

On Passive Models of Property Testing

Shivam Nadimpalli



On Passive Models of Property Testing

Shivam Nadimpalli

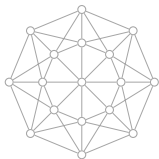


Property Testing

Q: Can we decide if a **massive object** has a desired property while inspecting a tiny fraction?

Property Testing

Q: Can we decide if a **massive object** has a desired property while inspecting a tiny fraction?

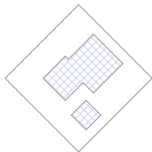
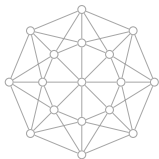


For example,

► graphs $\triangle$ -freeness, expansion, ...

Property Testing

Q: Can we decide if a **massive object** has a desired property while inspecting a tiny fraction?

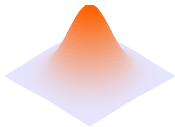
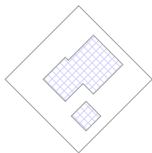
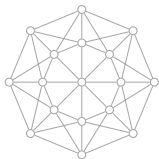


For example,

- ▶ graphs $\triangle$ -freeness, expansion, ...
- ▶ Boolean functions monotonicity, linearity, ...

Property Testing

Q: Can we decide if a **massive object** has a desired property while inspecting a tiny fraction?

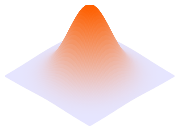
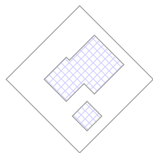
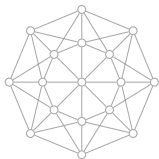


For example,

- ▶ graphs $\triangle$ -freeness, expansion, ...
- ▶ Boolean functions monotonicity, linearity, ...
- ▶ distributions uniformity, log-concavity, ...

Property Testing

Q: Can we decide if a **massive object** has a desired property while inspecting a tiny fraction?



For example,

- ▶ graphs $\triangle$ -freeness, expansion, ...
- ▶ Boolean functions monotonicity, linearity, ...
- ▶ distributions uniformity, log-concavity, ...
- ▶ matrices PSD-ness, sparsity, ...
- ▶ geometric objects convexity, halfspaces, ...

and many more...

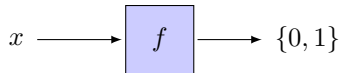
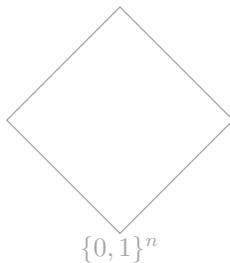
An Example

An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

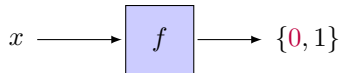
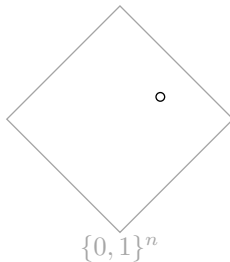
An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?



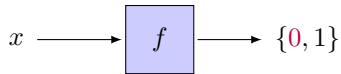
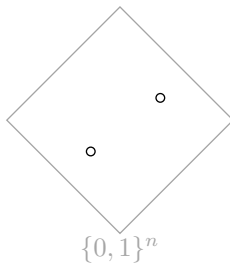
An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?



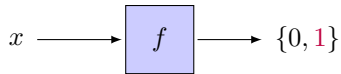
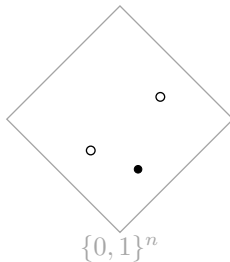
An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?



An Example

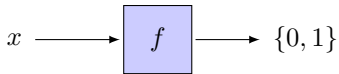
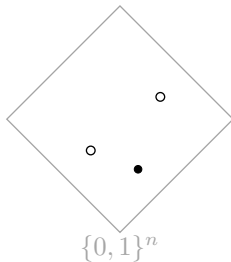
Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?



An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

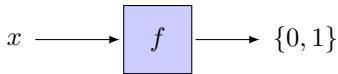
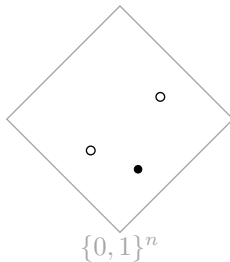
- ▶ Clearly $\Omega(2^n)$ queries in the worst case



An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

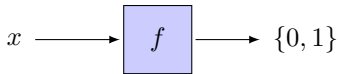
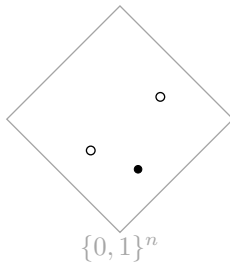
- ▶ Clearly $\Omega(2^n)$ queries in the worst case
- ▶ Natural relaxation:
 - Accept w.h.p. if $f \equiv 0$
 - Reject w.h.p. if f “far” from constant-0



An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

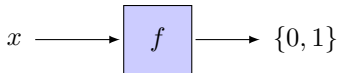
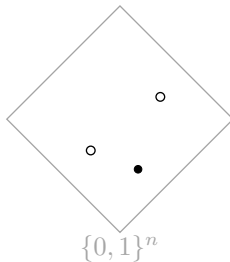
- ▶ Clearly $\Omega(2^n)$ queries in the worst case
- ▶ Natural relaxation:
 - Accept **w.h.p.** if $f \equiv 0$
 - Reject **w.h.p.** if f “far” from constant-0



An Example

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

- ▶ Clearly $\Omega(2^n)$ queries in the worst case
- ▶ Natural relaxation:
 - Accept w.h.p. if $f \equiv 0$
 - Reject w.h.p. if f “far” from constant-0



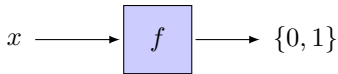
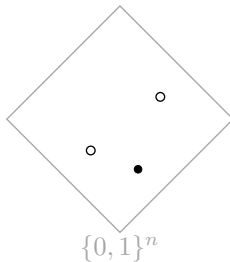
An Example

Q: Test if $f : \{0,1\}^n \rightarrow \{0,1\}$ is $f \equiv 0$?

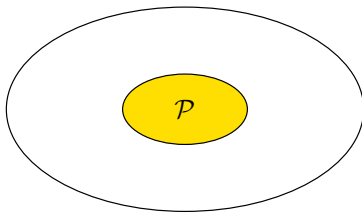
- ▶ Clearly $\Omega(2^n)$ queries in the worst case
- ▶ Natural relaxation:
 - Accept w.h.p. if $f \equiv 0$
 - Reject w.h.p. if f “far” from constant-0

$$\Pr[f(x) \neq 0] \geq \varepsilon$$

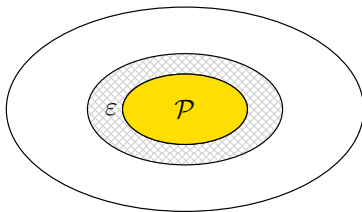
↑



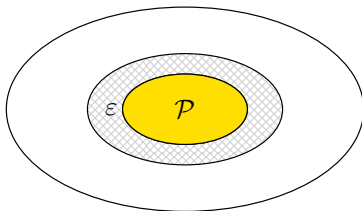
Property Testing $\equiv$ **Approximate** Decision Making



Property Testing $\equiv$ **Approximate** Decision Making



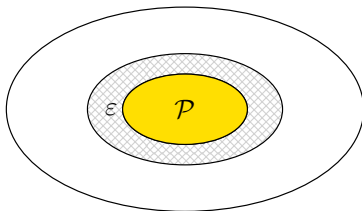
Property Testing $\equiv$ Approximate Decision Making



Definition: A tester for a property $\mathcal{P}$ is a randomized algorithm which, given access to an unknown function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, w.p. $2/3$:

- ▶ Accepts if $f \in \mathcal{P}$, and
- ▶ Rejects if f is ϵ -far from $\mathcal{P}$.

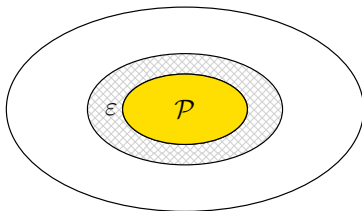
Property Testing $\equiv$ Approximate Decision Making



Definition: A tester for a property $\mathcal{P}$ is a **randomized algorithm** which, given access to an unknown function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, w.p. 2/3:

- ▶ Accepts if $f \in \mathcal{P}$, and
- ▶ Rejects if f is ϵ -far from $\mathcal{P}$.

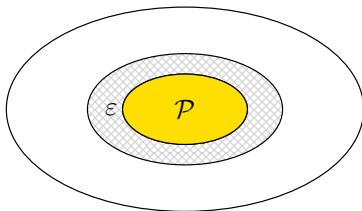
Property Testing $\equiv$ Approximate Decision Making



Definition: A tester for a property $\mathcal{P}$ is a randomized algorithm which, given access to an unknown function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, w.p. $2/3$:

- ▶ Accepts if $f \in \mathcal{P}$, and
- ▶ Rejects if f is ϵ -far from $\mathcal{P}$.

