Computational Lower Bounds for Statistical Estimation Problems

Ilias Diakonikolas (USC)

(joint with Daniel Kane (UCSD) and Alistair Stewart (USC))

Workshop on Local Algorithms, MIT, June 2018

THIS TALK

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

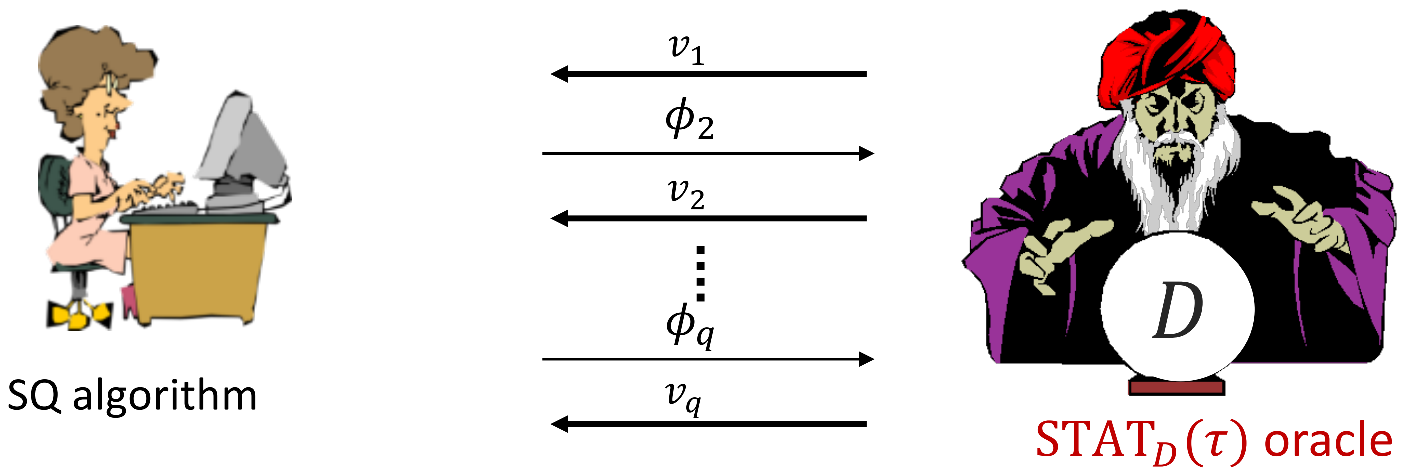
Concrete Applications of our Technique:

- Learning Gaussian Mixture Models (GMMs)
- Robustly Learning a Gaussian
- Robustly Testing a Gaussian
- Statistical-Computational Tradeoffs

STATISTICAL QUERIES [KEARNS' 93]

$x_1, x_2, \ldots, x_m \sim D$ 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 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 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.

THIS TALK

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

Concrete Applications of our Technique:

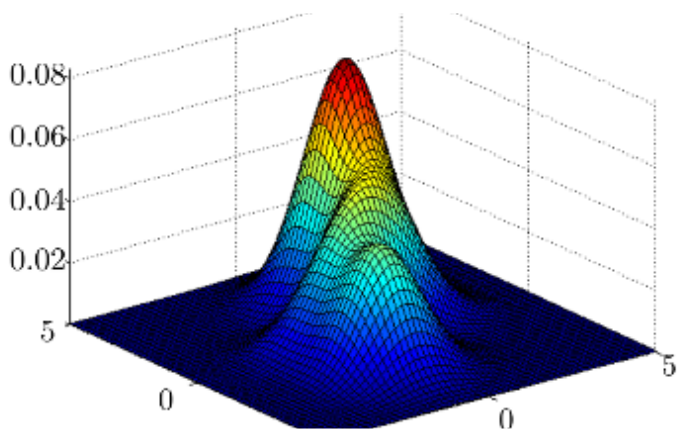
- Learning Gaussian Mixture Models (GMMs)
- Robustly Learning a Gaussian
- Robustly Testing a Gaussian
- Statistical-Computational Tradeoffs

GAUSSIAN MIXTURE MODEL (GMM)

- GMM: Distribution on $\mathbb{R}^d$ with probability density function

$$F = \sum_{i=1}^{k} w_i \mathcal{N}(\mu_i, \Sigma_i)$$

- Extensively studied in statistics and TCS



Karl Pearson (1894)

