Efficient Truncated Statistics with Unknown Truncation

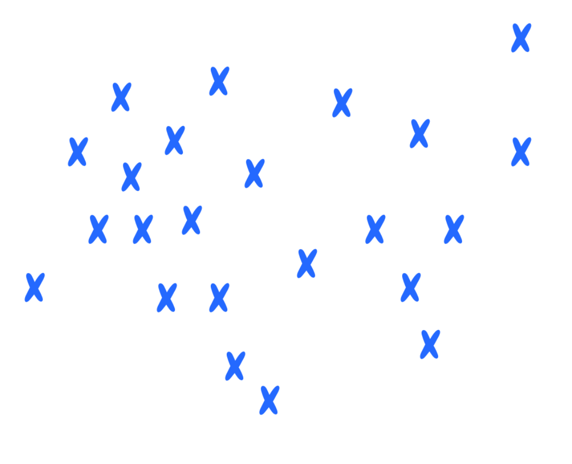
Vasilis Kontonis (UW-Madison)

Christos Tzamos (UW-Madison)

Manolis Zambetakis (MIT)

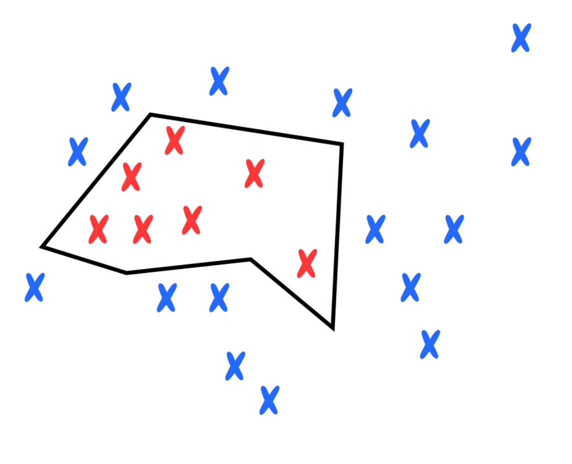
Truncated Data

We want to estimate the mean of a population.



Truncated Data

But we're given only data from a **subset** of space.



Poincare's Baker

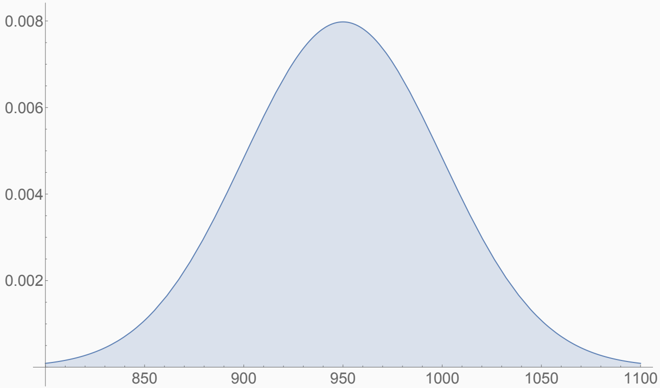
- Poincare's baker was advertising his loaves to be 1Kg.

Poincare's Baker

- Poincare's baker was advertising his loaves to be 1Kg.
- Poincare weighted the bread he bought.

Poincare's Baker

- Poincare's baker was advertising his loaves to be 1Kg.
- Poincare weighted the bread he bought.
- Average weight was 950 grams!



Next Year

- After another year of bread data...

Next Year

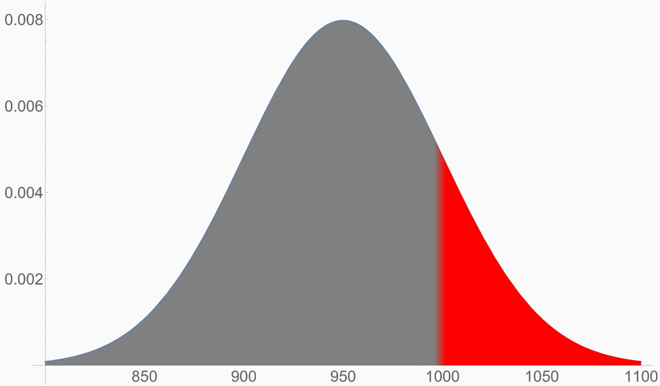
- After another year of bread data...
- All Poincare's loaves were above 1 Kg...