Property Testing $\equiv$ Approximate Decision Making



Definition: A tester for a property $\mathcal{P}$ is a randomized algorithm which, given access to an unknown function $f : \{0, 1\}^n \rightarrow \{0, 1\}$, w.p. $2/3$:

- ▶ Accepts if $f \in \mathcal{P}$, and
- ▶ Rejects if f is ϵ -far from $\mathcal{P}$.

Can often be done much more efficiently than learning f

A Glorious Field



- ▶ Applications throughout TCS
Coding theory, PCPs, hardness of approximation,
learning theory, quantum complexity, ...
- ▶ Rich set of techniques
Fourier analysis, regularity lemmas, CLTs,
communication complexity, ...

A Glorious Field and a Glorious Community



- ▶ Applications throughout TCS
Coding theory, PCPs, hardness of approximation, learning theory, quantum complexity, ...
- ▶ Rich set of techniques
Fourier analysis, regularity lemmas, CLTs, communication complexity, ...

This Workshop

June 25, 8:30am-11:00am

8:35-9:20	On Passive Models of Property Testing Shivam Nadimpalli (MIT)	▼
9:20-10:10	Testing high-dimensional distributions with subcube conditioning Erik Waingarten (University of Pennsylvania)	▼
10:10-11:00	Your data is blue, but can you prove it? Interactive proofs for distribution properties. Tal Herman (UC Berkeley and MIT)	▼

June 26, 8:30am-11:00am

8:30-9:20	Boolean function property testing: the next generation Rocco Servedio (Columbia University)	▼
9:20-10:10	Quality control in sublinear time Cassandra Marcussen (Harvard University)	▼
10:10-11:00	Testable Sanitization via Boundary Geometry Manolis Zampetakis (Yale University)	▼

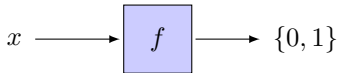
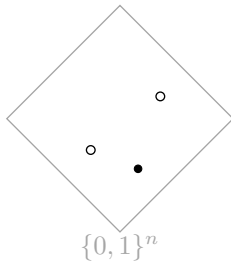
This Talk

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

- ▶ Clearly $\Omega(2^n)$ queries in the worst case
- ▶ Natural relaxation:
 - Accept w.h.p. if $f \equiv 0$
 - Reject w.h.p. if f “far” from constant-0

$$\Pr[f(x) \neq 0] \geq \varepsilon$$

↑

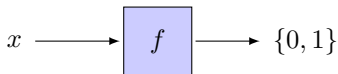
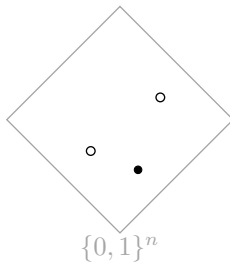


This Talk: **Active** vs. Passive Access

Q: Test if $f : \{0, 1\}^n \rightarrow \{0, 1\}$ is $f \equiv 0$?

- ▶ Clearly $\Omega(2^n)$ **queries** in the worst case
- ▶ Natural relaxation:
 - Accept w.h.p. if $f \equiv 0$
 - Reject w.h.p. if f “far” from constant-0

$$\Pr[f(x) \neq 0] \geq \varepsilon$$

This Talk: Three **Passive** Models

This Talk: Three **Passive** Models



Sample-Based Testing

This Talk: Three **Passive** Models



Sample-Based Testing



Truncated Statistics

This Talk: Three **Passive** Models



Sample-Based Testing



Truncated Statistics

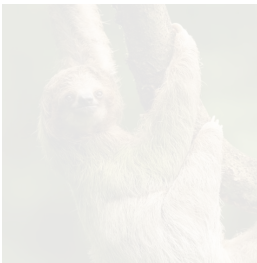


Testing Under Promises

This Talk: Three **Passive** Models



Sample-Based Testing



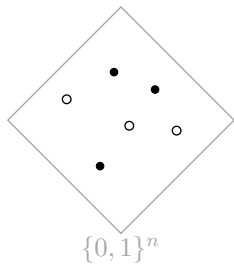
Truncated Statistics



Testing Under Promises

Sample-Based Testing

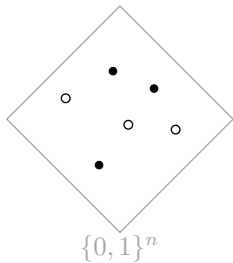
Sample-Based Testing



$\longrightarrow (x, f(x))$ for $x \sim \{0, 1\}^n$

Sample-Based Testing

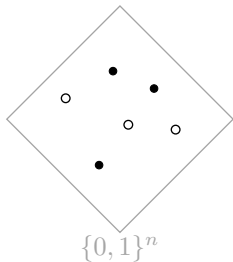
Recall that $\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \Pr_{x \sim \{0,1\}^n} [f \neq g]$



$\longrightarrow (x, f(x)) \text{ for } x \sim \{0,1\}^n$

Sample-Based Testing

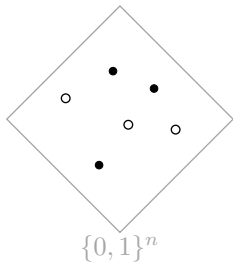
Recall that $\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \Pr_{x \sim \{0,1\}^n} [f \neq g]$



$\longrightarrow (x, f(x))$ for $x \sim \{0,1\}^n$

Can be **much harder** than standard testing:

Sample-Based Testing



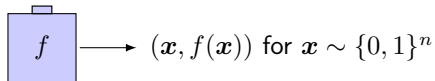
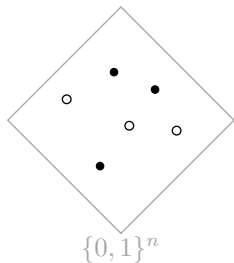
$\longrightarrow (x, f(x)) \text{ for } x \sim \{0, 1\}^n$

Recall that $\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \Pr_{x \sim \{0, 1\}^n} [f \neq g]$

Can be **much harder** than standard testing:

- **Monotonicity:** $2^{\Omega(n)}$ samples vs. $\Theta(\sqrt{n})$ queries
 $x \preceq y \implies f(x) \leq f(y)$

Sample-Based Testing

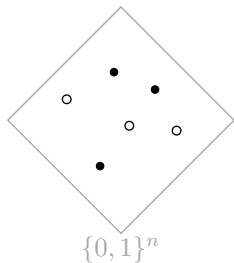


Recall that $\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \Pr_{x \sim \{0, 1\}^n} [f \neq g]$

Can be **much harder** than standard testing:

- ▶ **Monotonicity:** $2^{\Omega(n)}$ samples vs. $\Theta(\sqrt{n})$ queries
 $x \preceq y \implies f(x) \leq f(y)$
- ▶ **Juntas:** $\Omega(\log n)$ samples vs. $\Theta_k(1)$ queries
 f depends on only $k \ll n$ coordinates

Sample-Based Testing



→ $(x, f(x))$ for $x \sim \{0, 1\}^n$

Recall that $\text{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \Pr_{x \sim \{0, 1\}^n} [f \neq g]$

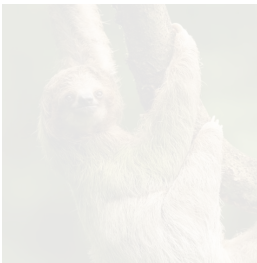
Can be **much harder** than standard testing:

- ▶ **Monotonicity:** $2^{\Omega(n)}$ samples vs. $\Theta(\sqrt{n})$ queries
 $x \preceq y \implies f(x) \leq f(y)$
- ▶ **Juntas:** $\Omega(\log n)$ samples vs. $\Theta_k(1)$ queries
 f depends on only $k \ll n$ coordinates
- ▶ **Linearity:** $\Omega(n)$ samples vs. $\Theta_\epsilon(1)$ queries
 $f(x \oplus y) = f(x) \oplus f(y)$

Talk Outline: Three **Passive** Models



Sample-Based Testing



Truncated Statistics

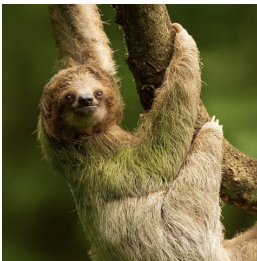


Testing Under Promises

Talk Outline: Three **Passive** Models



Sample-Based Testing



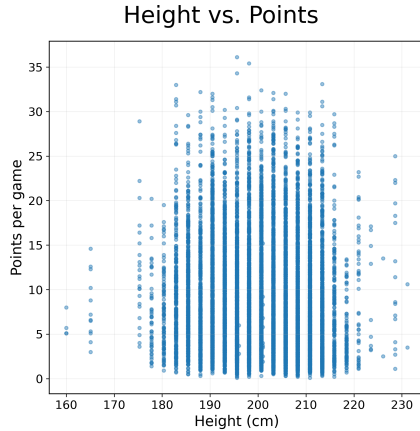
Truncated Statistics



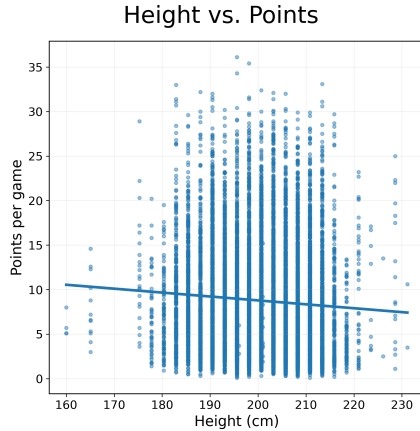
Testing Under Promises

An Appetizer

An Appetizer

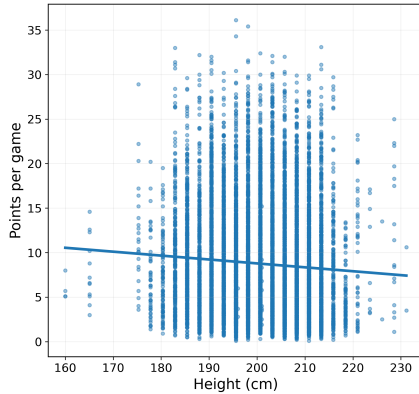


An Appetizer



An Appetizer

Height vs. Points in NBA



An Old Story:

An Old Story: Smallpox in the 18th Century



- ▶ 1760: No vaccines yet, only inoculation

An Old Story: Smallpox in the 18th Century



- ▶ 1760: No vaccines yet, only inoculation
- ▶ D. Bernoulli: Life expectancy if all inoculated?

