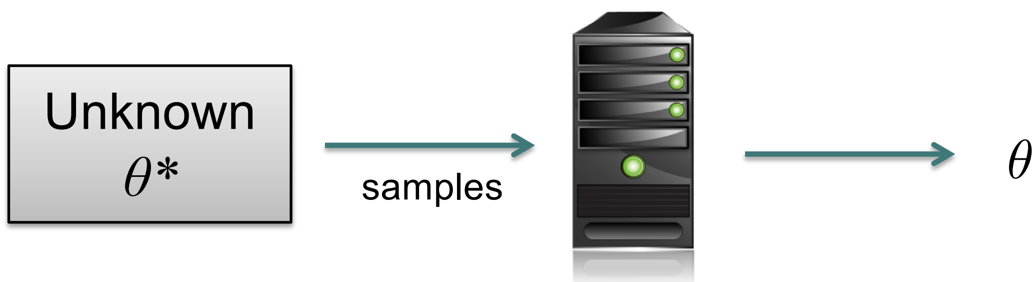
Computational-Statistical Tradeoffs and Open Problems

Ilias Diakonikolas (USC)

STOC 2019 Tutorial
June 2019

The Statistical Learning Problem



- *Input*: sample generated by a **probabilistic 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?

OUTLINE

Part I: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

- **Part II: Future Directions**

OUTLINE

Part I: Computational Limits to Robust Estimation

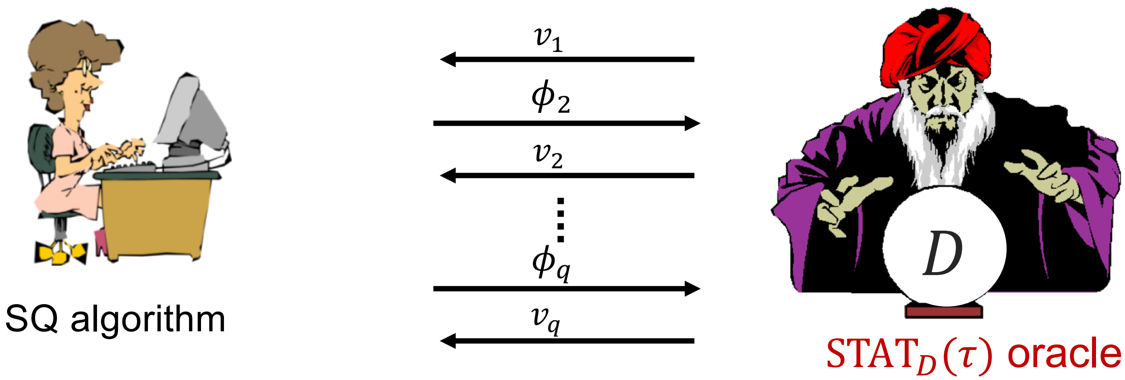
- **Statistical Query Learning Model**
- Our Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

- **Part II: Future Directions**

Statistical Queries [Kearns'93]

$x_1, x_2, \dots, x_m \sim D$ over X

Statistical Queries [Kearns'93]



$\phi_1 : X \to [-1,1] \qquad |v_1 - \mathbf{E}_{x\sim D}[\phi_1(x)]| \le \tau$

τ is tolerance of the query; $\tau = 1/\sqrt{m}$

Problem $P \in \mathrm{SQCompl}(q, m)$:
If exists a SQ algorithm that solves P using q queries to $\mathrm{STAT}_D(\tau = 1/\sqrt{m})$

Power of SQ Learning Algorithms

- **Restricted Model**: Hope to prove unconditional computational 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]

- **Only known exception**: Gaussian elimination over finite fields (e.g., learning parities).

- For all problems in this talk, strongest known algorithms are SQ.

METHODOLOGY FOR PROVING SQ LOWER BOUNDS

Statistical Query Dimension:

- Fixed-distribution PAC Learning
  [Blum-Furst-Jackson-Kearns-Mansour-Rudich'95; …]

- General Statistical Problems
  [Feldman-Grigorescu-Reyzin-Vempala-Xiao'13, …, Feldman'16]

- Pairwise correlation between D_1 and D_2 with respect to D:

$$\chi_D(D_1, D_2) := \int_{\mathbb{R}^d} D_1(x) D_2(x) / D(x) dx - 1$$

- **Fact**: Suffices to construct a large set of distributions that are *nearly* uncorrelated.