Next Year

- After another year of bread data...
- All Poincare's loaves were above 1 Kg...
- But Poincare complained again! Average weight was still 950 grams!

Next Year

- After another year of bread data...
- All Poincare's loaves were above 1 Kg...
- But Poincare complained again! Average weight was still 950 grams!



Truncated Normals

Truncated Normal Distribution

$$\mathcal{N}(\mu, \Sigma, S; x) = \frac{\mathbf{1}_S(x)}{\alpha} \mathcal{N}(\mu, \Sigma; x),$$

Truncated Normals

Truncated Normal Distribution

$$\mathcal{N}(\mu, \Sigma, S; x) = \frac{\mathbf{1}_S(x)}{\alpha} \mathcal{N}(\mu, \Sigma; x),$$

$$\alpha = \int \mathbf{1}_S(x) \mathcal{N}(\mu, \Sigma; x) \mathrm{d}x$$

We assume that the set S has (Gaussian) mass α at least 1%.

Truncated Normals

Truncated Normal Distribution

$$\mathcal{N}(\mu, \Sigma, S; x) = \frac{\mathbf{1}_S(x)}{\alpha} \mathcal{N}(\mu, \Sigma; x),$$

$$\alpha = \int \mathbf{1}_S(x) \mathcal{N}(\mu, \Sigma; x) \mathrm{d}x$$

We assume that the set S has (Gaussian) mass α at least 1%.

Estimation Problem

- Data $x_i \sim \mathcal{N}(\mu, \Sigma, S)$

Truncated Normals

Truncated Normal Distribution

$$\mathcal{N}(\mu, \Sigma, S; x) = \frac{\mathbf{1}_S(x)}{\alpha} \mathcal{N}(\mu, \Sigma; x),$$

$$\alpha = \int \mathbf{1}_S(x) \mathcal{N}(\mu, \Sigma; x) \mathrm{d}x$$

We assume that the set S has (Gaussian) mass α at least 1%.

Estimation Problem

- Data $x_i \sim \mathcal{N}(\mu, \Sigma, S)$
- Find $\widetilde{\mu}, \widetilde{\Sigma}$ such that

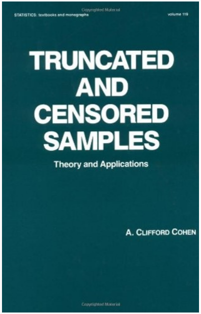
$$d_{\mathrm{tv}}(\mathcal{N}(\mu, \Sigma), \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})) \leq \varepsilon$$

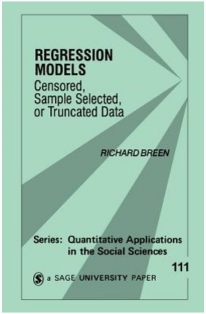
Previous Work

- Has **long history** in statistics that dates back to Galton and Pearson.

Previous Work

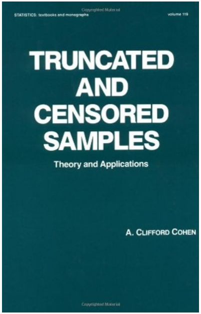
- Has **long history** in statistics that dates back to Galton and Pearson.

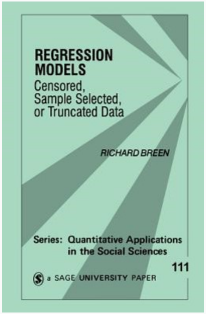




Previous Work

- Has **long history** in statistics that dates back to Galton and Pearson.





- **simple** truncation sets are considered: left or box truncation etc.

Previous Work

Daskalakis, Gouleakis, Tzamos, Zambetakis, FOCS 2018.

- Assume that the set S is known. Membership access to the set.
- $\widetilde{O}(d^2/\varepsilon^2)$ samples suffice to learn the parameters.

Previous Work

Daskalakis, Gouleakis, Tzamos, Zambetakis, FOCS 2018.

- Assume that the set S is known. Membership access to the set.
- $\widetilde{O}(d^2/\varepsilon^2)$ samples suffice to learn the parameters.
- S unknown?
  They construct a very complicated truncation set that makes it information theoretically impossible.

