Iterative Methods for Efficient Statistics in High Dimensions

Manolis Zampetakis
MIT

Constantinos Daskalakis
MIT

Themis Gouleakis
USC

Vasilis Kontonis
UW Madison

Christos Tzamos
UW Madison

NTUA 2019

Classical Estimation Problem

Given n samples from a Gaussian distribution in d dimensions estimate:

- the **mean** and
- the **covariance matrix**.

Classical Estimation Problem

Given n samples from a Gaussian distribution in d dimensions estimate:

- the **mean** and
- the **covariance matrix**.

Answer:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i, \qquad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T$$

Classical Estimation Problem

Given n samples from a Gaussian distribution in d dimensions estimate:

- the **mean** and
- the **covariance matrix**.

Answer:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

How large n should be?

Classical Estimation Problem

Theorem ([Folklore]). Let $x_1, x_2, \ldots, x_n$ i.i.d. samples from $\mathcal{N}(\mu^*, \Sigma^*)$, then the estimates

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

satisfy with probability at least 99%:

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu^*, \Sigma^*)) \leq \varepsilon$$

for $n = \Theta(d^2/\varepsilon^2)$.

Classical Estimation Problem

Theorem ([Folklore]). Let $x_1, x_2, \ldots, x_n$ i.i.d. samples from $\mathcal{N}(\mu^*, \Sigma^*)$, then the estimates

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$$

satisfy with probability at least 99%:

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu^*, \Sigma^*)) \leq \varepsilon$$

for $n = \Theta(d^2/\varepsilon^2)$.

Time and Sample Efficient in High Dimensions.

Contents

Contents

- ❑ Estimation Obstacles
  - ○ Truncated Samples
  - ○ Inhomogeneous Population

Optimization Problems

- ❖ Global Convergence of Iterative Methods

Contents

Contents

- ❑ Estimation Obstacles
  - ○ Truncated Samples
  - ○ Inhomogeneous Population
- ❖ Global Convergence of Iterative Methods

Optimization Problems

Convex Landscape

Non-Convex Landscape

Iterative Methods

Contents

- ❑ Estimation Obstacles
  - ○ **Truncated Samples**
  - ○ Inhomogeneous Population
- ❖ Global Convergence of Iterative Methods

Optimization Problems

Convex Landscape

Non-Convex Landscape

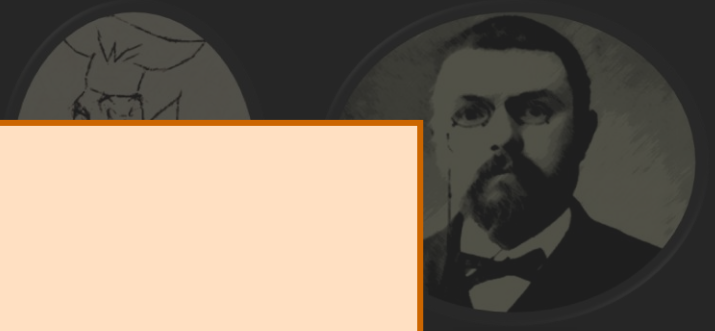
Iterative Methods

Poincare and his friend the Baker

Anecdotal Story [Gil Kalai's Blog Post]

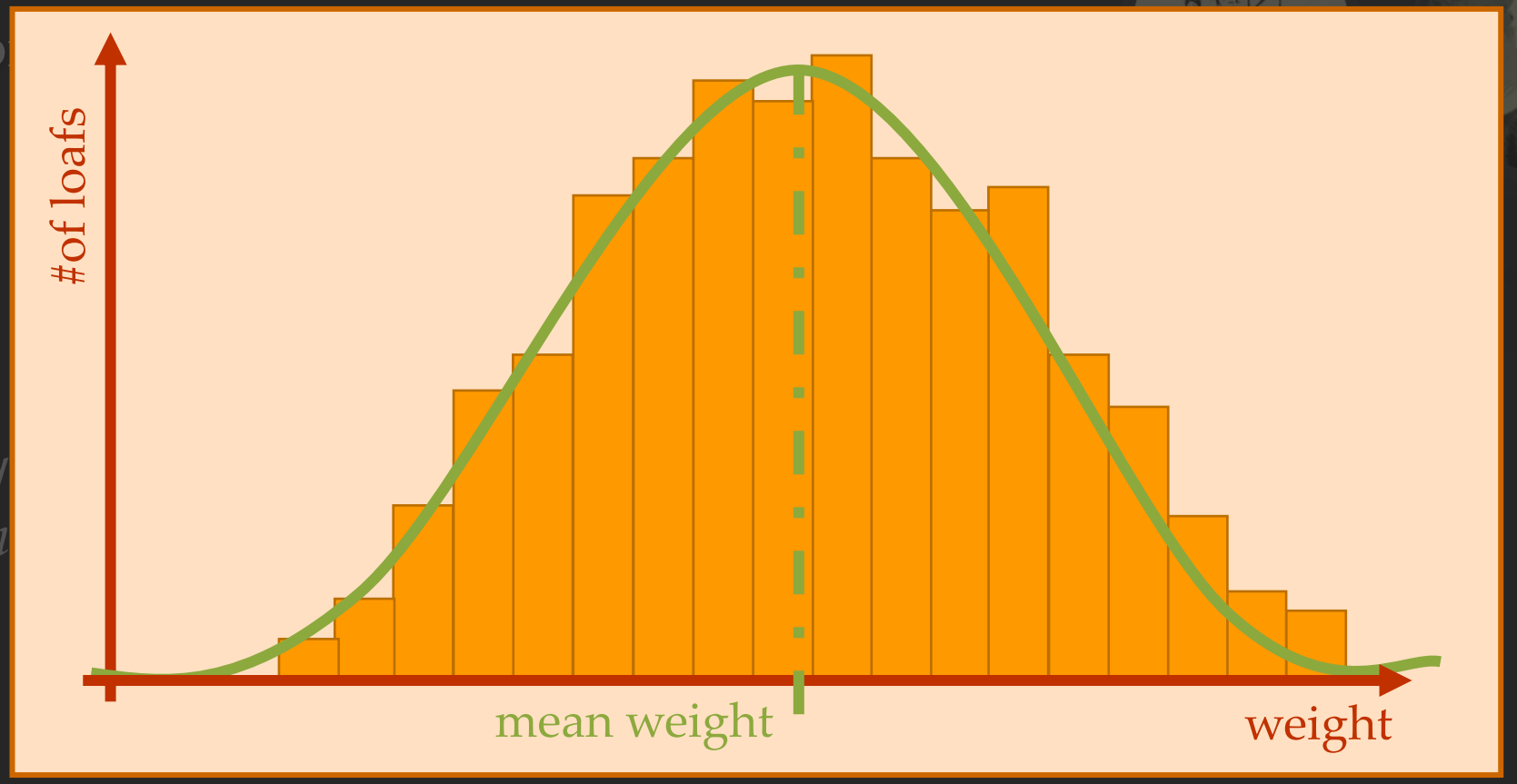
"My friend the baker," said Poincaré, *"I weighed every loaf of bread that I bought from you in the last year and the distribution is Gaussian with mean 950 grams. How can you claim that your average loaf is 1 kilogram?"*

Poincare and his friend the Baker



Anecdotal Sto

"My friend the
I bought from y
mean 950 gram



Poincare and his friend the Baker

Anecdotal Story (continued) [Gil Kalai's Blog Post]

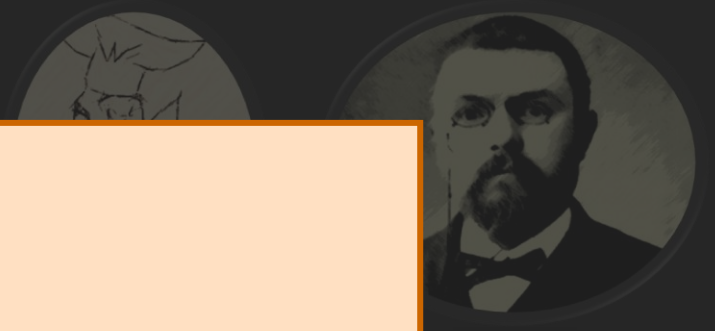
A year later the two pals meet again.

"How are you doing dear Henri" asked the baker *"are my bread loaves heavy enough for you?"*

"Yes, for me they are," answered Poincaré *"but when I weighed all the loaves last year I discovered that your mean value is still 950 grams."*

"How is this possible?" asked the baker.

Poincare and his friend the Baker



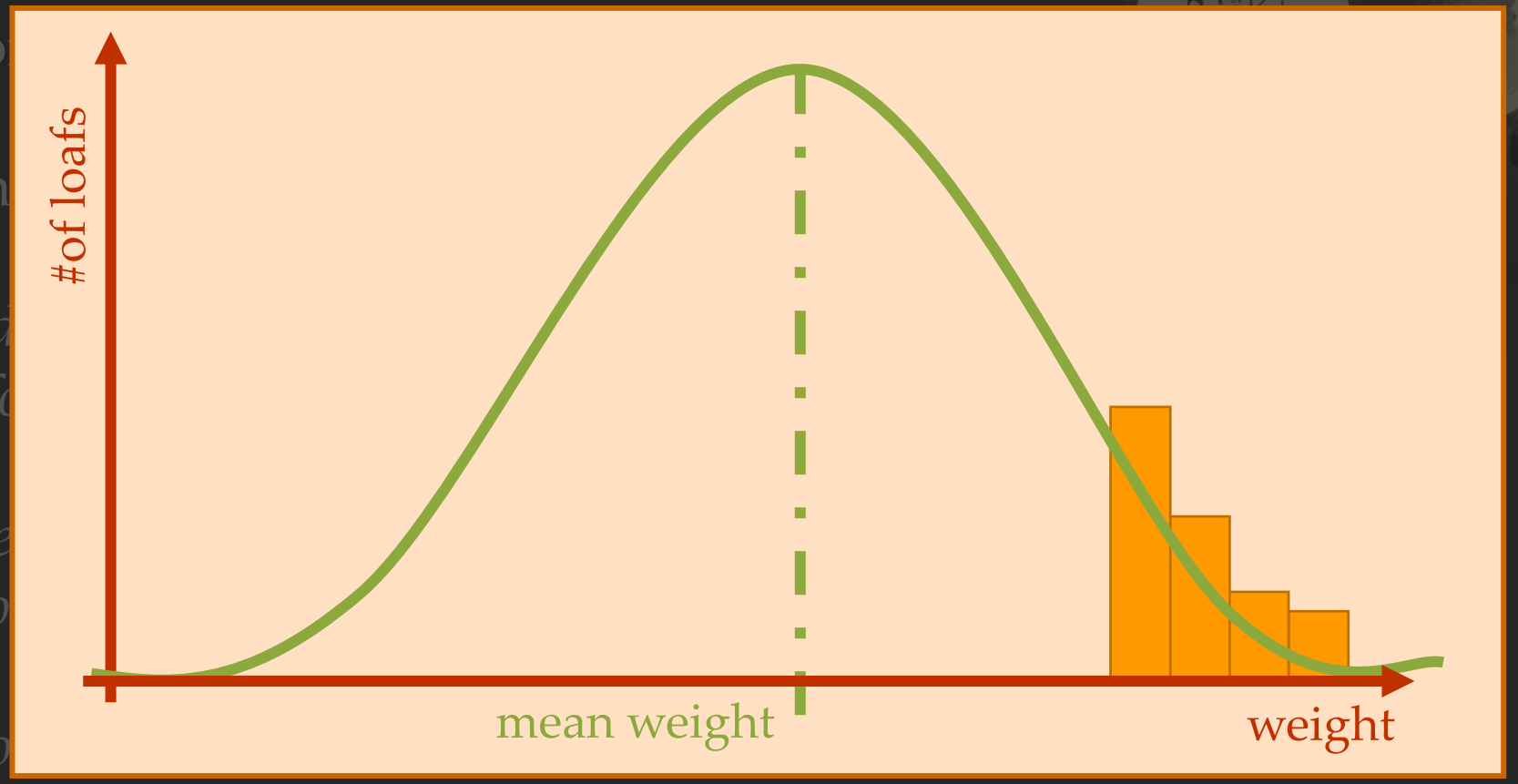
Anecdotal Sto

A year later th

"How are you d
heavy enough fo

"Yes, for me the
last year I disco

"How is this po



Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

Data Set:

(wage rate, income, education level – I.Q.)

Study: intelligence vs income for low wage rate workers

Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

Data Set:

(wage rate, income, education level – I.Q.)

Study: intelligence vs income for low wage rate workers

***Truncation!*:** samples collected only if income was less than 1.5 times the poverty level.

Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

Data Set:

(wage rate, income, education level – I.Q.)

Findings (e.g. [Hause 1971]*)*:
Intelligence and education level have mostly **no effect** on the income and sometimes even **negative effect**! (?)

Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

Data Set:

(wage rate, income, education level – I.Q.)

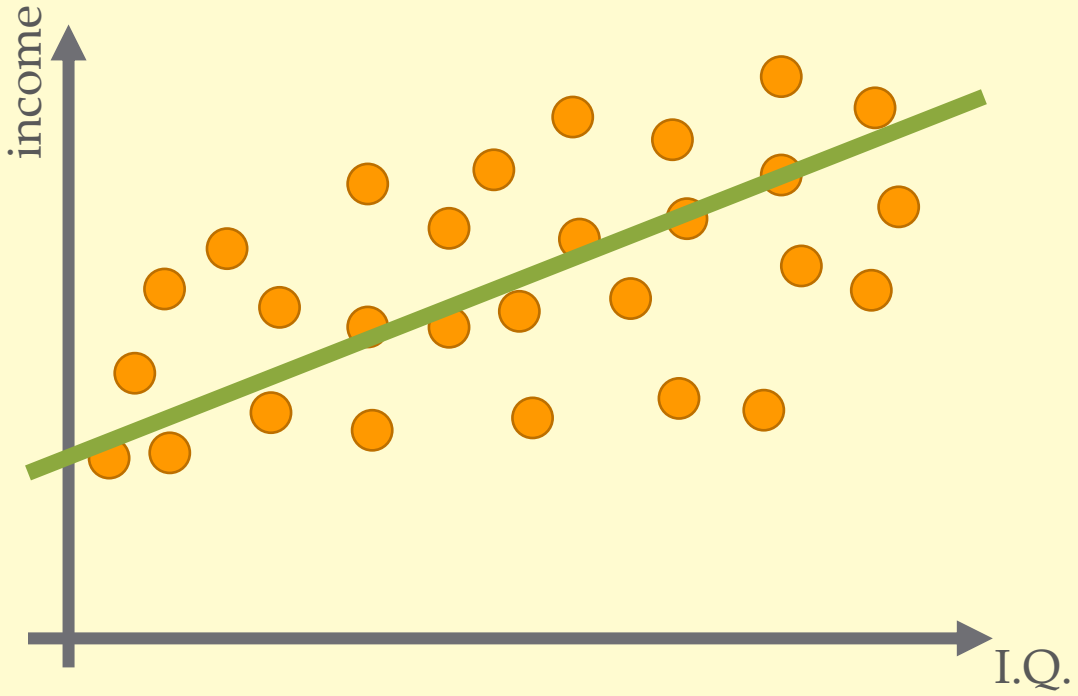
Findings (e.g. [Hause 1971]*)*:
Intelligence and education level have mostly **no effect** on the income and sometimes even **negative effect**! (?)

Explanation: Because of the **truncation bias**. Workers of the same rate excluded if they work more hours and have higher income.

Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

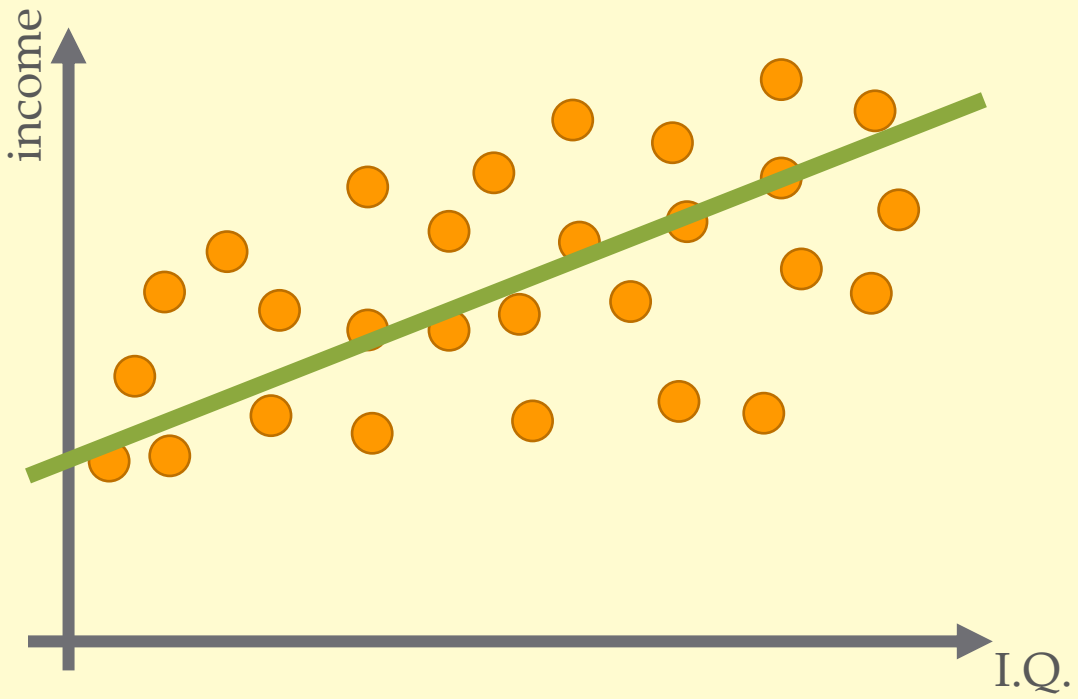
Explanation: Because of the **truncation bias**. Workers of the same rate excluded if they work more hours and have higher income.

