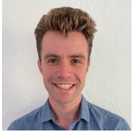
SoS Certifiability of Subgaussian Distributions & Its Algorithmic Applications

Ankit Pensia

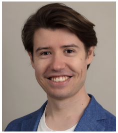
11th TOCA-SV, 2024
Stanford

Joint work with

Ilias Diakonikolas



Sam Hopkins



Stefan Tiegel

Overview

- **Motivation**
- Our Result
- Proof Sketch
- Conclusion

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\big(\widehat{\theta}, \theta^*(Q)\big)$ is small w.h.p.

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\big(\widehat{\theta}, \theta^*(Q)\big)$ is small w.h.p.

As $\mathcal{P}$ gets **nicer** (more structured),

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\left(\widehat{\theta}, \theta^*(Q)\right)$ is small w.h.p.

As $\mathcal{P}$ gets **nicer** (more structured),

- the information-theoretic optimal error gets **smaller**

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\left(\widehat{\theta}, \theta^*(Q)\right)$ is small w.h.p.

As $\mathcal{P}$ gets **nicer** (more structured),

- the information-theoretic optimal error gets **smaller**
- unfortunately, algorithmic guarantees often **do not improve...**

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\left(\widehat{\theta}, \theta^*(Q)\right)$ is small w.h.p.

As $\mathcal{P}$ gets **nicer** (more structured),

- the information-theoretic optimal error gets **smaller**
- unfortunately, algorithmic guarantees often **do not improve...** unless the **niceness** assumption on $\mathcal{P}$ is **algorithmically tractable**

Motivation: distributional assumptions

Generic estimation problem. Let $\mathcal{P}$ be a family of distributions over $\mathbb{R}^d$ and $\theta^* : \mathcal{P} \to \mathcal{Y}$ be the target parameter.

Input:	samples from (unknown) $Q \in \mathcal{P}$
Output:	$\widehat{\theta}$ such that distance$\left(\widehat{\theta}, \theta^*(Q)\right)$ is small w.h.p.

As $\mathcal{P}$ gets **nicer** (more structured),

- the information-theoretic optimal error gets **smaller**
- unfortunately, algorithmic guarantees often **do not improve...** unless the **niceness** assumption on $\mathcal{P}$ is **algorithmically tractable**

Understanding "niceness" versus "tractable niceness"

Motivation: **moment bounds**

- To arrive at the right notion of “niceness”, we note that . . .

Motivation: **moment bounds**

- To arrive at the right notion of "niceness", we note that . . .
- . . . the error is governed by the size of **moments of linear projections**

For all unit vectors $v \in \mathbb{R}^d$: $\quad \mathbb{E}_X \left[|\langle v, X - \mu \rangle|^p \right] \leq \text{Size}(p)$

Motivation: **moment bounds**

- To arrive at the right notion of “niceness”, we note that . . .
- . . . the error is governed by the size of **moments of linear projections**

For all unit vectors $v \in \mathbb{R}^d$: $\quad \mathbb{E}_X \left[|\langle v, X - \mu \rangle|^p \right] \leq \text{Size}(p)$

- Thus, the right notion of *niceness* is captured by moment bounds of X
  - Assuming X is gaussian is unrealistic
  - But assuming that the moments of X are smaller than those of a gaussian is both **more realistic** and often **sufficient**.

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is **subgaussian** if its moments grow slower than those of a **gaussian**. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X - \mu\rangle^p\right] \leq \mathbb{E}_{G\sim\mathcal{N}(0,I_d)}\left[\langle v, G\rangle^p\right]$

- Equivalently, $\mathbb{P}\left(\langle v, X - \mu\rangle > t\right) \ll \exp(-t^2)$

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is **subgaussian** if its moments grow slower than those of a **gaussian**. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X - \mu\rangle^p\right] \leq \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X - \mu\rangle > t\right) \ll \exp(-t^2)$

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is subgaussian if its moments grow slower than those of a **gaussian**. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X - \mu\rangle^p\right] \le \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X - \mu\rangle > t\right) \ll \exp(-t^2)$
- Thus, we relax the *unrealistic* gaussianity assumption

