star}$ instead of d for the constraint K!
- $k_1^\star = 0 \longrightarrow \sup_{\theta\in K}\|\theta\|_2 \le \frac{1}{\sqrt{N}}\sigma \longrightarrow$ the constraint K is insufficient rich for the test!

Main Results (Second Lower Bounds)

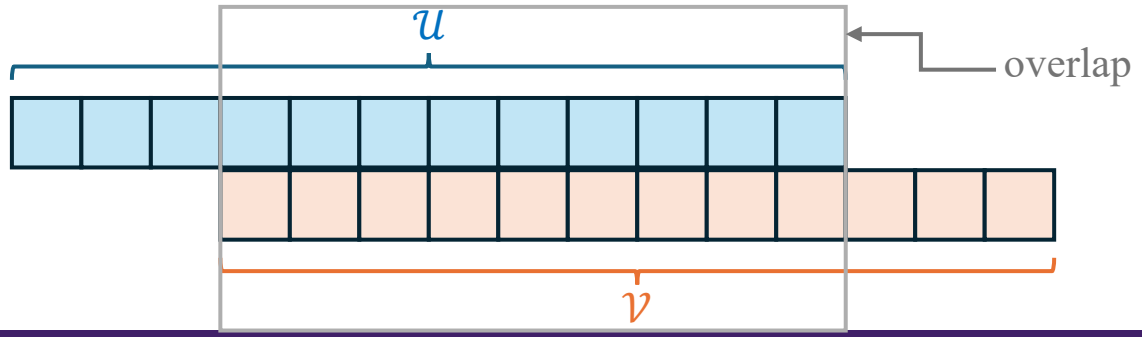
Important lemma. Suppose $\mathbf{X}, \mathbf{X}' \in \mathbb{R}^{\mathbb{N}\times\mathbb{d}}$. Assume the laws of $\mathbf{X}, \mathbf{X}'$ are $\mathcal{U}$ and $\mathcal{V}$, respectively. If there is a coupling between $\mathcal{U}$ and $\mathcal{V}$ such that the expectation of the Hamming distance between $\mathbf{X}$ and $\mathbf{X}'$ is $\mathcal{O}(\epsilon N)$:

$$\mathbb{E}_{(\mathbf{X},\mathbf{X}')\sim(\mathcal{U},\mathcal{V})} d_H(\mathbf{X}, \mathbf{X}') \lesssim \epsilon N.$$

Then, there is no robust test to distinguish between $\mathcal{U}$ and $\mathcal{V}$ with ϵ contamination and small errors.

Why is it useful?

A special coupling between $\mathbb{P}_{\mathbf{0}}^{\otimes\mathbb{N}}$ and $\mathbb{P}_{\mu}^{\otimes\mathbb{N}}$ ⟶ impossibility for a robust test!



Main Results (Second Lower Bounds)

Second lower bound. If $\rho \lesssim \frac{(k_2^\star)^{1/4}\sqrt{\epsilon}}{N^{1/4}}\sigma$, or $k_2^\star = 0$, then it is impossible to find a robust testing with uniformly small Type I and Type II errors. Here the omitted constant only depends on α.

High-Level Proof

- Interpolate between $\mathbb{P}_0^{\otimes \mathbb{N}}$ and $\mathbb{E}_{\mu\sim\gamma}\mathbb{P}_\mu^{\otimes \mathbb{N}}$ via a chain of optimal coupling
  (γ is any prior on the mean μ)
- Expectation of the Hamming distance → calculation of the total variance
- Bound the total variation by KL divergence
- Compute and control the KL divergence using the output vector θ

Main Results (Second Lower Bounds)

Second lower bound. If $\rho \lesssim \frac{(k_2^\star)^{1/4}\sqrt{\epsilon}}{N^{1/4}}\sigma$, or $k_2^\star = 0$, then it is impossible to find a robust testing with uniformly small Type I and Type II errors. Here the omitted constant only depends on α.

Remark:

- When $\epsilon = 0$: the lower bound collapses to zero
- Rate from Canonne, Narayanan, et al (2023): $\frac{d^{1/4}\sqrt{\epsilon}}{N^{1/4}}\sigma$
- Another intrinsic dimensional parameter is $\boldsymbol{k_2^\star}$!

Main Results (Third Lower Bounds)

Third lower bound. If $\epsilon \gtrsim \frac{1}{\sqrt{N}}$ and $\rho \lesssim \epsilon\sigma$, then it is impossible to find a robust testing with uniformly small Type I and Type II errors. Here the omitted constants only depend on α.

Remark:

- This lower bound comes from the counterpart estimation problem
- Irrelevant with sample size N
- The condition $\epsilon \gtrsim \frac{1}{\sqrt{N}}$ can be safely removed (See later discussion)

Main Results (Synthesized Lower Bounds)

Synthesized lower bound. Given the hypothesis testing problem previously formulated, the following condition is necessary for the existence of the desired robust testing:

$$\rho^2_{\text{critical}} \gtrsim \sigma^2 \max\left\{\underbrace{\frac{\sqrt{k_1^\star}}{N}, \epsilon\sqrt{\frac{k_2^\star}{N}}}_{\text{geometric terms}}, \underbrace{\epsilon^2}_{\text{corruption terms}}\right\}$$

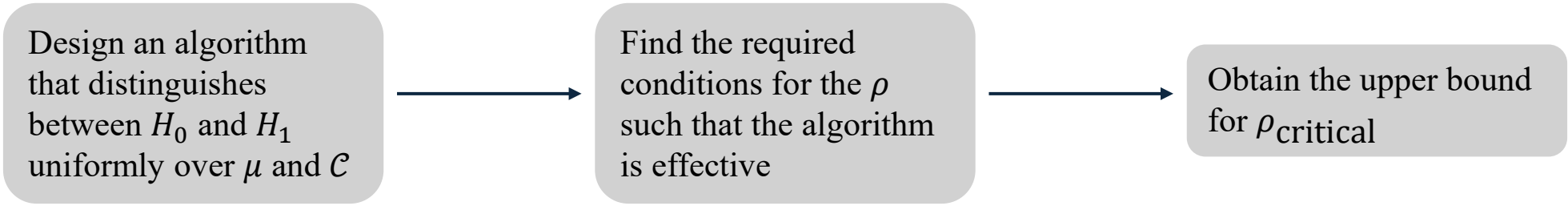
Remark:

- The first two terms involve the geometry of the constraint
- The last two terms involve the corruption process

Main Results (Upper Bound)

The upper bound should be derived constructively

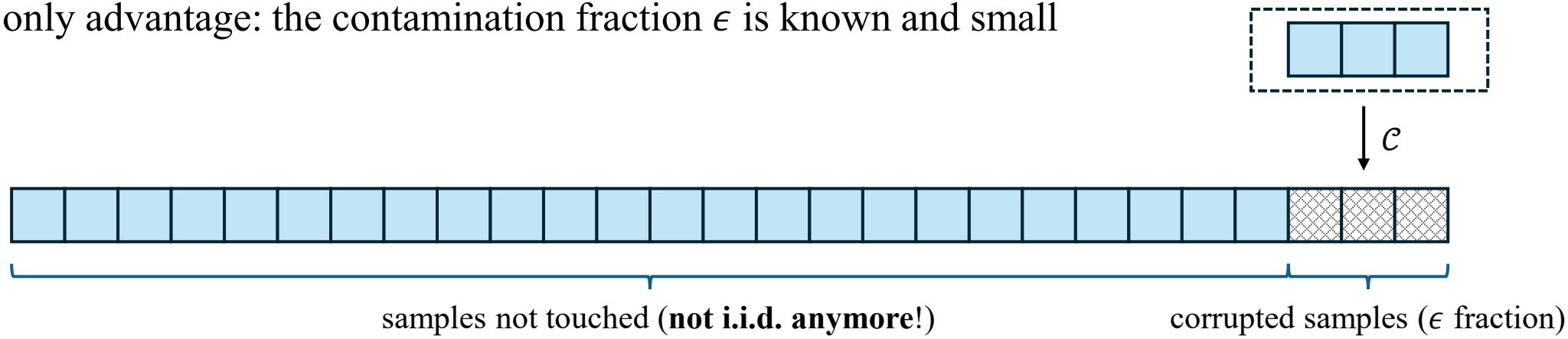
Overall roadmap:



Main Results (Theoretical Algorithm)

A natural motivation: conceal the influence of the adversary $\mathcal{C}$

- However, $\mathcal{C}$ is too powerful ⟶ not easy to work with the contaminated samples
- Our only advantage: the contamination fraction ϵ is known and small

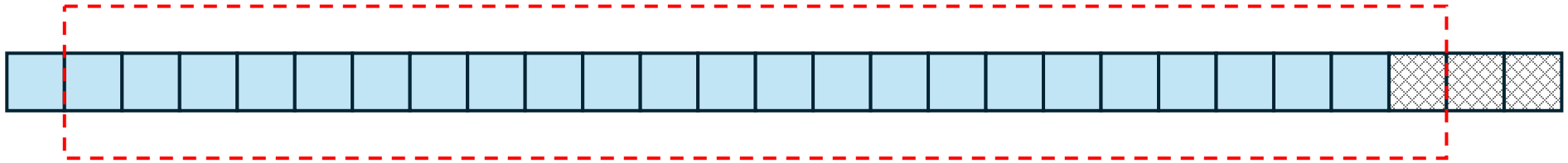


Main Results (Theoretical Algorithm)

Introduce "consistent subset"

Definition (consistent subset). A subset $S \subset \mathbf{X} = \{X_1, \dots, X_N\}$ is called a consistent subset regarding ϵ and a test $\phi: 2^S \to \{0,1\}$ if:

(i), $|S| \geq (1-\epsilon)N$,
(ii), $\phi(S') = \phi(S)$ for any $S' \subset S$ with $|S'| \geq (1-2\epsilon)N$.