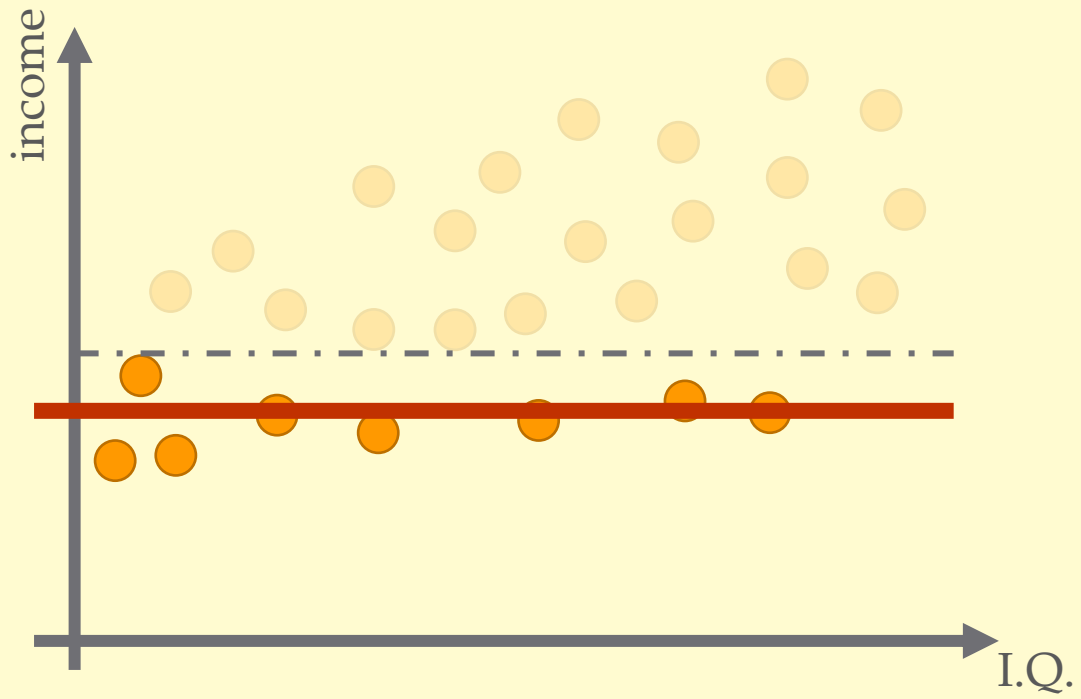


Truncated Regression

Example [Hausman, Wise (Econometrica) 1976]

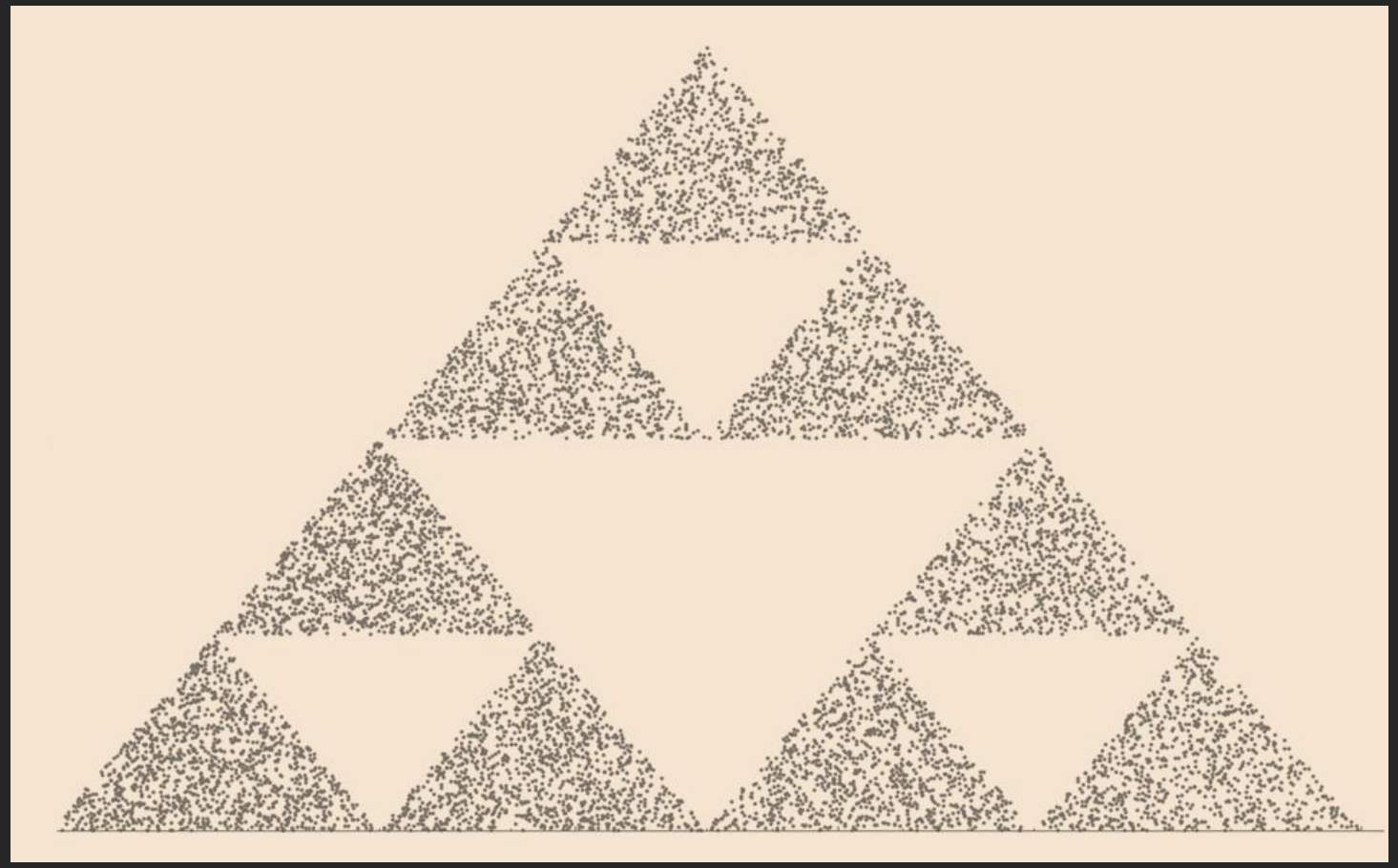
Explanation: Because of the **truncation bias**. Workers of the same rate excluded if they work more hours and have higher income.



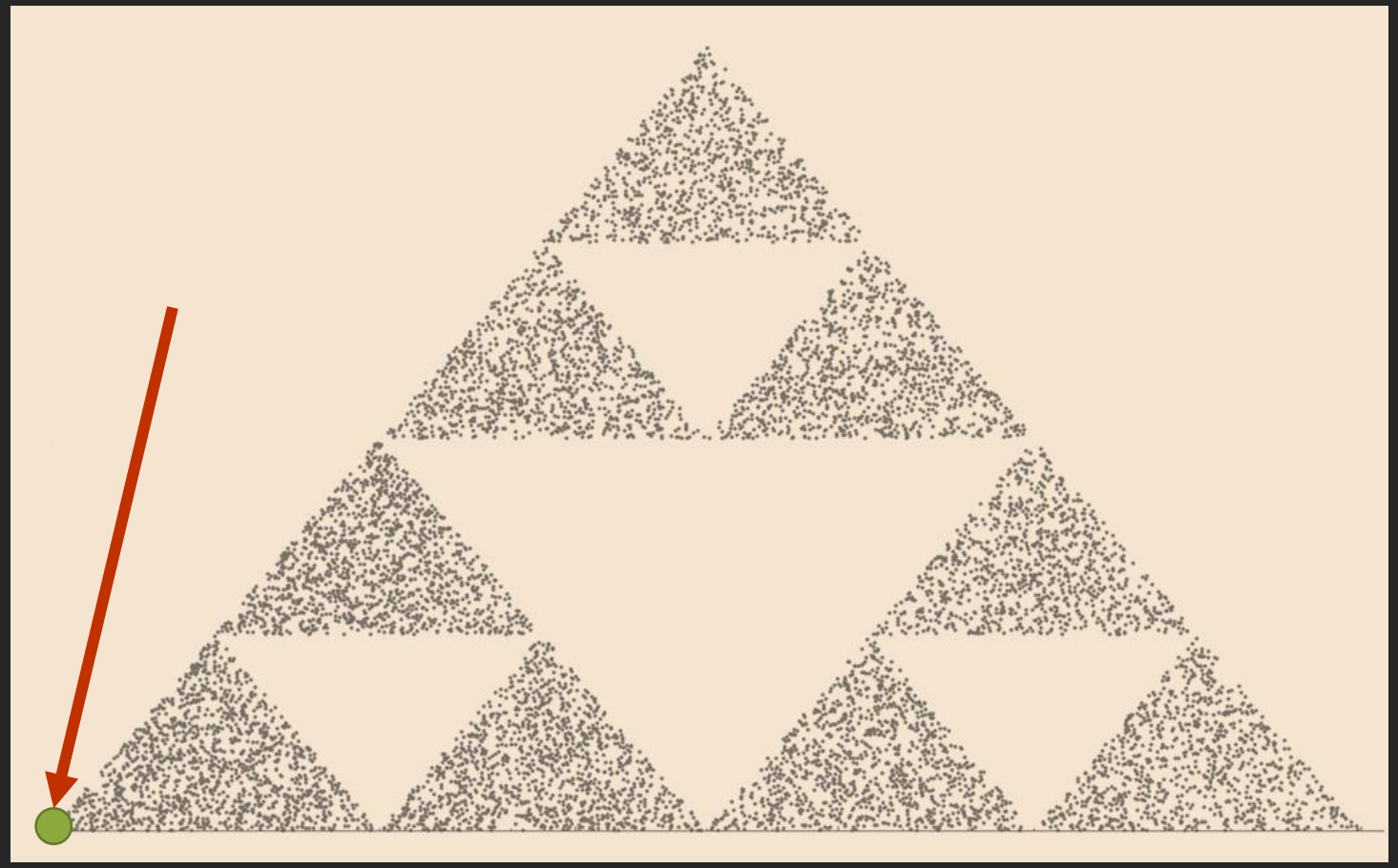


Multidimensional Example

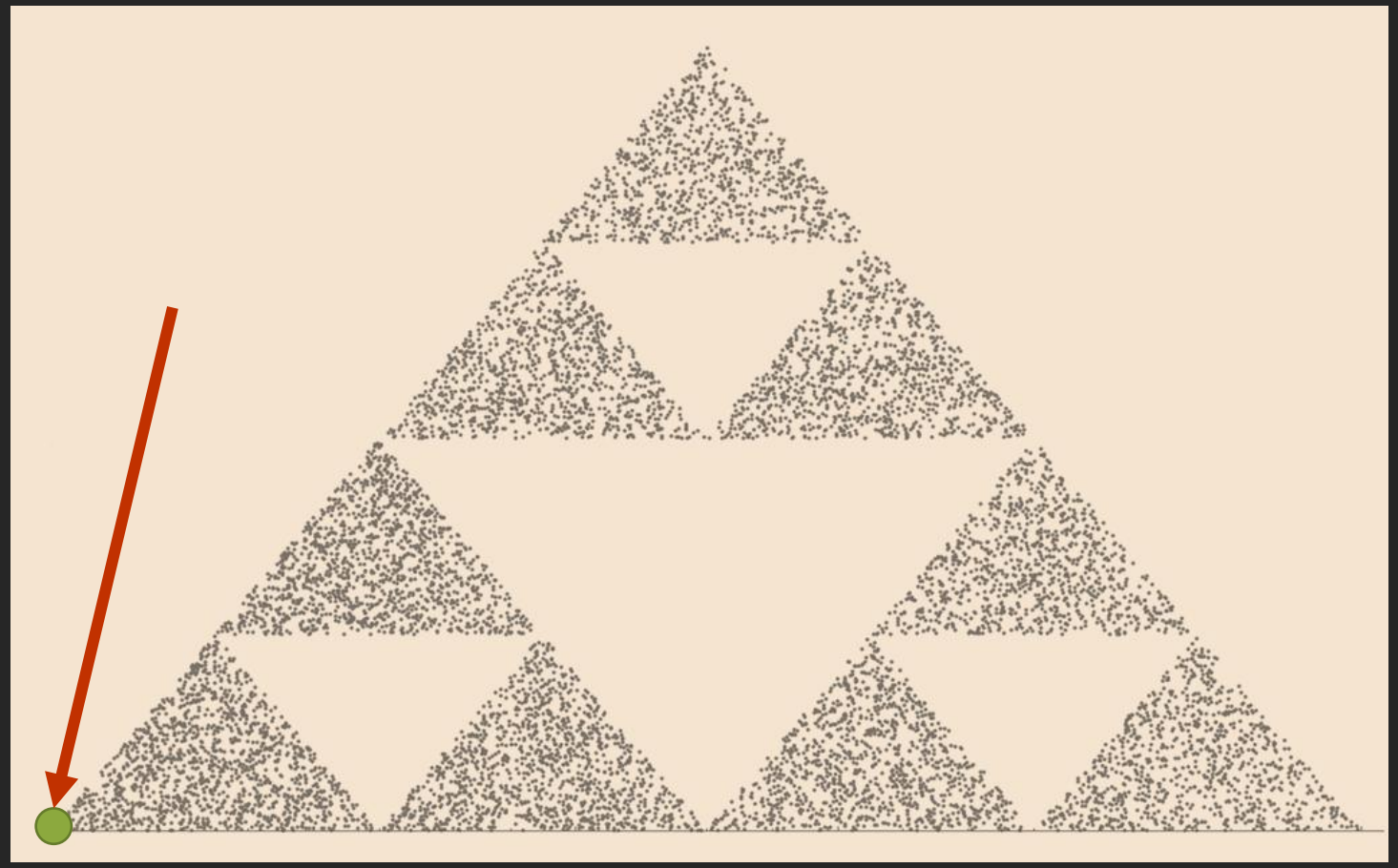
Truncated Samples - Multivariate Normal



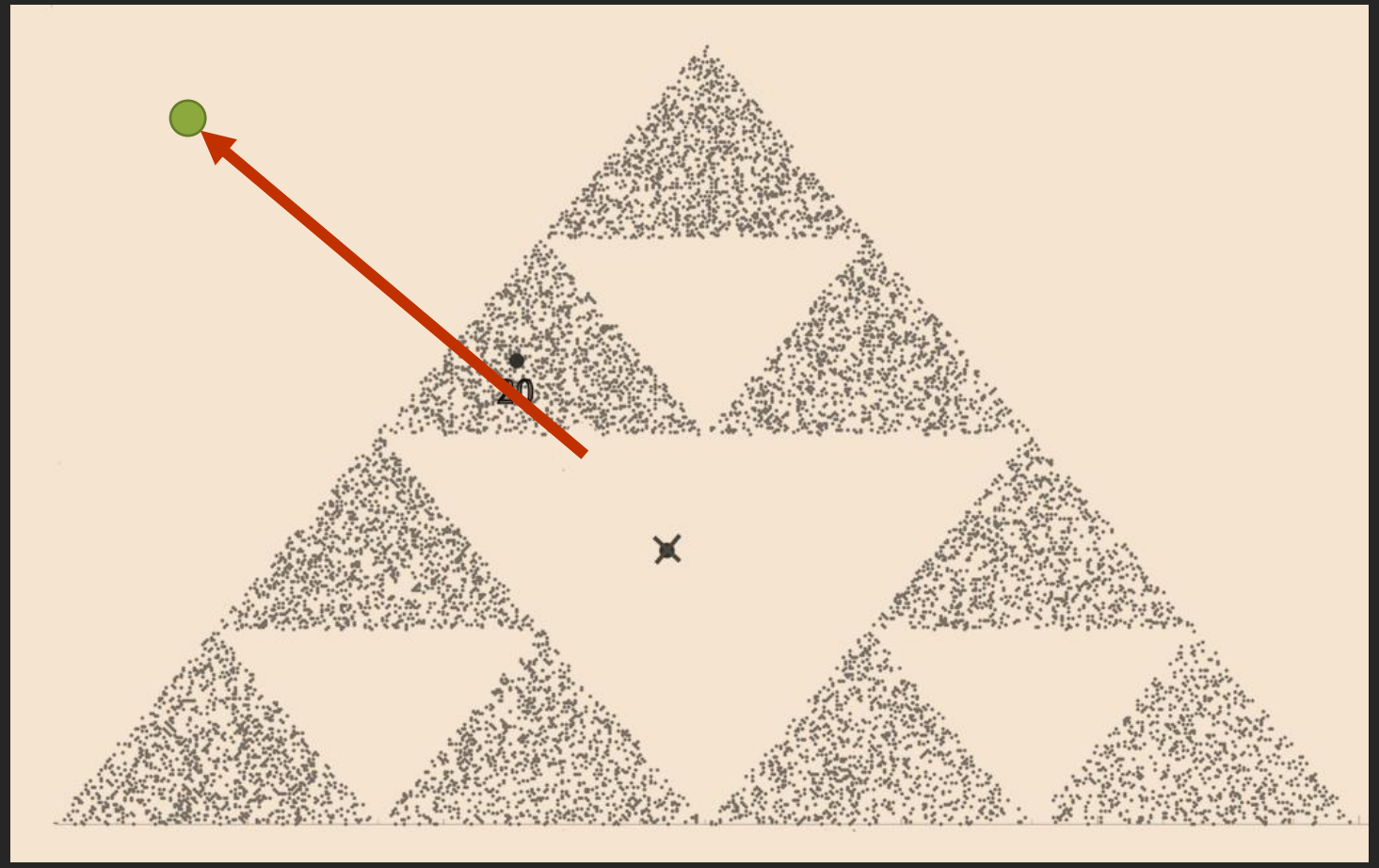
Truncated Samples - Multivariate Normal



Truncated Samples - Multivariate Normal



Truncated Samples - Multivariate Normal



Truncated Samples

Since Galton there has been an extensive literature on estimation from truncated samples, motivated by

- measurement limitations,

Truncated Samples

Since Galton there has been an extensive literature on estimation from truncated samples, motivated by

- measurement limitations,
- ethical considerations,

Truncated Samples

Since Galton there has been an extensive literature on estimation from truncated samples, motivated by

- measurement limitations,
- ethical considerations,
- privacy considerations,

Truncated Samples

Since Galton there has been an extensive literature on estimation from truncated samples, motivated by

- measurement limitations,
- ethical considerations,
- privacy considerations,
- …

For this reason truncated samples appear in:

- ❑ physics,
- ❑ economics,
- ❑ social sciences,
- ❑ psychological studies.

