Minimax risk of truncated series estimators over symmetric convex polytopes

Adel Javanmard (Stanford University)
with Li Zhang (Microsoft Research)

July 4, 2012

Motivation: function estimation

Consider a continuous function $f : [0,1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \quad \text{for } 1 \le i \le n.$$

Estimate $\{f(t_i)\}_{i=1}^n$ under linear inequality constraints:

$$x = (f(t_1), \cdots, f(t_n))$$

$$Ax \le b$$

- Lipschitz constraint:

$$|x_{i+1} - x_i| \le L|t_{i+1} - t_i| \text{ for } 1 \le i \le n-1.$$

- General convex constraints

What is a good estimator?

Motivation: function estimation

Consider a continuous function $f : [0,1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \qquad \text{for } 1 \le i \le n.$$

Estimate $\{f(t_i)\}_{i=1}^n$ under linear inequality constraints:

$$x = (f(t_1), \cdots, f(t_n))$$

$$Ax \le b$$

- Lipschitz constraint:

$$|x_{i+1} - x_i| \le L|t_{i+1} - t_i| \text{ for } 1 \le i \le n-1.$$

- General convex constraints

What is a good estimator?

Motivation: function estimation

Consider a continuous function $f : [0,1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \quad \text{for } 1 \le i \le n.$$

Estimate $\{f(t_i)\}_{i=1}^n$ under linear inequality constraints:

$$x = (f(t_1), \cdots, f(t_n))$$

$$Ax \le b$$

- Lipschitz constraint:

$$|x_{i+1} - x_i| \le L|t_{i+1} - t_i| \text{ for } 1 \le i \le n-1.$$

- General convex constraints

What is a good estimator?

Motivation: function estimation

Consider a continuous function $f : [0, 1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \quad \text{for } 1 \le i \le n.$$

Estimate $\{f(t_i)\}_{i=1}^n$ under linear inequality constraints:

$$x = (f(t_1), \cdots, f(t_n))$$

$$Ax \le b$$

- Lipschitz constraint:

$$|x_{i+1} - x_i| \le L|t_{i+1} - t_i| \text{ for } 1 \le i \le n-1.$$

- General convex constraints

What is a good estimator?

General problem

$$y = x + w\,,$$

$$x \in X \subseteq \mathbb{R}^n, \qquad w \sim \mathsf{N}(0, \sigma^2 I_{n\times n}).$$

For any estimator $M : \mathbb{R}^n \to \mathbb{R}^n$, define

$$R(M, X, \sigma) = \max_{x \in X} \mathbb{E}_{y \sim x+w} \|x - M(y)\|^2.$$

Minimax risk of a set

$$R(X, \sigma) = \min_M R(M, X, \sigma).$$

General problem

$$y = x + w\,,$$

$$x \in X \subseteq \mathbb{R}^n, \qquad w \sim \mathsf{N}(0, \sigma^2\, I_{n\times n}).$$

For any estimator $M : \mathbb{R}^n \to \mathbb{R}^n$, define

$$R(M, X, \sigma) = \max_{x\in X} \mathbb{E}_{y\sim x+w}\, \|x - M(y)\|^2.$$

Minimax risk of a set

$$R(X, \sigma) = \min_M R(M, X, \sigma).$$

Classes of estimators

- Nonlinear estimators:
  The estimator M can be generally nonlinear. $R(X, \sigma)$

Classes of estimators

- Nonlinear estimators:
  The estimator M can be generally nonlinear. $R(X, \sigma)$

- Linear estimators:
  When M is linear, we denote the minimax risk by $R_L(X, \sigma)$

Classes of estimators

- Nonlinear estimators:
  The estimator M can be generally nonlinear. $R(X, \sigma)$
- Linear estimators:
  When M is linear, we denote the minimax risk by $R_L(X, \sigma)$
- Truncated series estimators:
  Especial class of linear estimators given by orthogonal projections

$$M(y) = Py.$$

  The minimax risk is denoted by $R_T(X, \sigma)$.

