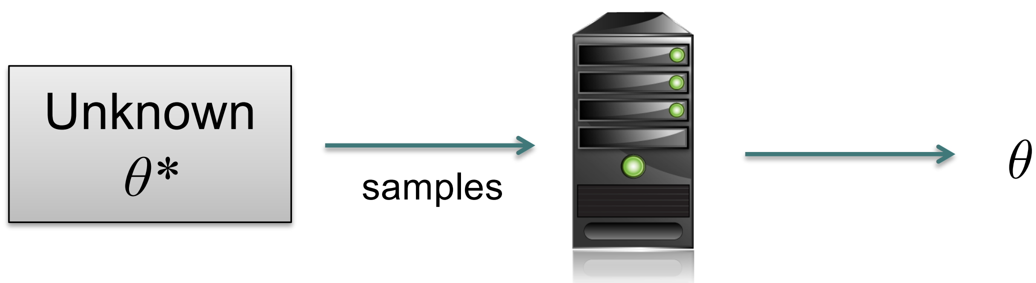
Information-Computation Tradeoffs *via* NGCA

Ilias Diakonikolas (UW Madison)

TTIC, June 2024

Can we develop learning algorithms that are ***robust*** to a ***constant*** fraction of ***corruptions*** in the data?

The Statistical Learning Problem



- *Input*: sample generated by a **statistical model** with unknown θ^*
- *Goal*: estimate parameters θ so that $\theta \approx \theta^*$

Question 1: Is there an *efficient* learning algorithm?

Main performance criteria:
- Sample size
- Running time
- **Robustness**

Question 2: Are there *tradeoffs* between these criteria?

(Outlier-) Robustness

Strong Contamination Model:

Let $\mathcal{F}$ be a family of statistical models.
We say that a set of N samples is ϵ -corrupted from $\mathcal{F}$ if it is generated as follows:

- N samples are drawn from an unknown $F \in \mathcal{F}$
- An omniscient adversary inspects these samples and changes arbitrarily an ϵ-fraction of them.

cf. Huber's contamination model [1964]

Observed Information-Computation (IC) Gaps

Problem 1: Robust Mean Estimation for $\mathcal{N}(\mu, I)$ in strong contamination model

- Information-theoretic: $O(\epsilon)$
- Computational: $O(\epsilon\sqrt{\log(1/\epsilon)})$ [D-Kane-Kamath-Li-Moitra-Stewart'16]

Problem 2: Robust *Sparse* Mean Estimation for $\mathcal{N}(\mu, I)$ in Huber's model

- Information-theoretic: $O(k \log(d)/\epsilon^2)$
- Computational: $O(k^2 \log(d)/\epsilon^2)$ [Li'17]

Problem 3: Robust covariance estimation for $\mathcal{N}(0, \Sigma)$ in spectral norm
- Information-theoretic: $O(d)$
- Computational: $\Omega(d^2)$ [D-Kane-Kamath-Li-Moitra-Stewart'16]

Are these observed information-computation gaps **inherent**?

How Do we Prove IC Tradeoffs?

- Unconditional hardness beyond reach. Need some assumptions.

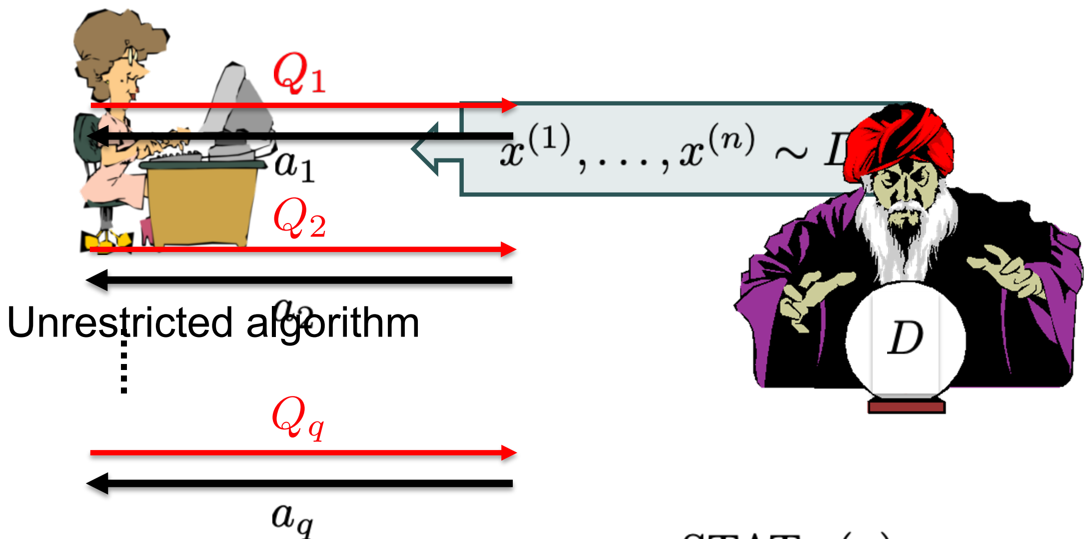
- **Reduction-based hardness**
  Efficient reduction from known "hard" problem
  General theory lacking for statistical problems