Contents

- Known Survival Set
  - Gaussian Estimation **[Daskalakis, Gouleakis, Tzamos, Z. ‘18]**
  - Linear Regression **[Daskalakis, Gouleakis, Tzamos, Z. ‘19]**
- Unknown Survival Set **[Kontonis, Tzamos, Z. ‘19]**

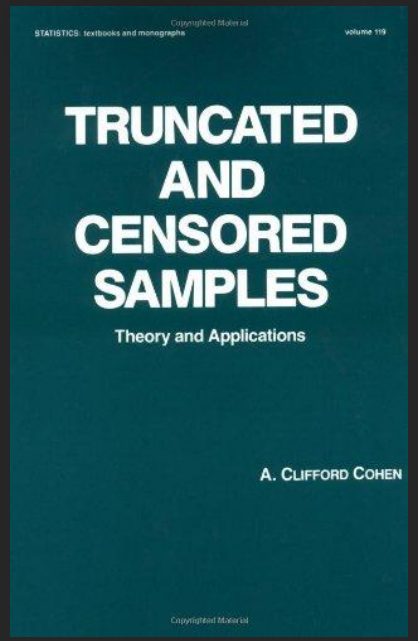
Contents

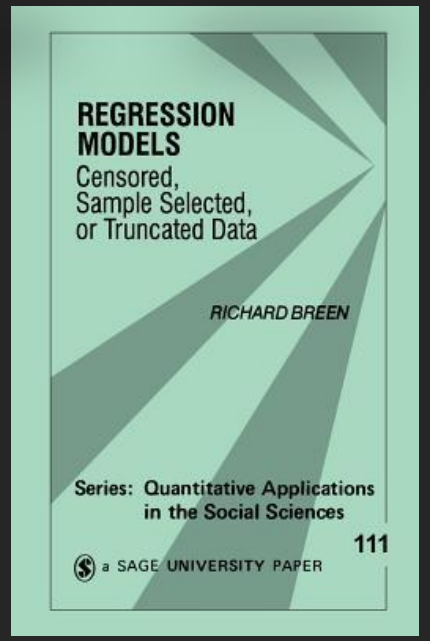
- **Known Survival Set**
  - Gaussian Estimation **[Daskalakis, Gouleakis, Tzamos, Z. ‘18]**
  - Linear Regression **[Daskalakis, Gouleakis, Tzamos, Z. ‘19]**
- Unknown Survival Set **[Kontonis, Tzamos, Z. ‘19]**

Contents

- **Known Survival Set**
  - Gaussian Estimation **[Daskalakis, Gouleakis, Tzamos, Z. '18]**
  - Linear Regression **[Daskalakis, Gouleakis, Tzamos, Z. '19]**
- Unknown Survival Set **[Kontonis, Tzamos, Z. '19]**

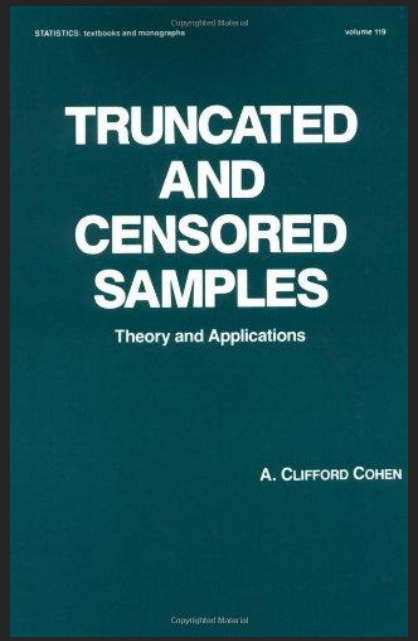
Truncated Statistics

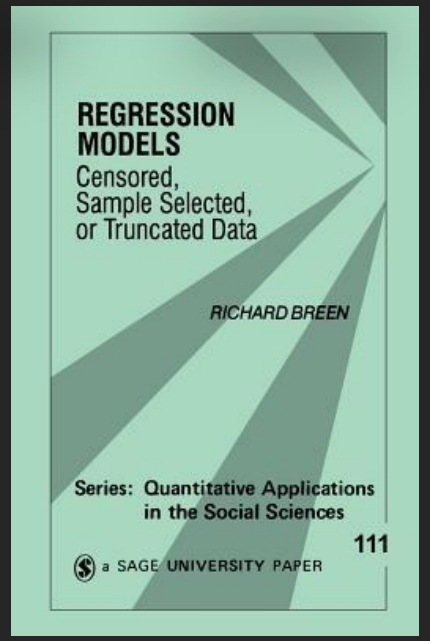




Truncated Statistics



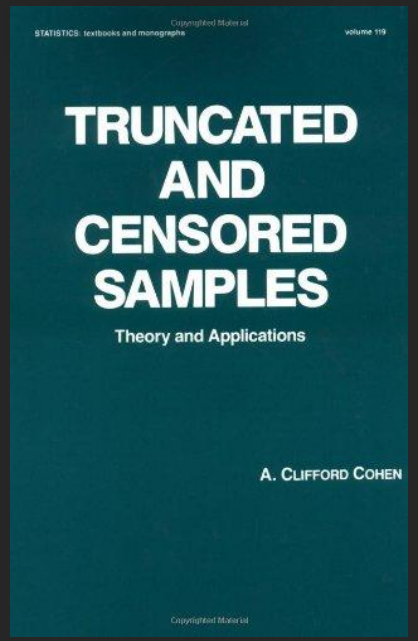


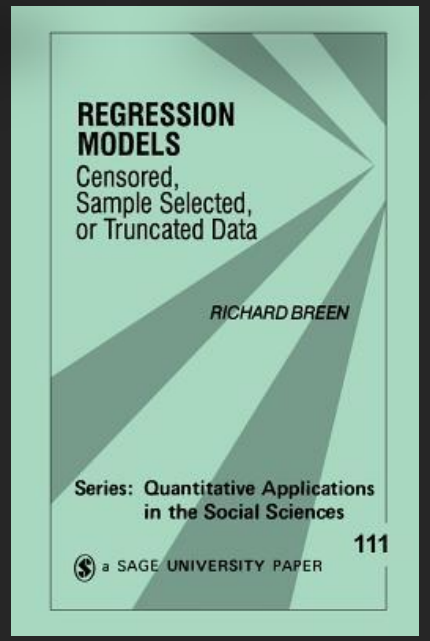


1. Only for known axes aligned box!

Truncated Statistics



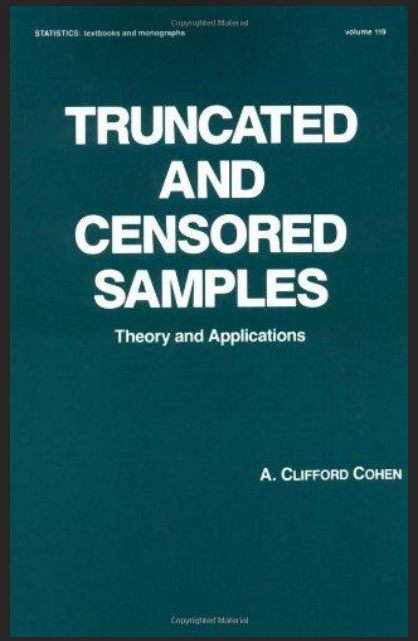


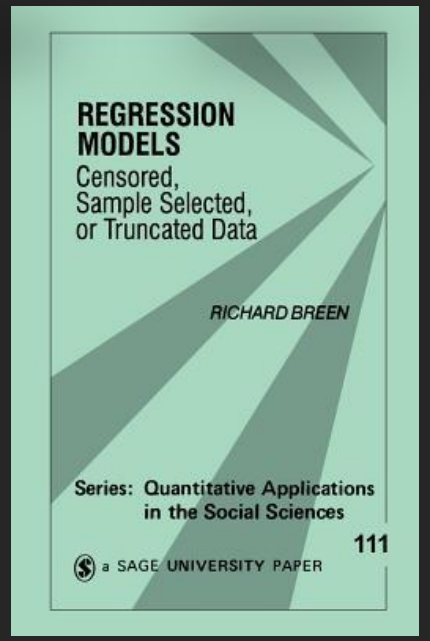


1. Only for known axes aligned box!
2. No sample complexity analysis.

Truncated Statistics



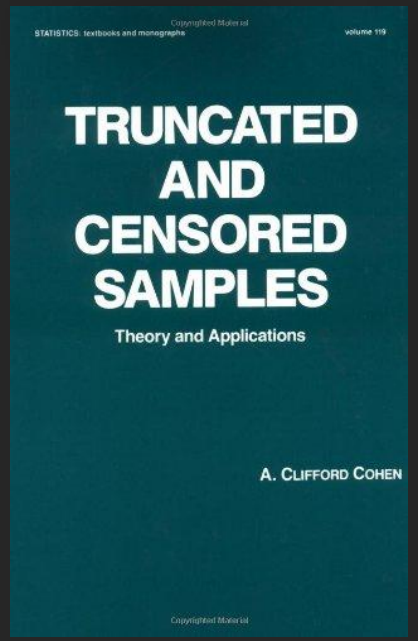


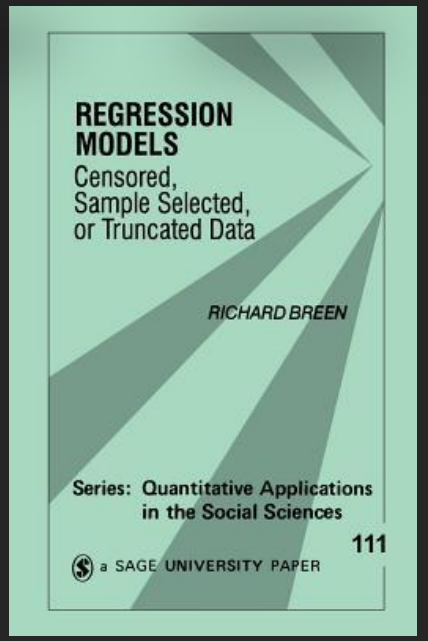


1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

Truncated Statistics







1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

Even for the case of estimating a multi-normal distribution.

Our Results – Overview

We provide efficient estimation for **truncated multi-normal distribution**.

1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

Our Results – Overview

We provide efficient estimation for **truncated multi-normal distribution**.

1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

The survival set should only have non-trivial mass!

Our Results – Overview

We provide efficient estimation for **truncated multi-normal distribution**.

1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

The survival set should only have non-trivial mass!

Can even be **unknown** under stronger assumptions.

Our Results – Overview

We provide efficient estimation for **truncated multi-normal distribution**.

1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

We get almost optimal # of samples.

Our Results – Overview

We provide efficient estimation for **truncated multi-normal distribution**.

1. Only for known axes aligned box!
2. No sample complexity analysis.
3. No efficient algorithm.

We provide a **convex** programming formulation which can be solved efficiently.

Estimation from Truncated Samples

Multivariate Gaussian Distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Estimation from Truncated Samples

Multivariate Gaussian Distribution $\mathcal{N}(\mu, \Sigma)$

$$\mathcal{N}(\mu, \Sigma; x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right).$$

Let $S \subseteq \mathbb{R}^d$, we define the mass of S with respect to measure $\mathcal{N}(\mu, \Sigma)$

$$\mathcal{N}(\mu, \Sigma; S) = \int_S \mathcal{N}(\mu, \Sigma; y) dy.$$

Estimation from Truncated Samples

Let $S \subseteq \mathbb{R}^d$, we define the *truncated normal distribution* $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S)$ as

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, S; \boldsymbol{x}) = \begin{cases} \dfrac{1}{\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; S)} \cdot \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \boldsymbol{x}) & \boldsymbol{x} \in S \\ 0 & \boldsymbol{x} \notin S \end{cases}.$$

Estimation from Truncated Samples

Let $S \subseteq \mathbb{R}^d$, we define the *truncated normal distribution* $\mathcal{N}(\mu, \Sigma, S)$ as

$$\mathcal{N}(\mu, \Sigma, S; x) = \begin{cases} \frac{1}{\mathcal{N}(\mu, \Sigma; S)} \cdot \mathcal{N}(\mu, \Sigma; x) & x \in S \\ 0 & x \notin S \end{cases}.$$

Given samples from $\mathcal{N}(\mu^*, \Sigma^*, S)$ compute estimates $\hat{\mu}$, $\hat{\Sigma}$ such that

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu^*, \Sigma^*)) \leq \varepsilon.$$

Estimation from Truncated Samples

Theorem ([Daskalakis Gouleakis Tzamos Z. '18]). Let $S \subseteq \mathbb{R}^d$ that satisfies Assumptions 1 and 2 and $x_1, x_2, \ldots, x_n$ i.i.d. samples from $\mathcal{N}(\mu^*, \Sigma^*, S)$, then we can efficiently compute estimates $\hat{\mu}$ and $\hat{\Sigma}$ that satisfy with probability at least 99%:

$$d_{TV}(\mathcal{N}(\hat{\mu}, \hat{\Sigma}), \mathcal{N}(\mu^*, \Sigma^*)) \leq \varepsilon$$

for $n = \tilde{\Theta}(d^2/\varepsilon^2)$.

Truncation Set

Assumption 1 (Constant Mass)

$$\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \alpha$$

Truncation Set

Assumption 1 (Constant Mass)

$$\mathcal{N}(\mu^*, \Sigma^*; S) \geq \alpha$$

If $\alpha \to 0$ then # of sample $\to \infty$.

Truncation Set

Assumption 1 (Constant Mass)

$$\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*; S) \geq \alpha$$

Assumption 2 (Oracle Access)

Given $\boldsymbol{x} \in \mathbb{R}^d$ we can answer if $\boldsymbol{x} \in S$ or not.

Truncation Set

Assumption 1 (Constant Mass)

$$\mathcal{N}(\mu^*, \Sigma^*; S) \geq \alpha$$

Assumption 2 (Oracle Access)

Given $x \in \mathbb{R}^d$ we can answer if $x \in S$ or not.

Impossible for unknown and **arbitrary** S!

In this talk

Theorem ([Daskalakis Gouleakis Tzamos Z. '18]). Let $S \subseteq \mathbb{R}^d$ that satisfies Assumptions 1 and 2 and $x_1, x_2, \ldots, x_n$ i.i.d. samples from $\mathcal{N}(\mu^*, I, S)$, then we can efficiently compute an estimate $\hat{\mu}$ that satisfies with probability at least 99%:

$$d_{TV}(\mathcal{N}(\hat{\mu}, I), \mathcal{N}(\mu^*, I)) \leq \varepsilon$$

for $n = \tilde{\Theta}(d/\varepsilon^2)$.

In this talk

Theorem ([Daskalakis Gouleakis Tzamos Z. '18]). Let $S \subseteq \mathbb{R}^d$ that satisfies Assumptions 1 and 2 and $x_1, x_2, \ldots, x_n$ i.i.d. samples from $\mathcal{N}(\mu^*, I, S)$, then we can efficiently compute an estimate $\hat{\mu}$ that satisfies with probability at least 99%:

$$\|\hat{\mu} - \mu^*\|_2 \leq \varepsilon$$

for $n = \tilde{\Theta}(d/\varepsilon^2)$.

The Estimation Algorithm

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\boldsymbol{\mu}^{(T)}$

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\boldsymbol{\mu}^{(T)}$

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\boldsymbol{\mu}^{(T)}$

The Estimation Algorithm

1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, S)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, S)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\mu^{(T)}$

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\boldsymbol{\mu}^{(T)}$

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\boldsymbol{\mu}^{(T)}$

The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, S)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{\mu}^{(t)}$

Convergence Analysis

Algorithm: Stochastic Gradient Descent in the population negative log-likelihood function.