Classes of estimators

- Nonlinear estimators:
  The estimator M can be generally nonlinear. $R(X, \sigma)$

- Linear estimators:
  When M is linear, we denote the minimax risk by $R_L(X, \sigma)$

- Truncated series estimators:
  Especial class of linear estimators given by orthogonal projections

$$M(y) = Py.$$

  The minimax risk is denoted by $R_T(X, \sigma)$.

$$R(X, \sigma) \le R_L(X, \sigma) \le R_T(X, \sigma).$$

$$X \subseteq Y \Rightarrow R(X, \sigma) \le R(Y, \sigma).$$

Challenges

- How to compute the minimax risk for arbitrary convex bodies?
- How to design the minimax optimal estimator?

Related work

- Minimax bounds have been developed for various families:
  - Orthosymmetric and quadratically convex objects:
    Hypercubes, ellipsoids, ℓ_p balls for $p \geq 2$.
    [Donoho, Liu, MacGibbon'90]
  - Class of Hölder balls, Sobolev balls, and Besov balls (continuity and energy conditions)
    [Tsybakov'09]
- Techniques for bounding the minimax risk
  [Donoho'90, Nemirovski'99, Yang, Barron'99]

Our main contributions

- A lower bound for minimax risk
  - based on a geometric quantity of the set (approximation radius)
  - depends on the volume of the set (intuitive and strong bounds)

- Optimality of truncated estimators over symmetric polytopes

information theory tools → geometrical understanding of minimax risks

Our main contributions

- A lower bound for minimax risk
  - based on a geometric quantity of the set (approximation radius)
  - depends on the volume of the set (intuitive and strong bounds)
- Optimality of truncated estimators over symmetric polytopes

information theory tools → geometrical understanding of minimax risks

Our main contributions

- A lower bound for minimax risk
  - based on a geometric quantity of the set (approximation radius)
  - depends on the volume of the set (intuitive and strong bounds)
- Optimality of truncated estimators over symmetric polytopes

information theory tools $\rightarrow$ geometrical understanding of minimax risks

Outline

1. Main result
2. An application
3. Proof techniques
4. Further comments

Main result

Notation

For $p > 0$ and $m, n \geq 1$, let

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}$$

$\mathcal{F}_\infty^{m,n}$: Family of symmetric polytopes defined by m hyperplanes

Definition

$$\beta(X) = \max_{\sigma > 0} \frac{R_T(X, \sigma)}{R(X, \sigma)}, \quad \beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \beta(X).$$

Notation

For $p > 0$ and $m, n \geq 1$, let

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}$$

$\mathcal{F}_\infty^{m,n}$: Family of symmetric polytopes defined by m hyperplanes

Definition

$$\beta(X) = \max_{\sigma > 0} \frac{R_T(X, \sigma)}{R(X, \sigma)}, \quad \beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \beta(X).$$

Notation

For $p > 0$ and $m, n \geq 1$, let

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}$$

$\mathcal{F}_\infty^{m,n}$: Family of symmetric polytopes defined by m hyperplanes

Definition

$$\beta(X) = \max_{\sigma > 0} \frac{R_T(X, \sigma)}{R(X, \sigma)}, \quad \beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \beta(X).$$

Optimality of truncated estimators

Theorem (Javanmard, Zhang '11)

If $n = \Omega(\log m)$, for some universal constant C we have

$$\beta_\infty^{m,n} \le C \log m.$$

Furthermore, $\beta_\infty^{m,n} = \Omega(\sqrt{\log m / \log\log m})$.

[Recall:
$a = \Omega(b)$ if a is bounded below by b (up to a constant factor) asymptotically]

An application

Function estimation

Consider a univariate Lipschitz function $f : [0, 1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \quad \text{for } 1 \le i \le n.$$

Goal: estimate $\{f(t_i)\}_{i=1}^n$ from measurements $\{y_i\}_{i=1}^n$.

Lipschitz condition (with constant L):

$$|f(t_{i+1}) - f(t_i)| \le L|t_{i+1} - t_i|, \text{ for } 1 \le i \le n-1.$$

Previous work shows near-optimality of truncated estimators for uniform sampling.