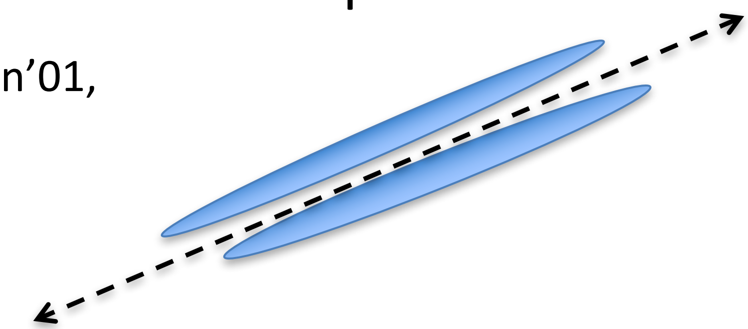
LEARNING GMMS - PRIOR WORK (I)

Two Related Learning Problems

Parameter Estimation: Recover model parameters.

- **Separation Assumptions**: Clustering-based Techniques

  [Dasgupta'99, Dasgupta-Schulman'00, Arora-Kanan'01, Vempala-Wang'02, Achlioptas-McSherry'05, **Brubaker-Vempala'08**]

  Sample Complexity: $\mathrm{poly}(d, k)$
  (Best Known) Runtime: $\mathrm{poly}(d, k)$

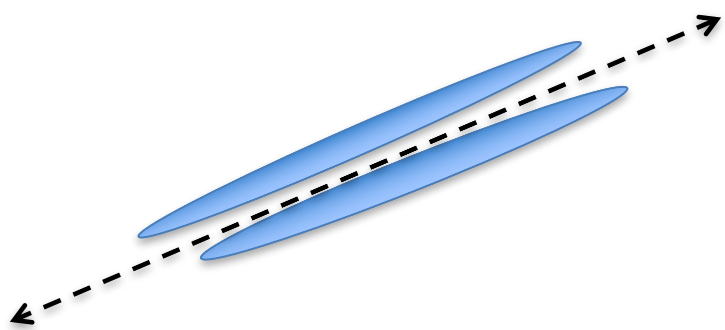
- **No Separation**: Moment Method

  [Kalai-Moitra-Valiant'10, Moitra-Valiant'10, Belkin-Sinha'10, Hardt-Price'15]

  Sample Complexity: $\mathrm{poly}(d) \cdot (1/\gamma)^{\Theta(k)}$
  (Best Known) Runtime: $(d/\gamma)^{\Omega(k)}$

SEPARATION ASSUMPTIONS

- Clustering is possible only when the components have very little overlap.
- Formally, we want the total variation distance between components to be close to 1.
- Algorithms for learning spherical GMMS work under this assumption.
- For non-spherical GMMs, known algorithms require stronger assumptions.



LEARNING GMMS - PRIOR WORK (II)

Density Estimation: Recover underlying distribution (within statistical distance ϵ).

[Feldman-O'Donnell-Servedio'05, Moitra-Valiant'10, Suresh-Orlitsky-Acharya-Jafarpour'14, Hardt-Price'15, Li-Schmidt'15]

Sample Complexity: $\mathrm{poly}(d, k, 1/\epsilon)$

(Best Known) Runtime: $(d/\epsilon)^{\Omega(k)}$

Fact: For separated GMMs, density estimation and parameter estimation are equivalent.

LEARNING GMMS – OPEN QUESTION

Summary: The sample complexity of density estimation for k-GMMs is $\mathrm{poly}(d, k)$. The sample complexity of parameter estimation for *separated* k-GMMs is $\mathrm{poly}(d, k)$.

Question: Is there a $\mathrm{poly}(d, k)$ ***time*** learning algorithm?

STATISTICAL QUERY LOWER BOUND FOR LEARNING 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 **dimension of latent space**.