OUTLINE

Part I: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- **Our Results**
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

- **Part II: Future Directions**

GENERIC SQ LOWER BOUND CONSTRUCTION

General Technique for SQ Lower Bounds:
Leads to Tight Lower Bounds
for a range of High-dimensional Estimation Tasks

Concrete Applications of our Technique:

- Robustly Learning Mean and Covariance
- Learning Gaussian Mixture Models (GMMs)
- Statistical-Computational Tradeoffs (e.g., sparsity)
- Robustly Testing a Gaussian

SQ Lower Bound for Robust Mean Estimation

Theorem: Suppose $d \geq \mathrm{polylog}(1/\epsilon)$. Any SQ algorithm that learns an ϵ - corrupted Gaussian $\mathcal{N}(\mu, I)$ in the strong contamination model within error

$$O(\epsilon\sqrt{\log(1/\epsilon)}/M)$$

requires either:

- SQ queries of accuracy $d^{-M/6}$

or

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

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

SQ Lower Bounds for Learning *Separated* GMMs

Theorem: Suppose that $d \geq \mathrm{poly}(k)$. Any SQ algorithm that learns *separated* k-GMMs over $\mathbb{R}^d$ to constant error requires either:

- SQ queries of accuracy $d^{-k/6}$

or

- At least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

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

APPLICATIONS: CONCRETE SQ LOWER BOUNDS

Learning Problem	Upper Bound	SQ Lower Bound
Robust Gaussian Mean Estimation	Error: $O(\epsilon \log^{1/2}(1/\epsilon))$ [DKKLMS'16]	Runtime Lower Bound: $d^{\mathrm{poly}(M)}$ for factor M improvement in error.
Robust Gaussian Covariance Estimation	Error: $O(\epsilon \log(1/\epsilon))$ [DKKLMS'16]	
Learning k-GMMs (without noise)	Runtime: $d^{g(k)}$ [MV'10, BS'10]	Runtime Lower Bound: $d^{\Omega(k)}$
Robust k-Sparse Mean Estimation	Sample size: $\tilde{O}(k^2 \log d)$ [BDLS'17]	If sample size is $O(k^{1.99})$ runtime lower bound: $d^{k^{\Omega(1)}}$
Robust Covariance Estimation in Spectral Norm	Sample size: $\tilde{O}(d^2)$ [DKKLMS'16]	If sample size is $O(d^{1.99})$ runtime lower bound: $2^{d^{\Omega(1)}}$

OUTLINE

Part I: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- **Generic Lower Bound Technique**
- Applications: Robust Mean Estimation & Learning GMMs

- **Part II: Future Directions**

GENERAL RECIPE FOR SQ LOWER BOUNDS

- **Step #1:** Construct distribution $\mathbf{P}_v$ that is standard Gaussian in all directions except v.
- **Step #2:** Construct the univariate projection in the v direction so that it matches the first *m* moments of $\mathcal{N}(0,1)$
- **Step #3:** Consider the family of instances $\mathcal{D} = \{\mathbf{P}_v\}_v$