Main Results (Theoretical Algorithm)

Definition (consistent subset). A subset $S \subset \mathbf{X} = \{X_1, \dots, X_N\}$ is called a consistent subset regarding ϵ and a test $\phi: 2^S \to \{0,1\}$ if:

(i), $|S| \geq (1-\epsilon)N$,
(ii), $\phi(S') = \phi(S)$ for any $S' \subset S$ with $|S'| \geq (1-2\epsilon)N$.

Why consistent subset? **It can help recover the testing result on the unobserved original samples $\widetilde{\mathbf{X}}$!**

Suppose the original samples $\widetilde{\mathbf{X}}$ is consistent w.r.t. a valid test ϕ_e, then:

- It is guaranteed that there exists at least one consistent subset within the observation $\mathbf{X}$! (At least $[N] \setminus C$)
- Therefore:

$$\phi_e(\mathbf{X_S}) = \phi_e(\mathbf{X_{S\cap[N]\setminus C}}) = \phi_e(\mathbf{X_{[N]\setminus C}}) = \phi_e(\widetilde{\mathbf{X}})$$

Main Results (Theoretical Algorithm)

Define the testing event $E_e := \left\{ \left\| P^\star \sum_{i \in T} X_i \right\|_2^2 - k^\star \, |T| \, \sigma^2 \geq c \, |T|^2 \, E^2(\epsilon, K, N, \sigma^2) \right\}$

any dimension $1 \leq k^\star \leq d$

optimal projection w.r.t. $k^\star$

Testing function $\phi_e = \mathbf{1}_{E_e}$

- $E^2(\epsilon, K, N, \sigma^2)$ is defined as $\sigma^2 \max \left\{ \frac{\sqrt{k}}{N}, \epsilon^2 \ln\left(\frac{1}{\epsilon}\right), \sqrt{\frac{\epsilon^2 \ln\left(\frac{1}{\epsilon}\right) k}{N}} \right\}$
- ϕ_e is a valid test with small errors and $\widetilde{\mathbf{X}}$ is consistent with ϕ_e w.h.p. when $\|P^\star \mu\|_2^2 \gtrsim E^2(\epsilon, K, N, \sigma^2)$

Requirement on ρ?

$$\|P^\star \mu\|_2^2 = \rho^2 - \|\mu - P^\star \mu\|_2^2 \geq \rho^2 - D_{k^\star}^2(K)$$

$$\rho^2 \gtrsim \underbrace{D_k^2(K)}_{\text{approximation gap}} + \underbrace{\sigma^2 \max \left\{ \frac{\sqrt{k}}{N}, \epsilon^2 \ln\left(\frac{1}{\epsilon}\right), \sqrt{\frac{\epsilon^2 \ln\left(\frac{1}{\epsilon}\right) k}{N}} \right\}}_{\text{testing gap}}$$

Main Results (Theoretical Algorithm)

Theoretical upper bound. For the hypothesis testing problem previously formulated, if the following condition is satisfied, then there exists a test ϕ_e to distinguish between H_0 and H_1 with uniformly small errors over $\mathcal{C}$ and μ.

$$\rho^2 \gtrsim D_k^2(K) + \sigma^2 \max\left\{\frac{\sqrt{k}}{N}, \epsilon^2 \ln\left(\frac{1}{\epsilon}\right), \sqrt{\frac{\epsilon^2 \ln\left(\frac{1}{\epsilon}\right) k}{N}}\right\}$$

Corollary. Optimizing over k in the upper bound above, we obtain the sharpest upper bound as:

$$\rho^2_{\text{critical}} \lesssim \sigma^2 \max\left\{\frac{\sqrt{\min\{k_1^\star, k_2^\star\}}}{N}, \epsilon^2 \ln\left(\frac{1}{\epsilon}\right), \sqrt{\frac{\epsilon^2 \ln\left(\frac{1}{\epsilon}\right) \min\{k_1^\star, k_2^\star\}}{N}}\right\}$$

where $k_1^\star$ and $k_2^\star$ are the first and second optimal dimensions.

Limitation: require scan of subset of X ⟶ not efficient!

Main Results (Polynomial-time Algorithm)

From first-order statistic to second-order statistic

- The theoretical algorithm is based on the sum (first-order)
- What if we base on the covariance (second-order)

Filtering technique (Diakonikolas et al., 2016)

Robust Estimators in High Dimensions
without the Computational Intractability

Ilias Diakonikolas* Gautam Kamath† Daniel M. Kane‡ Jerry Li § Ankur Moitra ¶
Alistair Stewart ‖

March 18, 2019

Abstract

We study high-dimensional distribution learning in an agnostic setting where an adversary is allowed to arbitrarily corrupt an ε-fraction of the samples. Such questions have a rich history spanning statistics, machine learning and theoretical computer science. Even in the most basic settings, the only known approaches are either computationally inefficient or lose dimension-dependent factors in their error guarantees. This raises the following question: Is high-dimensional agnostic distribution learning even possible, algorithmically?

In this work, we obtain the first computationally efficient algorithms with dimension-independent error guarantees for agnostically learning several fundamental classes of high-dimensional distributions: (1) a single Gaussian, (2) a product distribution on the hypercube, (3) mixtures of two product distributions (under a natural balancedness condition), and (4) mixtures of spherical Gaussians. Our algorithms achieve error that is independent of the dimension. and in many cases scales nearly-linearly with the fraction of adversarially corrupted samples. Moreover, we develop a general recipe for detecting and correcting corruptions in high-dimensions that may be applicable to many other problems.

*University of Southern California. Supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship. Part of this work was performed while the author was at the University of Edinburgh, supported in part by EPSRC grant EP/L021749/1 and a Marie Curie Career Integration Grant. diakonik@usc.edu

†Simons Institute for the Theory of Computing. Supported by NSF Award CCF-0953960 (CAREER) and ONR grant N00014-12-1-0999. This work was done in part while the author was an intern at Microsoft Research Cambridge, visiting the Simons Institute for the Theory of Computing, and a graduate student at MIT. g@csail.mit.edu

‡University of California, San Diego. Part of this work was performed while visiting the University of Edinburgh. dakane@cs.ucsd.edu

§Microsoft Research AI. Supposed by NSF CAREER Award CCF-1453261, a Google Faculty Research Award, and an NSF Fellowship. This work was done in part while the author was an intern at Microsoft Research Cambridge and a graduate student at MIT. jerrl@microsoft.com

¶Massachusetts Institute of Technology. Supported by NSF CAREER Award CCF-1453261, a grant from the MIT NEC Corporation, and a Google Faculty Research Award. moitra@mit.edu

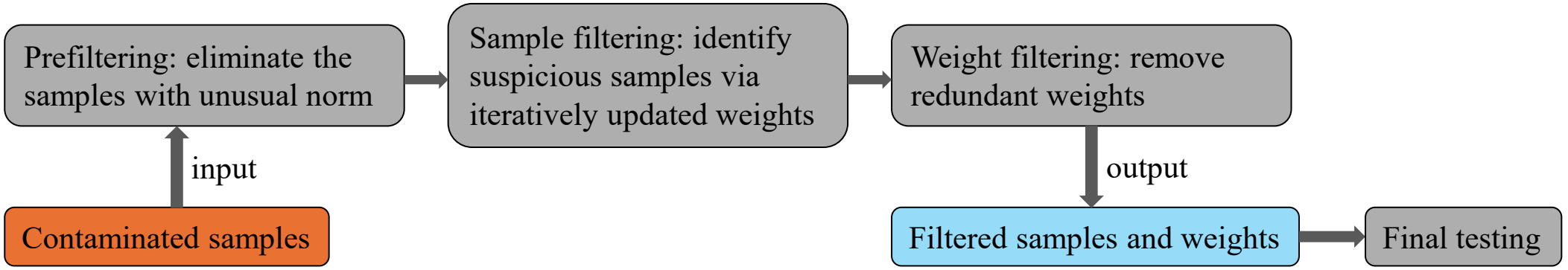
‖Web3 Foundation. Part of this work was performed while the author was at the University of Edinburgh and the University of Southern California. Research supported in part by EPSRC grant EP/L021749/1. stewart.al@gmail.com