- **Restricted Models of Computation**
  - Statistical Query (SQ) Model
  - Low-degree Polynomial Tests
  - Sum-of-Squares Algorithms

This talk: SQ Model

STATISTICAL QUERY (SQ) MODEL [KEARNS'93]

SQ algorithm



tolerance of the query

$\mathrm{STAT}_D(\tau)$ returns a_i :

$|a_i - \mathbf{E}_{x \sim D}[Q_i(x)]| \leq \tau$

Statistical Query: $Q_i : X \to [-1, 1]$

Complexity measures
- Number of queries: q — Runtime
- Query tolerance: τ — Sample complexity

Power of SQ Algorithms

- **Restricted Model**: Can prove unconditional lower bounds.

- **Powerful Model**: Wide range of algorithmic techniques in ML are implementable using SQs:
  - PAC Learning: AC^0, decision trees, linear separators, boosting
  - Unsupervised Learning: stochastic convex optimization, moment-based methods, *k*-means clustering, EM, … [Feldman-Grigorescu-Reyzin-Vempala-Xiao, JACM'17]

- **Exceptions**: Gaussian elimination, lattice basis-reduction [D-Kane'22, Zadik-Song-Wein-Bruna'22]

- SQ Model $\approx$ Low-degree Polynomial Tests [Brennan-Bresler-Hopkins-Li-Schramm'21]

INTERPRETATION OF SQ *LOWER BOUNDS*

Suppose we have proved:

Any SQ algorithm for problem P

- either requires queries of **tolerance** at most τ
- or makes at least q **queries**.

Then we can interpret:

Any SQ algorithm* for problem P

- either requires at least $1/\tau^2$ **samples**
- or has **runtime** at least q.

SQ Lower Bound for Robust Mean Estimation

Theorem: Any SQ algorithm that learns an ϵ - corrupted Gaussian $\mathcal{N}(\mu, I)$ in the strong contamination model within error

$$o(\epsilon\sqrt{\log(1/\epsilon)})$$

requires either:

- SQ queries of accuracy $d^{-\omega(1)}$

or

- at least $d^{\omega(1)}$ many SQ queries.

Take-away: Any asymptotic improvement in error guarantee over filtering algorithm requires super-polynomial time.

SQ Lower Bound for Robust *Sparse* Mean Estimation

Theorem: Any SQ algorithm that learns an ϵ - corrupted Gaussian $\mathcal{N}(\mu, I)$ where is k-sparse within constant error requires either:

- $\Omega(k^2)$ samples

or

- at least $d^{k^{\Omega(1)}}$ many SQ queries.

Minimax sample complexity is $\Theta(k \log(d/k)/\epsilon^2)$

Take-away: Any asymptotic improvement in error guarantee over known efficient algorithms [Li'17, DKKPS'19,…] requires super-polynomial time.

SQ Lower Bound for Learning GMMs

Theorem: Any SQ algorithm that learns GMMs on $\mathbb{R}^d$ to constant total variation error requires either:

- $d^{\Omega(k)}$ samples

or

- at least $2^{d^{\Omega(1)}}$ many SQ queries.

even if the components are pairwise separated in total variation distance.

Minimax sample complexity is $\mathrm{poly}(d, k)$

Take-away: Computational complexity of learning separated GMMs is inherently exponential in **number of components**.