An Old Story: Smallpox in the 18th Century

T A B L E I.

Âges par années.	Survivans filles M. Halley.	Néans pas eu la pet. vérole.	Ayant eu la pet. vérole.	Peu- ant la pet. vérole pendant l'année.	MORTS de la pet. vérole pendant chac. ann.	SOMME des morts de la pet. vérole.	MORTS par d'autres maladies pend. chac. année.
0	1500	1500	0				
1	1000	896	104	137	17.1	17.1	283
2	855	685	170	99	12.4	29.5	133
3	798	571	227	78	9.7	39.2	47
4	760	485	275	66	8.3	47.5	30
5	732	416	316	56	7.0	54.5	21
6	710	359	351	48	6.0	60.5	16
7	692	311	381	42	5.2	65.7	12.8
8	680	272	408	36	4.5	70.2	7.5
9	670	237	433	32	4.0	74.2	6
10	661	208	453	28	3.5	77.7	5.5
11	653	182	471	24.4	3.0	80.7	5
12	646	160	486	21.4	2.7	83.4	4.3
13	640	140	500	18.7	2.3	85.7	3.7
14	634	123	511	16.6	2.1	87.8	3.2
15	628	108	520	14.4	1.8	89.6	2.8
16	622	94	528	12.6	1.6	91.2	2.4
17	616	83	533	11.0	1.4	92.6	2.0
18	610	72	538	9.7	1.2	93.8	1.8
19	604	63	541	8.4	1.0	94.8	1.5
20	598	56	542	7.4	0.9	95.7	1.1
21	592	48.5	543	6.5	0.8	96.5	1.2
22	586	42.5	543	5.6	0.7	97.2	1.3
23	579	37	542	5.0	0.6	97.8	1.4
24	572	32.4	540	4.4	0.5	98.3	1.5

- ▶ 1760: No vaccines yet, only inoculation
- ▶ D. Bernoulli: Life expectancy if all inoculated?
- ▶ First example of inference from **truncated data**

Truncated Statistics Today

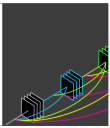
Truncated Statistics Today

- ▶ Biased & truncated data leads to **bad models**

Truncated Statistics Today

Reducing AI Bias using Truncated Statistics

Event Program
Location Room 116
Speaker(s) Constantinos Daskalakis
Date Wednesday, July 3, 2019
Time 11 a.m. - 12:30 p.m. PT



Reliable ML from Unreliable Data

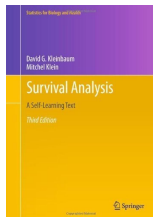
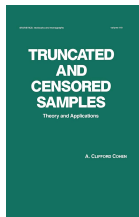
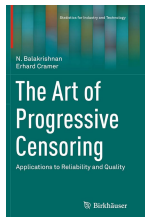
Workshop @ NeurIPS 2025

Upper Level Room 2, San Diego Convention Center | Dec 6, 2025

Official NeurIPS Workshop Page neurips.cc/virtual/2025/soc/san-diego/workshop/109580

- Biased & truncated data leads to **bad models**

Truncated Statistics Today



- ▶ Biased & truncated data leads to **bad models**
- ▶ Ongoing **explosion of work** on learning from biased, truncated & missing data

Truncated Statistics Today

Can SGD Select Good Fishermen? Local Convergence under Self-Selection Biases and Beyond

Alkis Kalivas Anay Mohotra Felix Zhou
Yale University Yale University Yale University
alkis.kalivas@yale.edu anaymohotra@gmail.com felix_zhou@yale.edu

Learning High-dimensional Gaussians from Censored Data

Arash Bhattacharyya Constantinos Daskalakis
The University of Warwick MIT
arash.bhattacharyya@warwick.ac.uk courtin@mit.edu
Thomas Goulis Yuhao Wang
Nanyang Technological University National University of Singapore
thomas.goulis@ntu.edu.sg yuhao.wang@nus.edu.sg

Sample-Efficient Linear Regression with Self-Selection Bias

Jaume Guitart Elchanan Mossel
Massachusetts Institute of Technology Massachusetts Institute of Technology
guitart@mit.edu mossel@mit.edu
February 23, 2024

Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians

Jane H. Lee Anay Mohotra Manolis Zampetakis
Yale University Yale University Yale University
jane.h.lee@yale.edu anaymohotra@gmail.com manolis.zampetakis@yale.edu

Truncated Linear Regression in High Dimensions

Constantinos Daskalakis Dhiraj Rohatgi Manolis Zampetakis
MIT MIT MIT
courtin@mit.edu drohatgi@mit.edu mzampet@mit.edu
July 30, 2020

- ▶ Biased & truncated data leads to **bad models**
- ▶ Ongoing **explosion of work** on learning from biased, truncated & missing data

Truncated Statistics Today



convex-truncated normal linear regression
If linear-truncated uniform then can do parameter learning

...

...

Truncated Statistics Today



convex-truncated normal linear regression
If linear-truncated uniform then can do parameter learning

...

...

Truncated Statistics Today



convex-truncated normal linear regression
If linear-truncated uniform then can do parameter learning

...

...

Q: Do **these** assumptions hold?

Setup

Setup

Question: Can you detect truncations of a distribution $\mathcal{D}$?

Setup

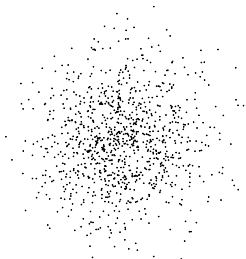
Question: Can you detect truncations of a distribution $\mathcal{D}$?

(For simplicity, suppose $\mathcal{D}$ is a standard Gaussian)

Setup

Question: Can you detect truncations of a distribution $\mathcal{D}$?

(For simplicity, suppose $\mathcal{D}$ is a standard Gaussian)

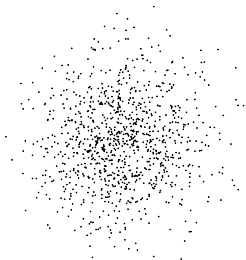


1000 samples from $N(0, 1)^2$

Setup

Question: Can you detect truncations of a distribution $\mathcal{D}$?

(For simplicity, suppose $\mathcal{D}$ is a standard Gaussian)



1000 samples from $N(0, 1)^2$

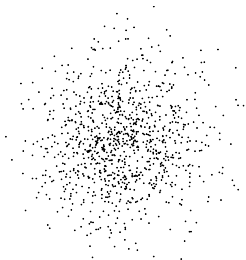


1000 samples from truncated $N(0, 1)^2$

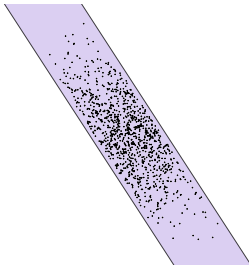
Setup

Question: Can you detect truncations of a distribution $\mathcal{D}$?

(For simplicity, suppose $\mathcal{D}$ is a standard Gaussian)



1000 samples from $N(0, 1)^2$



1000 samples from **truncated** $N(0, 1)^2$

A Natural Question: Testing Truncated Distributions

Question:

A Natural Question: Testing Truncated Distributions

Question: Given i.i.d. sample access to either

A Natural Question: Testing Truncated Distributions

Question: Given i.i.d. sample access to either

▶ $N(0, 1)^n$ or

A Natural Question: Testing Truncated Distributions

Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ **conditioned on some unknown set $K \subseteq \mathbb{R}^n$,**

A Natural Question: Testing Truncated Distributions

Question: Given i.i.d. sample access to either

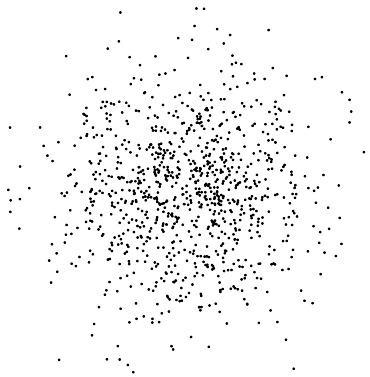
- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some unknown set $K \subseteq \mathbb{R}^n$,

how many samples needed to distinguish the two cases?