Main Results (Polynomial-time Algorithm)

Filtering technique (Diakonikolas et al., 2016)

High-level steps:

- Assign each observed sample with an initial equal weight
- Detect the disturbance in the sample covariance matrix to identify suspicious sample and update the weights
- Test based on the filtered samples and weights



Main Results (Polynomial-time Algorithm)

Polynomial-time upper bound. For the hypothesis testing problem previously formulated, if the following condition is satisfied, then there exists a test ϕ_e to distinguish between H_0 and H_1 with uniformly small errors over $\mathcal{C}$ and μ.

$$\rho^2 \gtrsim D_k^2(K) + \sigma^2 \max\left\{ \frac{\epsilon \ln\left(\frac{N}{\alpha}\right)}{\sqrt{N}}, \epsilon^2 \ln\left(\frac{N}{\alpha}\right), \sqrt{\frac{\epsilon^2 k \ln\left(\frac{N}{\alpha}\right)}{N}}, \frac{\sqrt{k}\ln\left(\frac{1}{\alpha}\right)}{N} \right\}$$

Corollary. Optimizing over k in the upper bound above, we obtain the sharpest upper bound as:

$$\rho^2_{\text{critical}} \lesssim \sigma^2 \max\left\{ \frac{\epsilon \ln\left(\frac{N}{\alpha}\right)}{\sqrt{N}}, \epsilon^2 \ln\left(\frac{N}{\alpha}\right), \sqrt{\frac{\epsilon^2 \min\{k_1^\star, k_2^\star\} \ln\left(\frac{N}{\alpha}\right)}{N}}, \frac{\sqrt{\min\{k_1^\star, k_2^\star\}}\ln\left(\frac{1}{\alpha}\right)}{N} \right\}$$

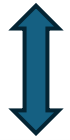
where again $k_1^\star$ and $k_2^\star$ are the first and second optimal dimensions.

Advantage: the filtering and testing process are finished in polynomial time!

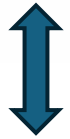
Main Results (Discussion)

Theorem. The synthesized lower bound, theoretical upper bound, and polynomial-time upper bound all match, except for a universal constant and logarithmic factors in $\alpha, N, \frac{1}{\epsilon}$.

Information-theoretic lower bound: $$\rho_c^2 \gtrsim \sigma^2 \max\left\{\frac{\sqrt{k_1^\star}}{N}, \epsilon\sqrt{\frac{k_2^\star}{N}}, \epsilon^2\right\}$$



Theoretical upper bound: $$\rho_c^2 \lesssim \sigma^2 \max\left\{\frac{\sqrt{\min\{k_1^\star, k_2^\star\}}}{N}, \epsilon^2 \ln\left(\frac{1}{\epsilon}\right), \sqrt{\frac{\epsilon^2 \ln\left(\frac{1}{\epsilon}\right)\min\{k_1^\star, k_2^\star\}}{N}}\right\}$$



Polynomial upper bound: $$\rho_c^2 \lesssim \sigma^2 \max\left\{\frac{\epsilon \ln\left(\frac{N}{\alpha}\right)}{\sqrt{N}}, \epsilon^2 \ln\left(\frac{N}{\alpha}\right), \sqrt{\frac{\epsilon^2 \min\{k_1^\star, k_2^\star\} \ln\left(\frac{N}{\alpha}\right)}{N}}, \frac{\sqrt{\min\{k_1^\star, k_2^\star\}} \ln\left(\frac{1}{\alpha}\right)}{N}\right\}$$

Main Results (Discussion)

Minimax rate for the constrained hypothesis testing problem with adversary:

$$\rho_c^2 \gtrsim \sigma^2 \max\left\{\frac{\sqrt{k_1^\star}}{N}, \epsilon\sqrt{\frac{k_2^\star}{N}}, \epsilon^2\right\}$$

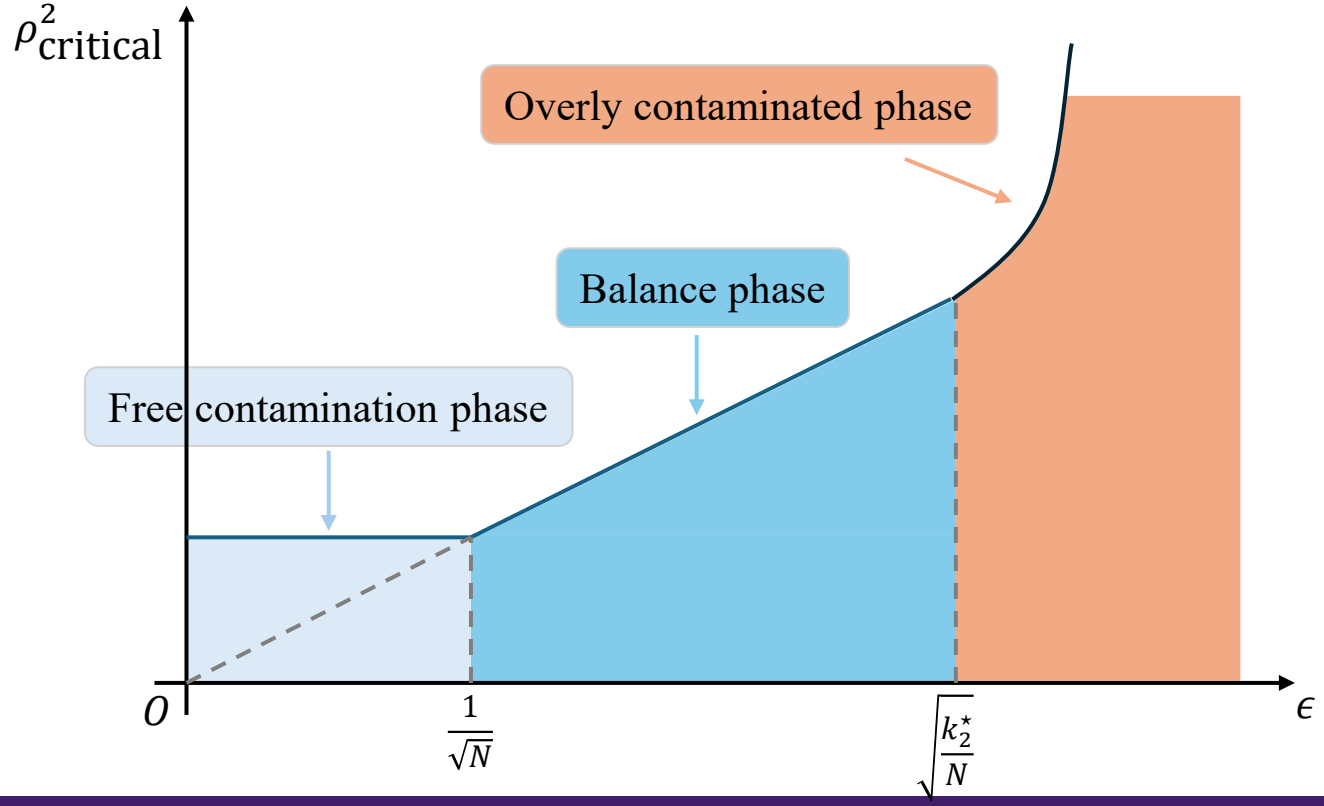
- Full dimension d does not appear --- completely replaced by $k_1^\star$ and $k_2^\star$!

- When $0 \le \epsilon \lesssim \frac{1}{\sqrt{N}}$: the first term dominates (Free contamination phase)
- When $\frac{1}{\sqrt{N}} \lesssim \epsilon \lesssim \sqrt{\frac{k_2^\star}{N}}$: the second term dominates (Balance phase)
- When $\epsilon \gtrsim \sqrt{\frac{k_2^\star}{N}}$: the third term dominates (Overly contaminated phase)

Main Results (Discussion)

Minimax rate for the constrained hypothesis testing problem with adversary:

$$\rho_c^2 \gtrsim \sigma^2 \max\left\{\frac{\sqrt{k_1^\star}}{N}, \epsilon\sqrt{\frac{k_2^\star}{N}}, \epsilon^2\right\}$$



Content

Introduction

- Problem Formulation
- Literature Review

Main Results

- Lower Bounds
- Upper Bounds
- Discussion

Experiments

- Numerical Simulation
- Discussion

Summary

- Summary
- Open Problems
- Plan for Future Works

Experiment

Experiment settings

- i.i.d. samples $\widetilde{\mathbf{X}} \coloneqq \{\widetilde{\mathrm{X}}_1, \ldots, \widetilde{\mathrm{X}}_\mathrm{N}\}$ from $\mathcal{N}(\mu, \mathbf{I_d})$
- Constraint $\mu \in K_e \coloneqq \left\{ v \middle| v \in \mathbb{R}^{\mathbb{d}}, \sum_{i=1}^{n} \frac{v_i^2}{a_i} \leq 1, a_i = \sqrt{d} \cdot i^{-2} \right\}$

Configuration for the parameters:

- Classical settings: $(N, d, \epsilon) \in \{200,2000\} \times \{5,10,20,50\} \times \{0.01,0.015,0.02,0.025\}$
- High-dimensional settings: $(N, d, \epsilon) \in \{100,500\} \times \{100,500\} \times \{0.01,0.015,0.02,0.025\}$

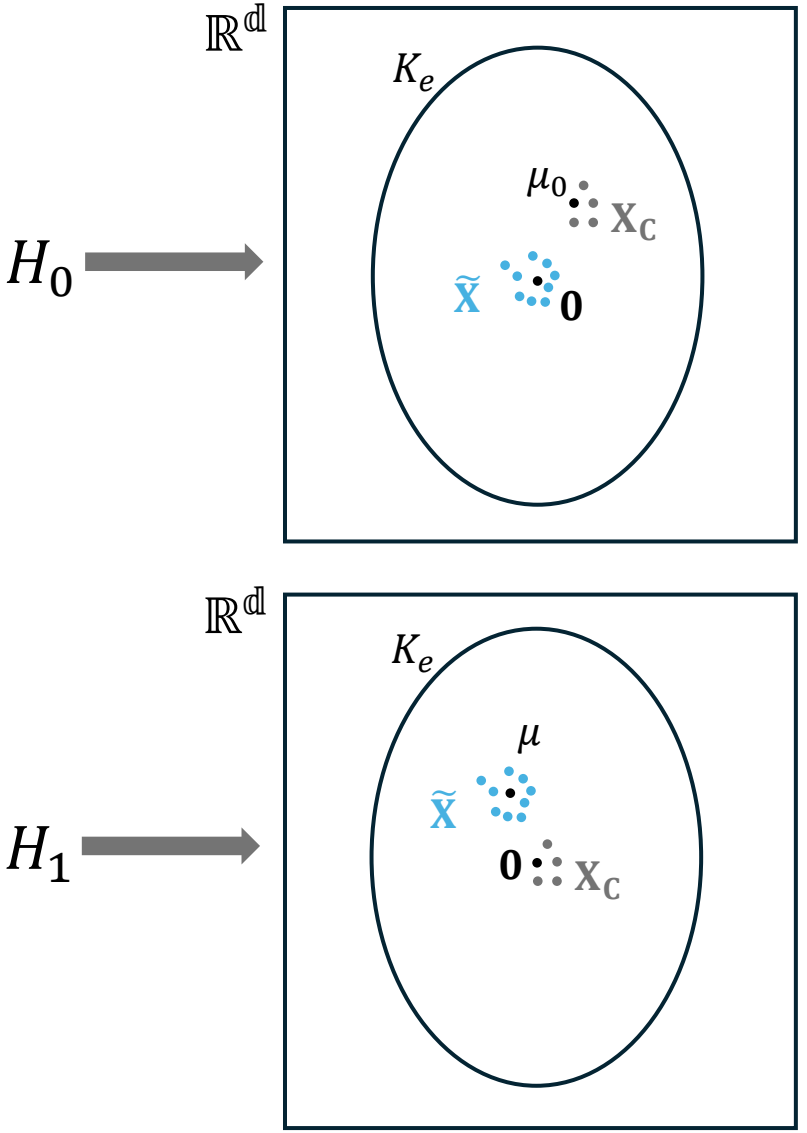
Experiment

Experiment settings

Configuration for the adversary $\mathcal{C}$:

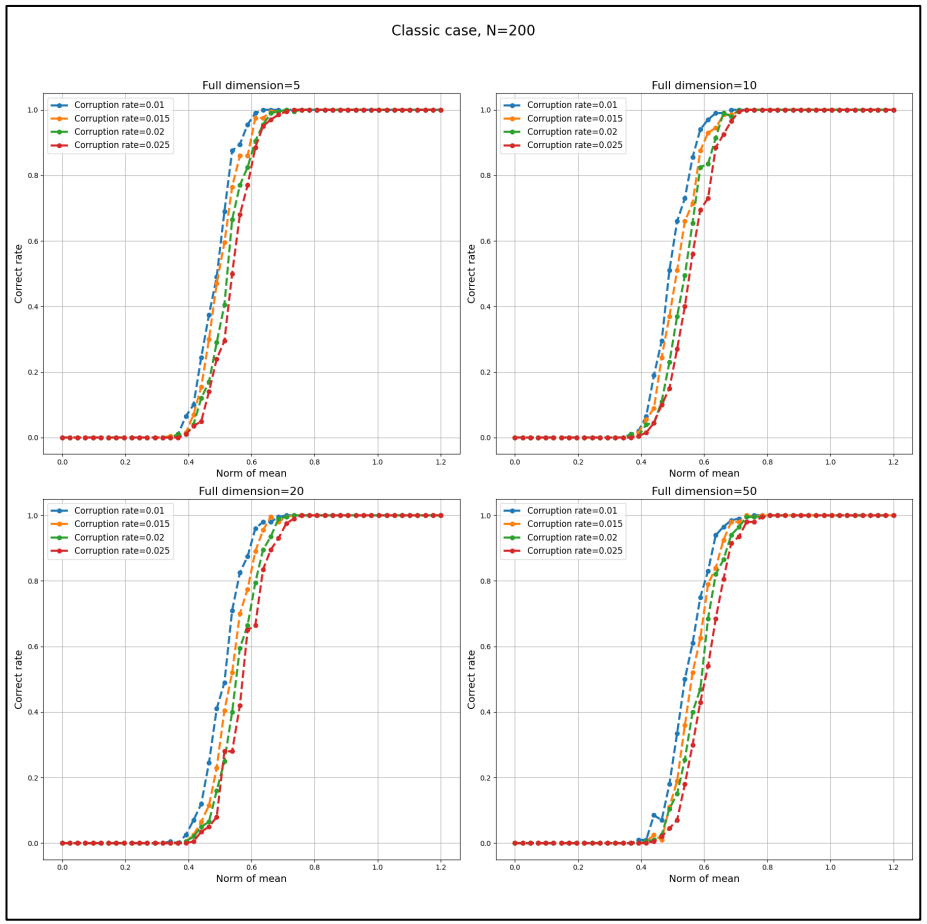
Algorithm 1: strategy 1 of the adversary $\mathcal{C}$

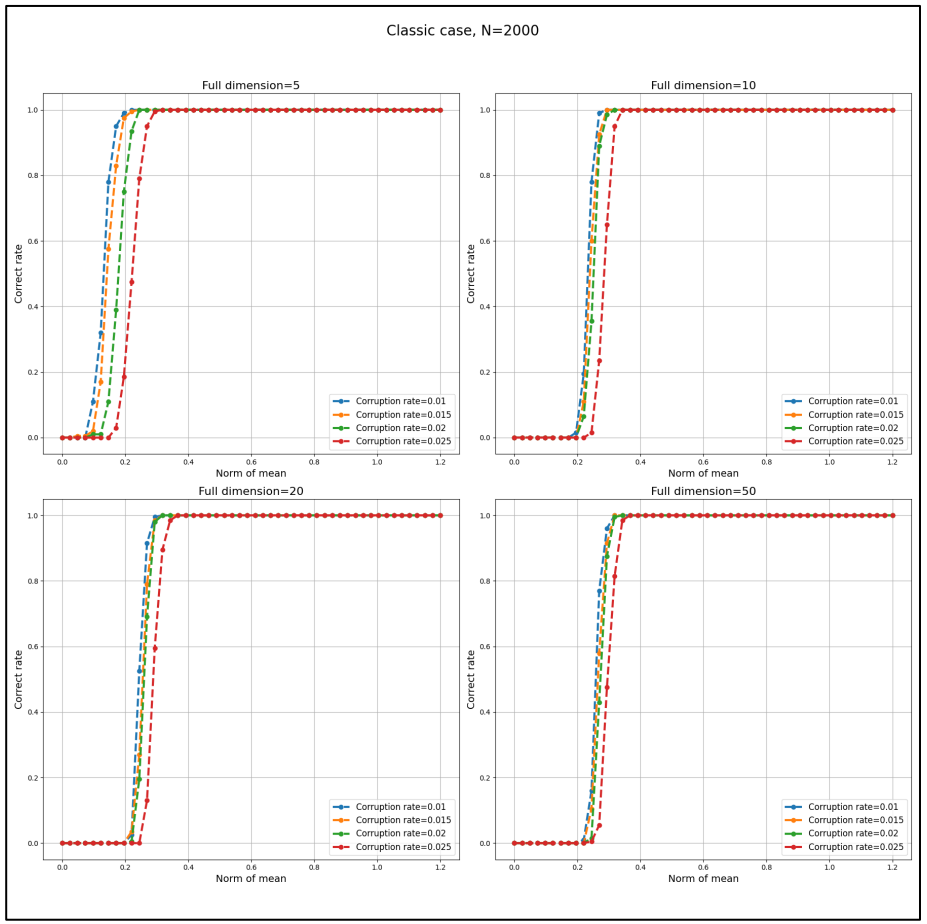
```
import X̃, N, d, σ = 1
Compute arr ∈ ℝ^N, arr_i = ‖X̃_i‖_2, 1 ≤ i ≤ N
if H_1 is true then
    C = indices of the largest εN entries in arr
    X_C = εN i.i.d. fake data from 𝒩(0, I_d)
    X̃_C = X_C
else
    import μ_0
    C = indices of the smallest εN entries in arr
    X_C = εN i.i.d. fake data from 𝒩(μ_0, I_d)
    X̃_C = X_C
return X = 𝒞(X̃)
```