Previous Work

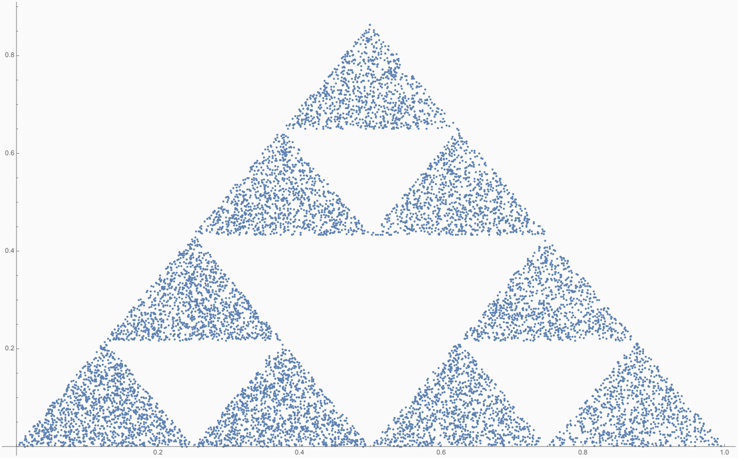
Daskalakis, Gouleakis, Tzamos, Zambetakis, FOCS 2018.

- Assume that the set S is known. Membership access to the set.
- $\widetilde{O}(d^2/\varepsilon^2)$ samples suffice to learn the parameters.
- S unknown?
  They construct a very complicated truncation set that makes it information theoretically impossible.

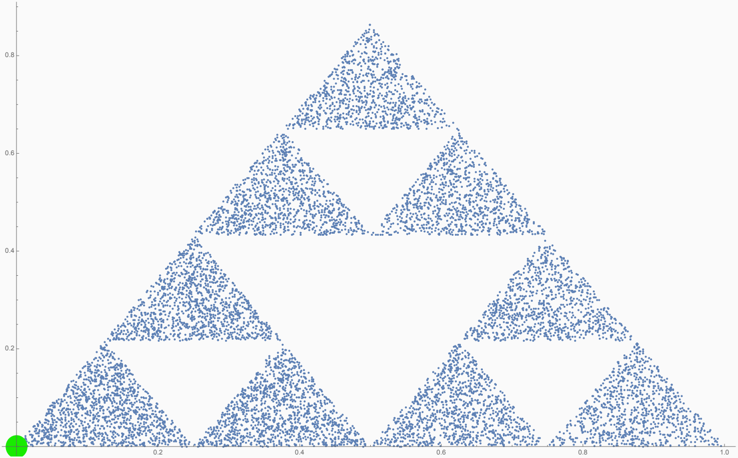
Main Open Problem

- Truncation S is unknown and of bounded "complexity".

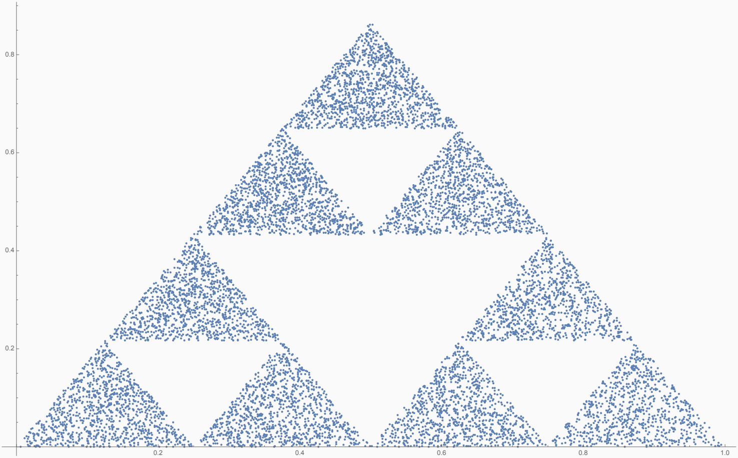
Can you find the mean?



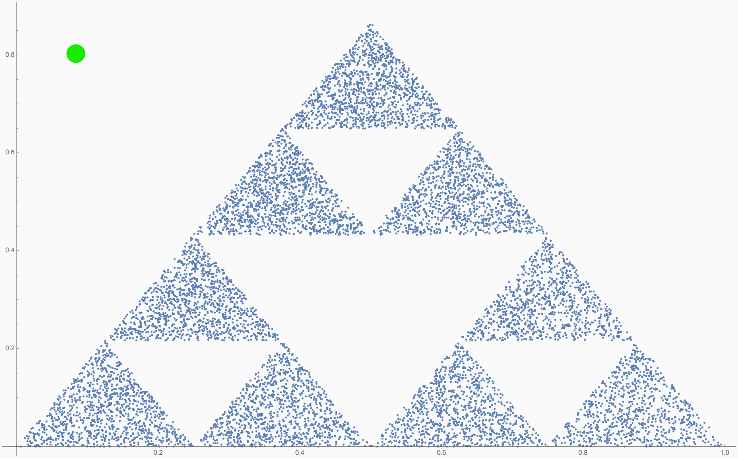
Here it is!