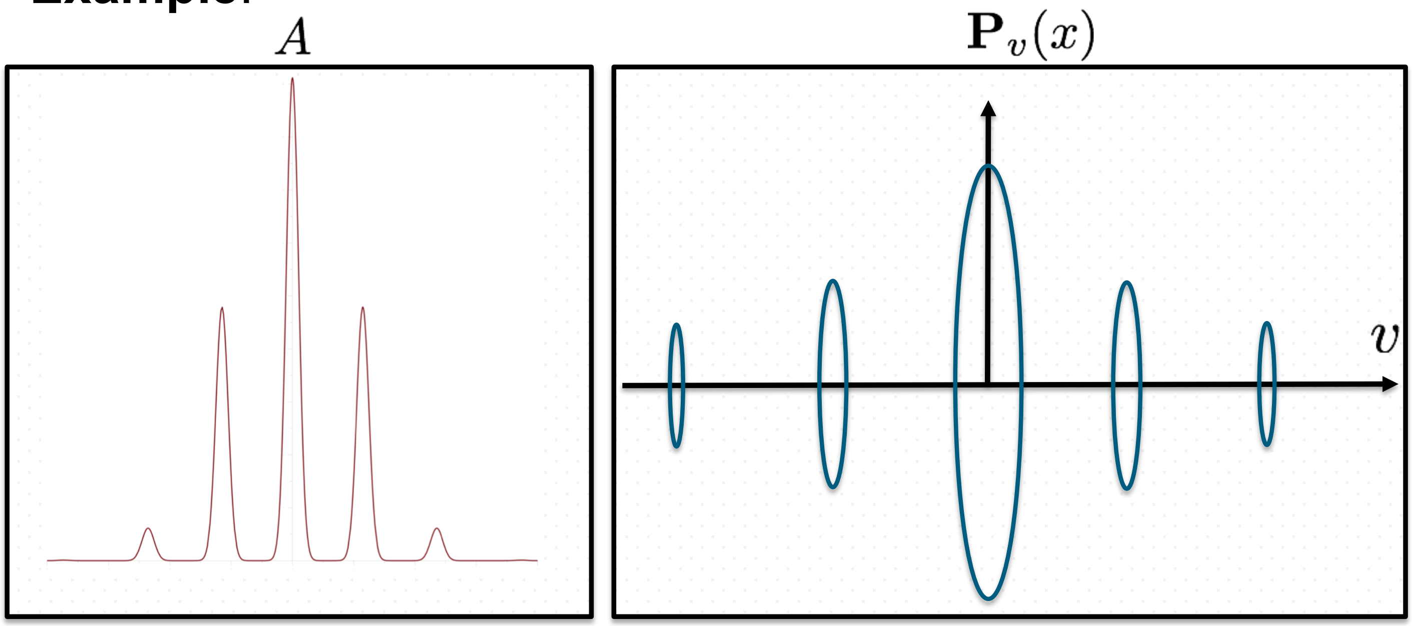
Non-Gaussian Component Analysis [Blanchard et al. 2006]

Hidden Direction Distribution

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Example:



Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Why is Finding a Hidden Direction Hard

Observation: Low-Degree Moments do not help.

- A matches the first m moments of $\mathcal{N}(0,1)$
- The first m moments of $\mathbf{P}_v$ are identical to those of $\mathcal{N}(0,I)$
- Degree-$(m+1)$ moment tensor has $\Omega(d^m)$ entries.

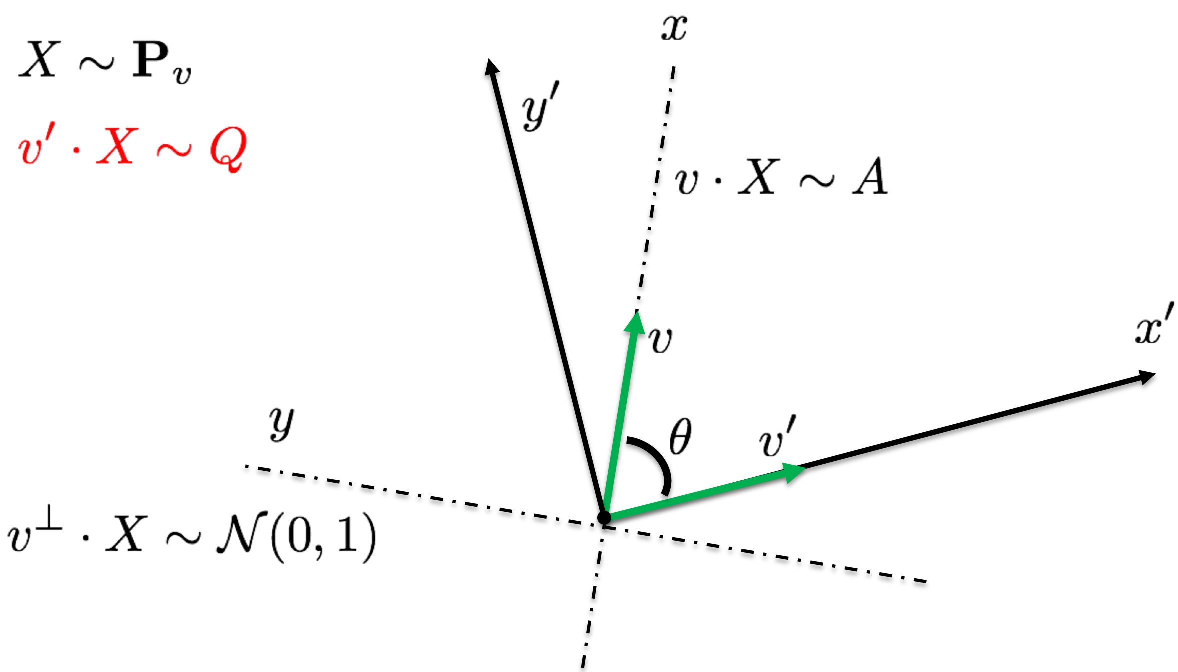
Claim: Random projections do not help.