[Nemirovski, Tsybakov'09]

Function estimation

Consider a univariate Lipschitz function $f : [0,1] \to \mathbb{R}$.
We have measurements

$$y_i = f(t_i) + w_i, \quad \text{for } 1 \le i \le n.$$

Goal: estimate $\{f(t_i)\}_{i=1}^n$ from measurements $\{y_i\}_{i=1}^n$.

Lipschitz condition (with constant L):

$$|f(t_{i+1}) - f(t_i)| \le L|t_{i+1} - t_i|, \text{ for } 1 \le i \le n-1.$$

Corollary of our theorem

The truncated series estimator is nearly optimal (within $O(\log n)$) for estimating Lipschitz function at arbitrary sample set.

Proof techniques

high-level idea

- We choose a family of obstruction objects for which we know tight lower bound of minimax risk.

- For $X \in \mathcal{F}_{\infty}^{m,n}$,we show that if X does not have a good truncated estimator, then it will have a "large" obstruction. Hence, no estimator can do well on X.

The difficulty is in choosing the right obstruction for the desired result.

high-level idea

- We choose a family of obstruction objects for which we know tight lower bound of minimax risk.
- For $X \in \mathcal{F}_{\infty}^{m,n}$,we show that if X does not have a good truncated estimator, then it will have a "large" obstruction. Hence, no estimator can do well on X.

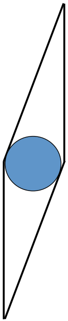
The difficulty is in choosing the right obstruction for the desired result.

Natural obstructions

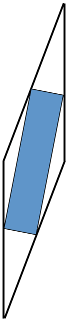
Euclidean balls Hyper-rectangles

Natural obstructions

Euclidean balls



Hyper-rectangles



Not suitable for skewed polytopes!

A remedy : notion of approximation radius

Obstructions: Family of objects which contain a "non-negligible" fraction of a "large" ball.

What does it mean formally?

For any $r > 0$, the volume ratio $\mathrm{vr}(X, r)$ is defined as

$$\mathrm{vr}(X, r) = \left(\frac{\mathrm{vol}(X \cap B_2^n(r))}{\mathrm{vol}(B_2^n(r))}\right)^{1/n}.$$

k-volume ratio $\mathrm{vr}_k(X, r)$:

$$\mathrm{vr}_k(X, r) = \max_{H \in \mathcal{H}_n^k} \mathrm{vr}(X \cap H, r).$$

A remedy : notion of approximation radius

Obstructions: Family of objects which contain a "non-negligible" fraction of a "large" ball.

What does it mean formally?

For any $r > 0$, the volume ratio $\mathrm{vr}(X, r)$ is defined as

$$\mathrm{vr}(X, r) = \left(\frac{\mathrm{vol}(X \cap B_2^n(r))}{\mathrm{vol}(B_2^n(r))} \right)^{1/n}.$$

k-volume ratio $\mathrm{vr}_k(X, r)$:

$$\mathrm{vr}_k(X, r) = \max_{H \in \mathcal{H}_n^k} \mathrm{vr}(X \cap H, r).$$

A remedy : notion of approximation radius

Obstructions: Family of objects which contain a "non-negligible" fraction of a "large" ball.

What does it mean formally?

For any $r > 0$, the volume ratio $\mathrm{vr}(X, r)$ is defined as

$$\mathrm{vr}(X, r) = \left(\frac{\mathrm{vol}(X \cap B_2^n(r))}{\mathrm{vol}(B_2^n(r))} \right)^{1/n}.$$

k-volume ratio $\mathrm{vr}_k(X, r)$:

$$\mathrm{vr}_k(X, r) = \max_{H \in \mathcal{H}_n^k} \mathrm{vr}(X \cap H, r).$$

Approximation radius and a lower bound

Definition (Approximation radius)

For $0 \le c \le 1$, and integer $1 \le k \le n$, the (c, k)-*approximation radius* of X, denoted by $z_{c,k}(X)$ is defined as

$$z_{c,k} = \sup\{r : \mathrm{vr}_k(X, r) \ge c\}.$$

Theorem (Javanmard, Zhang '11)

For any convex set X, and any $0 < c_* \le 1$,

$$R(X, \sigma) \ge C c_*^2 \max_{0 \le k \le n} \min\{z_{c_*,k}(X)^2, k\sigma^2\}.$$

Here, C is a universal constant.