This is a very different Gaussian



This time the mean is (0.1,0.8)



Our Results: Sample Complexity via VC-dimension

Theorem: Sample Complexity via VC dimension

If the class $\mathcal{S}$ of sets of $\mathbb{R}^d$ has VC-dimension $\mathrm{VC}(\mathcal{S})$ then with

$$\widetilde{O}\left(\frac{d^2}{\varepsilon^2} + \frac{\mathrm{VC}(\mathcal{S})}{\varepsilon}\right)$$

samples, we obtain $\widetilde{\mu}, \widetilde{\Sigma}$ such that $d_{\mathrm{tv}}(N(\mu, \Sigma), \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})) \leq \varepsilon$

Our Results: Sample Complexity via VC-dimension

Theorem: Sample Complexity via VC dimension
If the class $\mathcal{S}$ of sets of $\mathbb{R}^d$ has VC-dimension $\mathrm{VC}(\mathcal{S})$ then with

$$\widetilde{O}\left(\frac{d^2}{\varepsilon^2} + \frac{\mathrm{VC}(\mathcal{S})}{\varepsilon}\right)$$

samples, we obtain $\widetilde{\mu}, \widetilde{\Sigma}$ such that $d_{\mathrm{tv}}(N(\mu, \Sigma), \mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma})) \leq \varepsilon$

Theorem: Lower Bound
We construct a family $\mathcal{S}$ with $\mathrm{VC}(\mathcal{S}) = O(2^d)$ such that getting a $\widetilde{\boldsymbol{\mu}}$ with $\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_2 \leq 1$ requires $\Omega(2^{d/2})$ samples.

Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.

Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

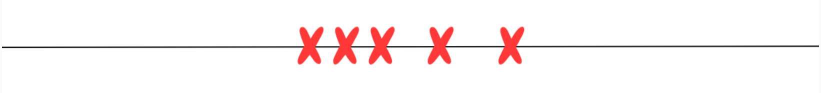
- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?

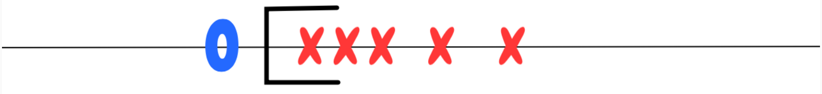


Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?

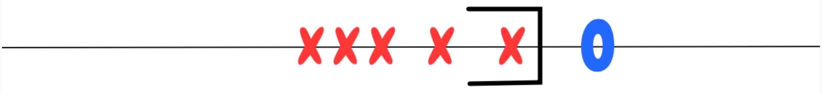


Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?

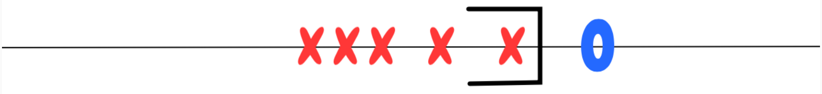


Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?



- We find $(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S})$ such that

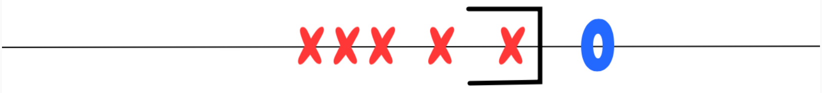
$$d_{\text{tv}}(\mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S}), \mathcal{N}(\mu, \Sigma, S)) \leq \varepsilon$$

Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?



- We find $(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S})$ such that

$$d_{\mathrm{tv}}(\mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S}), \mathcal{N}(\mu, \Sigma, S)) \leq \varepsilon$$

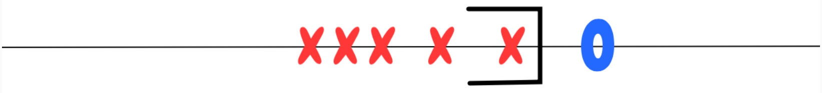
- Is this enough?