GENERAL RECIPE FOR (SQ) LOWER BOUNDS

Our generic technique for proving SQ Lower Bounds:

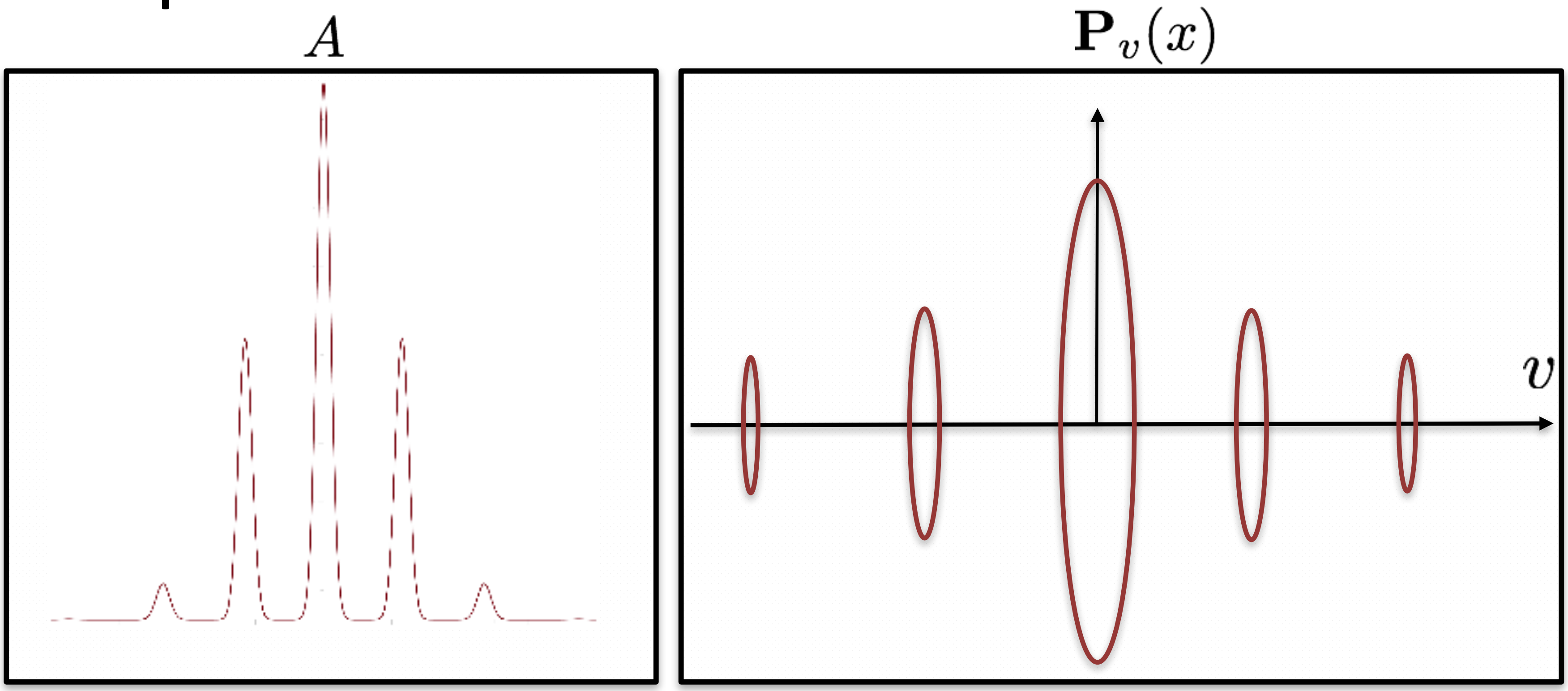
- **Step #1:** Construct distribution $\mathbf{P}_v$ that is standard Gaussian in all directions except v.
- **Step #2:** Construct the univariate projection in the v direction so that it matches the first m moments of $\mathcal{N}(0,1)$
- **Step #3:** Consider the family of instances $\mathcal{D} =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

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.