Experiment

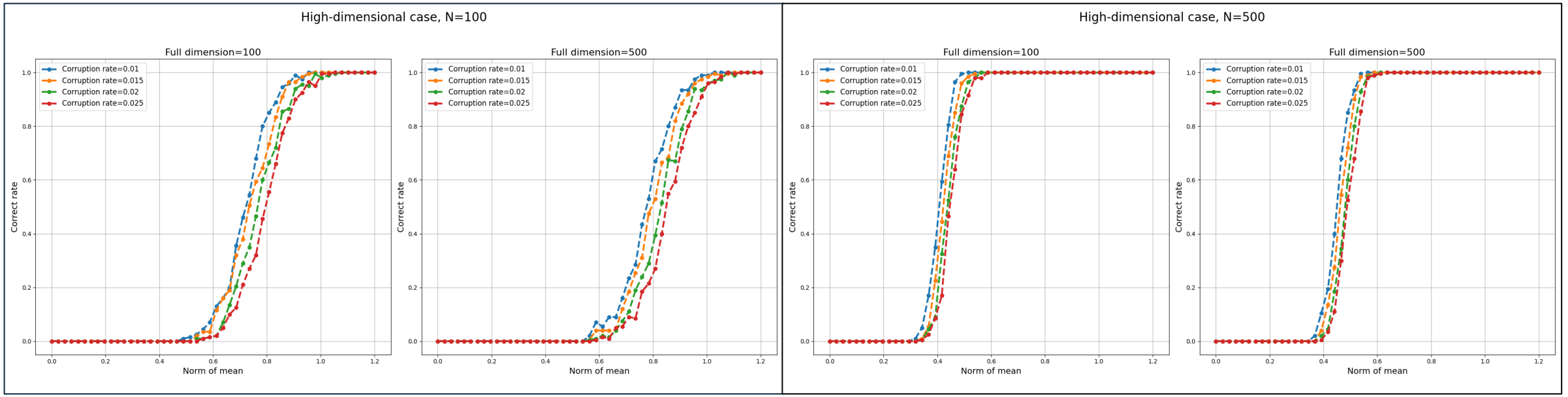
Experiment results (classic scenario)





Experiment

Experiment results (high-dimensional scenario)



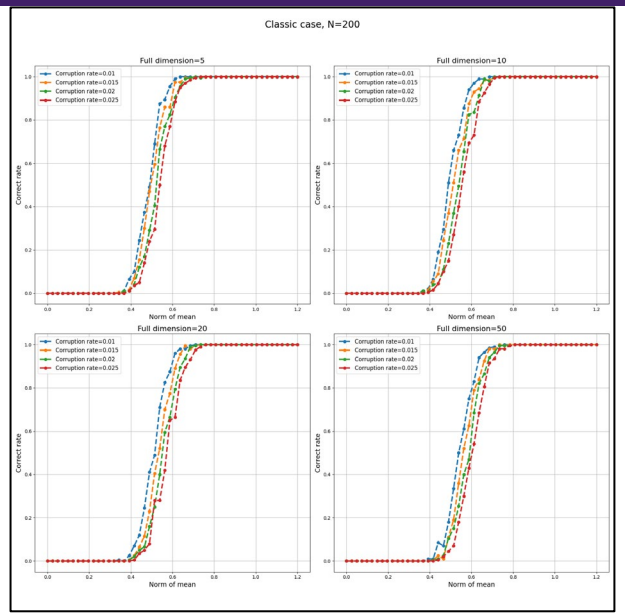
Experiment

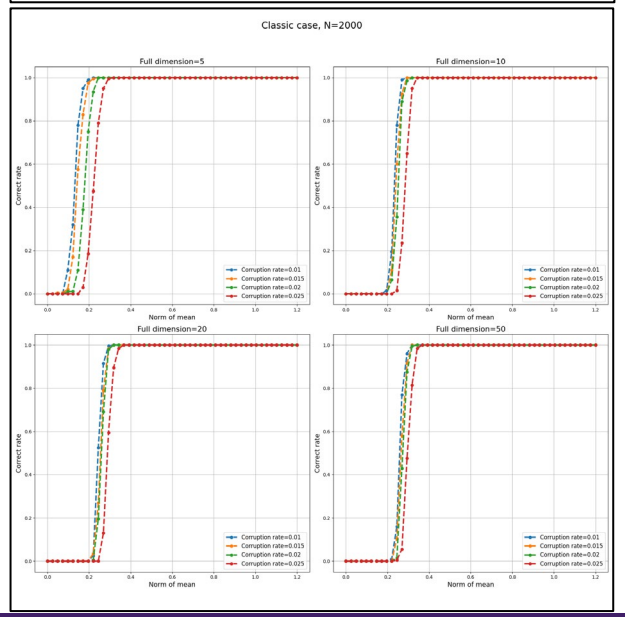
Discussion

- The algorithm performs reasonably well across different parameter configurations
- The strategy of the adversary is effective
- **The algorithm is less sensitive to d as d increases**

$$[5,10,20,50] \xrightarrow{\min\{k_1^\star, k_2^\star\}} [5,9,10,12]$$

- The algorithm is less sensitive to ϵ as N increases





Content

Introduction

- Problem Formulation
- Literature Review

Main Results

- Lower Bounds
- Upper Bounds
- Discussion

Experiments

- Numerical Simulation
- Discussion

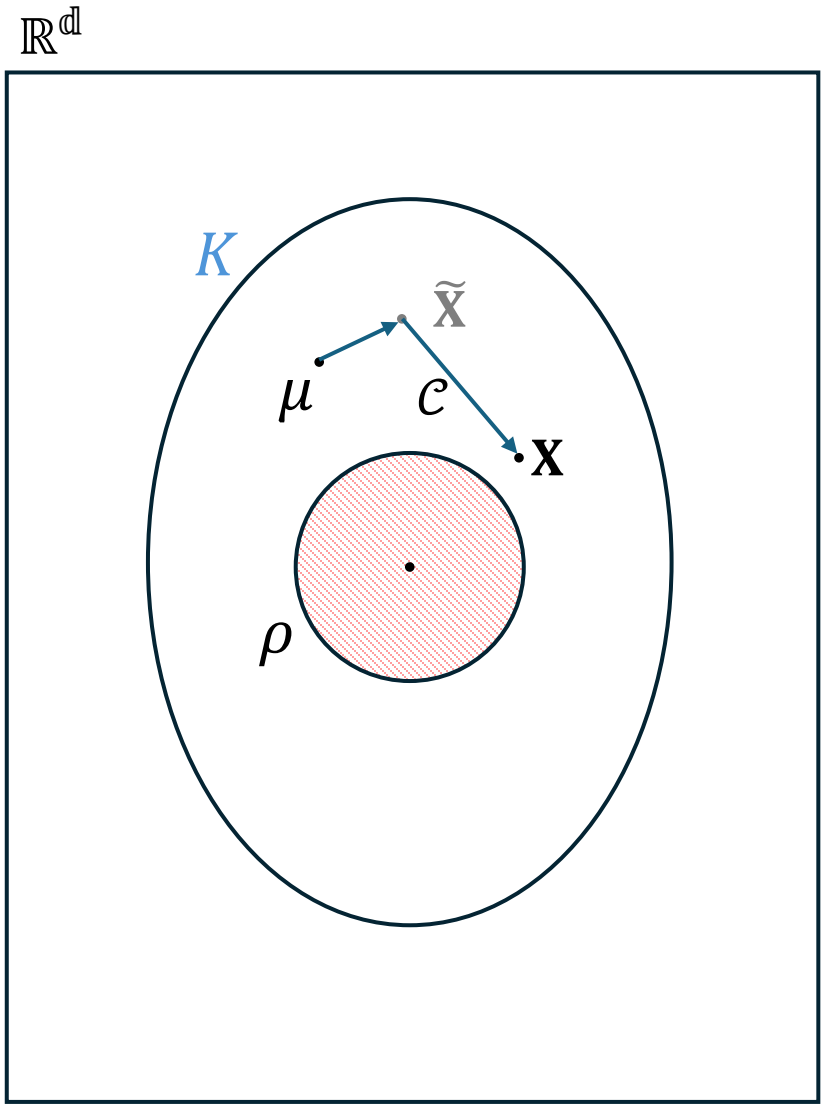
Summary

- Summary
- Open Problems
- Plan for Future Works

Summary

Conclusions & Takeaways

- Hypothesis testing problem with adversary and constraints
- Information-theoretic lower bounds
- Theoretical and polynomial-time algorithms
- Computational efficiency