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is subgaussian if its moments grow slower than those of a gaussian. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X-\mu\rangle^p\right] \leq \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X-\mu\rangle > t\right) \ll \exp(-t^2)$
- Thus, we relax the *unrealistic* gaussianity assumption, **while** keeping the same **information-theoretic** optimal error

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is subgaussian if its moments grow slower than those of a gaussian. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X - \mu\rangle^p\right] \leq \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X - \mu\rangle > t\right) \ll \exp(-t^2)$
- Thus, we relax the *unrealistic* gaussianity assumption, while keeping the same **information-theoretic** optimal error
- However, existing **computationally-efficient** algorithms seem unable to use subgaussianity, incurring large error

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is subgaussian if its moments grow slower than those of a gaussian. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X - \mu \rangle^p\right] \leq \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X - \mu \rangle > t\right) \ll \exp(-t^2)$
- Thus, we relax the *unrealistic* gaussianity assumption, while keeping the same **information-theoretic** optimal error
- However, existing **computationally-efficient** algorithms seem unable to use subgaussianity, incurring large error
    - Even though they *are able to* use gaussianity

Motivation: subgaussianity versus gaussianity

Definition. A distribution X over $\mathbb{R}^d$ is subgaussian if its moments grow slower than those of a gaussian. That is,

For all $v \in \mathbb{R}^d$ and even p: $\mathbb{E}\left[\langle v, X-\mu\rangle^p\right] \leq \sqrt{p}^p \|v\|_2^p$

- Equivalently, $\mathbb{P}\left(\langle v, X-\mu\rangle > t\right) \ll \exp(-t^2)$
- Thus, we relax the *unrealistic* gaussianity assumption, while keeping the same **information-theoretic** optimal error
- However, existing **computationally-efficient** algorithms seem unable to use subgaussianity, incurring large error
  - Even though they *are able to* use gaussianity

Algorithmically tractable notion of subgaussianity?

Motivation: "certifiable" subgaussianity

Definition. X is **certifiably-subgaussian** if its moments grow slower than those of a Gaussian AND are certifiable by a **sum-of-squares** proof.

Motivation: "certifiable" subgaussianity

Definition. X is **certifiably-subgaussian** if its moments grow slower than those of a Gaussian AND are certifiable by a **sum-of-squares** proof. That is, for all even p:

$$\sqrt{p}^p \|v\|_2^p - \mathbb{E}\left[\langle v, X - \mu \rangle^p\right] = \sum_i q_i^2(v)$$

for some polynomials $q_1(\cdot), \ldots,$

Motivation: “certifiable” subgaussianity

Definition. X is **certifiably-subgaussian** if its moments grow slower than those of a Gaussian AND are certifiable by a **sum-of-squares** proof. That is, for all even p:

$$\sqrt{p}^p \|v\|_2^p - \mathbb{E}\left[\langle v, X - \mu\rangle^p\right] = \sum_i q_i^2(v)$$

for some polynomials $q_1(\cdot), \ldots,$

Proposed in [KSS18; HL18] and hugely influential since then

[KSS18] P. K. Kothari, J. Steinhardt, D. Steurer. Robust Moment Estimation and Improved Clustering via Sum of Squares. *STOC*. 2018
[HL18] S. B. Hopkins, J. Li. Mixture Models, Robustness, and Sum of Squares Proofs. *STOC*. 2018

Motivation: "certifiable" subgaussianity

Definition. X is **certifiably-subgaussian** if its moments grow slower than those of a Gaussian AND are certifiable by a **sum-of-squares** proof. That is, for all even p:

$$\sqrt{p}^p \|v\|_2^p - \mathbb{E}\left[\langle v, X - \mu\rangle^p\right] = \sum_i q_i^2(v)$$

for some polynomials $q_1(\cdot), \ldots,$

Proposed in [KSS18; HL18] and hugely influential since then

- Leads to computationally-efficient (SDP-based) algorithms
  - robust estimation (mean, covariance, regression), clustering, privacy

---

[KSS18] P. K. Kothari, J. Steinhardt, D. Steurer. Robust Moment Estimation and Improved Clustering via Sum of Squares. *STOC*. 2018
[HL18] S. B. Hopkins, J. Li. Mixture Models, Robustness, and Sum of Squares Proofs. *STOC*. 2018