Our Results: Sample Complexity via VC-dimension

First learn the truncation set?

- The task is coupled with finding μ, Σ.
- Finding a set that contains all the samples is **not enough**.

Left or **right** truncation?



- We find $(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S})$ such that

$$d_{\mathrm{tv}}(\mathcal{N}(\widetilde{\mu}, \widetilde{\Sigma}, \widetilde{S}), \mathcal{N}(\mu, \Sigma, S)) \leq \varepsilon$$

- Is this enough?

  Yes!

Our Results: Sample Complexity via VC-dimension

Algorithm?

- We need to find a set that contains the samples.
- Not clear how to get generic algorithm for *all* sets of low VC-dimension.

Gaussian Surface Area (GSA)

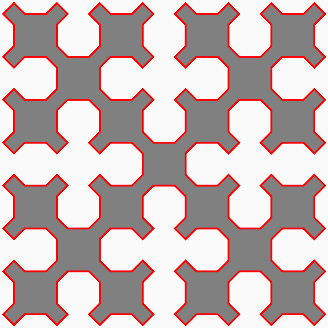
Gaussian Surface Area (GSA), $\Gamma(\mathcal{S})$

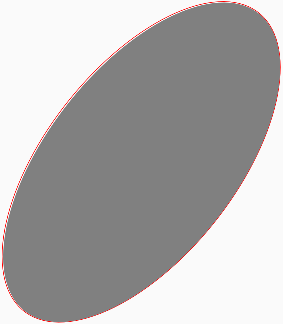
- Surface of S with respect to the Gaussian measure.

Gaussian Surface Area (GSA)

Gaussian Surface Area (GSA), $\Gamma(\mathcal{S})$

- Surface of S with respect to the Gaussian measure.





- $\Gamma(S) \leq \gamma$.

Our Results: Efficient Algorithm via Gaussian Surface Area

Theorem: Moment Matching
Two truncated Normals with $\kappa = O(\gamma^2/\varepsilon^8)$ "matching" moments are in TVD ε.

Our Results: Efficient Algorithm via Gaussian Surface Area

Theorem: Moment Matching

Two truncated Normals with $\kappa = O(\gamma^2/\varepsilon^8)$ "matching" moments are in TVD ε.

Theorem: Efficient Algorithm via GSA

With d^κ samples, in time $\mathrm{poly}(\#\mathrm{samples})$ we find $\widetilde{\boldsymbol{\mu}}$ such that $\|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2 \leq \varepsilon$.

Our Results: Efficient Algorithm via Gaussian Surface Area

Theorem: Moment Matching

Two truncated Normals with $\kappa = O(\gamma^2/\varepsilon^8)$ "matching" moments are in TVD ε.

Theorem: Efficient Algorithm via GSA

With d^κ samples, in time $\mathrm{poly}(\#\mathrm{samples})$ we find $\widetilde{\boldsymbol{\mu}}$ such that $\|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2 \le \varepsilon$.

Theorem: Lower Bound

We construct a family $\mathcal{S}$ with GSA $O(d)$ such that getting a $\widetilde{\boldsymbol{\mu}}$ with $\|\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}\|_2 \le 1$ requires $\Omega(2^{d/2})$ samples.

Performance of the Algorithm

Concept Class	GSA (γ)		Samples
degree k PTF	k	Kane '11	$d^{O(k^2)}$
inter. k halfspaces	$\sqrt{\log k}$	Klivans, O'Donnell, Servedio '08	$d^{O(\log k)}$
general convex sets	$d^{1/4}$	Ball '93	$d^{O(\sqrt{d})}$

Performance of the Algorithm

Concept Class	GSA (γ)		Samples
degree k PTF	k	Kane '11	$d^{O(k^2)}$
inter. k halfspaces	$\sqrt{\log k}$	Klivans, O'Donnell, Servedio '08	$d^{O(\log k)}$
general convex sets	$d^{1/4}$	Ball '93	$d^{O(\sqrt{d})}$

Main Ingredients of Algorithm

- Polynomial Approximation.
- Stochastic Gradient Descent.

Polynomial Approximation

- Hermite Polynomials

$$h_0(x) = 1,\ h_1(x) = x,\ h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \ldots$$

Polynomial Approximation

- Hermite Polynomials

$$h_0(x) = 1,\ h_1(x) = x,\ h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \ldots$$

- Orthonormal basis w.r.t $\mathcal{N}_0$.