Proof:
Fano's inequality and a lower bound established by Yang, Barron '99.

Approximation radius and a lower bound

Definition (Approximation radius)

For $0 \le c \le 1$, and integer $1 \le k \le n$, the (c, k)-*approximation radius* of X, denoted by $z_{c,k}(X)$ is defined as

$$z_{c,k} = \sup\{r : \mathrm{vr}_k(X, r) \ge c\}.$$

Theorem (Javanmard, Zhang '11)

For any convex set X, and any $0 < c_* \le 1$,

$$R(X, \sigma) \ge C c_*^2 \max_{0 \le k \le n} \min\{z_{c_*,k}(X)^2, k\sigma^2\}.$$

Here, C is a universal constant.

Proof:

Fano's inequality and a lower bound established by Yang, Barron '99.

Kolmogorov widths and truncated estimators

$$y = x + w, \qquad M(y) = Py.$$

Let $P \in \mathcal{P}_k$ (the set of k-dimensional projections), then

$$\begin{aligned}\mathbb{E}\,\|x - M(y)\|^2 &= \mathbb{E}(\|x - P(x)\|^2 + \|P(w)\|^2)\\ &= \|x - P(x)\|^2 + k\sigma^2.\end{aligned}$$

$k-$ Kolmogorov width:

$$d_k(X) = \min_{P \in \mathcal{P}_k} \max_{x \in X} \|x - Px\| = \min_{P \in \mathcal{P}_k} \max_{x \in X} \|P^{\perp} x\|.$$

$k-$ Kolmogorov widths characterize the risk!

$$R_T(X, \sigma) = \min_k (d_k(X)^2 + k\sigma^2).$$

Kolmogorov widths and truncated estimators

$$y = x + w, \qquad M(y) = Py.$$

Let $P \in \mathcal{P}_k$ (the set of k-dimensional projections), then

$$\begin{aligned}\mathbb{E}\,\|x - M(y)\|^2 &= \mathbb{E}(\|x - P(x)\|^2 + \|P(w)\|^2) \\ &= \|x - P(x)\|^2 + k\sigma^2.\end{aligned}$$

$k-$ Kolmogorov width:

$$d_k(X) = \min_{P \in \mathcal{P}_k} \max_{x \in X} \|x - Px\| = \min_{P \in \mathcal{P}_k} \max_{x \in X} \|P^{\perp} x\|.$$

$k-$ Kolmogorov widths characterize the risk!

$$R_T(X, \sigma) = \min_k (d_k(X)^2 + k\sigma^2).$$

Kolmogorov widths and truncated estimators

$$y = x + w, \qquad M(y) = Py.$$

Let $P \in \mathcal{P}_k$ (the set of k-dimensional projections), then

$$\begin{aligned}\mathbb{E}\,\|x - M(y)\|^2 &= \mathbb{E}(\|x - P(x)\|^2 + \|P(w)\|^2)\\ &= \|x - P(x)\|^2 + k\sigma^2.\end{aligned}$$

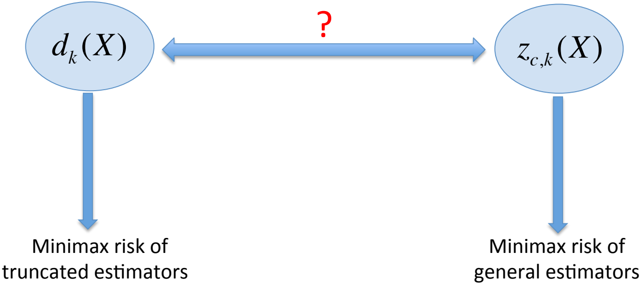
$k-$ Kolmogorov width:

$$d_k(X) = \min_{P\in\mathcal{P}_k} \max_{x\in X} \|x - Px\| = \min_{P\in\mathcal{P}_k} \max_{x\in X} \|P^{\perp}x\|.$$