- To distinguish between $\mathbf{P}_v$ and $\mathcal{N}(0,I)$, would need exponentially many random projections.

1-D PROJECTIONS ARE *ALMOST* STANDARD GAUSSIANS

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

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

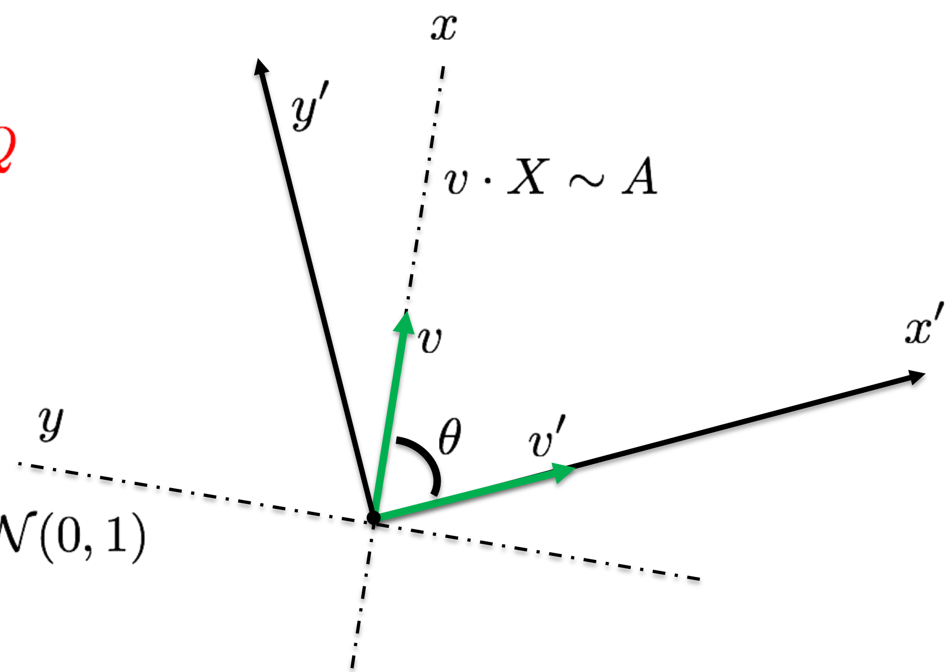


Proof of Key Lemma (I)