Non-Gaussian Component Analysis (NGCA)

Given samples from a distribution on $\mathbb{R}^d$, find a hidden "non-Gaussian" direction.

- Introduced in [Blanchard-Kawanabe-Sugiyama-Spokoiny-Muller'06].

- Studied extensively from algorithmic standpoint.
  [Kawanabe-Theis'06; Kawanabe-Sugiyama-Blanchard-Muller'07;
  Diederichs-Juditsky-Spokoiny-Schutte'10; Diederichs-Juditsky-Nemirovski-Spokoiny'13;
  Bean'14; Sasaki-Niu-Sugiyama'16; Virta-Nordhausen-Oja'16;
  Vempala-Xiao'11; Tan-Vershynin'18; Goyal-Shetty'19]

Non-Gaussian Component Analysis (NGCA): Definition

Definition: Let v be a unit vector in $\mathbb{R}^d$ and $A : \mathbb{R} \to \mathbb{R}_+$ be a pdf. We define $\mathbf{P}_v^A$ to be the distribution with v- projection equal to A and $v^\perp$- projection an independent standard Gaussian.

NGCA Problem: Given A that matches the first m moments with $\mathcal{N}(0,1)$:
Using i.i.d. samples from $\mathbf{P}_v^A$ where v is unknown, find the hidden direction v.

Generalizations: multi-dimensional, sparse, supervised, approximate moment-matching

NGCA captures interesting instances of several (robust) learning tasks

- Learning Gaussian Mixtures [D-Kane-Stewart’17, D-Kane-Pittas-Zarifis’23, D-Karmalkar-Pang-Potechin’24]
- Robust mean and covariance estimation [D-Kane-Stewart’17]
- Robust sparse mean estimation, sparse PCA [D-Kane-Stewart’17, D-Stewart’18]
- Robust linear regression [D-Kong-Stewart’19]
- List-decodable learning [D-Kane-Stewart’18, D-Kane-Pensia-Pittas-Stewart’21]
- Adversarially robust PAC learning [Bubeck-Price-Razenshteyn’18]
- Agnostic Learning [Goel-Gollakota-Klivans’20, D-Kane-Zarifis’20, D-Kane-Pittas-Zarifis’21]
- Learning LTFs with (Semi)-random Noise [D-Kane’20, Nasser-Tiegal’22, D-J.D.-Kane-Wang-Zarifis’23]
- Learning (Very Simple) NNs and Generative Models [D-Kane-Kontonis-Zarifis’20 Chen-Li-Li’22, Song’24]
- Learning Mixtures of LTFs [D-Kane-Sun’23]
- Learning Intersections of Halfspaces [Tiegel’24]
- Truncated statistics [D-Kane-Pittas-Zarifis’24]
- …

Informal Lower Bound Result

Fact: Non-Gaussian Component Analysis
- Can be solved with $\mathrm{poly}(d, m)$ samples.
- All known efficient algorithms require at least $d^{\Omega(m)}$ samples (and time).

Informal Theorem: For *any* "nice" univariate distribution A matching its first m moments with the standard Gaussian, any* algorithm that solves NGCA
- either draws at least $d^{\Omega(m)}$ samples
- or has runtime $2^{d^{\Omega(1)}}$.

*holds for any **Statistical Query (SQ)** algorithm

[D-Kane-Stewart, FOCS'17; …; D-Kane-Ren-Sun, NeurIPS'23]

General Methodology For SQ Lower Bounds

Hypothesis Testing Problem: Given access to a distribution D on $\mathbb{R}^d$ with promise that

- either $D = D_0$
- or D is selected randomly from $\mathcal{D} = \{D_u\}_{u \in S}$ according to prior μ

the goal is to distinguish between the two cases.

Pairwise correlation: $\chi_{D_0}(p, q) = \mathbf{E}_{x \sim D_0}[(p/D_0)(x)(q/D_0)(x)] - 1$

Theorem [FGRVX'17]: Suppose there exists a "large" set of distributions in $\mathcal{D}$ with "small" pairwise correlation with respect to D_0 . Then any SQ algorithm for hypothesis testing task:

- either requires at least one "high-accuracy" query
- or requires a "large" number of queries.

