Utilizing High Degree Moments

Daniel M. Kane

Departments of CS/Math
University of California, San Diego
dakane@ucsd.edu

June 23rd, 2019

Outline

- Obstacles to better list decoding
- Variance of polynomials method
- Sum of squares method

Last Talk

Robust List Decoding:

- Given samples *all but* an α-fraction are errors
- Return $\mathrm{poly}(1/\alpha)$ hypotheses at least one of which is close

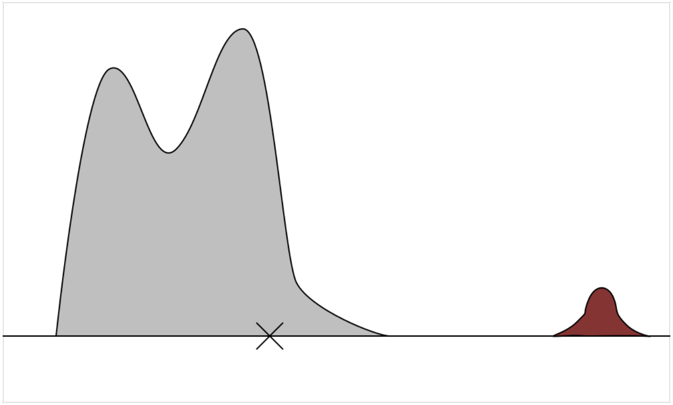
Error bounds:

- Lower bound: $\Omega(\sqrt{\log(1/\alpha)})$
- Upper bound: $\tilde{O}(\sqrt{1/\alpha})$

Can we do better?

Obstacle at $\alpha^{-1/2}$

Algorithm checked for directions of large variance. Unfortunately, this is not enough to ensure error better than $\alpha^{-1/2}$.



Idea

Bounds on the second moments are not enough to ensure concentration.

Idea

Bounds on the second moments are not enough to ensure concentration.
Fix: use higher moments.

Analysis

If for all unit vectors v,

$$\mathbb{E}[|v \cdot (X - \mu_X)|^{2d}] = O(1),$$

then

$$1 \gg \alpha |v \cdot (\mu - \mu_X)|^{2d},$$

so

$$|\mu - \mu_X| = O(\alpha^{-1/2d}).$$

Computational Difficulty

It is computationally intractable to determine whether or not there is a unit vector v for which $\mathbb{E}[(v \cdot X)^{2d}]$ is large when $d > 1$ [Hopkins-Li '19].

Computational Difficulty

It is computationally intractable to determine whether or not there is a unit vector v for which $\mathbb{E}[(v \cdot X)^{2d}]$ is large when $d > 1$ [Hopkins-Li '19].

Idea: Look at a relaxation of this problem.

1. Show that $\mathbb{E}[p(X)^2]$ is not too large for *every* degree-d polynomial p ([Diakonikolas-Kane-Stewart '18])
2. Use a Sum of Squares proof to show that $\mathbb{E}[(v \cdot X)^{2d}]$ is small for every unit vector v ([Hopkins-Li '18], [Kothari-Steinhardt-Steurer '18])

First Approach

Given a sample set S with sample mean $\hat{\mu}$ have two quadratic forms on degree-d polynomials p:

- $p \to \mathbb{E}[p(S)^2]$
- $p \to \mathbb{E}[p(\mathcal{N}(\hat{\mu}, I))^2]$

First Approach

Given a sample set S with sample mean $\hat{\mu}$ have two quadratic forms on degree-d polynomials p:

- $p \to \mathbb{E}[p(S)^2]$
- $p \to \mathbb{E}[p(\mathcal{N}(\hat{\mu}, I))^2]$

Can detect whether there is some p much bigger on one than the other.

Cases

If $\mathbb{E}[p(S)^2] \ll \mathbb{E}[p(\mathcal{N}(\hat{\mu}, I))^2]$ for all p:

- Take $p(x) = (v \cdot (x - \hat{\mu}))^d$
- $\mathbb{E}[p(S)^2] \gg \alpha(v \cdot (\mu - \hat{\mu}))^{2d}$
- Therefore $|\mu - \hat{\mu}| = O(\alpha^{-1/2d})$

Cases

If $\mathbb{E}[p(S)^2] \ll \mathbb{E}[p(\mathcal{N}(\hat{\mu}, I))^2]$ for all p:

- Take $p(x) = (v \cdot (x - \hat{\mu}))^d$
- $\mathbb{E}[p(S)^2] \gg \alpha(v \cdot (\mu - \hat{\mu}))^{2d}$
- Therefore $|\mu - \hat{\mu}| = O(\alpha^{-1/2d})$

If there is some p with $\mathbb{E}[p(S)^2]$ much larger

- p has larger empirical variance than it should
- (Multi)filter based on the values of p

Complication

If μ is unknown, so is $\mathrm{Var}(p(G))$. This makes it difficult to filter points based on the values of p

Complication

If μ is unknown, so is $\mathrm{Var}(p(G))$. This makes it difficult to filter points based on the values of p

Highly technical fix, using several facts about Gaussian polynomials.