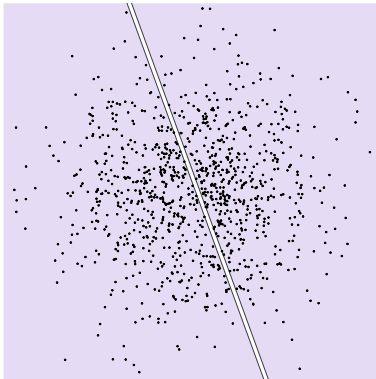
(As a function of dimension n)

Testing Truncated Distributions: Two Issues

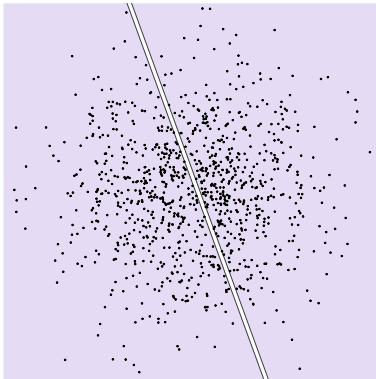
Testing Truncated Distributions: Two Issues



Testing Truncated Distributions: Two Issues



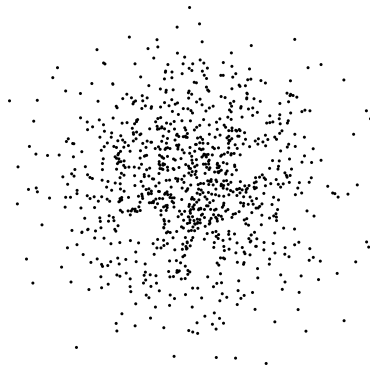
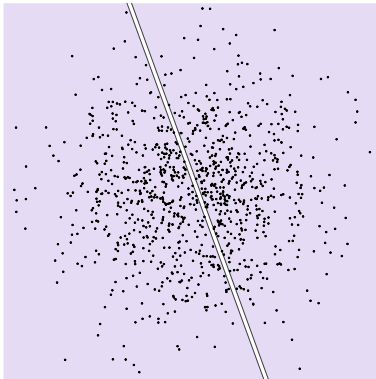
Testing Truncated Distributions: Two Issues



Truncation is “too small”

$$\Pr[x \in K] \approx 1$$

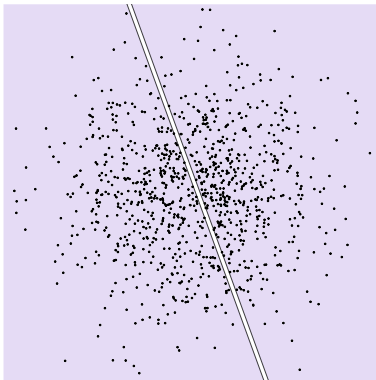
Testing Truncated Distributions: Two Issues



Truncation is “too small”

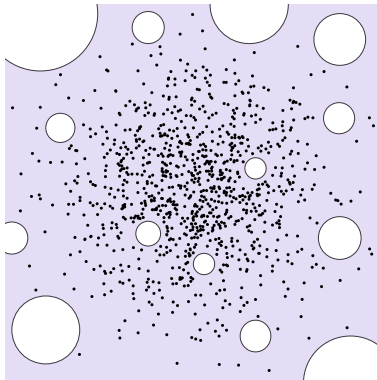
$$\Pr[x \in K] \approx 1$$

Testing Truncated Distributions: Two Issues

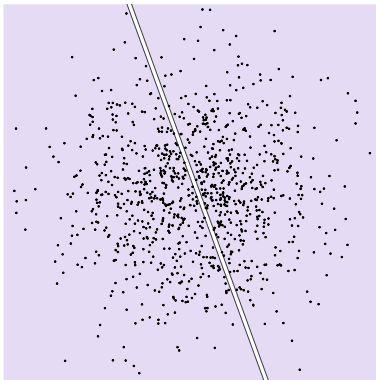


Truncation is “too small”

$$\Pr[x \in K] \approx 1$$

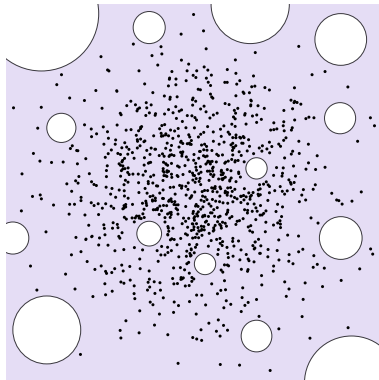


Testing Truncated Distributions: Two Issues



Truncation is “too small”

$$\Pr[x \in K] \approx 1$$



Truncation is “too complex”

Testing Truncated Distributions

[New Question:](#)

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

▶ $N(0, 1)^n$ or

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some unknown set $K \subseteq \mathbb{R}^n$

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some unknown set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some unknown set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$ and K is not “too complex”,

(As a function of dimension n)

Testing Truncated Distributions

New Question: Given i.i.d. sample access to either

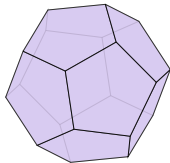
- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some unknown set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$ and K is **not “too complex”**,

how many samples needed to distinguish the two cases?

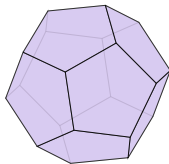
(As a function of dimension n)

Testing **Convex** Truncations

Testing **Convex** Truncations



Testing **Convex** Truncations

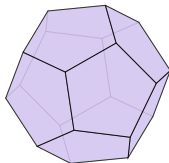


Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some **convex** set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$,

how many samples needed to distinguish the two cases?

Testing **Convex** Truncations



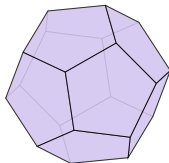
Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some **convex** set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$,

how many samples needed to distinguish the two cases?

[DNS, Math. Stat. Learn. '25]: $\Theta(n)$ samples necessary & sufficient

Testing **Convex** Truncations



Question: Given i.i.d. sample access to either

- ▶ $N(0, 1)^n$ or
- ▶ $N(0, 1)^n$ conditioned on some **convex** set $K \subseteq \mathbb{R}^n$ such that $\Pr[\mathbf{x} \in K] \leq 0.9$,

how many samples needed to distinguish the two cases?

[DNS, Math. Stat. Learn. '25]: $\Theta(n)$ **samples** necessary & sufficient

Testing Convex Truncations

Can SGD Select Good Fishermen? Local Convergence under Self-Selection Biases and Beyond

Alkis Kalarakis Anay Mehrotra Felix Zhou
Yale University Yale University Yale University
alkis.kalarakis@yale.edu anaymehrotra@gmail.com felix_zhou@yale.edu

Learning High-dimensional Gaussians from Censored Data

Arash Bhattaacharyya Constantine Daskalakis
The University of Warwick MIT
arash.bhattaacharyya@warwick.ac.uk courtin@mit.edu
Therese Gneibels Yukao Wang
Nanyang Technological University National University of Singapore
therese.gneibels@ntu.edu.sg yukao.wang@nus.edu.sg

Sample-Efficient Linear Regression with Self-Selection Bias

Jason Garivault Elchanan Mossel
Massachusetts Institute of Technology Massachusetts Institute of Technology
garivault@mit.edu elmos@mit.edu
February 23, 2024

Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians

Jane H. Lee Anay Mehrotra Manolis Zampetakis
Yale University Yale University Yale University
jane.h.lee@yale.edu anaymehrotra@gmail.com manolis.zampetakis@yale.edu

Truncated Linear Regression in High Dimensions

Constantinos Daskalakis Dhruv Rohatgi Manolis Zampetakis
MIT MIT MIT
courtin@mit.edu drohatgi@mit.edu mzampet@mit.edu
July 30, 2020

linear regression
If **convex-truncated normal** then parameter learning

...

Testing Convex Truncations

Can SGD Select Good Fishermen?
Local Convergence under Self-Selection Biases and Beyond

Alkis Kalarakis
Yale University
alkis.kalarakis@yale.edu

Anay Mehrotra
Yale University
anaymehrotra@gmail.com

Felix Zhou
Yale University
felix.zhou@yale.edu

Learning High-dimensional Gaussians from Censored Data

Arnab Bhattacharyya
The University of Warwick
arnab.bhattacharyya@warwick.ac.uk

Constantinos Daskalakis
MIT
cdaskal@mit.edu

Therisia Goulakakis
Nanyang Technological University
therisia.goulakakis@ntu.edu.sg

Yuhao Wang
National University of Singapore
yuhao.wang@nus.edu.sg

Sample-Efficient Linear Regression with Self-Selection Bias

Jason Garivault
Massachusetts Institute of Technology
garivault@mit.edu

Elchanan Mossel
Massachusetts Institute of Technology
elmossel@mit.edu

February 23, 2024

Efficient Statistics With Unknown Truncation,
Polynomial Time Algorithms, Beyond Gaussians

Jane H. Lee
Yale University
jane.h.lee@yale.edu

Anay Mehrotra
Yale University
anaymehrotra@gmail.com

Manolis Zampetakis
Yale University
manolis.zampetakis@yale.edu

Truncated Linear Regression in High Dimensions

Constantinos Daskalakis
MIT
cdaskal@mit.edu

Dhrub Kohli
MIT
dkohli@mit.edu

Manolis Zampetakis
MIT
mzampet@mit.edu

July 30, 2020

linear regression
If **convex-truncated normal** then parameter learning

...