Statistical Query Hardness of NGCA

Testing Version of NGCA: Given access to a distribution D on $\mathbb{R}^d$ with the promise that

- either $D = \mathcal{N}(0, I)$
- or $D = \mathbf{P}_v^A$, where v is a uniformly random unit vector

the goal is to distinguish between the two cases.

Main Theorem [D-Kane-Stewart'17]

Suppose that A matches its first m moments with $\mathcal{N}(0,1)$ and $\chi^2(A, \mathcal{N}(0,1)) < \infty$.

Any SQ algorithm for the testing version of NGCA:

- either requires a query of tolerance at most $d^{-\Omega(m)}\ \chi^2(A, \mathcal{N}(0,1))^{1/2}$
- or requires at least $2^{d^{\Omega(1)}}$ many queries.

Intuition: Why is NGCA "Hard"?

Claim 1: Low-degree moments do not help.

- Degree at most m moment tensor of $\mathbf{P}_v^A$ identical to that of $\mathcal{N}(\mathbf{0}, I_d)$

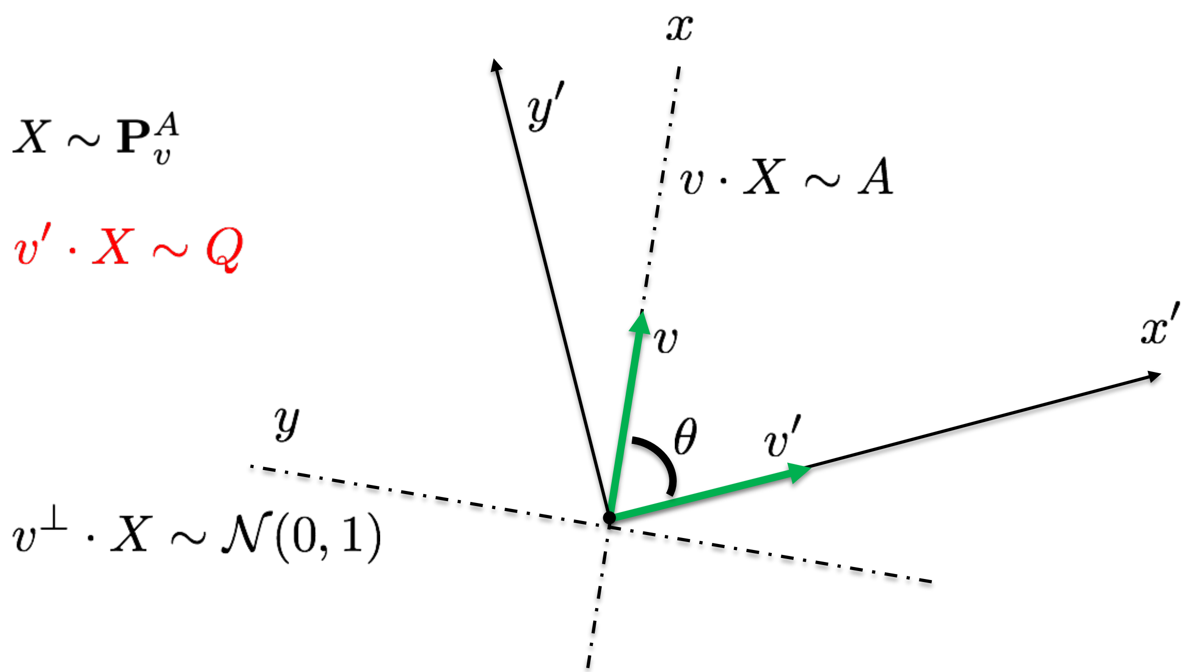
Claim 2: Random projections do not help.

Distinguishing requires exponentially many random projections.

KEY LEMMA: RANDOM PROJECTIONS ARE *ALMOST* GAUSSIAN

Key Lemma: Let Q be the distribution of $v' \cdot X$, where $X \sim \mathbf{P}_v^A$. Then, we have that:

$$\chi^2(Q, \mathcal{N}(0,1)) \leq (v \cdot v')^{2(m+1)} \chi^2(A, \mathcal{N}(0,1))$$



$$\boxed{Q = U_\theta(A)}$$

Ornstein-Uhlenbeck operator

SQ Lower Bound: Proof Overview

Want exponentially many $\mathbf{P}_v^A$'s that are nearly uncorrelated.