$$Q(x') = \int_{\mathbb{R}} A(x)G(y)dy'$$

$$X \sim \mathbf{P}_v$$

$$v' \cdot X \sim Q$$

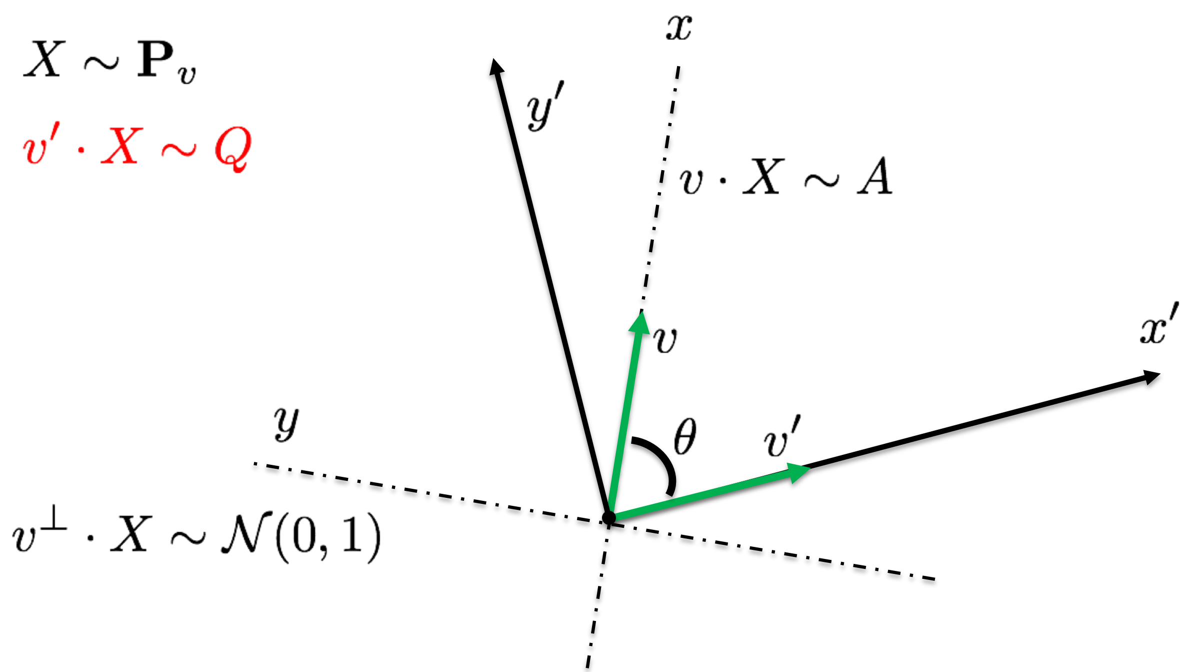


$$v^{\perp} \cdot X \sim \mathcal{N}(0,1)$$

Proof of Key Lemma (I)

$$Q(x') = \int_{\mathbb{R}} A(x)G(y)dy'$$

$$= \int_{\mathbb{R}} A(x'\cos\theta + y'\sin\theta)G(x'\sin\theta - y'\cos\theta)dy'$$



PROOF OF KEY LEMMA (II)

$$Q(x') = \int_{\mathbb{R}} A(x' \cos\theta + y' \sin\theta) G(x' \sin\theta - y' \cos\theta) dy'$$

$$= (U_\theta A)(x')$$

where U_θ is the operator over $f : \mathbb{R} \to \mathbb{R}$

$$U_\theta f(x) := \int_{y \in \mathbb{R}} f(x \cos\theta + y \sin\theta) G(x \sin\theta - y \cos\theta) dy$$

Gaussian Noise (Ornstein-Uhlenbeck) Operator

Eigen-decomposition of Ornstein-Uhlenbeck Operator

Linear Operator U_θ acting on functions $f : \mathbb{R} \to \mathbb{R}$

$$U_\theta f(x) := \int_{y \in \mathbb{R}} f(x \cos\theta + y \sin\theta) G(x \sin\theta - y \cos\theta) dy$$

Fact (Mehler'66): $U_\theta(He_i G)(x) = \cos^i(\theta) He_i(x) G(x)$

- $He_i(x)$ denotes the degree-i Hermite polynomial.
- Note that $\{He_i(x)G(x)/\sqrt{i!}\}_{i \geq 0}$ are orthonormal with respect to the inner product

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) / G(x) dx$$

Generic SQ Lower Bound

Definition: For a unit vector v and a univariate distribution with density A, consider the high-dimensional distribution

$$\mathbf{P}_v(x) = A(v \cdot x) \exp\left(-\|x - (v \cdot x)v\|_2^2/2\right) / (2\pi)^{(d-1)/2}.$$

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within δ error requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

OUTLINE

Part I: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- **Applications: Robust Mean Estimation & Learning GMMs**

- **Part II: Future Directions**

SQ Lower Bound for Robust Mean Estimation (I)

Want to show:

Theorem: Any SQ algorithm that learns an ϵ- corrupted Gaussian in the strong contamination model within error $\epsilon\sqrt{\log(1/\epsilon)}/M$ requires either SQ queries of accuracy $d^{-M/6}$ or at least $d^{\Omega(M^{1/2})}$ many SQ queries

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ Lower Bound for Robust Mean Estimation (II)

Proposition: Suppose that:
- *A* matches the first *m* moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Lemma: There exists a univariate distribution A that is ϵ - close to $\mathcal{N}(\mu, 1)$ such that:
- A agrees with $\mathcal{N}(0,1)$ on the first M moments.
- We have that $\mu = \Omega(\epsilon\sqrt{\log(1/\epsilon)}/M^2)$
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) = \Omega(\mu)$.