- Can **verify assumption** with $\Theta(n)$ samples
[DNS, Math. Stat. & Learn. '25]

Testing Convex Truncations

Can SGD Select Good Fishermen? Local Convergence under Self-Selection Biases and Beyond

Alkis Kalarakis Anay Mehrotra Felix Zhou
Yale University Yale University Yale University
alkis.kalarakis@yale.edu anaymehrotra@gmail.com felix_zhou@yale.edu

Learning High-dimensional Gaussians from Censored Data

Arash Bhattaicharyya Constantinos Daskalakis
The University of Warwick MIT
arash.bhattaicharyya@warwick.ac.uk cdaskal@mit.edu
Therisia Goulakakis Yukao Wang
Nanyang Technological University National University of Singapore
therisia.goulakakis@ntu.edu.sg yukao.wang@nus.edu.sg

Sample-Efficient Linear Regression with Self-Selection Bias

Jason Garivault Elchanan Mossel
Massachusetts Institute of Technology Massachusetts Institute of Technology
garivault@mit.edu elmossel@mit.edu

February 23, 2024

Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians

Jane H. Lee Anay Mehrotra Manolis Zampetakis
Yale University Yale University Yale University
jane.h.lee@yale.edu anaymehrotra@gmail.com manolis.zampetakis@yale.edu

Truncated Linear Regression in High Dimensions

Constantinos Daskalakis Dhruv Rohatgi Manolis Zampetakis
MIT MIT MIT
cdaskal@mit.edu [drohat@mit.edu](mailto:droh@mit.edu) mzampet@mit.edu
July 30, 2020

If convex-truncated normal then **linear regression**
parameter learning

...

- Can verify assumption with $\Theta(n)$ samples
[DNS, Math. Stat. & Learn. '25]

Testing Convex Truncations



linear regression
If convex-truncated normal then parameter learning

...

- ▶ Can verify assumption with $\Theta(n)$ samples
[DNS, Math. Stat. & Learn. '25]
- ▶ Previously $\exp(\sqrt{n})$ samples
[DGTZ, FOCS '18; KTZ, FOCS '19]

Testing Convex Truncations with $O(n)$ Samples

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

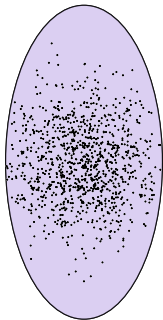
Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples



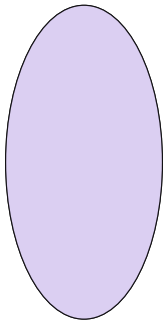
Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples



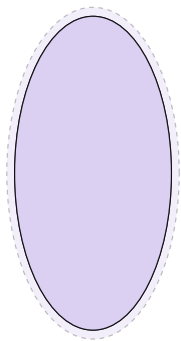
Testing Convex Truncations with $O(n)$ Samples

Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Testing Convex Truncations with $O(n)$ Samples

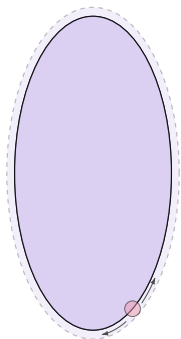
Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

Testing Convex Truncations with $O(n)$ Samples

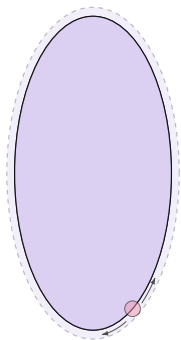
Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

Testing Convex Truncations with $O(n)$ Samples

Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

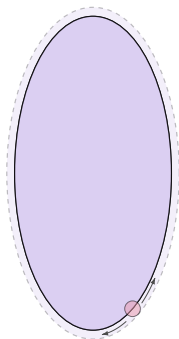


New complexity measure [DNS'23]

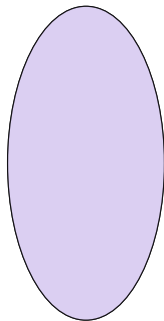


Testing Convex Truncations with $O(n)$ Samples

Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

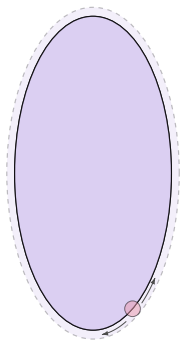


New complexity measure [DNS'23]

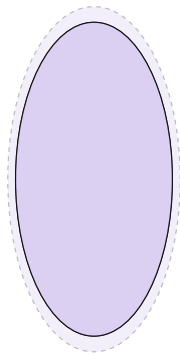


Testing Convex Truncations with $O(n)$ Samples

Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

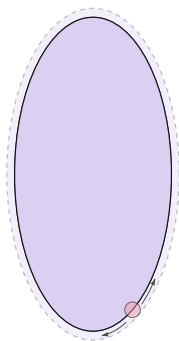


New complexity measure [DNS'23]

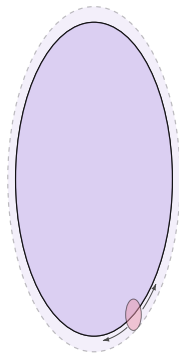


Testing Convex Truncations with $O(n)$ Samples

Idea: A **complexity measure** of conditioning set that can be estimated from one-sided samples



Surface area

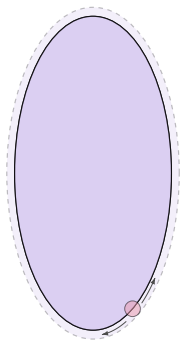


New complexity measure [DNS'23]

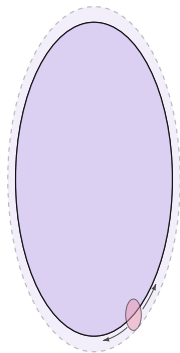


An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples



Surface area



New complexity measure [DNS'23]

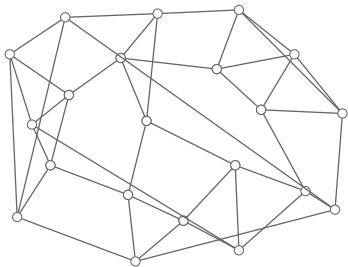


An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

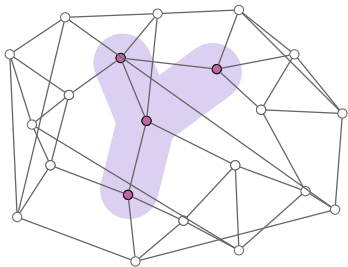
An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples



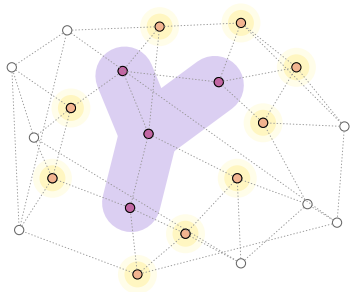
An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples



An Analogy with Graphs

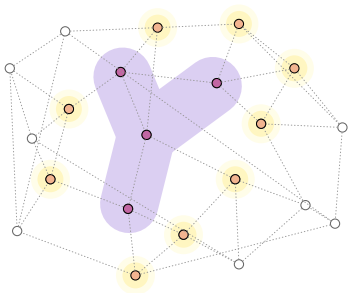
Idea: A complexity measure of conditioning set that can be estimated from one-sided samples



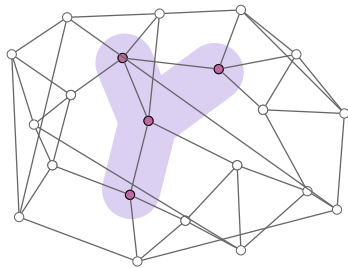
Vertex expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

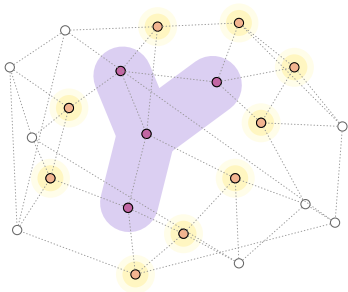


Vertex expansion

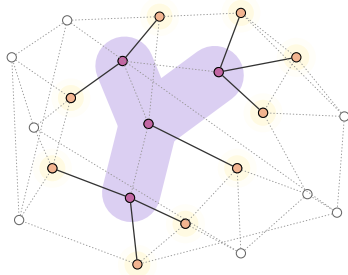


An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples



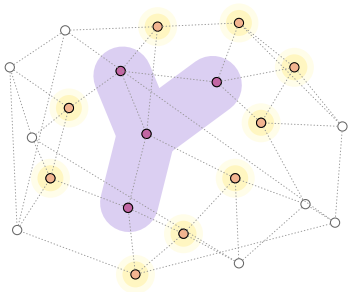
Vertex expansion



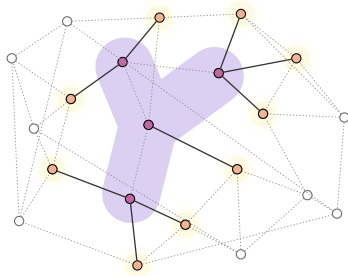
Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples