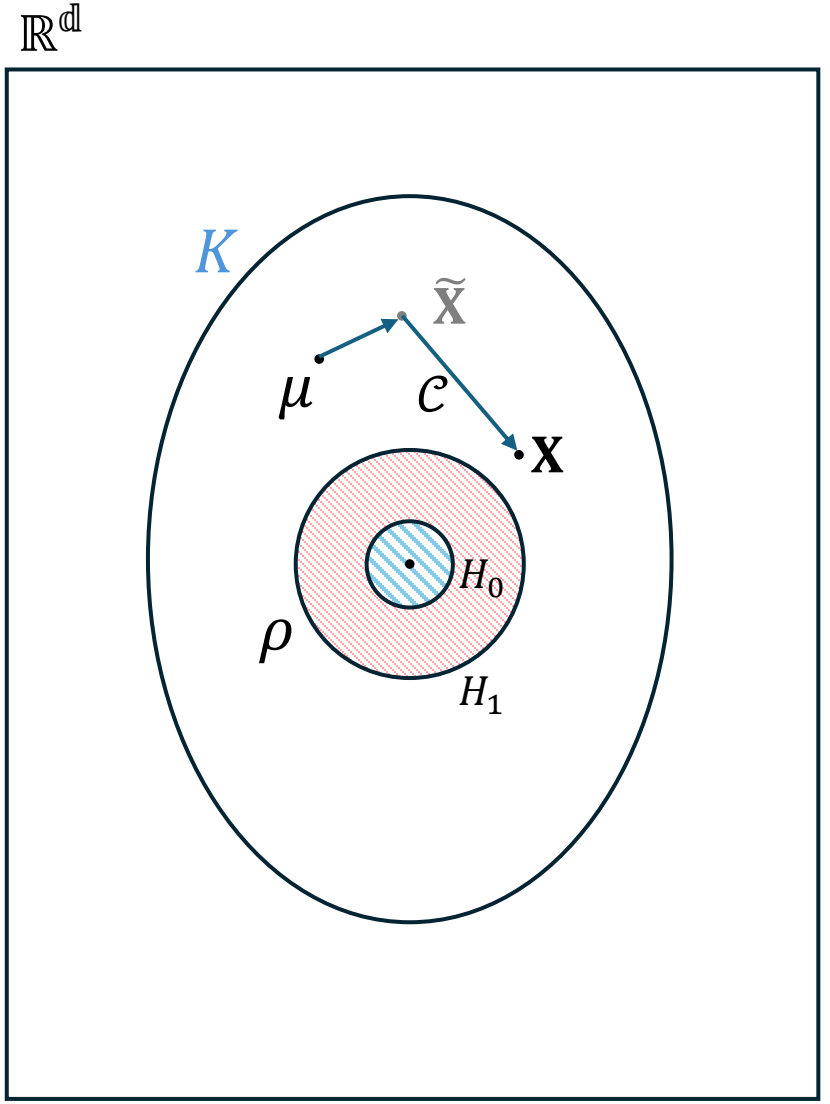
Summary

Open problems and possible future directions

- Kolmogorov width approximations (currently working)
- Imprecise H_0 (Kania, et al., 2025)

$$H_0: \|\mu\|_2 \leq \rho_0$$
$$H_1: \|\mu\|_2 \geq \rho_1$$

- ℓ_p norm separation ($p \geq 2$) (χ^p test)
- More general constraints (Convex? Star-shaped? …?)



Thank you!
Q & A

Appendix (ω-regularity)

Definition. Given a weight vector $\omega = (\omega_1, \dots, \omega_N) \in \mathbb{R}^{\mathbb{N}}$ and any $1 \le k^\star \le d$, samples $\mathbf{Y} \in \mathbb{R}^{\mathbb{N}\times\mathbb{d}}$ is said to be $(\epsilon, \beta_1, \beta_2)$-regular if for all subsets $S \in [N]$ with $|S| \le \epsilon N$, we have:

(i), $\left|\sum_{i\in S}\|Y_i\|_2^2 - |S|k^\star\right| \le c\beta_1$;

(ii), $\left|\left\|\sum_{i\in S}\sqrt{\omega_i}Y_i\right\|_2^2 - \|\omega_S\|_1 k^\star\right| \le c\beta_2$;

(iii), $\left|\left\langle\sum_{i\in S}\sqrt{\omega_i Y_i}, \sum_{j\in[N]}\sqrt{\omega_j Y_j}\right\rangle - \|\omega_S\|_1 k^\star\right| \le c\sqrt{N}\beta_1$

where $\omega_S \in \mathbb{R}^{\mathbb{N}}$ means the truncated ω on the subset S (fill truncated entries with zero).

Appendix (prefiltering, sample filtering, and weight filtering)

Algorithm 4: Prefiltering.

Set $\gamma_1 = c\left[\sqrt{k^\star \ln\left(\frac{N}{\alpha}\right)} + \ln\left(\frac{N}{\alpha}\right)\right]$, $\texttt{count} = 0, \texttt{i} = 0$
while $i < N$ **do**
  if $\left|\|Y_i\|_2^2 - k^\star\right| > \gamma_1$ **then**
    $\texttt{count} = \texttt{count} + 1$
    if $count > \epsilon N$ **then**
      return None
    Delete Y_i from $\mathbf{Y}$
  $\texttt{i} = \texttt{i} + 1$
return $\mathbf{Y}$

Appendix (prefiltering, sample filtering, and weight filtering)

Algorithm 5: Sample filtering when $N > k^\star$.

Set γ_2 as (C.13), and $\lambda = \left\|\mathbf{Y}^\top D(\omega)\mathbf{Y} - N\mathbf{I}_{k^\star}\right\|_2$. ($\omega$ is initialized as $\mathbf{1}$.)
while $\lambda \geq \gamma_2$ **do**
  Set v to be the unit singular vector associated with λ
  Compute $\tau_i = \langle v, Y_i\rangle^2 \mathbf{1}_{\{\omega_i > 0\}}$ for $1 \leq i \leq N$
  Set I be the smallest index such that $\sum_{i=1}^{I} \omega_i \geq 2\epsilon N$
  Update w according to (C.14)
  if $\|\omega\|_1 < N(1-2\epsilon)$ **then**
    return None
  Set $\lambda = \left\|\mathbf{Y}^\top D(\omega)\mathbf{Y} - N\mathbf{I}_{k^\star}\right\|_2$
return ω

$$\omega_i^{(t+1)} = \begin{cases} \left(1 - \frac{\tau_i^{(t)}}{\tau_1^{(t)}}\right)\omega_i^{(t)} & \text{if } i \leq I, \\ \omega_i^{(t)} & \text{if } i > I. \end{cases}$$

Appendix (prefiltering, sample filtering, and weight filtering)

Algorithm 6: Sample filtering when $N \le k^\star$.

Set γ_3 as (C.21), and $\lambda = \left\| \sqrt{D(\omega)}\mathbf{Y}\mathbf{Y}^\top\sqrt{D(\omega)} - k^\star D(\omega) \right\|_2$. ($\omega$ is initialized as $\mathbf{1}$.)
while $\lambda \ge \gamma_3$ **do**
  Set v to be the unit singular vector associated with λ
  Compute $\tau_i = \frac{v_i^2}{\omega_i}\mathbf{1}_{\{w_i>0\}}$
  Update w according to (C.22)
  if $\|\omega\|_1 < N(1-6\epsilon)$ **then**
    return None
  Set $\lambda = \left\| \sqrt{D(\omega)}\mathbf{Y}\mathbf{Y}^\top\sqrt{D(\omega)} - k^\star D(\omega) \right\|_2$
return ω

$$\omega_i = \left(1 - \frac{\tau_i}{\max_i \tau_i}\right)\omega_i, 1 \le i \le N$$

Appendix (prefiltering, sample filtering, and weight filtering)

Algorithm 7: Weight filtering.