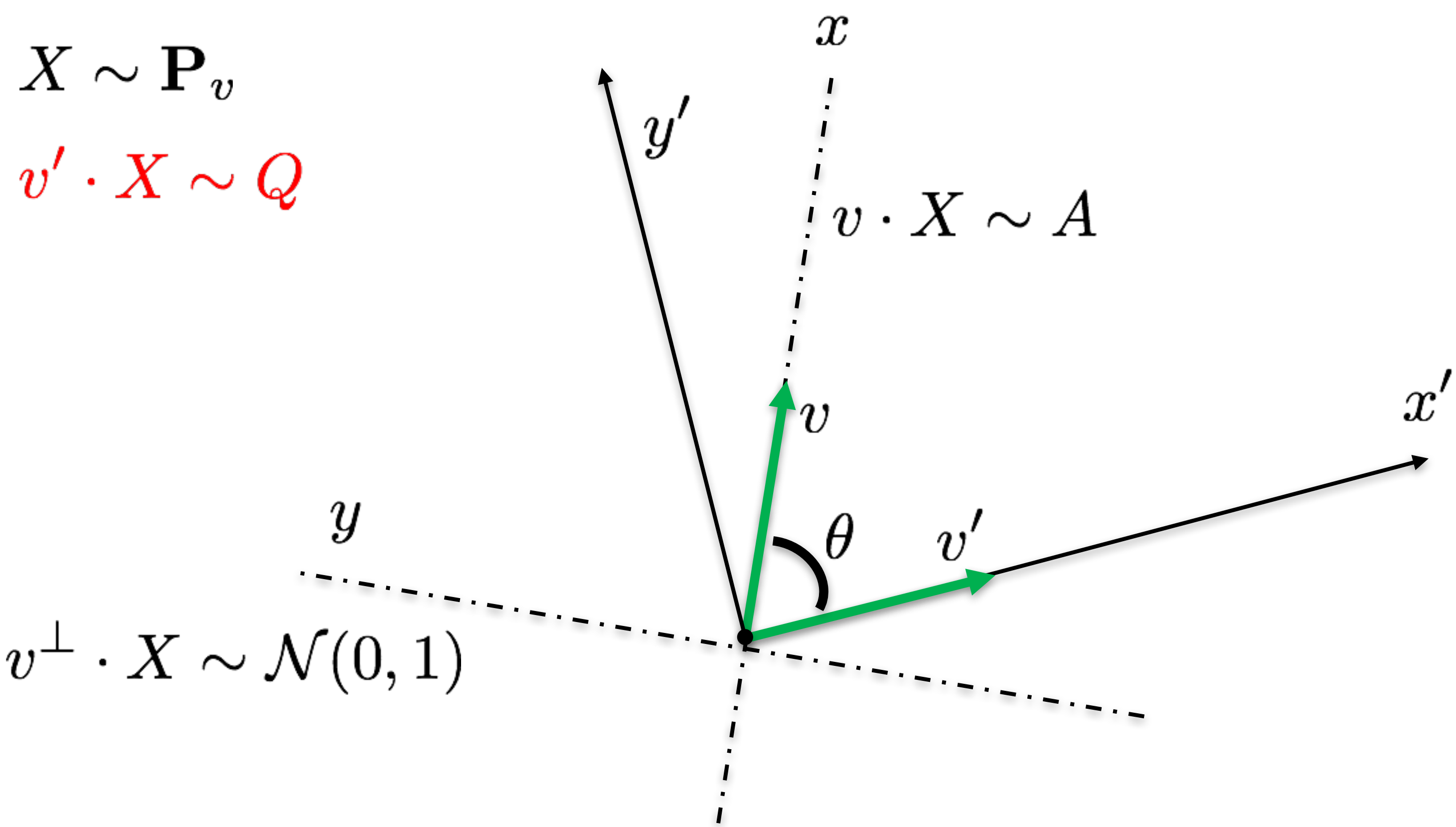
Claim: Random projections do not help.