- Pick set $\mathcal{V}$ of near-orthogonal unit vectors. Can get $|\mathcal{V}| = 2^{d^{\Omega(1)}}$
- Have

$$\chi_{\mathcal{N}(\mathbf{0},I_d)}(\mathbf{P}_v^A, \mathbf{P}_{v'}^A) = \chi_{\mathcal{N}(0,1)}(A, U_\theta A) \leq |\cos^{m+1}(\theta)|\chi^2(A, \mathcal{N}(0,1))$$

$\blacksquare$

RECIPE FOR SQ HARDNESS RESULTS

Main Theorem [D-Kane-Stewart'17]
Suppose that A matches its first m moments with $\mathcal{N}(0,1)$ and $\chi^2(A, \mathcal{N}(0,1)) < \infty$.
Any SQ algorithm for the testing version of NGCA:

- either requires a query of tolerance at most $d^{-\Omega(m)}\ \chi^2(A, \mathcal{N}(0,1))^{1/2}$
- or requires at least $2^{d^{\Omega(1)}}$ many queries.

Recipe. Encode Π as a NGCA instance:

- Construct moment-matching distribution A such that $\mathbf{P}_v^A$ is a **valid instance** of Π .
- Match **as many low-degree moments as possible**.

Moment-Matching for Robust Mean Estimation

Lemma: There exists a univariate distribution A such that:

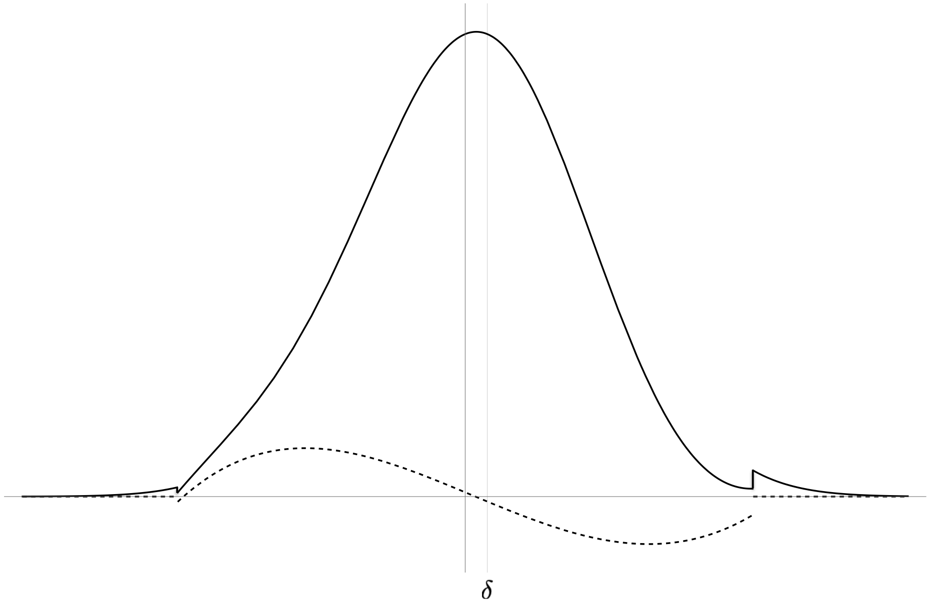
- A agrees with $\mathcal{N}(0,1)$ on the first m moments
- A satisfies $d_{\mathrm{TV}}(A, N(\delta,1)) \leq O(\delta m^2/\sqrt{\log(1/\delta)})$

Proof Idea:

- Take $C = \Theta(\sqrt{\log(1/\delta)})$
- Define

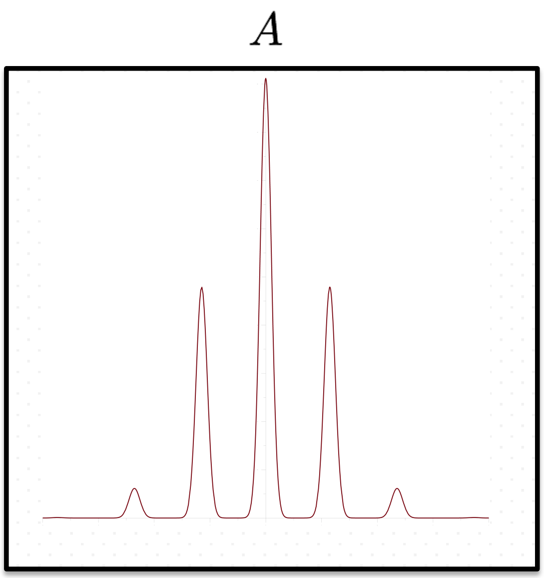
$$A(x) = \begin{cases} G(x-\delta), \ x \notin [-C, C] \\ G(x-\delta) + p(x), \ x \in [-C, C] \end{cases}$$

where p is degree-m moment-matching polynomial.



Moment-Matching for Learning GMMs

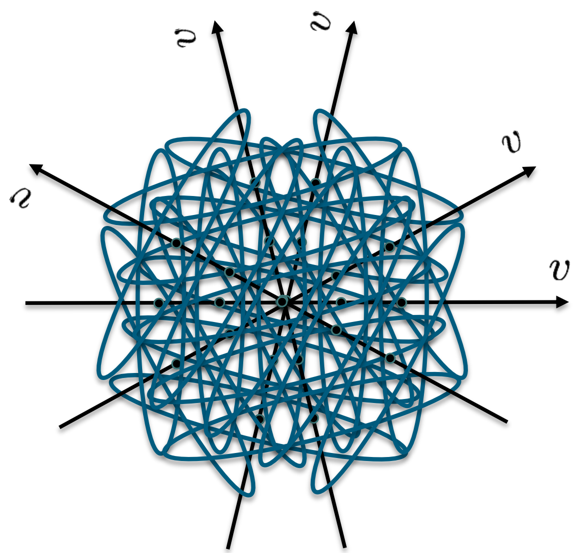
Lemma: There exists a univariate k-GMM A with nearly non-overlapping components such that: A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.



Proof Idea:

- Construct discrete distribution B with support k matching its first $2k$-1 moments with $\mathcal{N}(0,1)$.
- Rescale B and add a "skinny" Gaussian to get A.

SQ Hard Instances for GMMs: Parallel Pancakes



SQ Hardness For Wide Range of Problems

NGCA captures SQ hard instances of several well-studied learning tasks

- Learning Gaussian Mixtures [D-Kane-Stewart'17, D-Kane-Pittas-Zarifis'23, D-Karmalkar-Pang-Potechin'24]
- Robust mean and covariance estimation [D-Kane-Stewart'17]
- Robust sparse mean estimation, sparse PCA [D-Kane-Stewart'17, D-Stewart'18]
- Robust linear regression [D-Kong-Stewart'19]
- List-decodable learning [D-Kane-Stewart'18, D-Kane-Pensia-Pittas-Stewart'21]
- Adversarially robust PAC learning [Bubeck-Price-Razenshteyn'18]
- Agnostic Learning [Goel-Gollakota-Klivans'20, D-Kane-Zarifis'20, D-Kane-Pittas-Zarifis'21]
- Learning LTFs with (Semi)-random Noise [D-Kane'20, Nasser-Tiegal'22, D-J.D.-Kane-Wang-Zarifis'23]
- Learning (Very Simple) NNs and Generative Models [D-Kane-Kontonis-Zarifis'20 Chen-Li-Li'22, Song'24]
- Learning Mixtures of LTFs [D-Kane-Sun'23]
- Learning Intersections of Halfspaces [Tiegel'24]
- Truncated statistics [D-Kane-Pittas-Zarifis'24]
- …

OPEN PROBLEMS

NGCA leads to wide range of hardness results in SQ model

Open Problem 1: Alternative evidence of hardness?

Already known for special cases (reductions):

- Robust sparse mean estimation [Brennan-Bresler'20]
- Learning GMMs [Bruna-Regev-Song-Tang'21]
- Learning with Semi-random Noise [D-Kane-Panurangsi-Ren'22, D-Kane-Ren'23]

SQ hard instances are computationally hard

Open Problem 2: How general is this phenomenon?

Open Problem 3: Prove SoS lower bounds for NGCA.