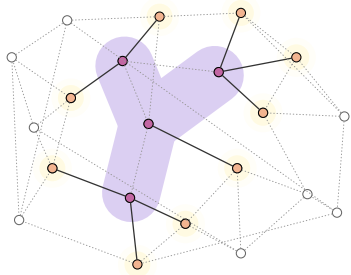
Vertex expansion



Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

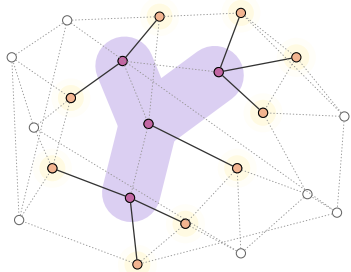


Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Adj} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}_{n \times n}$$



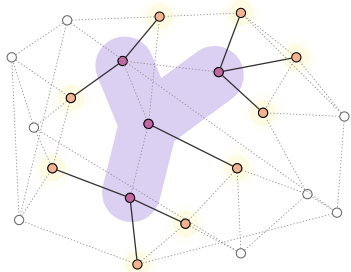
Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Adj} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}_{n \times n}$$

Eigenvalues of Adj



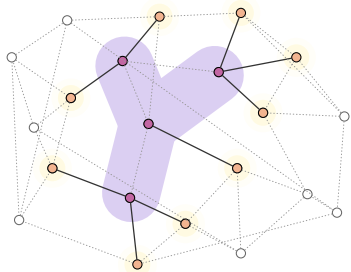
Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Adj} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}_{n \times n}$$

Eigenvalues of Adj



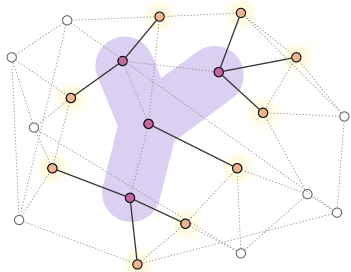
Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Adj} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}_{n \times n}$$

Eigenvalues of Adj



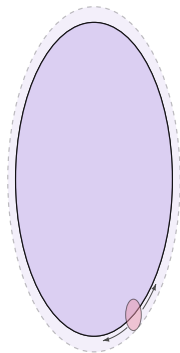
Edge expansion

An Analogy with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Adj} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}_{n \times n}$$

Eigenvalues of Adj



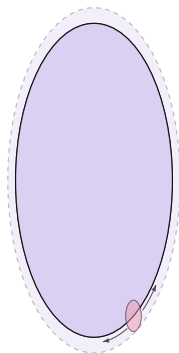
New complexity measure [DNS'23]

An **Analogy** with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

Eigenvalues of Cov



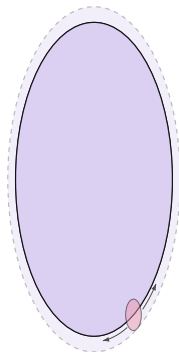
New complexity measure [DNS'23]

An **Analogy** with Graphs

Idea: A complexity measure of conditioning set that can be estimated from one-sided samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

Eigenvalues of Cov



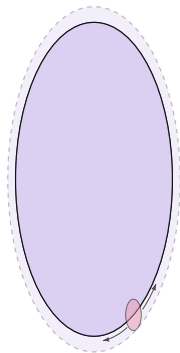
New complexity measure [DNS'23]

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

Eigenvalues of Cov



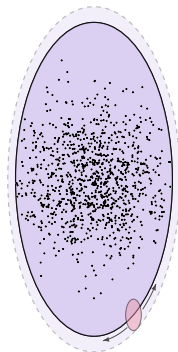
New complexity measure [DNS'23]

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

Eigenvalues of Cov



New complexity measure [DNS'23]

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

Eigenvalues of Cov

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

- Estimate **expansion** of truncation set from (empirical) spectrum of Cov
Expansion = $n - \text{tr}(\text{Cov})$

Eigenvalues of Cov

Testing Convex Truncations with $O(n)$ Samples

Idea: A complexity measure of conditioning set that can be estimated from **one-sided** samples

$$\text{Cov} = \left[\begin{array}{c} \mathbf{E}[x_i x_j] \text{ for} \\ x \sim \text{trunc. dist.} \end{array} \right]_{n \times n}$$

- ▶ Estimate **expansion** of truncation set from (empirical) spectrum of Cov

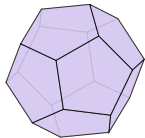
$$\text{Expansion} = n - \text{tr}(\text{Cov})$$

- ▶ Analysis requires quant. variant of Brascamp–Lieb inequality

$$\text{Var}_\mu[S] \leq \mathbf{E}_\mu \left[\nabla^\top S \cdot (\nabla^2 e^{-\mu})^{-1} \cdot \nabla S \right]$$

Eigenvalues of Cov

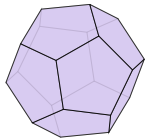
Testing Truncated Distributions,



Convex truncations of Gaussians

[DNS, Math. Stat. & Learn. '25]

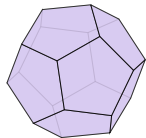
Testing Truncated Distributions, Optimally



Convex truncations of Gaussians

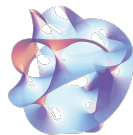
[DNS, Math. Stat. & Learn. '25]

Testing Truncated Distributions, Optimally



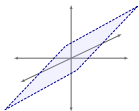
Convex truncations of Gaussians

[DNS, Math. Stat. & Learn. '25]



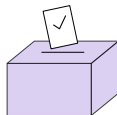
Low-degree trunks. of hypercontractive dists.

[DLNS, STOC '24]



Low-dim. truncations of product dists.

[BHK, SODA '26]



Unate truncations of product dists.

[LNQR, Manuscript '26]

Testing Truncated Distributions

Extensions in several directions & many open questions:

Testing Truncated Distributions

Extensions in several directions & many open questions:

[LMZ, STOC '24]

[LMZ, FOCS '24]

[GKK, SODA '24]

[KMG, STOC '25]

[CMZZ, NeurIPS '25]

[CGWY, AISTATS '25]

[BHK, SODA '26]

and more . . .

Testing Truncated Distributions

Extensions in several directions & many open questions:

[LMZ, STOC '24]

[LMZ, FOCS '24]

[GKK, SODA '24]

[KMG, STOC '25]

[CMZZ, NeurIPS '25]

[CGWY, AISTATS '25]

[BHK, SODA '26]

and more . . .

► More **general distributions**

Beyond Gaussians and product distributions

Testing Truncated Distributions

Extensions in several directions & many open questions:

[LMZ, STOC '24]

[LMZ, FOCS '24]

[GKK, SODA '24]

[KMG, STOC '25]

[CMZZ, NeurIPS '25]

[CGWY, AISTATS '25]

[BHK, SODA '26]

and more ...

- ▶ More general distributions

Beyond Gaussians and product distributions

- ▶ Quantitative **refinements**

Runtime optimality, refined analyses

Testing Truncated Distributions

Extensions in several directions & many open questions:

[LMZ, STOC '24]

[LMZ, FOCS '24]

[GKK, SODA '24]

[KMG, STOC '25]

[CMZZ, NeurIPS '25]

[CGWY, AISTATS '25]

[BHK, SODA '26]

and more ...

- ▶ More general distributions
Beyond Gaussians and product distributions
- ▶ Quantitative refinements
Runtime optimality, refined analyses
- ▶ **Connections** with other algorithmic tasks
Sampling, learning theory, sample-based testing, ...

Testing Truncated Distributions

Extensions in several directions & many open questions:

[LMZ, STOC '24]

[LMZ, FOCS '24]

[GKK, SODA '24]

[KMG, STOC '25]

[CMZZ, NeurIPS '25]

[CGWY, AISTATS '25]

[BHK, SODA '26]

and more ...

- ▶ **More general distributions**

Beyond Gaussians and product distributions

- ▶ **Quantitative refinements**

Runtime optimality, refined analyses

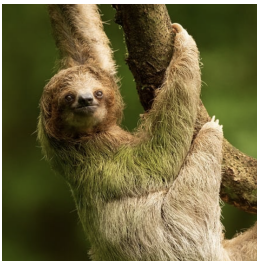
- ▶ **Connections with other algorithmic tasks**

Sampling, learning theory, sample-based testing, ...