Convergence Analysis

Algorithm: Stochastic Gradient Descent in the population negative log-likelihood function.

Conditions for Fast Convergence of SGD

1. $\mathbb{E}\left[\boldsymbol{r} - \hat{\boldsymbol{r}}\right] = \nabla \bar{\ell}(\boldsymbol{\mu})$,
2. $\bar{\ell}(\boldsymbol{\mu})$ is *strongly* convex,
3. $\mathbb{E}\left[\|\boldsymbol{r} - \hat{\boldsymbol{r}}\|_2^2\right]$ is bounded.

Convergence Analysis

Algorithm: Stochastic Gradient Descent in the population negative log-likelihood function.

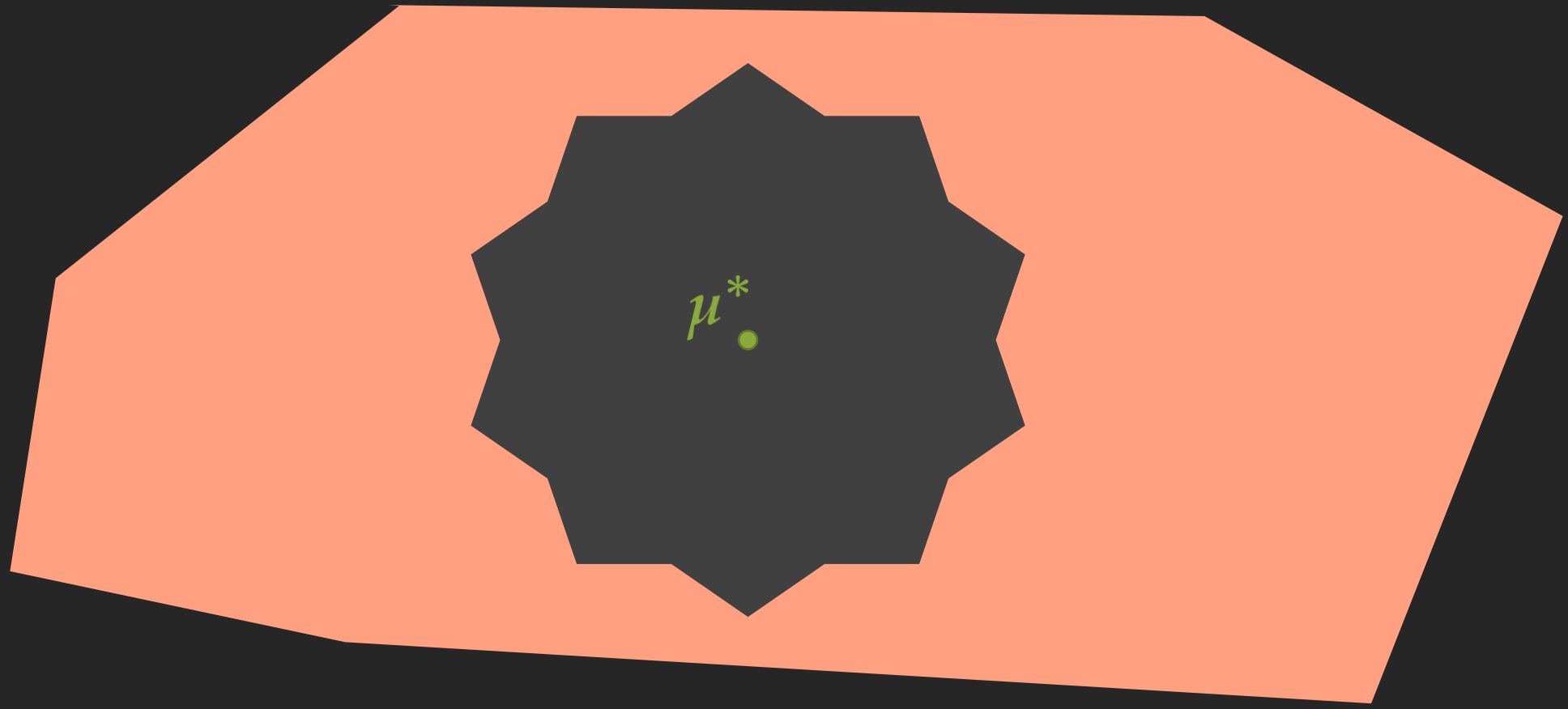
Conditions for Fast Convergence of SGD

1. $\mathbb{E}\left[\boldsymbol{r} - \hat{\boldsymbol{r}}\right] = \nabla \bar{\ell}(\boldsymbol{\mu})$,
2. $\bar{\ell}(\boldsymbol{\mu})$ is *strongly* convex,
3. $\mathbb{E}\left[\|\boldsymbol{r} - \hat{\boldsymbol{r}}\|_2^2\right]$ is bounded.

Convergence Analysis

$\bar{\ell}(\boldsymbol{\mu})$ is strongly convex.

Does not hold for any $\boldsymbol{\mu} \in \mathbb{R}^d$.



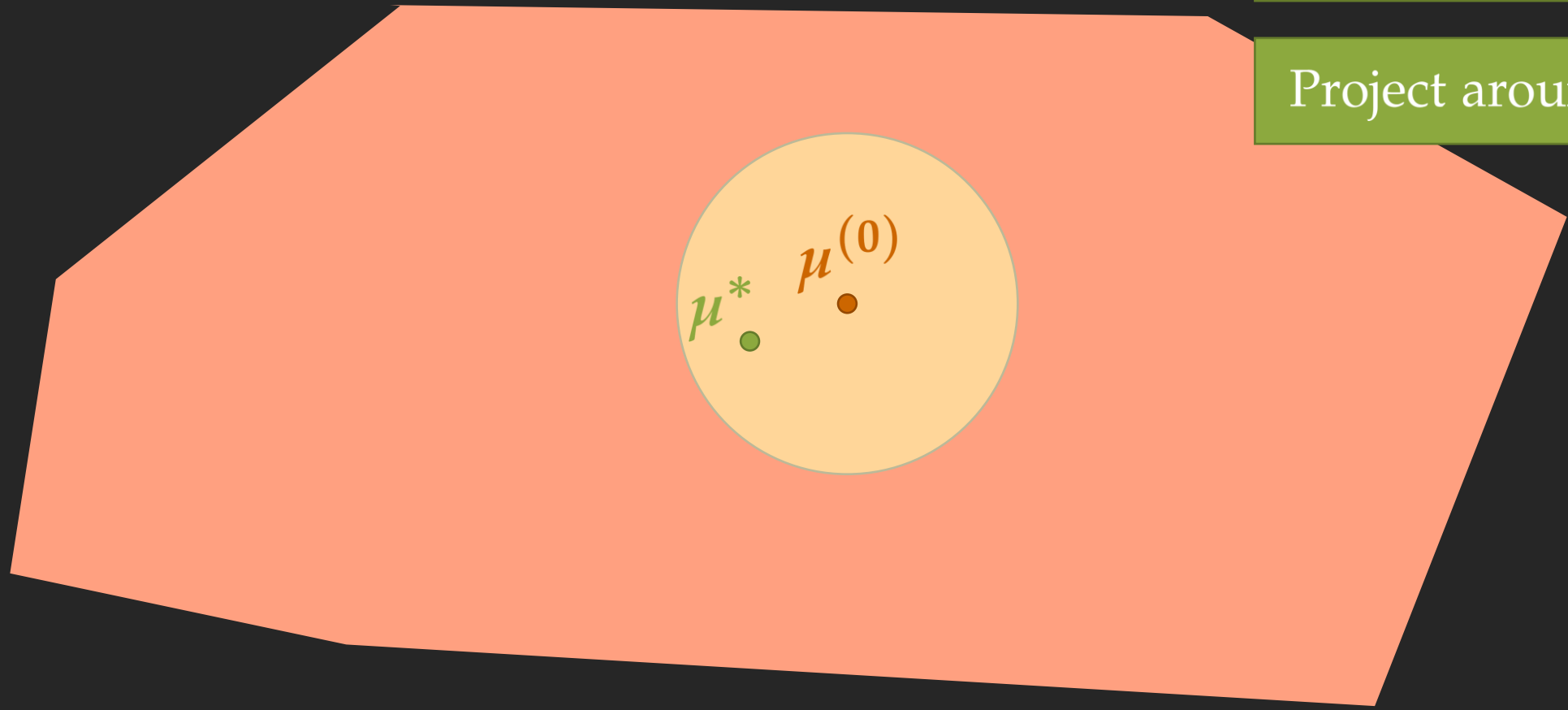
Convergence Analysis

$\bar{\ell}(\boldsymbol{\mu})$ is strongly convex.

Does not hold for any $\boldsymbol{\mu} \in \mathbb{R}^d$.

Project to a convex region that holds!

Project around $\boldsymbol{\mu}^{(0)}$!



The Estimation Algorithm

1. estimate $\boldsymbol{\mu}^{(0)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad \boldsymbol{r} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, C)$,
4. $\quad \hat{\boldsymbol{r}} \leftarrow$ Sample $\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}, C)$
5. $\quad \boldsymbol{v}^{(t)} \leftarrow \boldsymbol{\mu}^{(t-1)} - \eta_t(\boldsymbol{r} - \hat{\boldsymbol{r}})$
6. $\quad \boldsymbol{\mu}^{(t)} \leftarrow$ project $\boldsymbol{v}^{(t)}$ to the ball $B = \left\{\boldsymbol{x} \mid \left\|\boldsymbol{x} - \boldsymbol{\mu}^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{\mu}^{(t)}$

Convergence Analysis

Algorithm: Stochastic Gradient Descent in the population negative log-likelihood function.

Conditions for Fast Convergence of SGD

1. $\mathbb{E}\left[\boldsymbol{r} - \hat{\boldsymbol{r}}\right] = \nabla \bar{\ell}(\boldsymbol{\mu})$,
2. $\bar{\ell}(\boldsymbol{\mu})$ is *strongly* convex,
3. $\mathbb{E}\left[\|\boldsymbol{r} - \hat{\boldsymbol{r}}\|_2^2\right]$ is bounded.

Contents

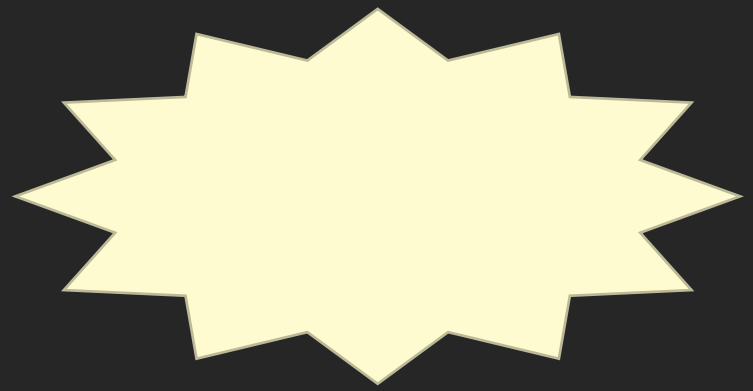
- Known Survival Set
  - Gaussian Estimation **[Daskalakis, Gouleakis, Tzamos, Z. '18]**
  - Linear Regression **[Daskalakis, Gouleakis, Tzamos, Z. '19]**
- **Unknown Survival Set** **[Kontonis, Tzamos, Z. '19]**

Contents

- Known Survival Set
  - Gaussian Estimation [Daskalakis, Gouleakis, Tzamos, Z. '18]
  - Linear Regression [Daskalakis, Gouleakis, Tzamos, Z. '19]
- **Unknown Survival Set** [Kontonis, Tzamos, Z. '19]

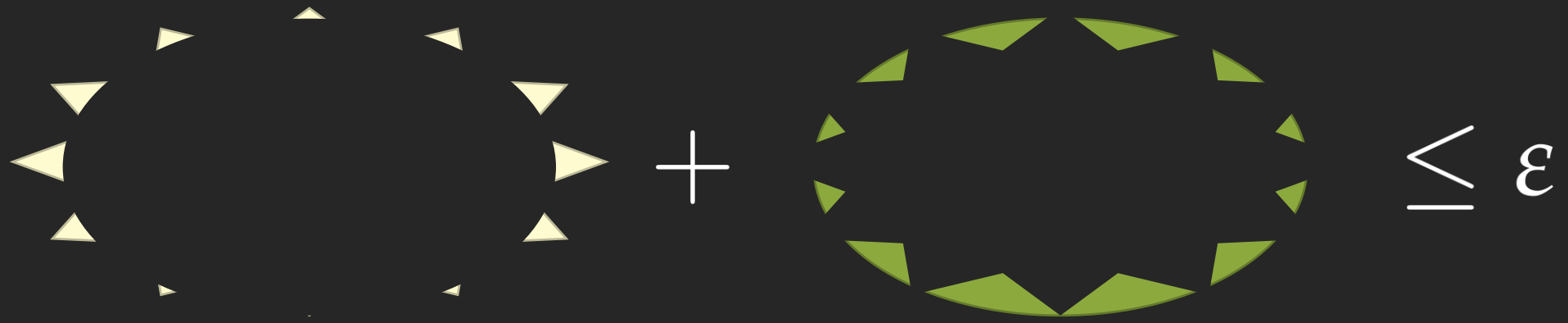
Goal: Learn the parameters of the Gaussian distribution and the set!

What learning the set means?



What learning the set means?

What learning the set means?



Gaussian Surface Area

Gaussian Surface Area

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

"Surface of S with respect to the Gaussian measure."

$$\Gamma(S) = \lim_{\delta \to 0} \frac{\mathcal{N}_0(S + B(\mathbf{0}, \delta)) - \mathcal{N}_0(S)}{\delta}$$

Gaussian Surface Area

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

"Surface of S with respect to the Gaussian measure."

$$\Gamma(S) = \lim_{\delta \to 0} \frac{\mathcal{N}_0(S + B(\mathbf{0}, \delta)) - \mathcal{N}_0(S)}{\delta} \qquad \Gamma(\mathcal{S}) = \max_{S \in \mathcal{S}} \Gamma(S)$$

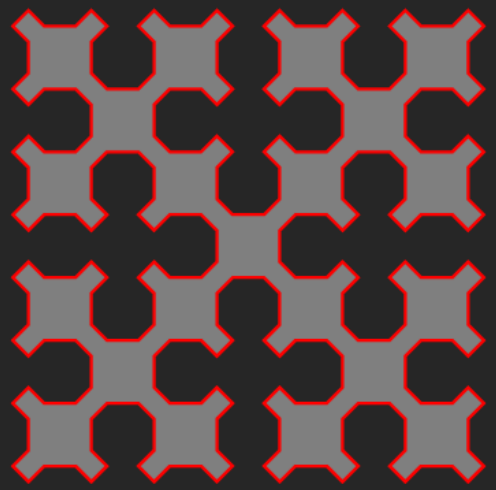
Gaussian Surface Area

Gaussian Surface Area: $\Gamma(S)$

"Surface of S with respect to the Gaussian measure."

$$\Gamma(S) = \lim_{\delta \to 0} \frac{\mathcal{N}_0(S + B(\mathbf{0}, \delta)) - \mathcal{N}_0(S)}{\delta}$$

$$\Gamma(\mathcal{S}) = \max_{S \in \mathcal{S}} \Gamma(S)$$



In this talk

Theorem ([Kontonis, Tzamos, Z. '19]). Let $S \subseteq \mathbb{R}^d$ unknown and $S \in \mathcal{S}$ with Gaussian Surface Area $\Gamma(\mathcal{S})$ that satisfies Assumptions 1 and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ i.i.d. samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$, then we can efficiently compute an estimate $\hat{\boldsymbol{\mu}}$ that satisfies with probability at least 99%:

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2 \leq \varepsilon$$

for $n = d^{\tilde{\Theta}(\Gamma^2(S)/\varepsilon^8)}$.

In this talk

Theorem ([Kontonis, Tzamos, Z. '19]). Let $S \subseteq \mathbb{R}^d$ unknown and $S \in \mathcal{S}$ with Gaussian Surface Area $\Gamma(\mathcal{S})$ that satisfies Assumptions 1 and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ i.i.d. samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$, then we can efficiently compute an estimate $\hat{\boldsymbol{\mu}}$ that satisfies with probability at least 99%:

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2 \leq \varepsilon$$

for $n = d^{\tilde{\Theta}(\Gamma^2(S)/\varepsilon^8)}$.

Concept Class	Surface Area	Sample Complexity
PTFs of degree k	$O(k)$ [Kane '11]	$d^{O(k^2)}$
Intersections of k halfspaces	$O(\sqrt{\log k})$ [KOS '08]	$d^{O(\log k)}$
General convex sets	$O(d^{1/4})$ [Ball 1993]	$d^{O(\sqrt{d})}$

Hermite Polynomials

- **Univariate** Hermite Polynomials $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots,$
- **Multivariate**: For $V = (v_1, \ldots, v_d)$ $H_V(x) = \prod_{i=1}^{d} h_{v_i}(x_i)$

Hermite Polynomials

- **Univariate** Hermite Polynomials $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots,$
- **Multivariate**: For $V = (v_1, \ldots, v_d)$ $H_V(x) = \prod_{i=1}^{d} h_{v_i}(x_i)$

Orthonormal basis under the Gaussian measure!