Upshot

Theorem (Informal Statement)

Given $(nd)^{O(d)} poly(1/\alpha)$ samples and $(nd/\alpha)^{O(d)}$ time there is an algorithm returning $poly(1/\alpha)$ hypotheses at least one of which is within $O_d(\alpha^{-1/2d})$ of μ.

With superconstant d can get polylog error in quasipolynomial time/samples.

Sum of Squares

How else might we try to show that our set has bounded central moments?

Sum of Squares

How else might we try to show that our set has bounded central moments?

Sum of squares proofs: We can show that $f \geq 0$ if we can write f as a sum of squares of lower degree polynomials. There is a convex program to determine if this is possible!

Pseudoexpectations

- Want to know if polynomial f is always non-negative.
  - Find an x with $f(x)$ small.
  - Instead consider evaluation function $g \to g(x)$.
  - Take convex relaxation.

Pseudoexpectations

- Want to know if polynomial f is always non-negative.
  - Find an x with $f(x)$ small.
  - Instead consider evaluation function $g \to g(x)$.
  - Take convex relaxation.
- See if there is a pseudoexpectation $\tilde{\mathbb{E}} : \{\text{degree-}d \text{ polynomials}\} \to \mathbb{R}$ so that:
  - $\tilde{\mathbb{E}}[1] = 1$
  - $\tilde{\mathbb{E}}[p^2] \geq 0$ for any p
  - $\tilde{\mathbb{E}}[f]$ as small as possible

Pseudoexpectations

- Want to know if polynomial f is always non-negative.
  - Find an x with $f(x)$ small.
  - Instead consider evaluation function $g \to g(x)$.
  - Take convex relaxation.
- See if there is a pseudoexpectation $\tilde{\mathbb{E}} : \{\text{degree-}d \text{ polynomials}\} \to \mathbb{R}$ so that:
  - $\tilde{\mathbb{E}}[1] = 1$
  - $\tilde{\mathbb{E}}[p^2] \geq 0$ for any p
  - $\tilde{\mathbb{E}}[f]$ as small as possible
- $\tilde{\mathbb{E}}$ behaves like the expectation over solutions.
- There is a convex program to find $\tilde{\mathbb{E}}$.

Idea

Need a distribution where the good samples have a SoS proof of bounded central moments.

- Try to find SoS proof of bounded central moments for sample set. Either:
  - Succeed. You have bounded central moments, so sample mean is good.
  - Fail. Find a pseudoexpectation. Behaves like distribution over directions with bad moments. Can use to construct (multi)filter.

Comparison

The SoS technique gets results qualitatively similar to the other one but:

- The use of convex programming likely means that it will be practically slower.
- Works for *any* distribution with a SoS proof of bounded central moments (Gaussians, rotations of product distributions,. . .).

Other Applications of SoS

To get better than $O(\sqrt{\epsilon})$ robust mean estimation, we generally need both:

- An accurate approximation to $\mathrm{Cov}(X)$
- Tail bounds

What if we only have the latter?

Only Tail Bounds

- Information-theoretically, tail bounds are enough.
- Can estimate the mean in any direction by truncated mean.
- Difficult to figure out which direction to filter in.
- For Gaussians can do better: Approximate $\mathrm{Cov}(X)$ using relation between 2nd and 4th moments.

SoS

Suppose that distribution had bounded d^{th} central moments provable by SoS. Filter by trying to find such a proof.

- If it works, bounded moments implies small effect $O(\epsilon^{1-1/d})$ of errors on mean
- If not, pseudoexpectation gives "direction" to filter in

Can learn to error $O(\epsilon^{1-1/d})$ with just (provable) bounded moments.

Conclusion

There are several instances were better errors in robust mean estimation can be obtained by considering higher moments.

- Ilias Diakonikolas, Daniel M. Kane, Alistair Stewart *List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians*, STOC 2018.
- Samuel B. Hopkins, Jerry Li *How Hard Is Robust Mean Estimation?*, COLT 2019.
- Samuel B. Hopkins, Jerry Li *Mixture Models, Robustness, and Sum of Squares Proofs*, STOC 2018.
- Pravesh Kothari, Jacob Steinhardt, David Steurer *Robust Moment Estimation and Improved Clustering via Sum of Squares*, STOC 2018.