Polynomial Approximation

- Hermite Polynomials

$$h_0(x) = 1,\ h_1(x) = x,\ h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \ldots$$

- Orthonormal basis w.r.t $\mathcal{N}_0$.
- **Approximation** of a function f.

$$p_\kappa(x) = \sum_{V:|V|\leq\kappa} \widehat{f}(V) H_V(x) \qquad \widehat{f}(V) = \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}[H_V(x) f(x)]$$

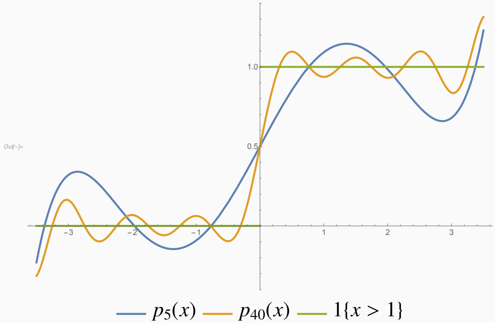
Polynomial Approximation

- Hermite Polynomials

$$h_0(x) = 1,\ h_1(x) = x,\ h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \ldots$$

- Orthonormal basis w.r.t $\mathcal{N}_0$.
- **Approximation** of a function f.

$$p_{\kappa}(x) = \sum_{V:|V|\le\kappa} \widehat{f}(V) H_V(x) \qquad \widehat{f}(V) = \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}[H_V(x) f(x)]$$



Learning a Weighted Indicator

Idea

- Let's see what we learn if we **evaluate** the Hermite polynomials on the samples.

Learning a Weighted Indicator

Idea

- Let's see what we learn if we **evaluate** the Hermite polynomials on the samples.
- Draw $x_1, \ldots, x_m$ from $\mathcal{N}(\mu, \mathrm{I}, S)$

$$\widetilde{c_V} = \frac{1}{m} \sum_{i=1}^{m} H_V(x_i)$$

Learning a Weighted Indicator

Idea

- Let's see what we learn if we **evaluate** the Hermite polynomials on the samples.
- Draw $x_1, \dots, x_m$ from $\mathcal{N}(\mu, \mathrm{I}, S)$

$$\widetilde{c_V} = \frac{1}{m} \sum_{i=1}^{m} H_V(x_i)$$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}(\mu,\mathrm{I},S)} [\widetilde{c_V}] = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_0} \left[H_V(x) \frac{\mathbf{1}_S(x)}{\alpha} \frac{\mathcal{N}(\mu, \mathrm{I}; x)}{\mathcal{N}_0(x)} \right]$$

Learning a Weighted Indicator

Idea

- Let's see what we learn if we **evaluate** the Hermite polynomials on the samples.
- Draw $x_1, \dots, x_m$ from $\mathcal{N}(\mu, \mathrm{I}, S)$

$$\widetilde{c_V} = \frac{1}{m}\sum_{i=1}^{m} H_V(x_i) \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}[H_V(x) f(x)]$$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(\mu,\mathrm{I},S)}[\widetilde{c_V}] = \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}\left[H_V(x)\underbrace{\frac{\mathbf{1}_S(x)}{\alpha}\frac{\mathcal{N}(\mu,\mathrm{I};x)}{\mathcal{N}_0(x)}}_{\psi(x)}\right]$$

Learning a Weighted Indicator

Idea

- Let's see what we learn if we **evaluate** the Hermite polynomials on the samples.
- Draw $x_1, \dots, x_m$ from $\mathcal{N}(\mu, \mathrm{I}, S)$

$$\widetilde{c_V} = \frac{1}{m}\sum_{i=1}^{m} H_V(x_i) \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}[H_V(x) f(x)]$$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(\mu,\mathrm{I},S)}[\widetilde{c_V}] = \mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}\left[H_V(x)\underbrace{\frac{\mathbf{1}_S(x)}{\alpha}\frac{\mathcal{N}(\mu,\mathrm{I};x)}{\mathcal{N}_0(x)}}_{\psi(x)}\right]$$

We can learn a function of μ and S!