Talk Outline: Three **Passive** Models



Sample-Based Testing



Truncated Statistics

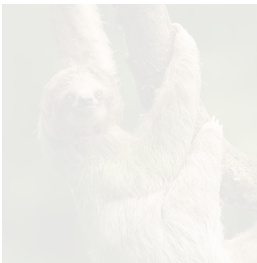


Testing Under Promises

Talk Outline: Three **Passive** Models



Sample-Based Testing

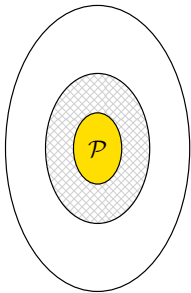


Truncated Statistics

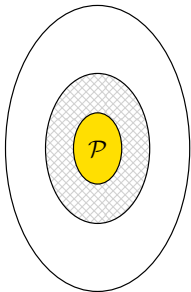


Testing Under Promises

Motivation

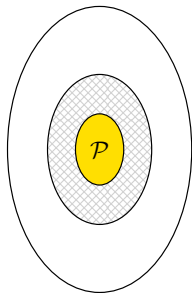


Motivation



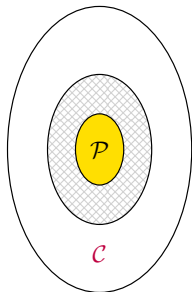
- ▶ Usual scenario: unknown f is any function

Motivation



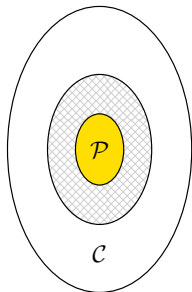
- ▶ Usual scenario: unknown f is any function
- ▶ But often f is not arbitrary...

Motivation



- ▶ Usual scenario: unknown f is any function
- ▶ But often f is not arbitrary...
- ▶ Given the **promise** that $f \in \mathcal{C}$, determine whether:
 - f has some property $\mathcal{P} \subseteq \mathcal{C}$, or
 - f is far from every function in $\mathcal{P}$.

Motivation



- ▶ Usual scenario: unknown f is any function
- ▶ But often f is not arbitrary...
- ▶ Given the **promise** that $f \in \mathcal{C}$, determine whether:
 - f has some property $\mathcal{P} \subseteq \mathcal{C}$, or
 - f is far from every function in $\mathcal{P}$.

Q: Can restricting to a **promise class** $\mathcal{C}$ make passive testing efficient?

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$

$$\begin{aligned}w_1 - w_2 + \cdots - w_n &= 7 \\-w_1 + w_2 + \cdots + w_n &= -2 \\&\vdots \\w_1 + w_2 - \cdots + w_n &= 11\end{aligned}$$

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$

$$\begin{aligned}w_1 - w_2 + \cdots - w_n &= 7 \\-w_1 + w_2 + \cdots + w_n &= -2 \\&\vdots \\w_1 + w_2 - \cdots + w_n &= 11\end{aligned}$$

Samples of form $(\mathbf{x}, \mathbf{y} := w \cdot \mathbf{x})$ where $\mathbf{x} \sim \{\pm 1\}^n$

Goal: Recover w .

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$

$$\begin{aligned}w_1 - w_2 + \cdots - w_n &= 7 \\-w_1 + w_2 + \cdots + w_n &= -2 \\&\vdots \\w_1 + w_2 - \cdots + w_n &= 11\end{aligned}$$

Samples of form $(x, y := w \cdot x)$ where $x \sim \{\pm 1\}^n$

Goal: Recover w .



$O(n)$ samples
Gaussian elimination

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$ promised to be k -sparse

$$\begin{aligned}w_1 - w_2 + \cdots - w_n &= 7 \\-w_1 + w_2 + \cdots + w_n &= -2 \\&\vdots \\w_1 + w_2 - \cdots + w_n &= 11\end{aligned}$$

Samples of form $(x, y := w \cdot x)$ where $x \sim \{\pm 1\}^n$

Goal: Recover w .

A Concrete Example: Sparse Recovery

Unknown vector $w \in \mathbb{R}^n$ promised to be k -sparse

$$\begin{aligned}w_1 - w_2 + \cdots - w_n &= 7 \\ -w_1 + w_2 + \cdots + w_n &= -2 \\ &\vdots \\ w_1 + w_2 - \cdots + w_n &= 11\end{aligned}$$

Samples of form $(x, y := w \cdot x)$ where $x \sim \{\pm 1\}^n$



$O(k \log n)$ samples + $\text{poly}(n)$ time
[Donoho, Candès–Romberg–Tao '05]

Goal: Recover w .

Testing Sparsity

Testing Sparsity of Linear Functions

Testing Sparsity of Linear Functions

Q: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector.

Testing Sparsity of Linear Functions

Q: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector. Given samples of the form $(x, y = x \cdot w)$ for $x \sim \{\pm 1\}^n$,

Testing Sparsity of Linear Functions

Q: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector. Given samples of the form $(x, y = x \cdot w)$ for $x \sim \{\pm 1\}^n$, distinguish between the case that

- ▶ the vector w is k -sparse, vs.
- ▶ for any k -sparse v , we have $\|w - v\|_2 \geq \varepsilon$.

Testing Sparsity of Linear Functions

Q: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector. Given samples of the form $(x, y = x \cdot w)$ for $x \sim \{\pm 1\}^n$, distinguish between the case that

- ▶ the vector w is k -sparse, vs.
- ▶ for any k -sparse v , we have $\|w - v\|_2 \geq \varepsilon$.

[Chen–De–Servedio '20]: $(k/\varepsilon)^{O(k)}$ samples suffice

The Full Result: Testing **Noisy** Linear Functions for Sparisty

The Full Result: Testing Noisy Linear Functions for Sparsity

[Chen–De–Servedio '20]: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector.

The Full Result: Testing Noisy Linear Functions for Sparsity

[Chen–De–Servedio '20]: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector.
Given samples

$$(\mathbf{x}, \mathbf{x} \cdot w) + \eta \quad \mathbf{x} \sim \mathcal{D}^n,$$

where $\mathcal{D}$ is an i.i.d. product distribution with moments of all orders and is not Gaussian

The Full Result: Testing Noisy Linear Functions for Sparsity

[Chen–De–Servedio '20]: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector.
Given samples

$$(\mathbf{x}, \mathbf{x} \cdot w) + \boldsymbol{\eta} \quad \mathbf{x} \sim \mathcal{D}^n,$$

where $\mathcal{D}$ is an i.i.d. product distribution with moments of all orders and is not Gaussian and $\boldsymbol{\eta}$ is a known noise distribution,

The Full Result: Testing Noisy Linear Functions for Sparsity

[Chen–De–Servedio '20]: Suppose $w \in \mathbb{S}^{n-1}$ is an unknown unit vector. Given samples

$$(\mathbf{x}, \mathbf{x} \cdot w) + \boldsymbol{\eta} \quad \mathbf{x} \sim \mathcal{D}^n,$$

where $\mathcal{D}$ is an i.i.d. product distribution with moments of all orders and is not Gaussian and $\boldsymbol{\eta}$ is a known noise distribution, one can distinguish

- ▶ the vector w is k -sparse, vs.
- ▶ for any k -sparse v , we have $\|w - v\|_2 \geq \varepsilon$

with $O_{\mathcal{D},k,\varepsilon,\eta}(1)$ samples.

Some Quick Remarks

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

- ▶ If $\mathcal{D}$ is Gaussian, then $\Omega(\log n)$ samples necessary even for $k = 1$

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

- ▶ If $\mathcal{D}$ is Gaussian, then $\Omega(\log n)$ samples necessary even for $k = 1$
- ▶ If **coordinates are product but not i.i.d.** then $\Omega(\log n)$ samples are necessary

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

- ▶ If $\mathcal{D}$ is Gaussian, then $\Omega(\log n)$ samples necessary even for $k = 1$
- ▶ If coordinates are product but not i.i.d. then $\Omega(\log n)$ samples are necessary
- ▶ If the **noise distribution is unknown**, then $\Omega(\log n)$ samples are necessary

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

- ▶ If $\mathcal{D}$ is Gaussian, then $\Omega(\log n)$ samples necessary even for $k = 1$
- ▶ If coordinates are product but not i.i.d. then $\Omega(\log n)$ samples are necessary
- ▶ If the noise distribution is unknown, then $\Omega(\log n)$ samples are necessary

Compared to $O(k \log n)$ recovery, **no dependence on ambient dimension n .**

Some Quick Remarks

Each hypothesis rules out a dimension-free sample complexity:

- ▶ If $\mathcal{D}$ is Gaussian, then $\Omega(\log n)$ samples necessary even for $k = 1$
- ▶ If coordinates are product but not i.i.d. then $\Omega(\log n)$ samples are necessary
- ▶ If the noise distribution is unknown, then $\Omega(\log n)$ samples are necessary

Compared to $O(k \log n)$ recovery, no dependence on ambient dimension n .

The [CDS'20] Approach:

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n$

Polynomial functions of moments:

$$\kappa_1(X) = \mathbf{E}[X]$$

$$\kappa_2(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

$$\kappa_3(X) = \mathbf{E}[X^3] - 3\mathbf{E}[X^2]\mathbf{E}[X] + 2\mathbf{E}[X]^3$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n$

Polynomial functions of moments:

$$\kappa_1(X) = \mathbf{E}[X]$$

$$\kappa_2(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

$$\kappa_3(X) = \mathbf{E}[X^3] - 3\mathbf{E}[X^2]\mathbf{E}[X] + 2\mathbf{E}[X]^3$$

and so on...

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n$

Polynomial functions of moments:

$$\kappa_1(X) = \mathbf{E}[X]$$

$$\kappa_2(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

$$\kappa_3(X) = \mathbf{E}[X^3] - 3\mathbf{E}[X^2]\mathbf{E}[X] + 2\mathbf{E}[X]^3$$

and so on...