1 Compute $\tau_i = \left| \left\langle \sqrt{\omega_i} Y_i, \sum_{j=1}^{N} \sqrt{\omega_j} Y_j \right\rangle - \omega_i k^\star \right| \mathbf{1}_{\{\omega_i > 0\}}, 1 \le i \le N$

2 Sort τ_i by desreasing order and find the indices $\{i_1, i_2, \ldots, i_{\epsilon N}\}$ corresponding to the first ϵN maximal τ_i

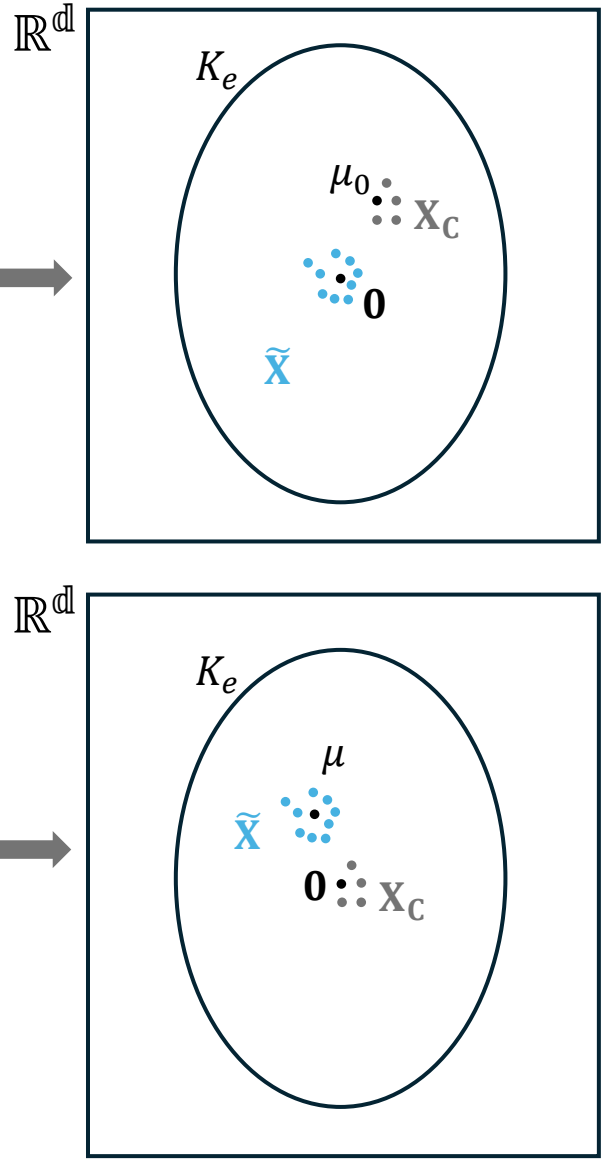
3 Set $\omega_{i_1}, \omega_{i_2}, \ldots, \omega_{i_{\epsilon N}}$ to zero

4 **return** ω

Appendix (alternative strategy for $\mathcal{C}$)

Algorithm 2: strategy 2 of the adversary $\mathcal{C}$

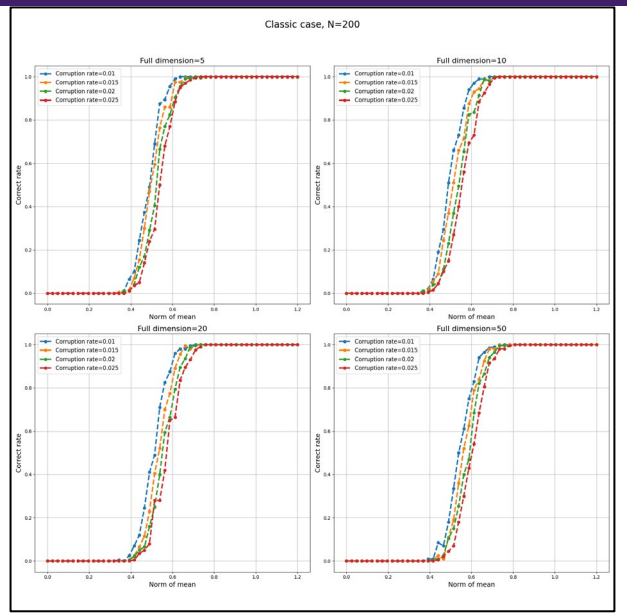
```
import X̃, N, d, σ = 1
if H_1 is true then
    Compute dist ∈ R^N, dist_i = ||X̃_i||_2, 1 ≤ i ≤ N
    C = indices of the largest εN entries in dist
    X_C = εN i.i.d. fake data from N(0, I_d)
    X̃_C = X_C
else
    Compute μ_0 according to the constraint type
    Compute dist ∈ R^N, dist_i = ||X̃_i − μ_0||_2
    C = indices of the largest εN entries in dist
    X_C = εN i.i.d. fake data from N(μ_0, I_d)
    X̃_C = X_C
return X = C(X̃)
```

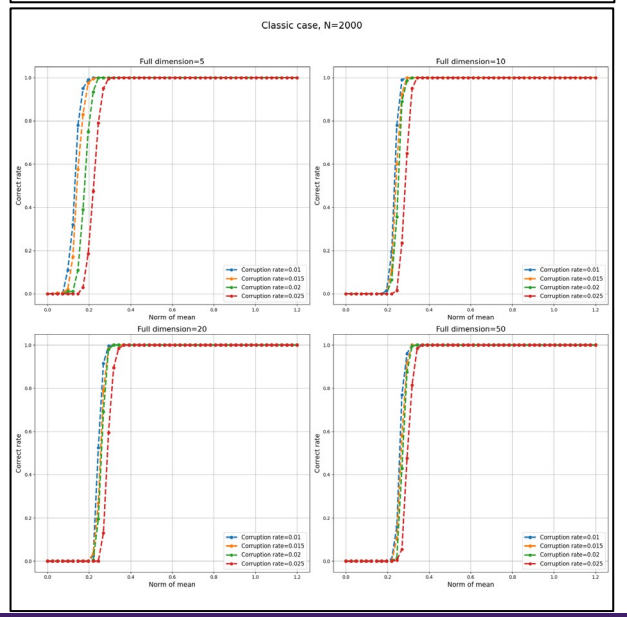


Experiment

Limitations

- Unable to verify the match between the lower and upper bounds
- Optimality of the adversary?



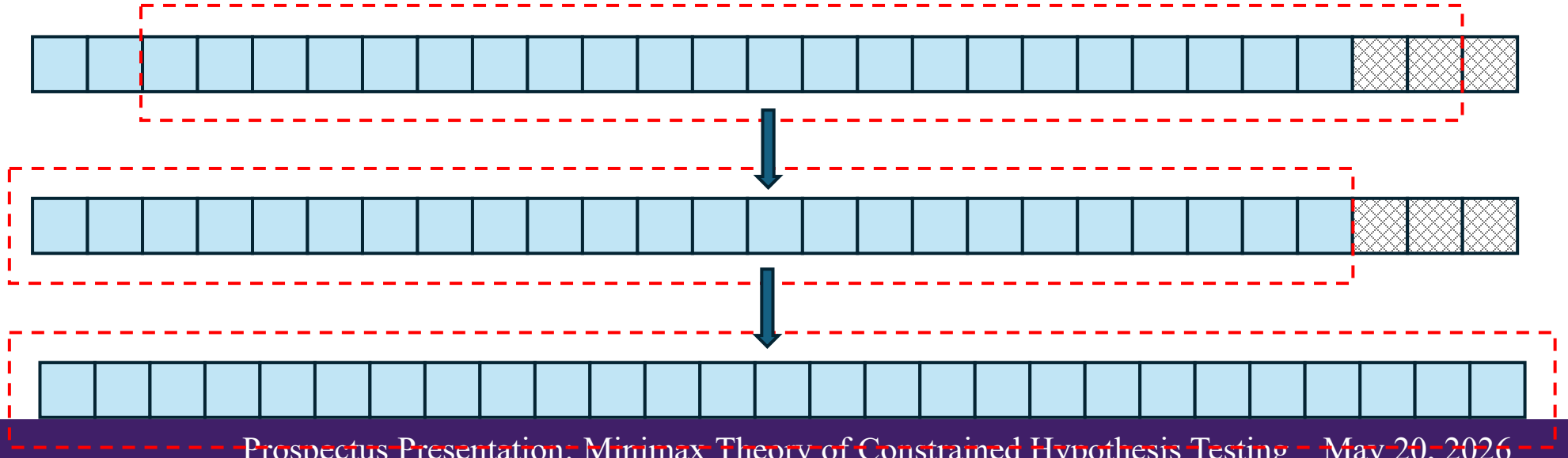


Main Results (Theoretical Algorithm)

Definition (consistent subset). A subset $S \subset \mathbf{X} = \{X_1, \ldots, X_N\}$ is called a consistent subset regarding ϵ and a test $\phi: 2^S \to \{0,1\}$ if:

(i), $|S| \geq (1-\epsilon)N$,
(ii), $\phi(S') = \phi(S)$ for any $S' \subset S$ with $|S'| \geq (1-2\epsilon)N$.

Motivation: consistent subset can help recover the testing result on the unobserved original samples!



What is Minimax Theory and Risk?

We interact with the general environment.