Motivation: "certifiable" subgaussianity

Definition. X is **certifiably-subgaussian** if its moments grow slower than those of a Gaussian AND are certifiable by a **sum-of-squares** proof. That is, for all even p:

$$\sqrt{p}^p \|v\|_2^p - \mathbb{E}\left[\langle v, X - \mu\rangle^p\right] = \sum_i q_i^2(v)$$

for some polynomials $q_1(\cdot), \ldots,$

Proposed in [KSS18; HL18] and hugely influential since then

- Leads to computationally-efficient (SDP-based) algorithms
  - robust estimation (mean, covariance, regression), clustering, privacy

Which distributions are certifiably subgaussian anyway?

[KSS18] P. K. Kothari, J. Steinhardt, D. Steurer. Robust Moment Estimation and Improved Clustering via Sum of Squares. *STOC*. 2018
[HL18] S. B. Hopkins, J. Li. Mixture Models, Robustness, and Sum of Squares Proofs. *STOC*. 2018

Prior work: certifiable subgaussianity vs. subgaussianity

- Known **sufficient** conditions for certifiability are rather strong

Prior work: certifiable subgaussianity vs. subgaussianity

- Known **sufficient** conditions for certifiability are rather strong
  - Rotational invariance, or
  - Independent coordinates, or
  - Log-Sobolev Distributions (more generally, Poincare) [KSS18], or
  - Simple transformations thereof (simple within SoS)

[KSS18] P. K. Kothari, J. Steinhardt, D. Steurer. Robust Moment Estimation and Improved Clustering via Sum of Squares. *STOC*. 2018

Prior work: **certifiable subgaussianity vs. subgaussianity**

- Known **sufficient** conditions for certifiability are rather strong
  - Rotational invariance, or
  - Independent coordinates, or
  - Log-Sobolev Distributions (more generally, Poincare) [KSS18], or
  - Simple transformations thereof (simple within SoS)
- However, many subgaussians **do not** satisfy these sufficient conditions

[KSS18] P. K. Kothari, J. Steinhardt, D. Steurer. Robust Moment Estimation and Improved Clustering via Sum of Squares. *STOC*. 2018

Prior work: certifiable subgaussianity vs. subgaussianity

- Known **sufficient** conditions for certifiability are rather strong
  - Rotational invariance, or
  - Independent coordinates, or
  - Log-Sobolev Distributions (more generally, Poincare) [KSS18], or
  - Simple transformations thereof (simple within SoS)
- However, many subgaussians **do not** satisfy these sufficient conditions
- A plausible belief was that there might be some subgaussian distributions that are **not** certifiably subgaussian

Prior work: certifiable subgaussianity vs. subgaussianity

- Known **sufficient** conditions for certifiability are rather strong
  - Rotational invariance, or
  - Independent coordinates, or
  - Log-Sobolev Distributions (more generally, Poincare) [KSS18], or
  - Simple transformations thereof (simple within SoS)
- However, many subgaussians **do not** satisfy these sufficient conditions
- A plausible belief was that there might be some subgaussian distributions that are **not** certifiably subgaussian
  - Indeed, there is a huge gap between bounded & certifiably bounded
  - Under SSEH, there exist distributions that have bounded 1000-moments, but do not have certifiably-bounded 4-th moments [HL19]

[HL19] S. B. Hopkins, J. Li. How Hard Is Robust Mean Estimation? *COLT*. 2019

Prior work: certifiable subgaussianity vs. subgaussianity

- Known **sufficient** conditions for certifiability are rather strong
  - Rotational invariance, or
  - Independent coordinates, or
  - Log-Sobolev Distributions (more generally, Poincare) [KSS18], or
  - Simple transformations thereof (simple within SoS)
- However, many subgaussians **do not** satisfy these sufficient conditions
- A plausible belief was that there might be some subgaussian distributions that are **not** certifiably subgaussian
  - Indeed, there is a huge gap between bounded & certifiably bounded
  - Under SSEH, there exist distributions that have bounded 1000-moments, but do not have certifiably-bounded 4-th moments [HL19]

Can we characterize certifiable subgaussianity?