SQ Lower Bound for Learning GMMs (I)

Want to show:

Theorem: Any SQ algorithm that learns separated k-GMMs over $\mathbb{R}^d$ to constant error requires either SQ queries of accuracy $d^{-k/6}$ or at least $2^{\Omega(d^{1/8})} \geq d^{2k}$ many SQ queries.

by using our generic proposition:

Proposition: Suppose that:

- A matches the first m moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as v, v' are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

SQ LOWER BOUND FOR LEARNING GMMS (II)

Proposition: Suppose that:

- *A* matches the first *m* moments of $\mathcal{N}(0,1)$
- We have $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) > 2\delta$ as long as *v, v'* are *nearly* orthogonal.

Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

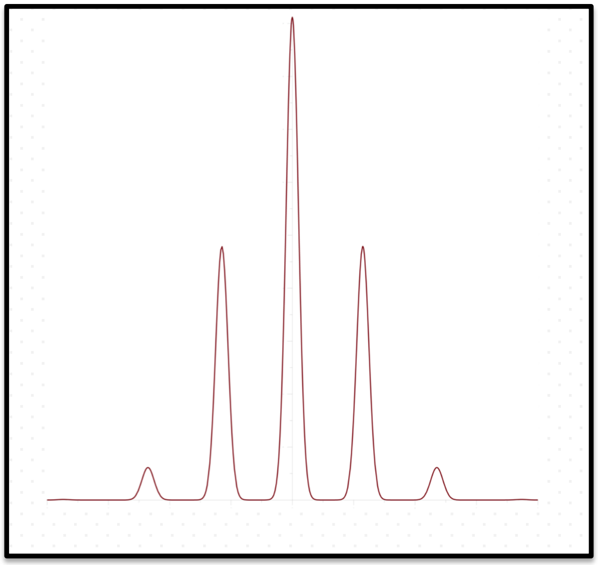
- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

SQ Lower Bound for Learning GMMs (III)

Lemma: There exists a univariate distribution A that is a k-GMM with components A_i such that:

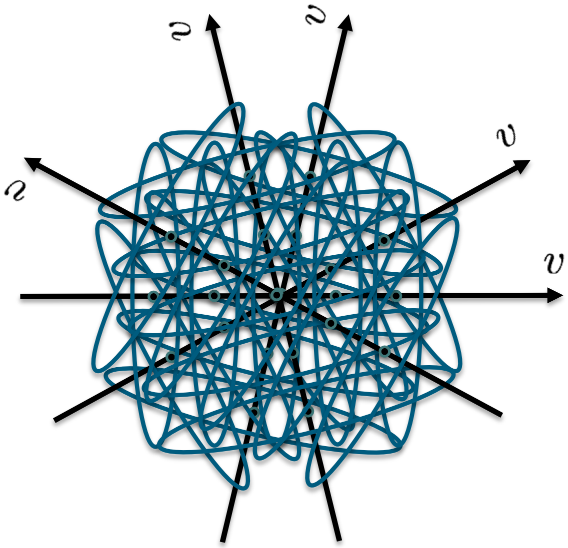
- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

A



SQ LOWER BOUND FOR LEARNING GMMS (IV)

High-Dimensional Distributions $\mathbf{P}_v$ look like "parallel pancakes":



Efficiently learnable for k=2. [Brubaker-Vempala'08]

OUTLINE

Part I: Computational Limits to Robust Estimation

- Statistical Query Learning Model
- Our Results
- Generic Lower Bound Technique
- Applications: Robust Mean Estimation & Learning GMMs

- **Part II: Future Directions**

Future Directions: Computational Lower Bounds

- General Technique to Prove SQ Lower Bounds
- Robustness can make high-dimensional estimation harder computationally and information-theoretically.