- Our decision or action yields rewards, penalties, etc

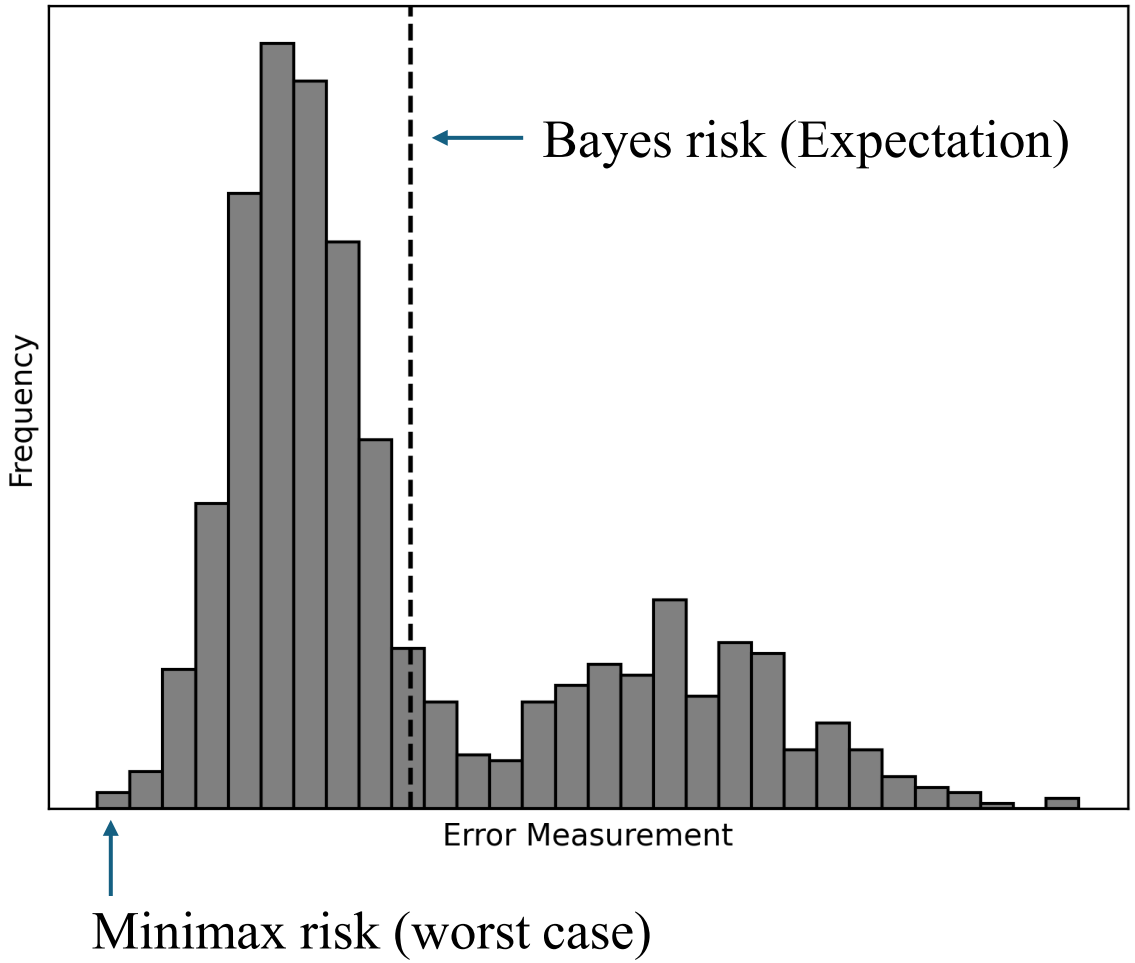
The randomness exists extensively.

- The response is not fixed

How to measure our performance?

- Model the response with a distribution, and calculate the average (Bayes risk)
- Only evaluate from the worst case (**minimax risk**)

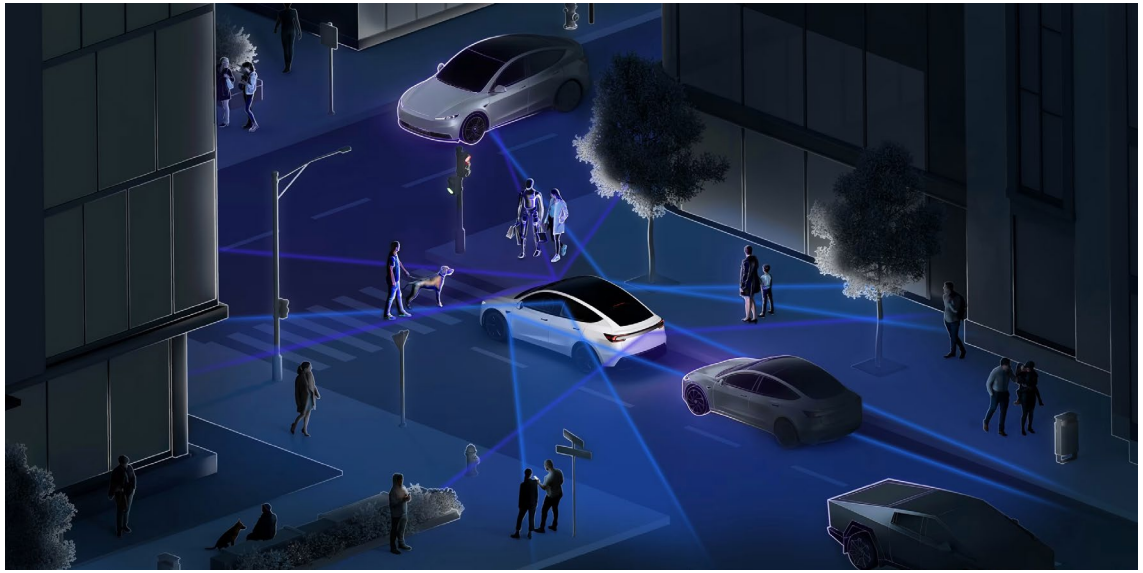
***In the worst case**, we will…*



Minimax risk is ubiquitous in real-world applications.

Robust Control

- LQR controller: noise is well-behaved and predictable
- Robust control (H_∞ control): zero-sum game with nature
- Nature's goal: find worst-case disturbance (wind, friction loss, sensor noise) that will **maximize** system's error.
- Our goal: design a control policy that **minimizes** the system's error, **even when** Nature plays its best move.
- Example: Autonomous driving, aerospace and flight control, etc.



Source: https://www.tesla.com/fsd/safety

Minimax risk is ubiquitous in real-world applications.

Robust Finance

- Should we assume that the market has some "desired" properties, or its behaviors can be very "wild"?
- The famous Black-Scholes model assumes that the volatility σ of a stock is constant and known. If volatility spikes unexpectedly, this can be disastrous
- In contrast, modern derivative dealers assume volatility is unknown but bounded within a range $[\sigma_{min}, \sigma_{max}]$. We want to find a dynamic hedging strategy that **minimizes the maximum possible hedging**

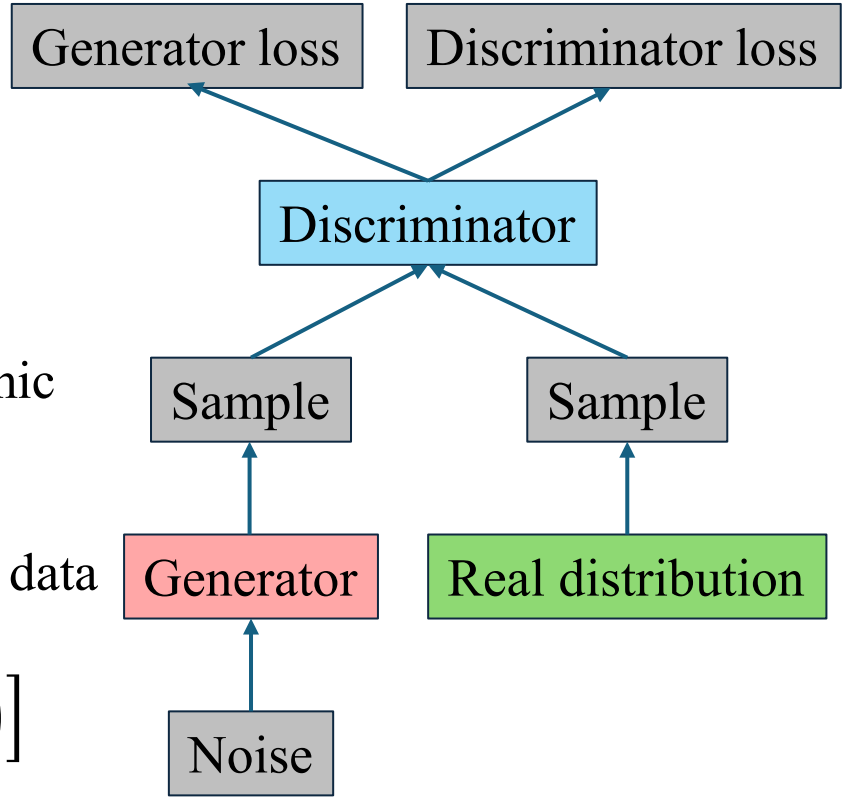
The front page of the *New York Times* on September 16, 2008. (Flickr/kbaird)

Source: https://www.americanprogress.org/article/think-again-the-crisis-from-nowhere/

Minimax risk is ubiquitous in real-world applications.

Generative Adversarial Network (GAN)

- Introduced by Ian Goodfellow et al. in 2014, GAN is a landmark in generative artificial intelligence
- The **generator** learns to generate plausible data from pure noise to mimic the real data
- The **discriminator** learns to distinguish between the real data and fake data
- $\min_G \max_D V(D,G) = \mathbb{E}_{x\sim\mathbb{P}_{data}}[\log D(x)] + \mathbb{E}_{z\sim\mathbb{P}_z}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$



Source: https://developers.google.com/machine-learning/gan/gan_structure