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

ONE-DIMENSIONAL PROJECTIONS ARE ALMOST GAUSSIAN

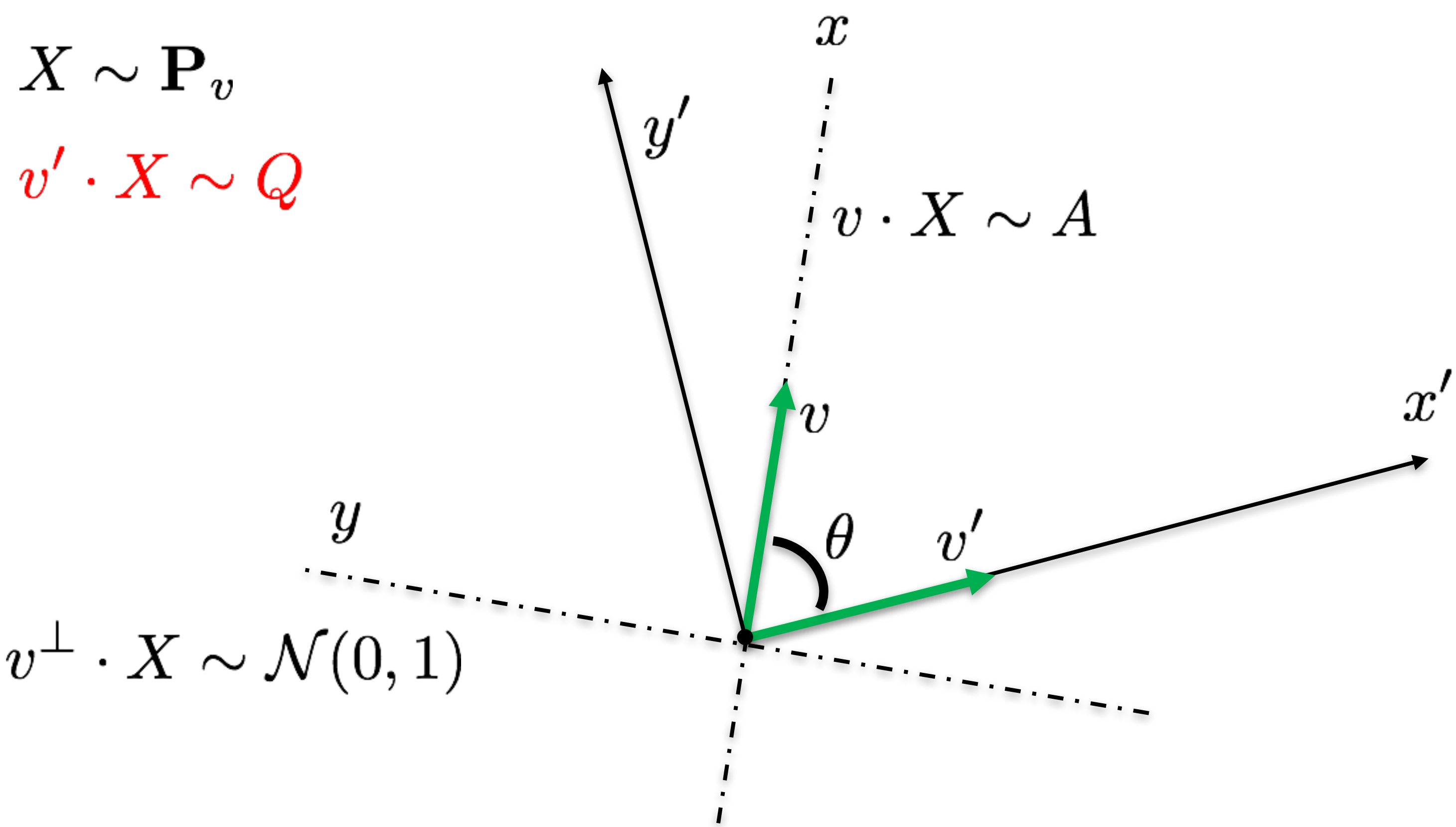
Key Lemma: Let Q be the distribution of $v' \cdot X$, where $X \sim \mathbf{P}_v$.
Then, we have that:

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



PROOF OF KEY LEMMA (I)

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



PROOF OF KEY LEMMA (I)