Overview

- Motivation
- **Our Result**
- Proof Sketch
- Conclusion

Our result: certifiability of subgaussian distributions

Theorem: [DHPT24]

For all $d \in \mathbb{N}$, subgaussian distributions X on $\mathbb{R}^d$, and even $p \in \mathbb{N}$,

Our result: certifiability of subgaussian distributions

Theorem: [DHPT24]

For all $d \in \mathbb{N}$, subgaussian distributions X on $\mathbb{R}^d$, and even $p \in \mathbb{N}$,

$$(C\sqrt{p})^p \|v\|_2^p - \mathbb{E}_X[\langle v, X - \mu \rangle^p]$$

is a sum of square polynomials.

- Hence, all subgaussian distributions are certifiably so (upto an absolute constant C)

Our result: certifiability of subgaussian distributions

Theorem: [DHPT24]

For all $d \in \mathbb{N}$, subgaussian distributions X on $\mathbb{R}^d$, and even $p \in \mathbb{N}$,

$$(C\sqrt{p})^p \|v\|_2^p - \mathbb{E}_X[\langle v, X - \mu\rangle^p]$$

is a sum of square polynomials.

- Hence, all subgaussian distributions are certifiably so (upto an absolute constant C)
- **New algorithmic implications** for **subgaussian** data:
  - Robust statistics (next slide)
  - Clustering and mixture models
  - Sparse PCA, Distortion of a subspace, Hypercontractivity, . . .
  - Likely, more in the future

Our result: certifiability of subgaussian distributions

Theorem: [DHPT24]

For all $d \in \mathbb{N}$, subgaussian distributions X on $\mathbb{R}^d$, and even $p \in \mathbb{N}$,

$$(C\sqrt{p})^p \|v\|_2^p - \mathbb{E}_X[\langle v, X - \mu\rangle^p]$$

is a sum of square polynomials.

- Hence, all subgaussian distributions are certifiably so (upto an absolute constant C)
- New algorithmic implications for subgaussian data:
  - Robust statistics (next slide)
  - Clustering and mixture models
  - Sparse PCA, Distortion of a subspace, Hypercontractivity, . . .
  - Likely, more in the future
- **Evidence of computational hardness** for **gaussian** data certification

Our result: new algorithmic guarantees for robust statistics

Robust Estimation Task	Inlier Distribution	Information-theoretic Error	Previous Best Guarantee in Polynomial Time	New Guarantees
Mean estimation: Euclidean norm	subgaussian	$\widetilde{\Theta}(\epsilon)$	$\sqrt{\epsilon}$	$\epsilon^{1-1/m}$
List-decodable mean estimation	subgaussian	$\widetilde{\Theta}(\epsilon)$	$\sqrt{\frac{1}{1-\epsilon}}$	$\left(\frac{1}{1-\epsilon}\right)^{-\Omega(\frac{1}{m})}$
Mixture models: Mixture of k Δ-separated components	each component is subgaussian	$\Delta \gtrsim \sqrt{\log k}$	$\Delta \gtrsim k^{\Omega(1)}$	$\Delta \gtrsim k^{O(\frac{1}{m})}$
Mean estimation: Mahalanobis norm	hypercontractive subgaussian	$\widetilde{\Theta}(\epsilon)$	No general algorithm	$\epsilon^{1-1/m}$
Covariance estimation: Relative spectral norm	hypercontractive subgaussian	$\widetilde{\Theta}(\epsilon)$	No general algorithm	$\epsilon^{1-2/m}$
Linear regression: Arbitrary noise	hypercontractive subgaussian	$\widetilde{\Theta}(\epsilon)$	No general algorithm	$\epsilon^{1-\frac{2}{m}}$

Overview

- Motivation
- Our Result
- **Proof Sketch**
- Conclusion

Proof Sketch

- (**SoS Duality**) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \leq$ small

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \leq$ small
- (**Ghost samples** & empirical process) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \leq$ small
- (Ghost samples & **empirical process**) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$
  - Control $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i f(X_i)$, where $\mathcal{F} := \left\{x \mapsto \frac{\widetilde{\mathbb{E}}[\langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]} : \widetilde{\mathbb{E}} \text{ is degree-}p\right\}$