Hermite Polynomials

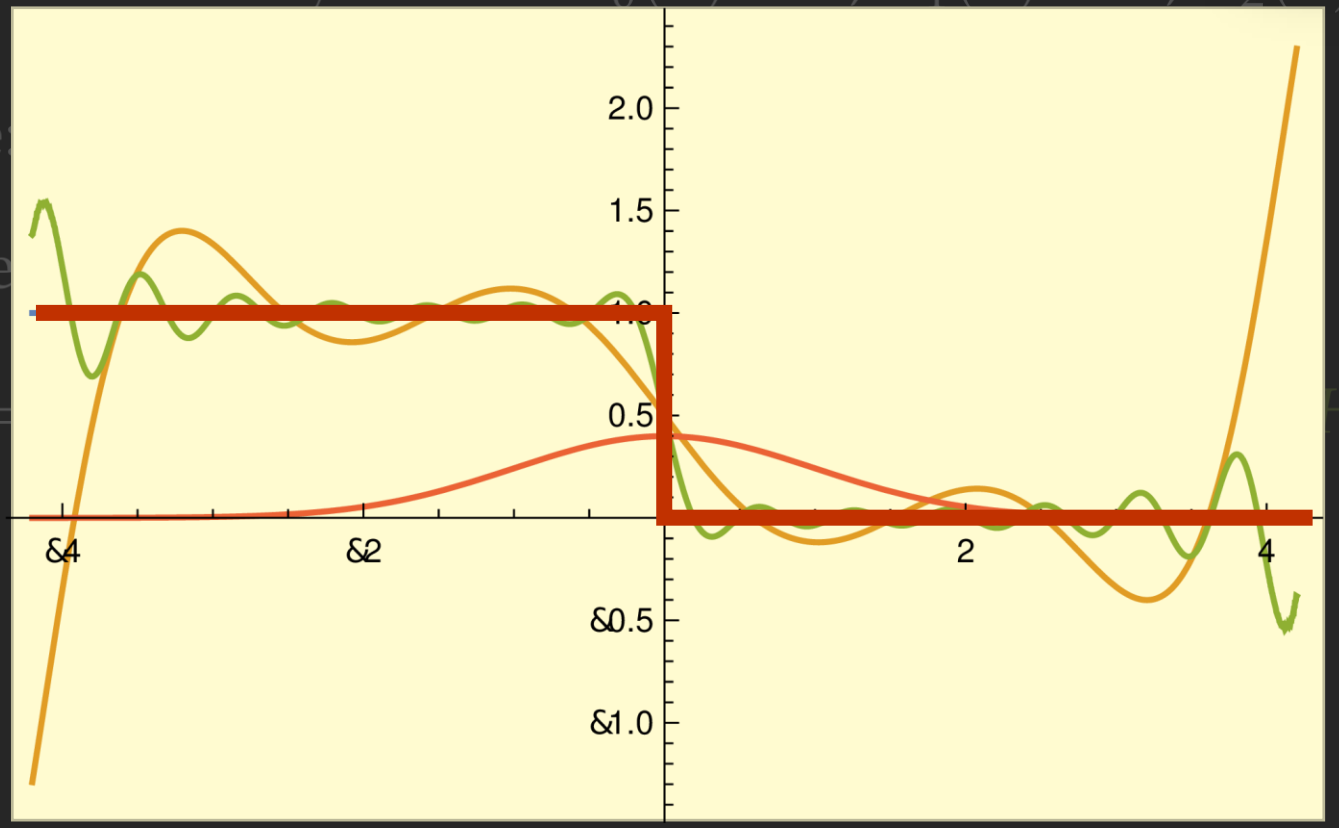
- **Univariate** Hermite Polynomials $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots,$
- **Multivariate**: For $V = (v_1, \ldots, v_d)$ $H_V(x) = \prod_{i=1}^{d} h_{v_i}(x_i)$
- Approximate a function f using Hermite polynomials

$$f_m(x) = \sum_{V:|V|\leq m} \hat{f}(V) \cdot H_V(x) \qquad \hat{f}(V) = \mathbb{E}_{x\sim\mathcal{N}_0}[f(x)H_V(x)]$$

Hermite Polynomials

- **Univariate** Hermite Polynomials $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = \frac{x^2-1}{\sqrt{2}}, \ldots,$
- **Multivariate**
- Approximate

$$f_m(x) = [H_V(x)]$$



Unknown Set – Unknown Gaussian

$\psi(x) = 1_S(x)w(x)$
weighted characteristic function of S

$\Gamma(S) =$ Gaussian surface area of S
$\alpha =$ Gaussian mass of S

Unknown Set – Unknown Gaussian

$\psi(x) = 1_S(x)w(x)$
weighted characteristic function of S

$\Gamma(S) =$ Gaussian surface area of S
$\alpha =$ Gaussian mass of S

Theorem ([Kontonis, Tzamos, Z. '19]). *If* $m \geq O\left(\frac{\Gamma(S)^2}{\varepsilon^4}\right)$ *then*

$$\mathbb{E}_{x\sim\mathcal{N}_0}\left[(\psi(x) - \psi_m(x))^2\right] \leq \varepsilon.$$

Unknown Set – Unknown Gaussian

$\psi(x) = 1_S(x)w(x)$
weighted characteristic function of S

$\Gamma(S) =$ Gaussian surface area of S
$\alpha =$ Gaussian mass of S

Theorem ([Kontonis, Tzamos, Z. '19]). *If* $m \geq O\left(\frac{\Gamma(S)^2}{\varepsilon^4}\right)$ *then*

$$\mathbb{E}_{x\sim\mathcal{N}_0}\left[(\psi(x) - \psi_m(x))^2\right] \leq \varepsilon.$$

We need #samples $n = d^{O(m)}$.

Optimization Problem

$$M_{\psi,h}(\mu) = \mathbb{E}_{x \sim \mathcal{N}_S^*}\left[e^{h(\mu;x)}\psi(x)\mathcal{N}_0(x)\right]$$

Optimization Problem

$$M_{\psi,h}(\mu) = \mathbb{E}_{x\sim\mathcal{N}_S^*}\left[e^{h(\mu;x)}\psi(x)\mathcal{N}_0(x)\right]$$

Theorem ([Kontonis, Tzamos, Z. '19]).

- $M_{\psi,h}(\mu)$ *is a convex function of* μ.
- $\nabla_\mu M_{\psi,h}(\mu) = 0 \Leftrightarrow \mu = \mu^*$.
- *We have sample access to* $\nabla_\mu M_{\psi,h}(\mu)$.(*)

$\Rightarrow$ We can run SGD!

Main Result

Theorem ([Kontonis, Tzamos, Z. '19]). Let $S \subseteq \mathbb{R}^d$ unknown and $S \in \mathcal{S}$ with Gaussian Surface Area $\Gamma(\mathcal{S})$ that satisfies Assumptions 1 and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ i.i.d. samples from $\mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{I}, S)$, then we can efficiently compute an estimate $\hat{\boldsymbol{\mu}}$ that satisfies with probability at least 99%:

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^*\|_2 \leq \varepsilon$$

for $n = d^{\tilde{\Theta}(\Gamma^2(S)/\varepsilon^8)}$.

Contents

- ❑ Estimation Obstacles
  - ○ Truncated Samples
  - ○ **Inhomogeneous Population**
- ❖ Global Convergence of Iterative Methods

Optimization Problems

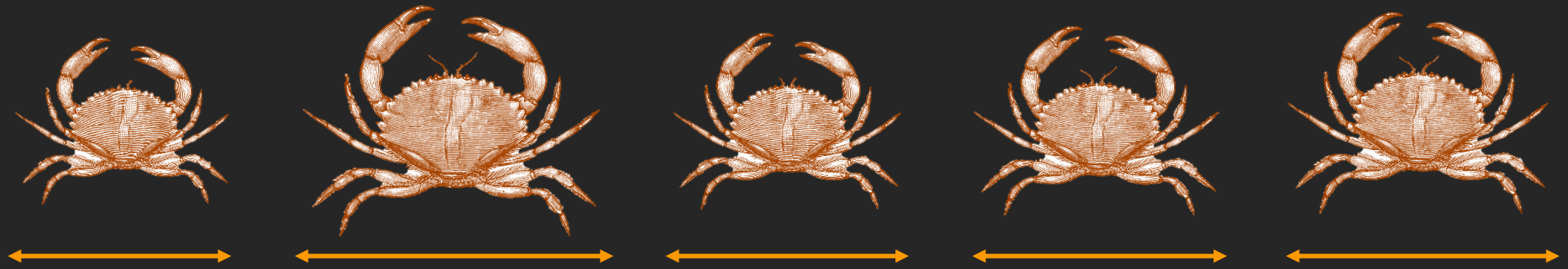
Convex Landscape

Non-Convex Landscape

Iterative Methods

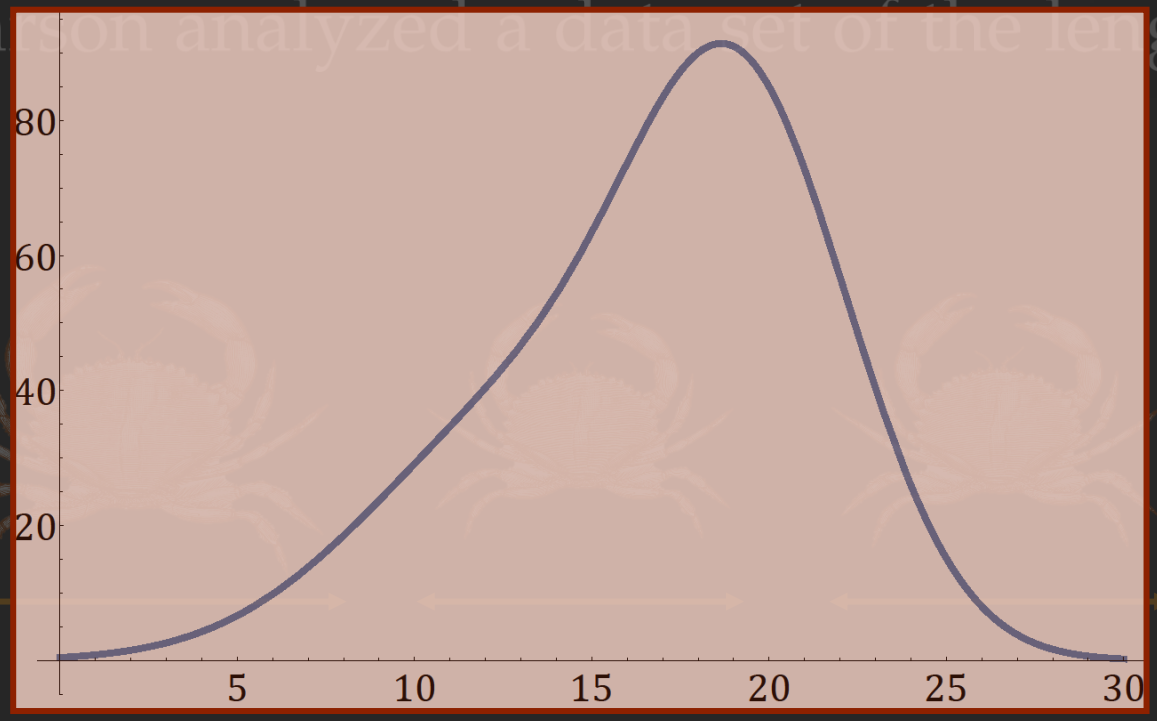
Mixtures of Two Gaussians

In 1894 Karl Pearson analyzed a data set of the lengths of 1000 crabs.



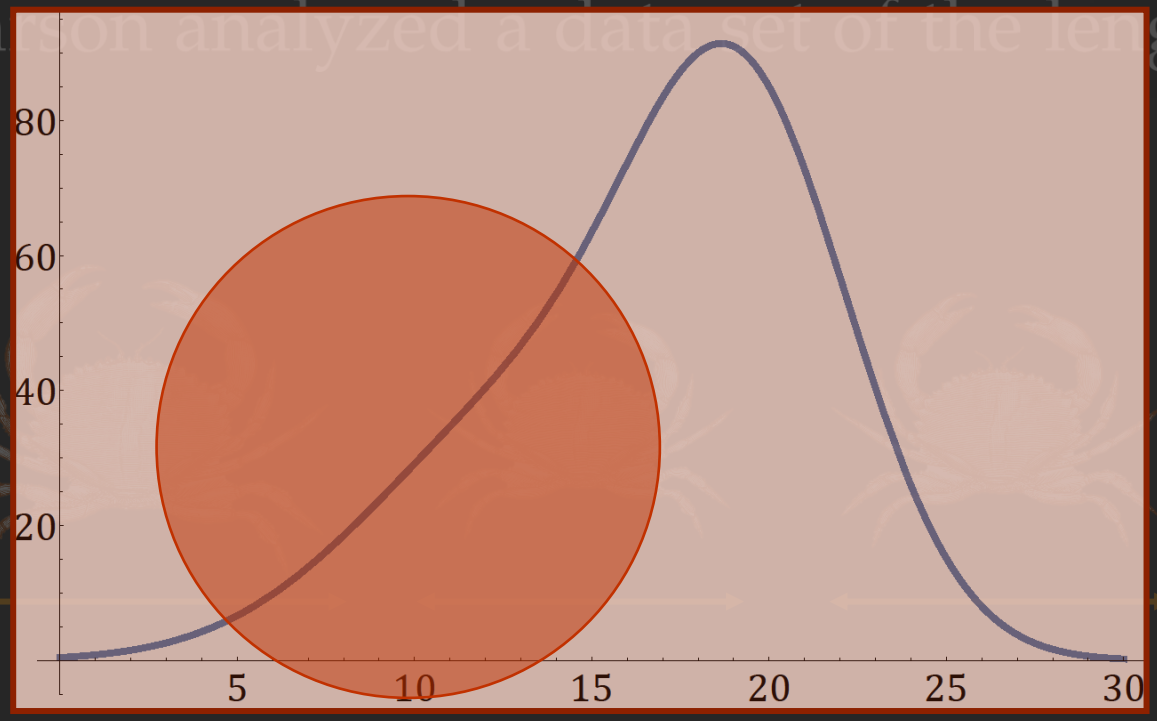
Mixtures of Two Gaussians

In 1894 Karl Pearson analyzed a data set of the lengths of 1000 crabs.



Mixtures of Two Gaussians

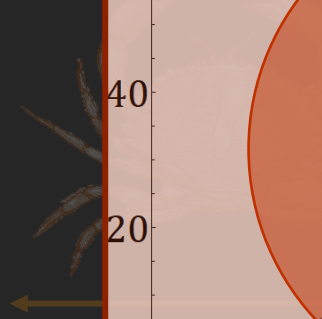
In 1894 Karl Pearson analyzed a data set of the lengths of 1000 crabs.



Mixtures of Two Gaussians

How we can explain this asymmetry?

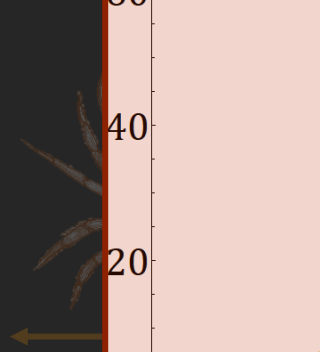
In 1894 Karl Pearson analyzed a data set of the lengths of 1000 crabs.



Mixtures of Two Gaussians

Inhomogeneous population!

In 1894 Karl Pea gths of 1000 crabs.



Mixtures of Two Gaussians

Pearson [Pearson 1894] used the theoretically rigorous **method of moments** in order to estimate the means and variances of the mixture of normal distributions but:

Mixtures of Two Gaussians

Pearson [Pearson 1894] used the theoretically rigorous **method of moments** in order to estimate the means and variances of the mixture of normal distributions but:

1. only asymptotic results,

2. only single dimensional.

Expectation – Maximization Algorithm

EM algorithm was proposed by [Dempster Lair Rubin 1977] as a general technique to solve estimation problems when likelihood is **non-convex**.

Expectation – Maximization Algorithm

EM algorithm was proposed by [Dempster Lair Rubin 1977] as a general technique to solve estimation problems when likelihood is **non-convex**.

...with more than 50,000 citations according to Google Scholar!

Mixtures of Two Gaussians

We assume that we have access to samples from the following mixture of two multidimensional normal distributions

$$p_{\mu_1,\mu_2}(x) = \frac{1}{2}\mathcal{N}(\mu_1, I; x) + \frac{1}{2}\mathcal{N}(\mu_2, I; x).$$

Goal: Estimate the means μ_1, μ_2.

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

We assume that length follows a Gaussian distribution with unit variance.

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

Initially we set $\mu_1 = 1.6$, $\mu_2 = 1.7$.

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

Initially we set $\mu_1 = 1.6$, $\mu_2 = 1.7$.