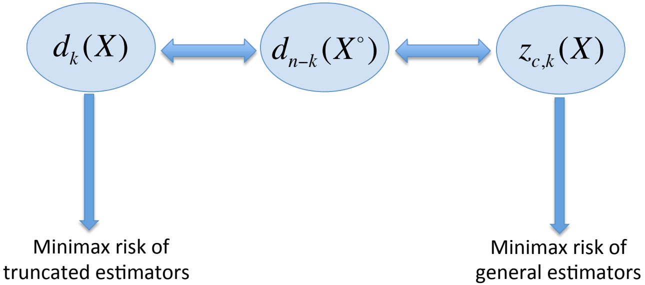
$k-$ Kolmogorov widths characterize the risk!

$$R_T(X,\sigma) = \min_k (d_k(X)^2 + k\sigma^2).$$

Schematic


$d_k(X)$
?
$z_{c,k}(X)$
Minimax risk of truncated estimators
Minimax risk of general estimators

Schematic


$d_k(X)$
$d_{n-k}(X^\circ)$
$z_{c,k}(X)$
Minimax risk of
truncated estimators
Minimax risk of
general estimators

A duality relationship

Lemma (Javanmard, Zhang '11)

For any convex symmetric $X \subset \mathbb{R}^n$ and any $0 \le k \le n$ and $0 < \epsilon < 1$,

$$d_k(X) d_{n-(1-\epsilon)k}(X^\circ) \le c_1 \sqrt{\frac{k}{\epsilon}},$$

where $c_1 > 0$ is a universal constant.

X contains a k-dimensional ball with radius $1/d_{n-k}(X^\circ)$.

$$d_k(X) \le c_1 \sqrt{\frac{k}{\epsilon}} \frac{1}{d_{n-(1-\epsilon)k}(X^\circ)}$$

$\Longrightarrow$ gives a fairly weak bound on $R_T(X, \sigma)$:(

A duality relationship

Lemma (Javanmard, Zhang '11)

For any convex symmetric $X \subset \mathbb{R}^n$ and any $0 \leq k \leq n$ and $0 < \epsilon < 1$,

$$d_k(X) d_{n-(1-\epsilon)k}(X^\circ) \leq c_1 \sqrt{\frac{k}{\epsilon}},$$

where $c_1 > 0$ is a universal constant.

X contains a k-dimensional ball with radius $1/d_{n-k}(X^\circ)$.

$$d_k(X) \leq c_1 \sqrt{\frac{k}{\epsilon}} \frac{1}{d_{n-(1-\epsilon)k}(X^\circ)}$$

$\Longrightarrow$ gives a fairly weak bound on $R_T(X, \sigma)$:(

A duality relationship

Lemma (Javanmard, Zhang '11)

For any convex symmetric $X \subset \mathbb{R}^n$ and any $0 \le k \le n$ and $0 < \epsilon < 1$,

$$d_k(X) d_{n-(1-\epsilon)k}(X^\circ) \le c_1 \sqrt{\frac{k}{\epsilon}},$$

where $c_1 > 0$ is a universal constant.

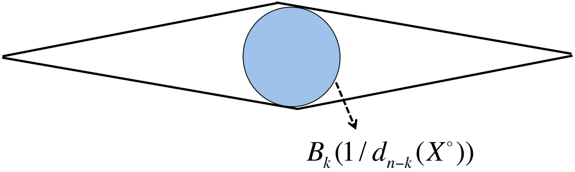
X contains a k-dimensional ball with radius $1/d_{n-k}(X^\circ)$.

$$d_k(X) \le c_1 \sqrt{\frac{k}{\epsilon}} \frac{1}{d_{n-(1-\epsilon)k}(X^\circ)}$$

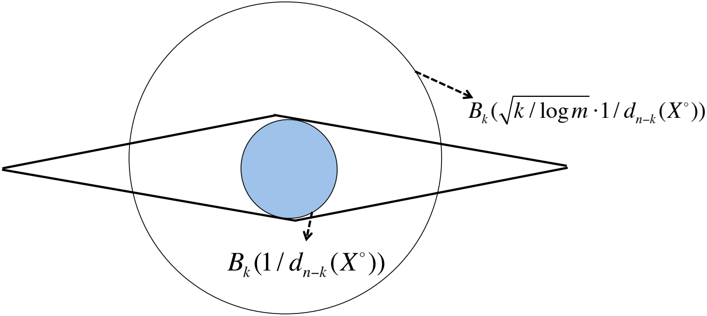
$\Longrightarrow$ gives a fairly weak bound on $R_T(X, \sigma)$:(

… and relation to the approximation radius

Suppose $X \in \mathcal{F}_\infty^{m,n}$.