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}\text{:degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \leq$ small
- (Ghost samples & empirical process) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}\text{:degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$
  - Control $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i f(X_i)$, where $\mathcal{F} := \left\{x \mapsto \frac{\widetilde{\mathbb{E}}[\langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]} : \widetilde{\mathbb{E}} \text{ is degree-}p\right\}$
- (**Generic Chaining**) For any **linear** function class $\mathcal{F}_{\text{lin}}$

$$\mathbb{E}\left[\sup_{(f_1,\dots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(X_i)\right] \lesssim \mathbb{E}\left[\sup_{(f_1,\dots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(G_i)\right]$$

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \leq$ small
- (Ghost samples & empirical process) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$
  - Control $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i f(X_i)$, where $\mathcal{F} := \left\{x \mapsto \frac{\widetilde{\mathbb{E}}[\langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]} : \widetilde{\mathbb{E}} \text{ is degree-}p\right\}$
- (Generic Chaining) For any **linear** function class $\mathcal{F}_{\text{lin}}$

$$\mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(X_i)\right] \lesssim \mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(G_i)\right]$$

  - Sadly, our $\mathcal{F}$ is **nonlinear** (subset of degree-p polynomials)

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \le$ small
- (Ghost samples & empirical process) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$
  - Control $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i f(X_i)$, where $\mathcal{F} := \left\{x \mapsto \frac{\widetilde{\mathbb{E}}[\langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]} : \widetilde{\mathbb{E}} \text{ is degree-}p\right\}$
- (Generic Chaining) For any linear function class $\mathcal{F}_{\text{lin}}$

$$\mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(X_i)\right] \lesssim \mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(G_i)\right]$$

  - Sadly, our $\mathcal{F}$ is **nonlinear** (subset of degree-p polynomials)
- (**Linearization**) Luckily, $\widetilde{\mathbb{E}}[\sum_i \langle v, X_i\rangle^p]$ has a **linear-ish formualation**
  - Just like $\sum_i y_i^p = \|y\|_p^p = \sup_{v\in\mathbb{B}_q} \langle v, y\rangle^p$ by Hölder's inequality

Proof Sketch

- (SoS Duality) Equivalent to showing $\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\mathbb{E}_X[\langle v, X\rangle^p]]}{\widetilde{\mathbb{E}}[\|v\|^p]} \le$ small
- (Ghost samples & empirical process) $\mathbb{E}\left[\sup_{\widetilde{\mathbb{E}}:\text{degree-}p} \frac{\widetilde{\mathbb{E}}_v[\frac{1}{n}\sum_{i=1}^n \langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]}\right]$
  - Control $\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_i f(X_i)$, where $\mathcal{F} := \left\{x \mapsto \frac{\widetilde{\mathbb{E}}[\langle v, X_i\rangle^p]}{\widetilde{\mathbb{E}}[\|v\|^p]} : \widetilde{\mathbb{E}} \text{ is degree-}p\right\}$
- (Generic Chaining) For any linear function class $\mathcal{F}_{\text{lin}}$

$$\mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(X_i)\right] \lesssim \mathbb{E}\left[\sup_{(f_1,\ldots,f_n)\in\mathcal{F}_{\text{lin}}} \frac{1}{n}\sum_{i=1}^n f_i(G_i)\right]$$

  - Sadly, our $\mathcal{F}$ is nonlinear (subset of degree-p polynomials)
- (Linearization) Luckily, $\widetilde{\mathbb{E}}[\sum_i \langle v, X_i\rangle^p]$ has a linear-ish formualation
  - Just like $\sum_i y_i^p = \|y\|_p^p = \sup_{v\in\mathbb{B}_q}\langle v, y\rangle^p$ by Hölder's inequality
- **Duality**+**Linearization**+**Chaining** $\to$ Proof

Overview

- Motivation
- Our Result
- Proof Sketch
- Conclusion

Conclusion

- We showed that all subgaussian distributions are certifiably so
- Algorithmic implications (both upper and lower bounds)
- Many **open problems**:
  - As a starting point, what about subexponential distributions?
    - Generic chaining does not apply as is
  - Can we characterize the penalty of certifiably bounded moments?
    - Huge gaps between lower (SSEH-based hardness) & upper bounds
  - Can we develop faster algorithms without solving large SDPs?

Thank you and happy to chat more!