Approximating a weighted Characteristic function

- **Klivans, O'Donell, Servedio '08 with degree** $\kappa = O(\gamma^2/\varepsilon^2)$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}\left(\mathbf{1}_S(x) - q_\kappa(x)\right)^2 \leq \varepsilon$$

Approximating a weighted Characteristic function

- **Klivans, O'Donell, Servedio '08 with degree** $\kappa = O(\gamma^2/\varepsilon^2)$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}_0} (\mathbf{1}_S(x) - q_\kappa(x))^2 \leq \varepsilon$$

- **This work with degree** $\kappa = O(\gamma^2/\varepsilon^4)$

$$\mathop{\mathbb{E}}_{x \sim \mathcal{N}_0} (\psi(x) - p_\kappa(x))^2 \leq \varepsilon.$$

Approximating a weighted Characteristic function

- **Klivans, O'Donell, Servedio '08 with degree** $\kappa = O(\gamma^2/\varepsilon^2)$

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}(\mathbf{1}_S(x) - q_\kappa(x))^2 \le \varepsilon$$

- **This work with degree** $\kappa = O(\gamma^2/\varepsilon^4)$

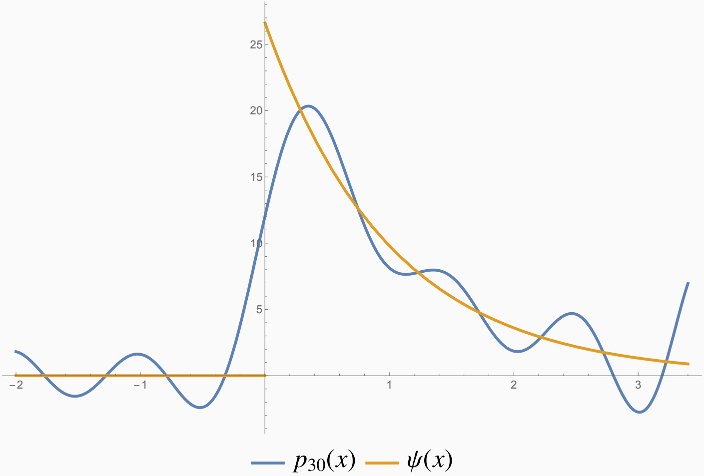
$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}_0}(\psi(x) - p_\kappa(x))^2 \le \varepsilon.$$

- **#samples** $= d^\kappa$

ψ and its approximation

The true Gaussian has mean -1, and the set is $x > 0$.

Out[▫]=



The Convex Objective

SGD objective

$$L(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_S^*} [???]$$

The Convex Objective

SGD objective

$$L(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_S^*}[h(u; x)\psi(x)]$$

- correction function $h(u; x)$ such that

The Convex Objective

SGD objective

$$L(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_S^*}[h(u;x)\psi(x)]$$

- correction function $h(u;x)$ such that
- $L(u)$ is **convex** and **the minimizer is μ!**

The Convex Objective

SGD objective

$$L(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_S^*} [h(u; x) p_\kappa(x)]$$

- correction function $h(u; x)$ such that
- $L(u)$ is still **convex** and if $\kappa = \gamma^2 / \varepsilon^8$ then **the minimizer is ε-close to μ!**

The Convex Objective

SGD objective

$$L(u) = \mathop{\mathbb{E}}_{x \sim \mathcal{N}_S^*}[h(u;x)p_\kappa(x)]$$

- correction function $h(u;x)$ such that
- $L(u)$ is still **convex** and if $\kappa = \gamma^2/\varepsilon^8$ then **the minimizer is ε-close to μ!**
- L is **strongly convex**.
- The variance of the update is bounded.

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

First efficient algorithm for truncated statistics with **unknown truncation sets**.

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

First efficient algorithm for truncated statistics with **unknown truncation sets**.

Open Problems

- Truncated statistics **beyond Gaussian**?

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

First efficient algorithm for truncated statistics with **unknown truncation sets**.

Open Problems

- Truncated statistics **beyond Gaussian**?
- **Improve the runtime** for specific classes.

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

First efficient algorithm for truncated statistics with **unknown truncation sets**.

Open Problems

- Truncated statistics **beyond Gaussian**?
- **Improve the runtime** for specific classes.
- Depend **polynomially** on the **accuracy** $1/\varepsilon$.

Recap and Open Problems

Our Results

Nearly tight sample complexity bounds with respect to VC-dimension and GSA.

First efficient algorithm for truncated statistics with **unknown truncation sets**.

Open Problems

- Truncated statistics **beyond Gaussian**?
- **Improve the runtime** for specific classes.
- Depend **polynomially** on the **accuracy** $1/\varepsilon$.

Thank You!