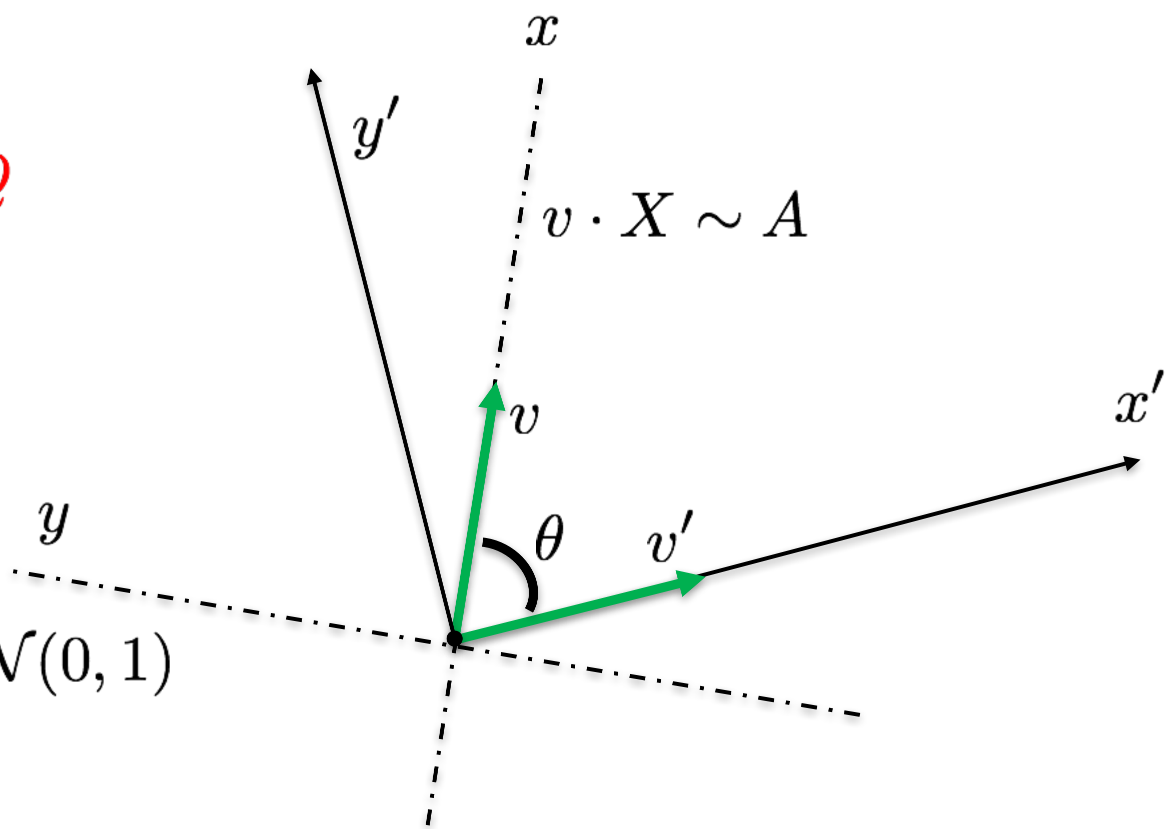
$$Q(x') = \int_{\mathbb{R}} A(x)G(y)dy'$$

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

$X \sim \mathbf{P}_v$

$v' \cdot X \sim Q$

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



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

EIGENFUNCTIONS 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 error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

PROOF OF GENERIC SQ LOWER BOUND

- Suffices to construct a large set of distributions that are *nearly* uncorrelated.
- Pairwise correlation between D_1 and D_2 with respect to D:

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

Two Main Ingredients:

Correlation Lemma:

$$|\chi_{N(0,I)}(\mathbf{P}_v, \mathbf{P}_{v'})| \leq |v \cdot v'|^{m+1} \chi^2(A, N(0,1))$$

Packing Argument: There exists a set S of $2^{\Omega(d^{1/4})}$ unit vectors on $\mathbb{R}^d$ with pairwise inner product $O(1/d^{1/4})$

APPLICATION: SQ LOWER BOUND FOR 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.

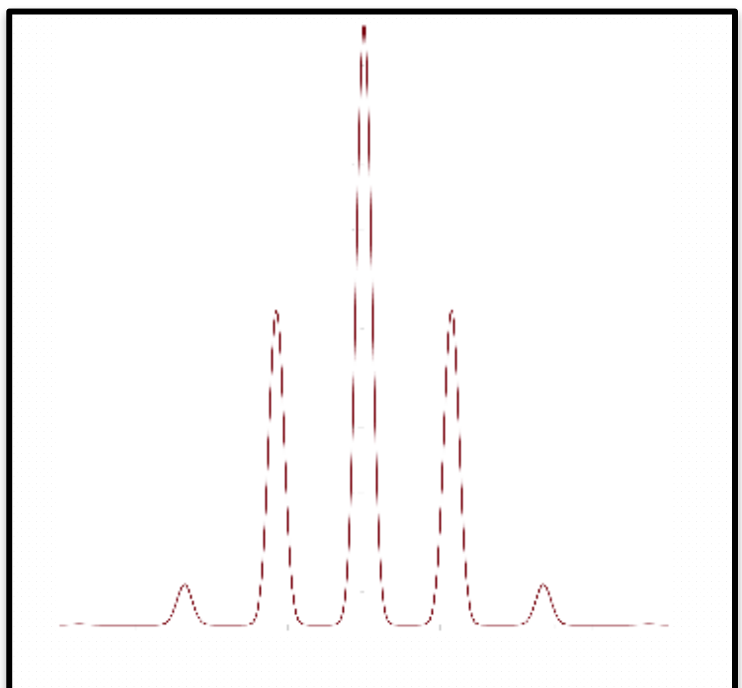
Then any SQ algorithm that learns an unknown $\mathbf{P}_v$ within error δ requires either queries of accuracy d^{-m} or $2^{d^{\Omega(1)}}$ many queries.