- **Linearity:** if $\mathbf{a}, \mathbf{b}$ independent, then
$$\kappa_j(\mathbf{a} + \mathbf{b}) = \kappa_j(\mathbf{a}) + \kappa_j(\mathbf{b})$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(\mathbf{x}, \mathbf{y} = \mathbf{x} \cdot \mathbf{w} + \boldsymbol{\eta})$ for $\mathbf{x} \sim \{\pm 1\}^n$

Polynomial functions of moments:

$$\kappa_1(X) = \mathbf{E}[X]$$

$$\kappa_2(X) = \mathbf{E}[X^2] - \mathbf{E}[X]^2$$

$$\kappa_3(X) = \mathbf{E}[X^3] - 3\mathbf{E}[X^2]\mathbf{E}[X] + 2\mathbf{E}[X]^3$$

and so on...

- ▶ **Linearity:** if $\mathbf{a}, \mathbf{b}$ independent, then
$$\kappa_j(\mathbf{a} + \mathbf{b}) = \kappa_j(\mathbf{a}) + \kappa_j(\mathbf{b})$$
- ▶ **Homogeneity:** For scalar c ,
$$\kappa_j(c\mathbf{x}) = c^j \kappa_j(\mathbf{x})$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n = \mathcal{D}^n$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(\mathbf{x}, \mathbf{y} = \mathbf{x} \cdot \mathbf{w} + \boldsymbol{\eta})$ for $\mathbf{x} \sim \{\pm 1\}^n = \mathcal{D}^n$

So we have:

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(x, y = x \cdot w + \eta)$ for $x \sim \{\pm 1\}^n = \mathcal{D}^n$

So we have:

$$\kappa_j(\mathbf{y})$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(\mathbf{x}, \mathbf{y} = \mathbf{x} \cdot \mathbf{w} + \boldsymbol{\eta})$ for $\mathbf{x} \sim \{\pm 1\}^n = \mathcal{D}^n$

So we have:

$$\kappa_j(\mathbf{y}) = \sum_{i=1}^j \kappa_j(w_i \mathbf{x}_i) + \kappa_j(\boldsymbol{\eta}) \quad \text{linearity}$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(\mathbf{x}, \mathbf{y} = \mathbf{x} \cdot \mathbf{w} + \boldsymbol{\eta})$ for $\mathbf{x} \sim \{\pm 1\}^n = \mathcal{D}^n$

So we have:

$$\begin{aligned}\kappa_j(\mathbf{y}) &= \sum_{i=1}^j \kappa_j(w_i \mathbf{x}_i) + \kappa_j(\boldsymbol{\eta}) && \text{linearity} \\ &= \left(\sum_{i=1}^j w_i^j \right) \kappa_j(\mathcal{D}) + \kappa_j(\boldsymbol{\eta}), && \text{homogeneity}\end{aligned}$$

The [CDS'20] Approach: Cumulants

Let's focus on the special case $(\mathbf{x}, \mathbf{y} = \mathbf{x} \cdot \mathbf{w} + \boldsymbol{\eta})$ for $\mathbf{x} \sim \{\pm 1\}^n = \mathcal{D}^n$

So we have:

$$\begin{aligned}\kappa_j(\mathbf{y}) &= \sum_{i=1}^j \kappa_j(w_i \mathbf{x}_i) + \kappa_j(\boldsymbol{\eta}) && \text{linearity} \\ &= \left(\sum_{i=1}^j w_i^j \right) \kappa_j(\mathcal{D}) + \kappa_j(\boldsymbol{\eta}), && \text{homogeneity}\end{aligned}$$

or in other words $\|\mathbf{w}\|_j^j = \kappa_j(\mathcal{D})^{-1} \cdot (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$.

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

- ▶ We know $\mathcal{D}, \boldsymbol{\eta}$; estimate $\kappa_j(\mathbf{y})$.

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

- ▶ We know $\mathcal{D}, \boldsymbol{\eta}$; estimate $\kappa_j(\mathbf{y})$.
- ▶ If $|\kappa_j(\mathcal{D})| > 0$, estimate $\|w\|_j$.

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

- ▶ We know $\mathcal{D}, \boldsymbol{\eta}$; estimate $\kappa_j(\mathbf{y})$.
- ▶ If $|\kappa_j(\mathcal{D})| > 0$, estimate $\|w\|_j$.
- ▶ Since $\lim_{j \rightarrow \infty} \|w\|_j = \|w\|_\infty$, can choose $j_1, \dots, j_k$ to **estimate k -largest entries** of w .

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

- ▶ We know $\mathcal{D}, \boldsymbol{\eta}$; estimate $\kappa_j(\mathbf{y})$.
- ▶ If $|\kappa_j(\mathcal{D})| > 0$, estimate $\|w\|_j$.
- ▶ Since $\lim_{j \rightarrow \infty} \|w\|_j = \|w\|_\infty$, can choose $j_1, \dots, j_k$ to estimate k -largest entries of w .

The [CDS'20] Approach: Cumulants

$$\text{So far, we have } \|w\|_j^j = \frac{1}{\kappa_j(\mathcal{D})} (\kappa_j(\mathbf{y}) - \kappa_j(\boldsymbol{\eta}))$$

- ▶ We know $\mathcal{D}, \boldsymbol{\eta}$; estimate $\kappa_j(\mathbf{y})$.
- ▶ If $|\kappa_j(\mathcal{D})| > 0$, estimate $\|w\|_j$.
- ▶ Since $\lim_{j \rightarrow \infty} \|w\|_j = \|w\|_\infty$, can choose $j_1, \dots, j_k$ to estimate k -largest entries of w .

Fact: For Gaussian $\mathcal{D}$, $\kappa_j(\mathcal{D}) = 0$ for all $j > 2$.

The [CDS'20] Approach: Cumulants

The [CDS'20] Approach: Cumulants

Happily, this is the only thing that can go wrong 😊

The [CDS'20] Approach: Cumulants

Happily, this is the only thing that can go wrong 😊

[Marcinkiewicz '39]: If $\mathcal{D}$ has a finite number of non-zero cumulants, then it must be Gaussian.

The [CDS'20] Approach: Cumulants

Happily, this is the only thing that can go wrong ☺

[Marcinkiewicz '39]: If $\mathcal{D}$ has a finite number of non-zero cumulants, then it must be Gaussian.

- ▶ So for non-Gaussian $\mathcal{D}$, we have sequence of $j_1, \dots, j_k$ to estimate top- k entries of w .

The [CDS'20] Approach: Cumulants

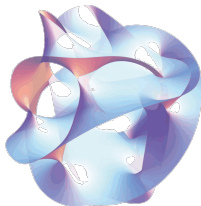
Happily, this is the only thing that can go wrong ☺

[Marcinkiewicz '39]: If $\mathcal{D}$ has a finite number of non-zero cumulants, then it must be Gaussian.

- ▶ So for non-Gaussian $\mathcal{D}$, we have sequence of $j_1, \dots, j_k$ to estimate top- k entries of w .
- ▶ Need to prove (first) **quantitative** strengthening of [M'39].

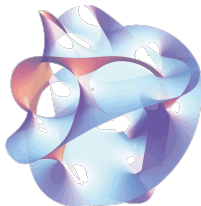
Some Follow-Up Work

Some Follow-Up Work

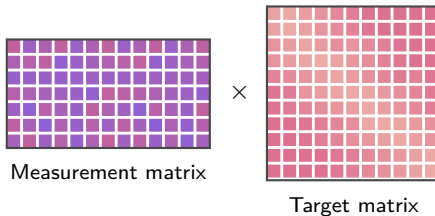


Testing **Low-Degree Polynomials** for Sparsity
[BDNSW, STOC'26]

Some Follow-Up Work



Testing **Low-Degree Polynomials** for Sparsity
[BDNSW, STOC'26]

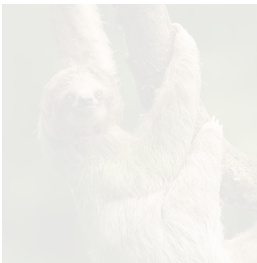


Testing **Matrix Rank**
[BDNSSW, In Prep.]

Talk Outline: Three **Passive** Models



Sample-Based Testing



Truncated Statistics



Testing Under Promises

Talk Outline: Three **Passive** Models



Sample-Based Testing



Truncated Statistics



Testing Under Promises

Summary



Summary



- ▶ Passive access in many “beyond worst-case” settings

Summary



- ▶ Passive access in many “beyond worst-case” settings
- ▶ New **modeling** as well as **mathematical** challenges:

Summary



- ▶ Passive access in many “beyond worst-case” settings
- ▶ New **modeling** as well as **mathematical** challenges:
 - Right question depends on the problem instance

Summary



- ▶ Passive access in many “beyond worst-case” settings
- ▶ New **modeling** as well as **mathematical** challenges:
 - Right question depends on the problem instance
 - Connections across analysis, complexity, probability, & statistics

Summary



- ▶ Passive access in many “beyond worst-case” settings
- ▶ New **modeling** as well as **mathematical** challenges:
 - Right question depends on the problem instance
 - Connections across analysis, complexity, probability, & statistics

Thank you! Questions?