(E-step)

Species 1 Posterior

0.505	0.504	0.502	0.496	0.492	0.499	0.495

Species 2 Posterior

0.495	0.496	0.498	0.504	0.508	0.501	0.505

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

(M-step)

Species 1 Mean Length

$$\frac{0.505 \cdot 1.5 + 0.504 \cdot 1.52 + 0.502 \cdot 1.57 + 0.496 \cdot 1.72 + 0.492 \cdot 1.83 + 0.499 \cdot 1.64 + 0.495 \cdot 1.75}{0.505 + 0.504 + 0.502 + 0.496 + 0.492 + 0.499 + 0.495} =$$

$$= 1.63$$

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

(M-step)

Species 2 Mean Length

$$\frac{0.495 \cdot 1.5 + 0.496 \cdot 1.52 + 0.498 \cdot 1.57 + 0.504 \cdot 1.72 + 0.508 \cdot 1.83 + 0.501 \cdot 1.64 + 0.505 \cdot 1.75}{0.495 + 0.496 + 0.498 + 0.504 + 0.508 + 0.501 + 0.505} =$$

$$= 1.67$$

Numerical Example

We are given the following samples

1.5	1.52	1.57	1.72	1.83	1.64	1.75

and we are looking for an estimation of the mean length of species 1, μ_1 and the mean length of species 2 μ_2.

(M-step)

Species 2 Mean Length

$$\frac{0.495\cdot 1.5+0.496\cdot 1.52+0.498\cdot 1.57+0.504\cdot 1.72+0.508\cdot 1.83+0.501\cdot 1.64+0.505\cdot 1.75}{0.495+0.496+0.498+0.504+0.508+0.501+0.505} =$$

$=1.67$

Repeat…

The Estimation Algorithm

1. pick $\boldsymbol{\mu}^{(0)}$ at random,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\dfrac{\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I};\, \boldsymbol{x}_i)}{\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I};\, \boldsymbol{x}_i) + \mathcal{N}(-\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I};\, \boldsymbol{x}_i)}$,
4. $\quad \boldsymbol{\mu}^{(t)} \leftarrow \dfrac{\sum_i p_i \boldsymbol{x}_i}{\sum_i p_i}$
5. **output** $\boldsymbol{\mu}^{(T)}$

Our Result

Theorem ([Daskalakis Tzamos Z. '17]). Given access to $n = \tilde{\Theta}(d/\varepsilon^2)$ samples from p_{μ} the EM algorithm gives an estimation $\hat{\mu}$ s.t. with probability 99 %

$$d_{TV}\left(p_{\mu}, p_{\hat{\mu}}\right) \leq \varepsilon.$$

Our Result

Theorem ([Daskalakis Tzamos Z. '17]). Given access to $n = \tilde{\Theta}(d/\varepsilon^2)$ samples from p_μ the EM algorithm gives an estimation $\hat{\mu}$ s.t. with probability 99 %

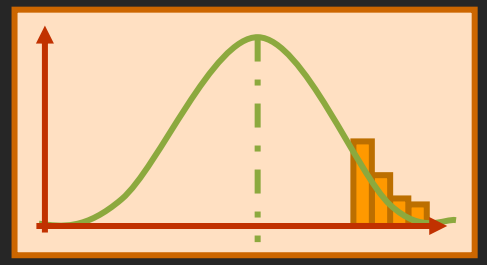
$$d_{TV}\left(p_\mu, p_{\hat{\mu}}\right) \leq \varepsilon.$$

First global convergence guarantees for a non-trivial instance of EM algorithm since its definition in 1977!

Concurrent and independent work by [Xu Hsu Maleki '16].

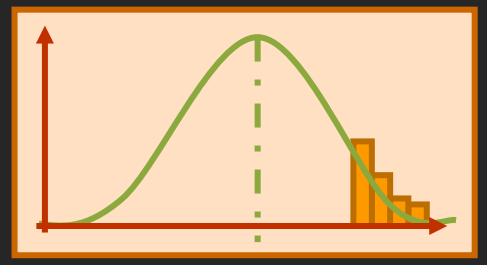
Truncated Samples – Inhomogeneous Population

Truncated Samples



Truncated Samples – Inhomogeneous Population

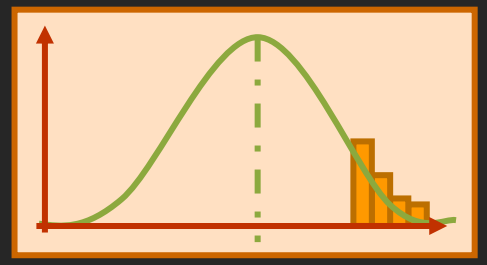
Truncated Samples



1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

Truncated Samples – Inhomogeneous Population

Truncated Samples

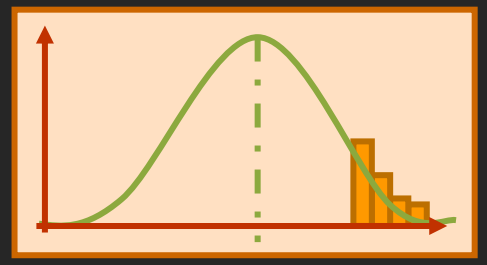


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

Proof of Convergence

Truncated Samples – Inhomogeneous Population

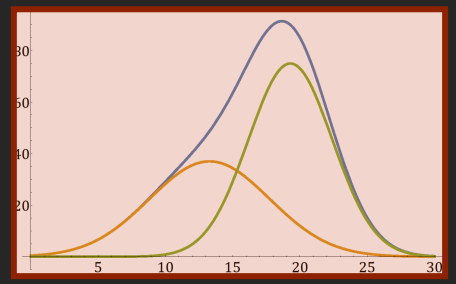
Truncated Samples



1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

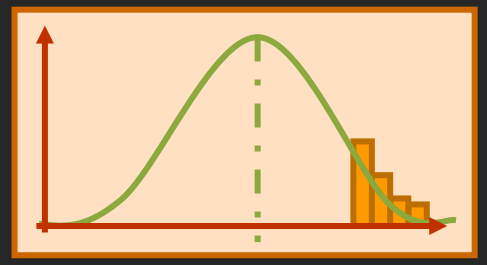
Proof of Convergence

Inhomogeneous



Truncated Samples – Inhomogeneous Population

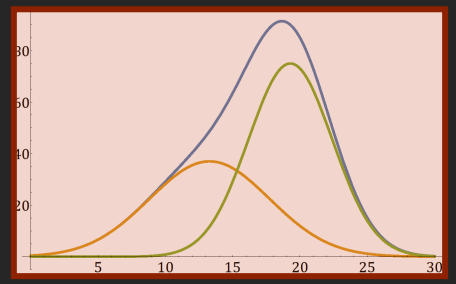
Truncated Samples



1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

Proof of Convergence

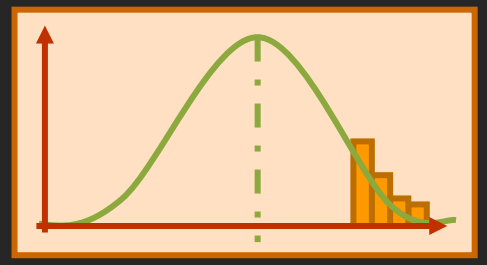
Inhomogeneous



1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Truncated Samples – Inhomogeneous Population

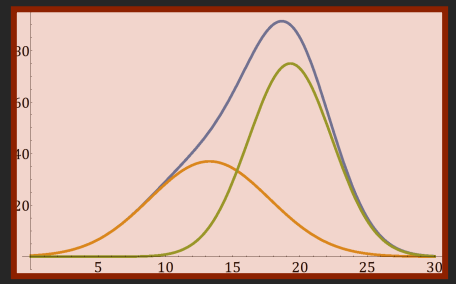
Truncated Samples



1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

Proof of Convergence

Inhomogeneous

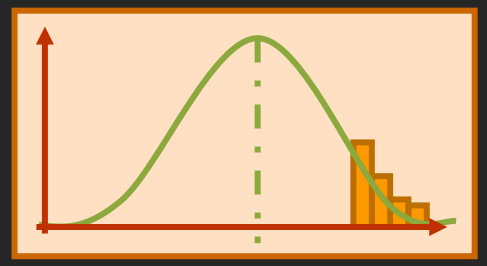


1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Proof of Convergence

Truncated Samples – Inhomogeneous Population

Truncated Samples

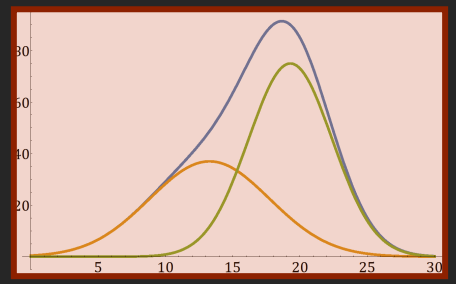


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$

Proof of Convergence

Convex Landscape

Inhomogeneous



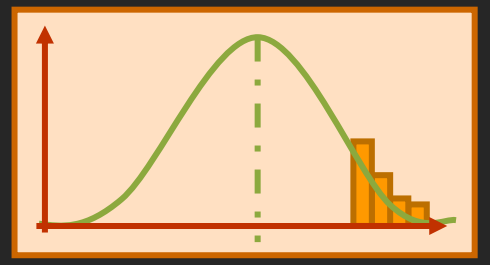
1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Proof of Convergence

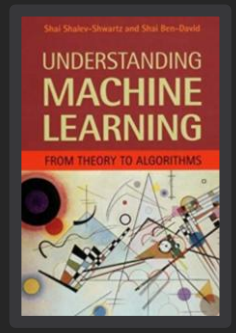
Non-Convex Landscape

Truncated Samples – Inhomogeneous Population

Truncated Samples

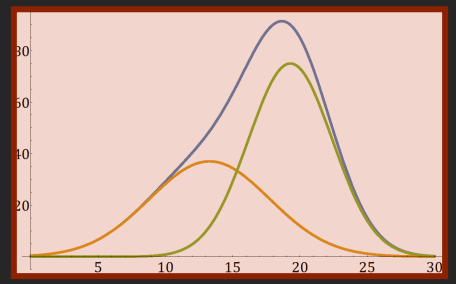


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$



$$\mathcal{H}_{\bar{\ell}} \succeq \lambda I$$

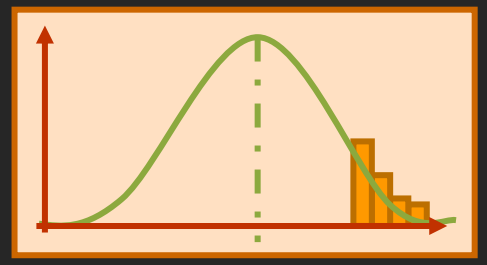
Inhomogeneous



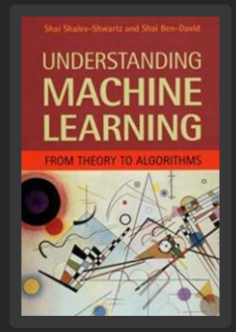
1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Truncated Samples – Inhomogeneous Population

Truncated Samples

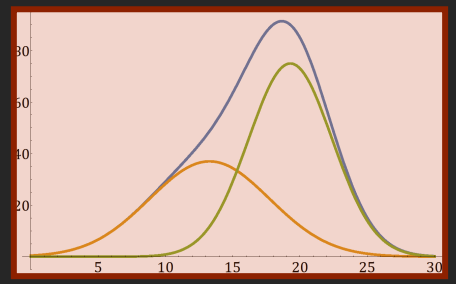


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$



$$\mathcal{H}_{\bar{\ell}} \succeq \lambda I$$

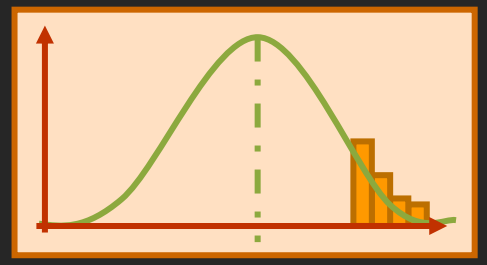
Inhomogeneous



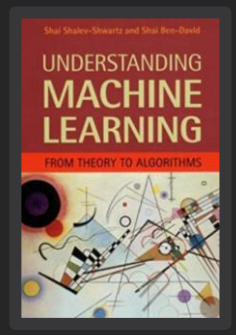
1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Truncated Samples – Inhomogeneous Population

Truncated Samples

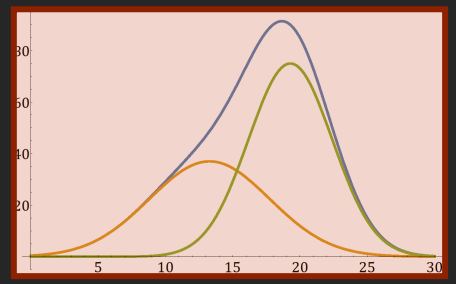


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$



$$\mathcal{H}_{\bar{\ell}} \succeq \lambda I$$

Inhomogeneous

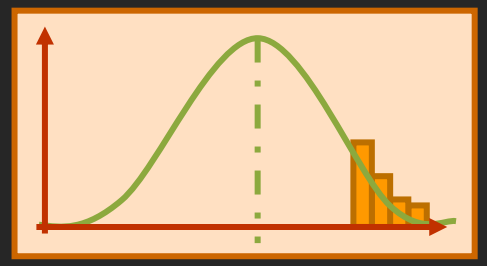


1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

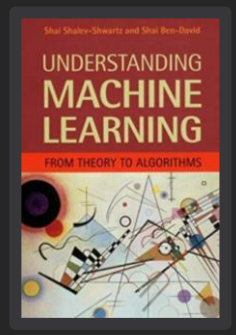
Try to find potential function.

Truncated Samples – Inhomogeneous Population

Truncated Samples

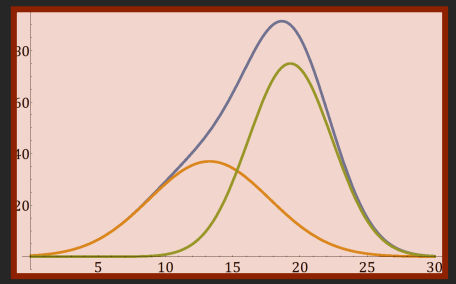


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$



$$\mathcal{H}_{\bar{\ell}} \succeq \lambda I$$

Inhomogeneous



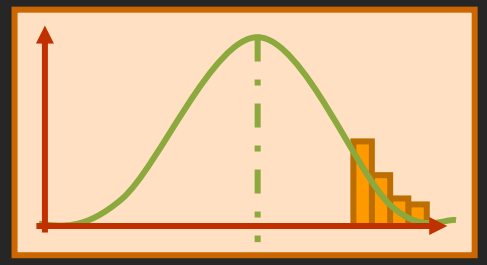
1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Try to find potential function.

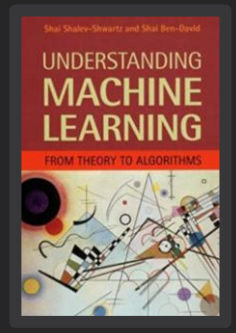
[Lyapunov] Universal tool but only local convergence.

Truncated Samples – Inhomogeneous Population

Truncated Samples

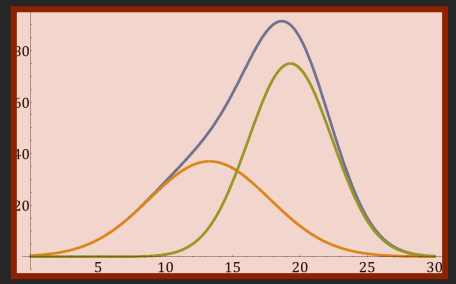


1. estimate $\mu^{(0)} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. **for** $t = 1 \dots T$ **do**
3. $\quad r \leftarrow$ Sample $\mathcal{N}(\mu^*, I, C)$,
4. $\quad \hat{r} \leftarrow$ Sample $\mathcal{N}(\mu^{(t-1)}, I, C)$
5. $\quad v^{(t)} \leftarrow \mu^{(t-1)} - \eta_t(r - \hat{r})$
6. $\quad \mu^{(t)} \leftarrow$ project $v^{(t)}$ to the ball $B = \left\{x \mid \left\|x - \mu^{(0)}\right\| \leq R\right\}$
7. **output** $\frac{1}{T+1}\sum_{t=0}^{T} \mu^{(t)}$