Future Directions:

- Further Applications of our Framework
  - List-Decodable Mean Estimation [D-Kane-Stewart'18]
  - Robust Regression [D-Kong-Stewart'18]
  - Adversarial Examples [Bubeck-Price- Razenshteyn'18]
  - Discrete Distributions [D-Gouleakis-Kane-Stewart'19]

- Alternative Evidence of Computational Hardness?
  - ❖ SoS Lower Bounds
  - ❖ Reductions from Average-Case Problems (e.g., Planted Clique, R-3SAT)
  - ❖ Reductions from Worst-case Problems? First step: [Hopkins-Li, COLT'19]

Future Directions: Algorithms

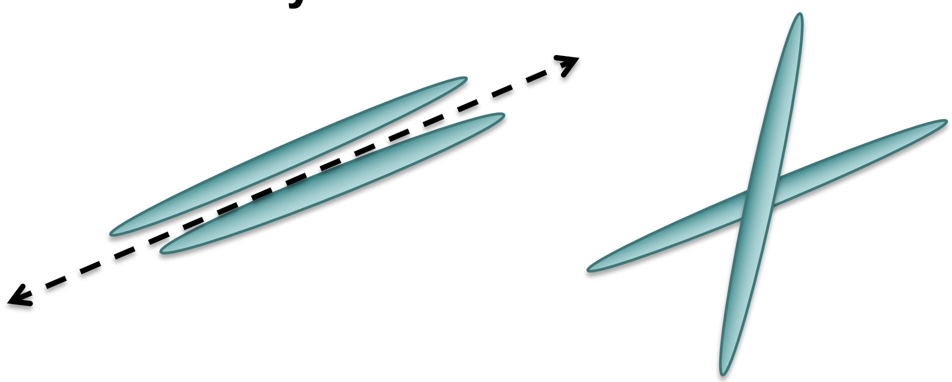
- Pick your favorite high-dimensional probabilistic model for which a (non-robust) efficient learning algorithm is known.
- Make it robust!

Concrete Algorithmic Open Problems

Open Problem 1: Robustly Estimating Gaussian *Covariance* Within Error $O(\epsilon)$ **in Additive Contamination Model (Huber's Model)**

Currently Best Known Algorithm [DKKLMS'18] runs in time $\mathrm{poly}(d) \cdot (1/\epsilon)^{\mathrm{polylog}(1/\epsilon)}$.

Open Problem 2: Robustly Learn a Mixture of 2 *Arbitrary* Gaussians



Spherical components: [Diakonikolas-Kane-Stewart'18, Hopkins-Li'18, Kothari-Steinhardt'18]

FAST / NEAR-LINEAR TIME ALGORITHMS

Filtering for robust mean estimation is practical, but runtime is *super-linear* $\tilde{\Theta}(Nd^2)$.

Question: Can we design near-linear time algorithms*?*

- Robust Mean Estimation:
  - ❖ [Cheng-D-Ge, SODA'19] $\tilde{\Theta}(Nd/\mathrm{poly}(\epsilon))$.
  - ❖ [Depersin-Lecue, Arxiv-June 2019] $\tilde{\Theta}(Nd)$.
  - ❖ [Dong-Hopkins-Li, upcoming] $\tilde{\Theta}(Nd)$.

- How about more general estimation tasks?
  - ❖ [Cheng-D-Ge-Woodruff, COLT'19]
  - ❖ Robust *Sparse* Estimation?
  - ❖ List-Decodable Learning?

Broader Research Directions

General Algorithmic Theory of Robustness

How can we robustly learn rich representations of data, based on natural hypotheses about the structure in data?

Can we robustly *test* our hypotheses about structure in data before learning?

Broader Challenges:

- Richer Families of Problems and Models
- Connections to Non-convex Optimization, Adversarial Examples, GANs, …
- Relation to Related Notions of Algorithmic Stability
  (Differential Privacy, Adaptive Data Analysis)
- Further Applications (ML Security, Computer Vision, …)
- Other notions of robustness?

Thank you!
Questions?