APPLICATION: SQ LOWER BOUND FOR GMMS (II)

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

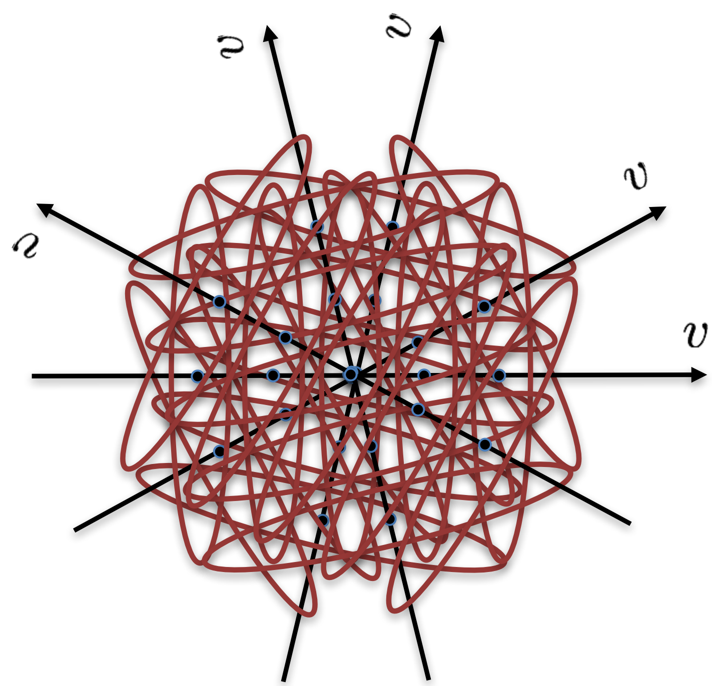
- A agrees with $\mathcal{N}(0,1)$ on the first $2k$-1 moments.
- Each pair of components are separated.
- Whenever v and v' are nearly orthogonal $d_{\mathrm{TV}}(\mathbf{P}_v, \mathbf{P}_{v'}) \geq 1/2$.

A



APPLICATION: SQ LOWER BOUND FOR GMMS (III)

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



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

FURTHER RESULTS

Unified technique yielding a range of applications.

SQ Lower Bounds:

- Learning GMMs
- Robustly Learning a Gaussian

  "Error guarantee of [DKK+16] are optimal for all poly time algorithms."
- Robust Covariance Estimation in Spectral Norm:

  "Any efficient SQ algorithm requires $\Omega(d^2)$ samples."
- Robust k-Sparse Mean Estimation:

  "Any efficient SQ algorithm requires $\Omega(k^2 + k \log d)$ samples."

Sample Complexity Lower Bounds

- Robust Gaussian Mean Testing
- Testing Spherical 2-GMMs:

"Distinguishing between $\mathcal{N}(0, I)$ and $(1/2)\mathcal{N}(\mu_1, I) + (1/2)\mathcal{N}(\mu_2, I)$ requires $\Omega(d)$ samples."

- Sparse Mean Testing

SAMPLE COMPLEXITY OF ROBUST TESTING

High-Dimensional Hypothesis Testing

Gaussian Mean Testing
Distinguish between:

- Completeness: $D = \mathcal{N}(0, I)$
- Soundness: $D = \mathcal{N}(\mu, I)$ with $\|\mu\|_2 \geq \epsilon$

Simple mean-based algorithm with $O(\sqrt{d}/\epsilon^2)$ samples.

Suppose we add corruptions to soundness case at rate $\delta \ll \epsilon$.

Theorem
Sample complexity of robust Gaussian mean testing is $\Omega(d)$.

Take-away: Robustness can dramatically increase the sample complexity of an estimation task.

SUMMARY AND FUTURE DIRECTIONS

- General Technique to Prove SQ Lower Bounds
- Implications for a Range of Unsupervised Estimation Problems

Future Directions:

- Further Applications of our Framework

  Discrete Setting [D-Kane-Stewart'18],
  Robust Regression [D-Kong-Stewart'18],
  Adversarial Examples [Bubeck-Price- Razenshteyn'18]
  …

- Alternative Evidence of Computational Hardness?

Thanks! Any Questions?