$$\mathcal{H}_{\bar{\ell}} \succeq \lambda I$$

Inhomogeneous



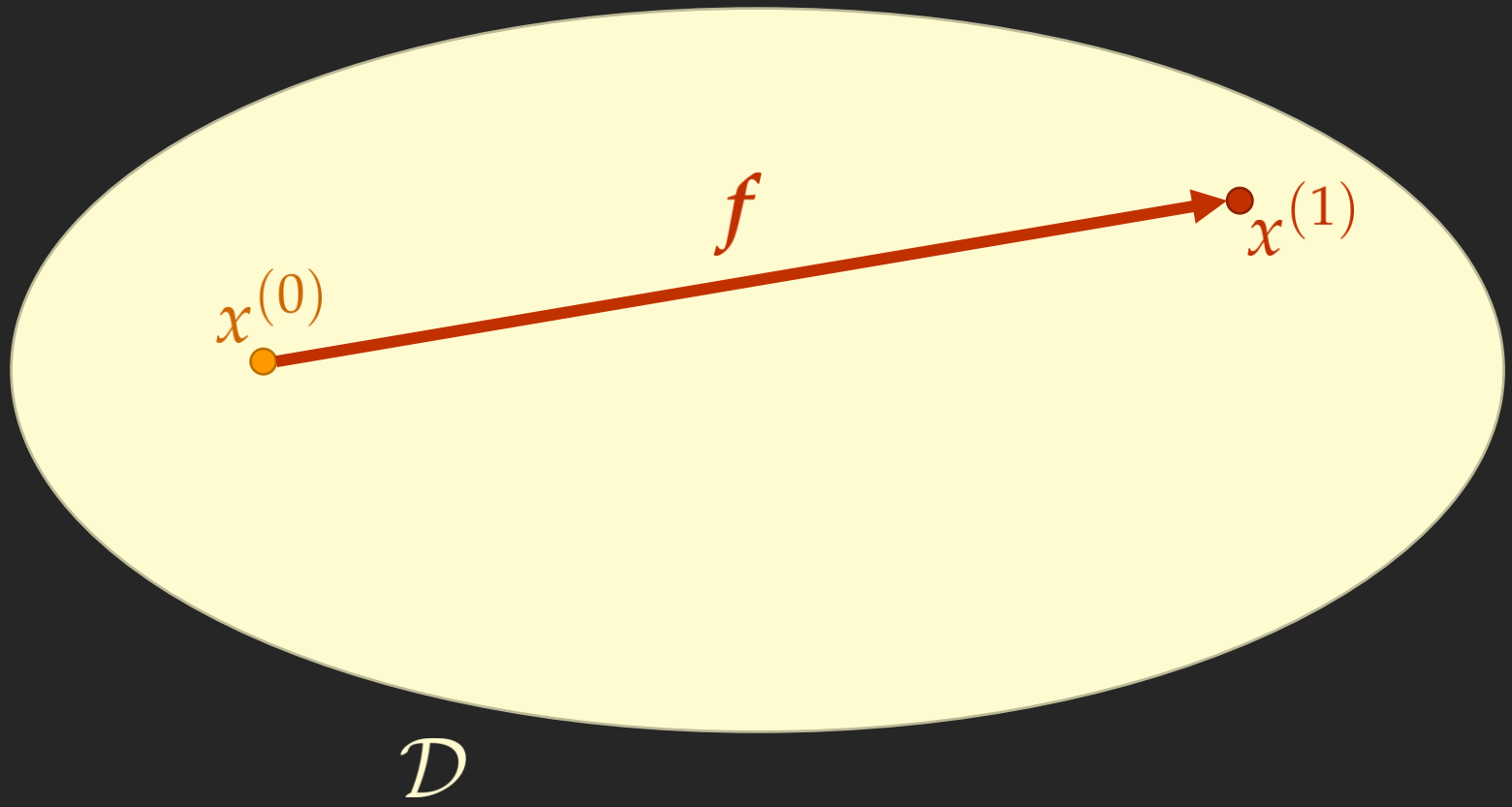
1. pick $\mu^{(0)}$ at random,
2. **for** $t = 1 \dots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\mu^{(t-1)}, I;\, x_i)}{\mathcal{N}(\mu^{(t-1)}, I;\, x_i) + \mathcal{N}(-\mu^{(t-1)}, I;\, x_i)}$,
4. $\quad \mu^{(t)} \leftarrow \frac{\sum_i p_i x_i}{\sum_i p_i}$
5. **output** $\mu^{(T)}$

Try to find potential function.

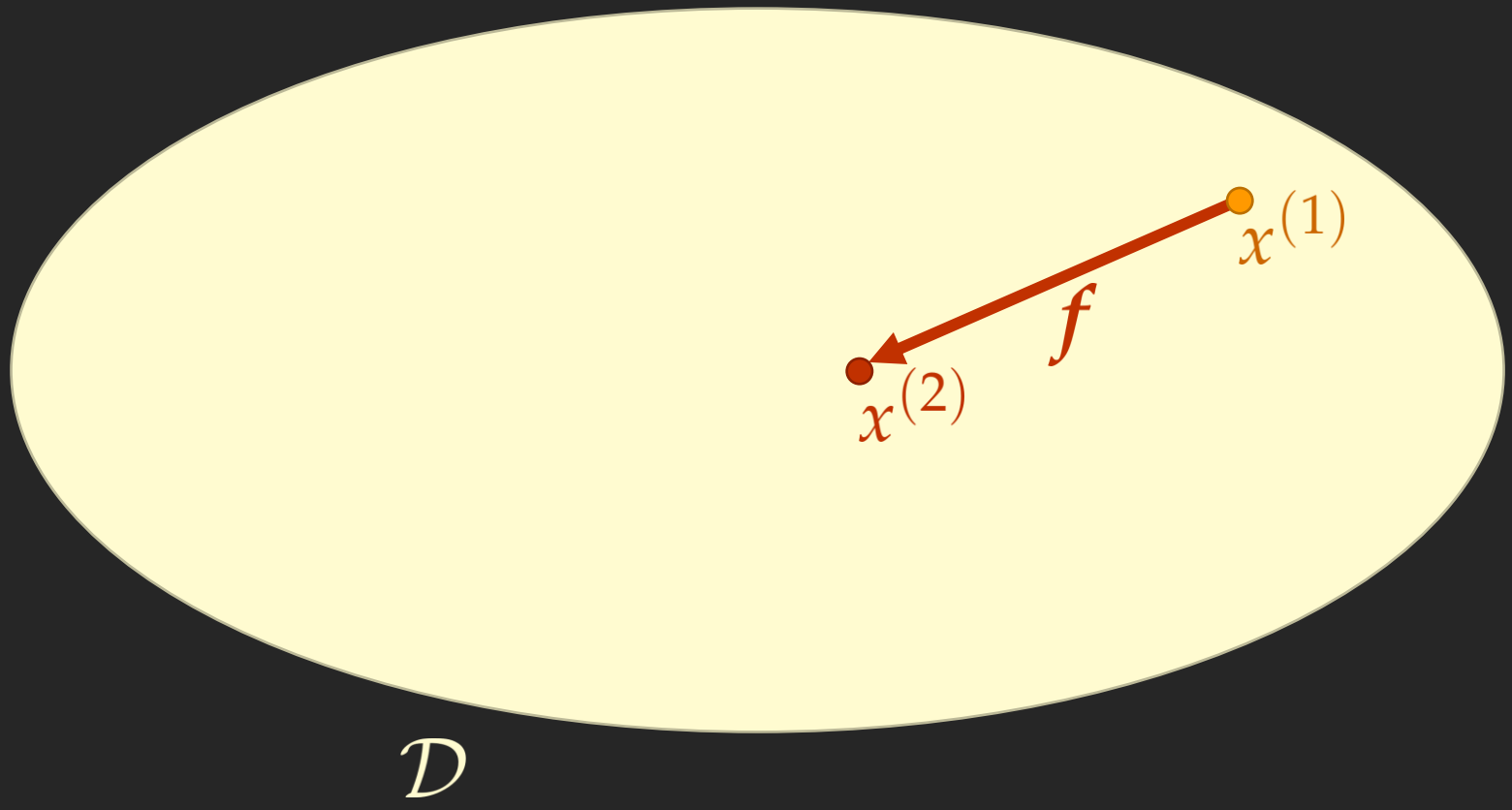
Global Convergence?

…versal tool but only local convergence.

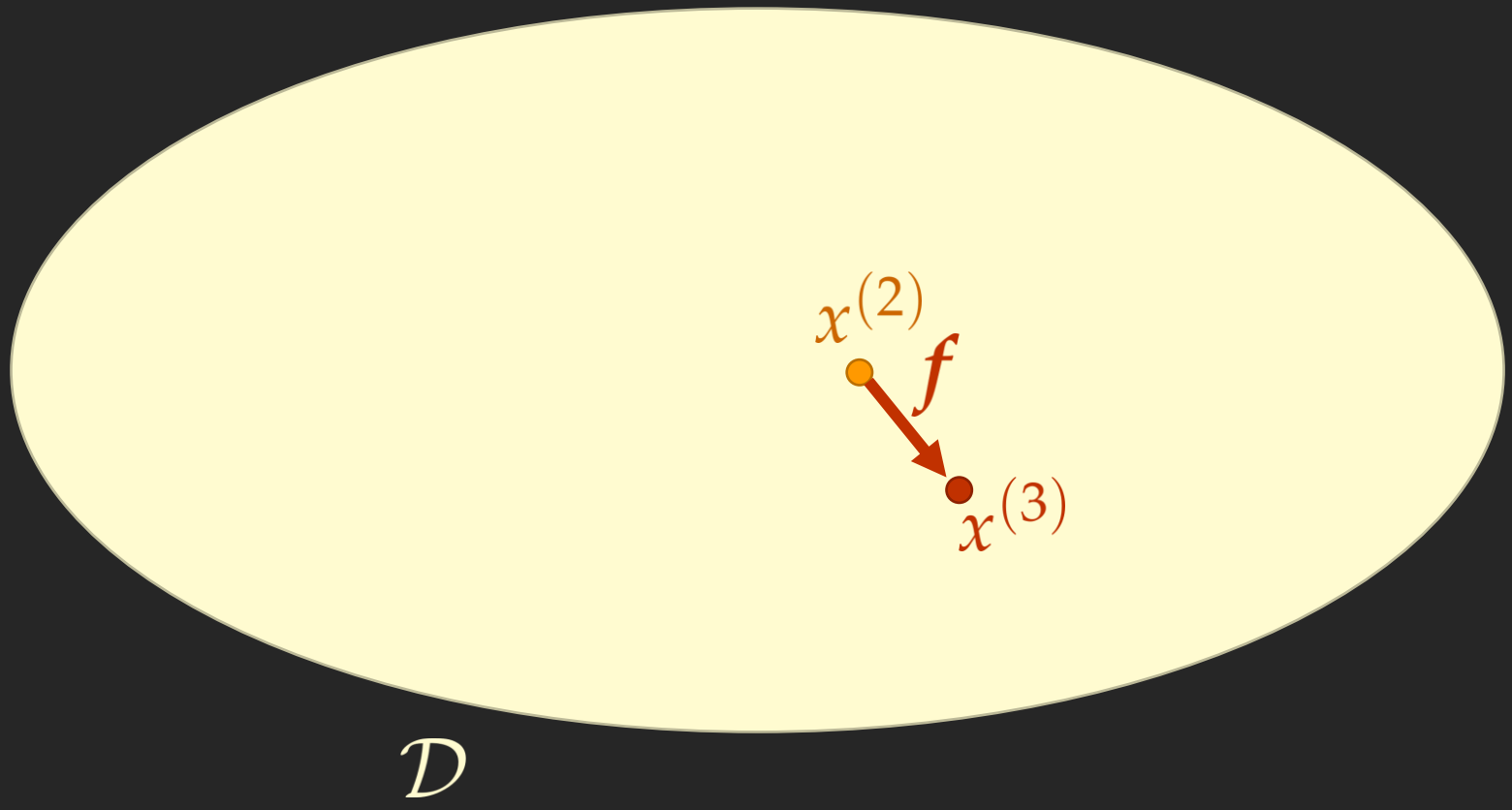
Iterative Methods



Iterative Methods



Iterative Methods



Solution Concept

ε-Approximate Fixed Point
Let d^* be a distance metric,

$$d^*(\hat{x}, f(\hat{x})) \leq \varepsilon.$$

Solution Concept

ε-Approximate Fixed Point
Let d^* be a distance metric,

$$d^*(\hat{x}, f(\hat{x})) \leq \varepsilon.$$

How to prove **global convergence** to approximate fixed points?

Contraction Maps

A function $f : \mathcal{D} \to \mathcal{D}$ is a **contraction map** with respect to d if

$$d(f(x), f(y)) \leq c \cdot d(x, y) \quad \forall x, y \in \mathcal{D}.$$

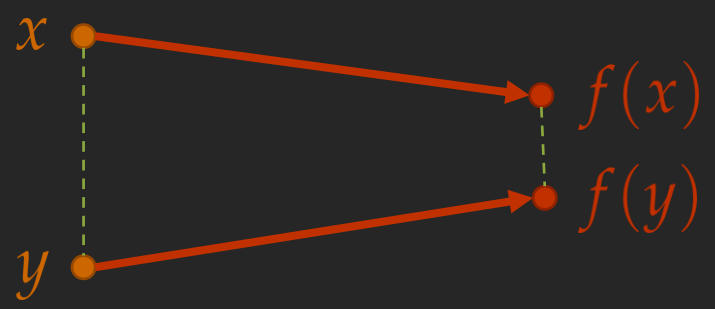
and $c < 1$.

Contraction Maps

A function $f : \mathcal{D} \to \mathcal{D}$ is a **contraction map** with respect to d if

$$d(f(x), f(y)) \leq c \cdot d(x, y) \quad \forall x, y \in \mathcal{D}.$$

and $c < 1$.



Contraction Maps

Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

Contraction Maps

Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

Contraction Maps

Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

Contraction Maps

Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

Contraction Maps

Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

Contraction Maps

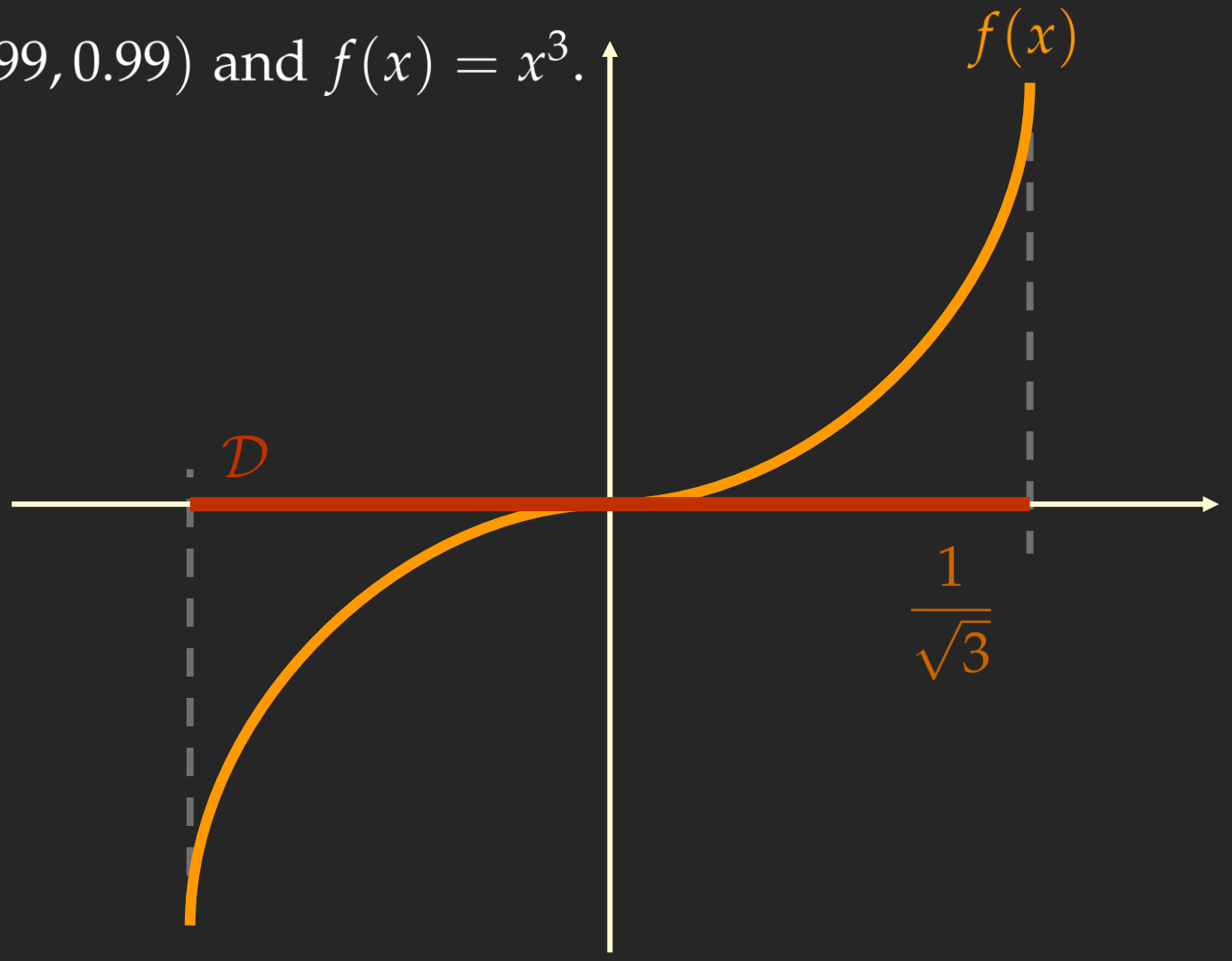
Banach's Fixed Point Theorem
Let $f : \mathcal{D} \to \mathcal{D}$ be a contraction map, then f **has a unique fixed point** x^*. Moreover f convergences globally and geometrically to x^*.

How general is Banach's Theorem?

$\{x^*\}$

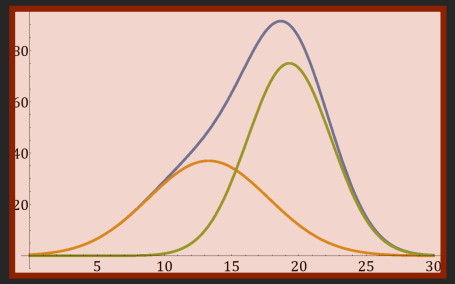
Example 1

Let $\mathcal{D} = (-0.99, 0.99)$ and $f(x) = x^3$.