… and relation to the approximation radius



Larger ball still has non-negligible fraction of its volume inside X.

Final step

Lemma (Javanmard, Zhang '11)

For any $X \in \mathcal{F}_{\infty}^{m,n}$, $0 < c_* \leq 0.2$, and $0 < k \leq n$,

$$z_{c_*,k}(X) \geq c_2 \sqrt{\frac{k}{\log m}} \cdot \frac{1}{d_{n-k}(X^\circ)},$$

where c_2 is a universal constant.

Combining the lemmas,

$$d_k(X) \leq c_1 \sqrt{\frac{k}{\epsilon}} \cdot \frac{1}{d_{n-(1-\epsilon)k}(X^\circ)} \leq c_1 c_2 \sqrt{\frac{\log m}{\epsilon}} z_{c_*,k}(X).$$

Skinny objects always have a small shadow!

Final step

Lemma (Javanmard, Zhang '11)

For any $X \in \mathcal{F}_{\infty}^{m,n}$, $0 < c_* \leq 0.2$, and $0 < k \leq n$,

$$z_{c_*,k}(X) \geq c_2 \sqrt{\frac{k}{\log m}} \cdot \frac{1}{d_{n-k}(X^\circ)},$$

where c_2 is a universal constant.

Combining the lemmas,

$$d_k(X) \leq c_1 \sqrt{\frac{k}{\epsilon}} \cdot \frac{1}{d_{n-(1-\epsilon)k}(X^\circ)} \leq c_1 c_2 \sqrt{\frac{\log m}{\epsilon}} z_{c_*,k}(X).$$

Skinny objects always have a small shadow!

Further comments

What about $\beta_p^{m,n}$?

[Recall:

$$\beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \max_{\sigma > 0} \frac{R_T(X,\sigma)}{R(X,\sigma)},$$

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}.$$

]

Corollary

For $p \geq 2$, $\beta_p^{m,n} = O(\min(n^{1-2/p}, m^{2/p} \log m))$.

Conjecture: For any $p \geq 2$, there exists a constant $C = C(p)$, such that $\beta_p^{m,n} \leq C \log m$.

What about $\beta_p^{m,n}$?

[Recall:

$$\beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \max_{\sigma > 0} \frac{R_T(X,\sigma)}{R(X,\sigma)},$$

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}.$$

]

Corollary

For $p \geq 2$, $\beta_p^{m,n} = O(\min(n^{1-2/p}, m^{2/p} \log m))$.

Conjecture: For any $p \geq 2$, there exists a constant $C = C(p)$, such that $\beta_p^{m,n} \leq C \log m$.

What about $\beta_p^{m,n}$?

[Recall:

$$\beta_p^{m,n} = \max_{X \in \mathcal{F}_p^{m,n}} \max_{\sigma > 0} \frac{R_T(X,\sigma)}{R(X,\sigma)},$$

$$\mathcal{F}_p^{m,n} = \{X : X = \{x : \|Ax\|_p \leq 1\}, \text{ for } A \in \mathbb{R}^{m \times n}\}.$$

]

Corollary

For $p \geq 2$, $\beta_p^{m,n} = O(\min(n^{1-2/p}, m^{2/p} \log m))$.

Conjecture: For any $p \geq 2$, there exists a constant $C = C(p)$, such that $\beta_p^{m,n} \leq C \log m$.

Design problem

How to design optimal truncated series estimators for symmetric polytopes?

- NP-hard in general $\rightarrow$ SDP formulations to solve a relaxation

Thanks!

Design problem

How to design optimal truncated series estimators for symmetric polytopes?

- NP-hard in general $\rightarrow$ SDP formulations to solve a relaxation

Thanks!