Example 2

Inhomogeneous



1. pick $\boldsymbol{\mu}^{(0)}$ at random,
2. **for** $t = 1 \ldots T$ **do**
3. $\quad p_i \leftarrow$ evaluate $\frac{\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}; \boldsymbol{x}_i)}{\mathcal{N}(\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}; \boldsymbol{x}_i) + \mathcal{N}(-\boldsymbol{\mu}^{(t-1)}, \boldsymbol{I}; \boldsymbol{x}_i)}$,
4. $\quad \boldsymbol{\mu}^{(t)} \leftarrow \frac{\sum_i p_i \boldsymbol{x}_i}{\sum_i p_i}$
5. **output** $\boldsymbol{\mu}^{(T)}$

Not a contraction map!

Contraction Maps

A function $f : \mathcal{D} \to \mathcal{D}$ is a **contraction map** with respect to d if

$$d(f(x), f(y)) \leq c \cdot d(x, y) \quad \forall x, y \in \mathcal{D}.$$

and $c < 1$.

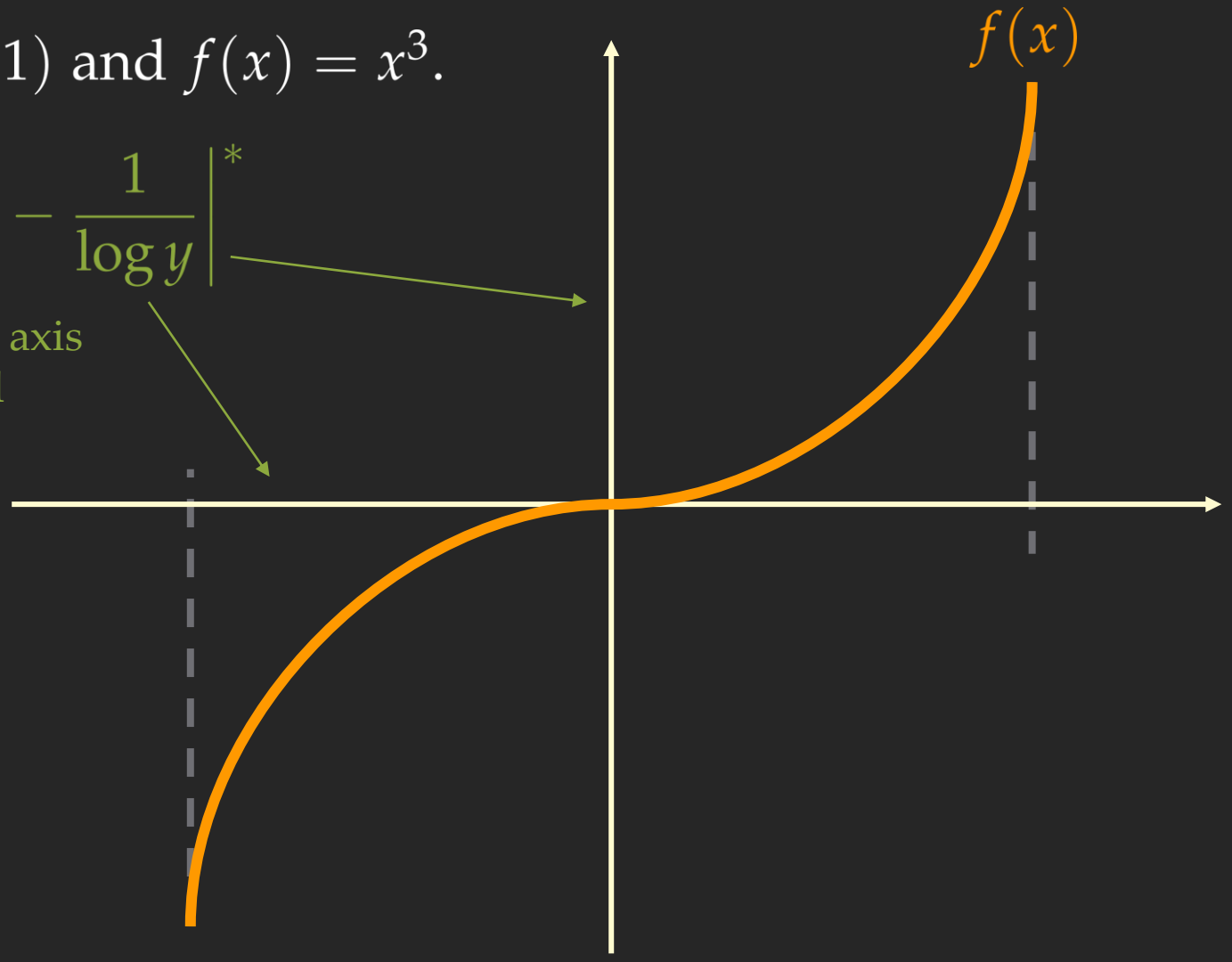
We are free to choose any distance metric d!

Example 1

Let $\mathcal{D} = (-1, 1)$ and $f(x) = x^3$.

$$d(x,y) = \left| \frac{1}{\log x} - \frac{1}{\log y} \right|^*$$

we rescale the axis according to d

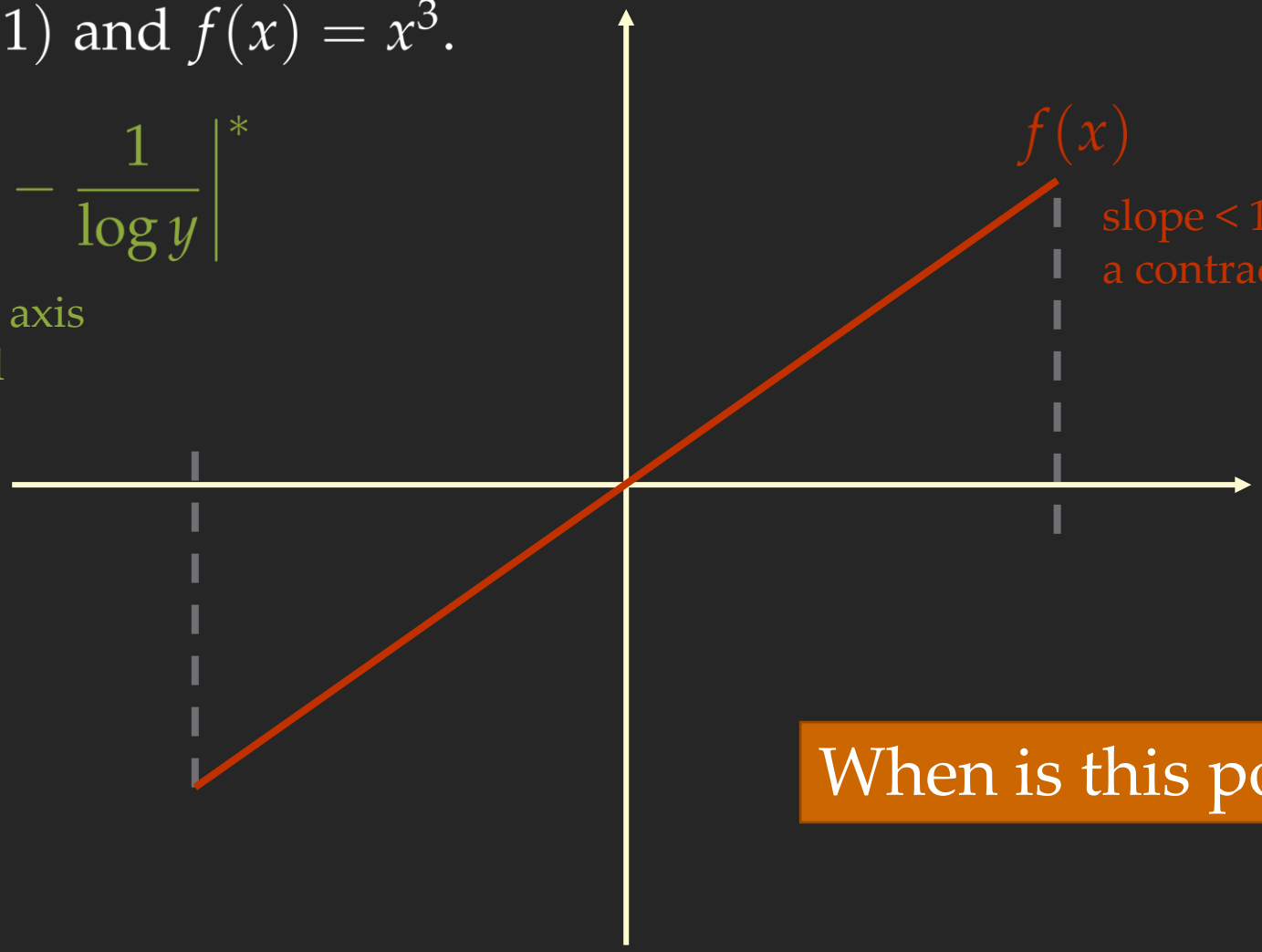


Example 1

Let $\mathcal{D} = (-1, 1)$ and $f(x) = x^3$.

$$d(x, y) = \left| \frac{1}{\log x} - \frac{1}{\log y} \right|^*$$

we rescale the axis according to d



When is this possible?

Converse Fixed Point Theorems!

Bessaga's and Meyers's Converse Fixed Point Theorems
Let $f : \mathcal{D} \to \mathcal{D}$ with a unique fixed point x^*, then there exists a distance metric d' such that f is a contraction map with respect to d'.

Our Converse Fixed Point Theorems

Theorem ([Daskalakis Tzamos Z. '18]). Let $f : \mathcal{D} \to \mathcal{D}$ with a unique fixed point x^*. If f converges globally and geometrically to x^* then this can be shown through a contraction map argument.

Our Converse Fixed Point Theorems

Theorem ([Daskalakis Tzamos Z. '18]). Let $f : \mathcal{D} \to \mathcal{D}$ with a unique fixed point x^*. If f converges globally and geometrically to x^* then this can be shown through a contraction map argument.

Fast Convergence

Our Converse Fixed Point Theorems

Theorem ([Daskalakis Tzamos Z. '18]). Let $f : \mathcal{D} \to \mathcal{D}$ with a unique fixed point x^*. If f converges globally and geometrically to x^* then this can be shown through a contraction map argument.

Fast Convergence

Hardness

Our Converse Fixed Point Theorems

Theorem ([Daskalakis Tzamos Z. '18]). Let $f : \mathcal{D} \to \mathcal{D}$ with a unique fixed point x^*. If f converges globally and geometrically to x^* then this can be shown through a contraction map argument.

Fast Convergence

Hardness

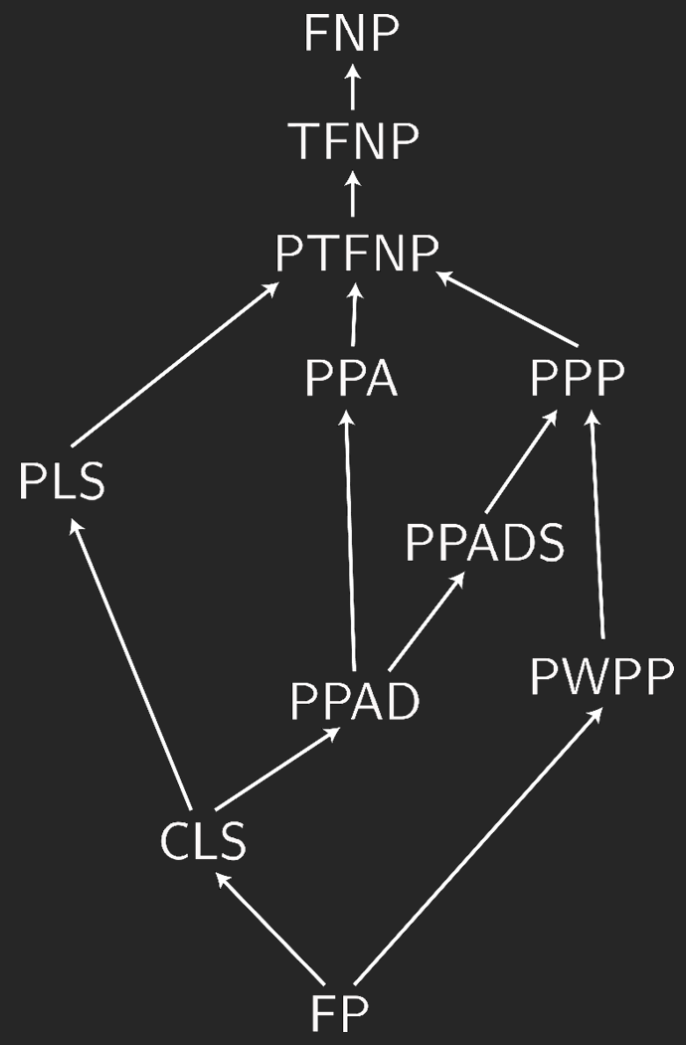
Complexity of Finding Fixed Points?

Complexity of Total Search Problems

FNP: class of search problems whose decision version is in NP.

TFNP: class of total search problems of FNP, i.e. a solution always exists
[Megiddo Papadimitriou '91]

Subclasses of TFNP introduced by
[Johnson Papadimitriou Yannakakis '88],
[**Papadimitriou '94**],
[Daskalakis Papadimitriou '11], [Jerabek '16]

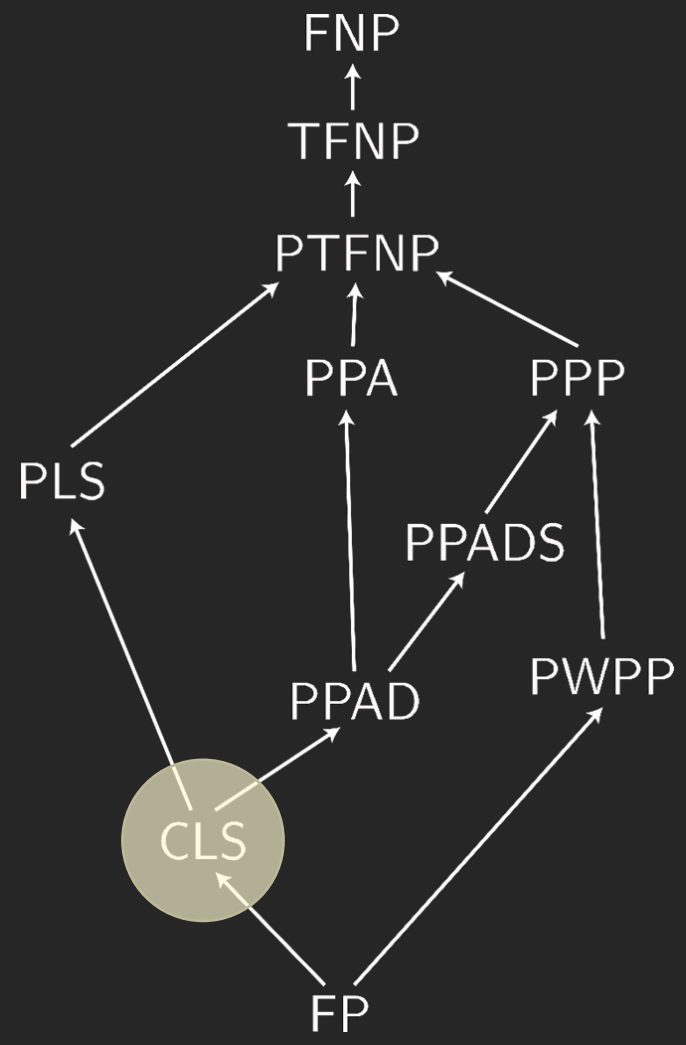


Complexity of Total Search Problems

FNP: class of search problems whose decision version is in NP.

TFNP: class of total search problems of FNP, i.e. a solution always exists
[Megiddo Papadimitriou '91]

Subclasses of TFNP introduced by
[Johnson Papadimitriou Yannakakis '88],
[**Papadimitriou '94**],
[Daskalakis Papadimitriou '11], [Jerabek '16]



Summary

- ❑ Estimation Obstacles
  - ○ Truncated Samples
  - ○ Inhomogeneous Population

Optimization Problems

Convex Landscape

Non-Convex Landscape

Iterative Methods

- ❖ Global Convergence of Iterative Methods

Future Directions

- Truncated estimation in other models? [Daskalakis Ilyas Z. ’20]

Future Directions

- Truncated estimation in other models? [Daskalakis Ilyas Z. ’20]
- Estimation in truncated + Inhomogeneous Population?

Future Directions

- Truncated estimation in other models? [Daskalakis Ilyas Z. ’20]
- Estimation in truncated + Inhomogeneous Population?
- ➢ (known estimation problem) + (censoring)

Future Directions

- Truncated estimation in other models? [Daskalakis Ilyas Z. ’20]
- Estimation in truncated + Inhomogeneous Population?

➢ (known estimation problem) + (censoring)

➢ Better understanding of complexity of Fixed Points, e.g. relations to cryptography? [Bitanski Paneth Rosen ‘15], [Hubacek Yogev ‘17] [Sotiraki Zirdelis Z. ‘18]

Future Directions

- Truncated estimation in other models? [Daskalakis Ilyas Z. '20]
- Estimation in truncated + Inhomogeneous Population?

➢ (known estimation problem) + (censoring)

➢ Better understanding of complexity of Fixed Points, e.g. relations to cryptography? [Bitanski Paneth Rosen '15], [Hubacek Yogev '17] [Sotiraki Zirdelis Z. '18]

Thank you! ☺