Optimal Estimation of a Nonsmooth Functional

T. Tony Cai
Department of Statistics
The Wharton School
University of Pennsylvania
http://stat.wharton.upenn.edu/~tcai

Joint work with Mark Low

Question

Suppose we observe $\boldsymbol{X \sim N(\mu,\ 1)}$. What is the best way to estimate $|\boldsymbol{\mu}|$?

Question

Suppose we observe $\boldsymbol{X_i \overset{ind.}{\sim} N(\theta_i,\ 1),\ i = 1, ..., n}$.

How to optimally estimate

$$\mathbf{T}(\theta) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} |\theta_{\mathbf{i}}| \ ?$$

Outline

- **Introduction & Motivation**
- **Approximation Theory**
- **Optimal Estimator & Minimax Upper Bound**
- **Testing Fuzzy Hypotheses & Minimax Lower Bound**
- **Discussions**

Introduction & Motivation

Introduction

Estimation of functionals occupies an important position in the theory of nonparametric function estimation.

- **Gaussian Sequence Model:**

$$y_i = \theta_i + \sigma z_i, \quad z_i \overset{iid}{\sim} N(0,1), \quad i = 1, 2, \ldots$$

- **Nonparametric regression:**

$$y_i = f(t_i) + \sigma z_i, \quad z_i \overset{iid}{\sim} N(0,1), \quad i = 1, \cdots, n.$$

- **Density Estimation:**

$$X_1,\ X_2,\ \cdots, X_n \overset{i.i.d.}{\sim} f.$$

Estimate: $\boldsymbol{L(\theta) = \sum c_i \theta_i}$, $\boldsymbol{L(f) = f(t_0)}$, $\boldsymbol{Q(\theta) = \sum c_i \theta_i^2}$, $\boldsymbol{Q(f) = \int f^2}$, etc.

Linear Functionals

- **Minimax estimation over convex parameter spaces:** Ibragimov and Hasminskii (1984), Donoho and Liu (1991) and Donoho (1994). The minimax rate of convergence is determined by a **modulus of continuity**.

- **Minimax estimation over nonconvex parameter spaces:** C. & L. (2004).

- **Adaptive estimation over convex parameter spaces:** C. & L. (2005). The key quantity is a between-class modulus of continuity,

$$\omega(\epsilon, \Theta_1, \Theta_2) = \sup\{|L(\theta_1) - L(\theta_2)| : \|\theta_1 - \theta_2\|_2 \leq \epsilon, \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}.$$

Confidence intervals, adaptive confidence intervals/bands, ...

$$\Downarrow$$

Estimation of linear functionals is now well understood.

Quadratic Functionals

- **Minimax estimation over orthosymmetric quadratically convex parameter spaces:** Bickel and Ritov (1988), Donoho and Nussbaum (1990), Fan (1991), and Donoho (1994). **Elbow phenomenon**.
- **Minimax estimation over parameter spaces which are not quadratically convex:** C. & L. (2005).
- **Adaptive estimation over L_p and Besov spaces:** C. & L. (2006).

Estimating quadratic functionals is closely related to **signal detection (nonparametric hypothesis testing):**

$$H_0: \ f = f_0 \quad vs. \quad H_1: \|f - f_0\|_2 \geq \epsilon,$$

risk/loss estimation, adaptive confidence balls, ...

$$\Downarrow$$

Estimation of quadratic functionals is also well understood.

Smooth Functionals

Linear and quadratic functionals are the most important examples in the class of smooth functionals.

In these problems, minimax lower bounds can be obtained by testing hypotheses which have relatively simple structures. (More later.)

Construction of rate-optimal estimators is also relatively well understood.

Nonsmooth Functionals

Recently some non-smooth functionals have been considered. A particularly interesting paper is Lepski, Nemirovski and Spokoiny (1999) which studied the problem of estimating the L_r norm:

$$T(f) = (\int |f(x)|^r dx)^{1/r}$$

- **The behavior of the problem depends strongly on whether or not r is an even integer.**
- For the lower bounds, one needs to consider testing between two composite hypotheses where **the sets of values of the functional on these two hypotheses are interwoven.** These are called **fuzzy hypotheses** in the language of Tsybakov (2009).

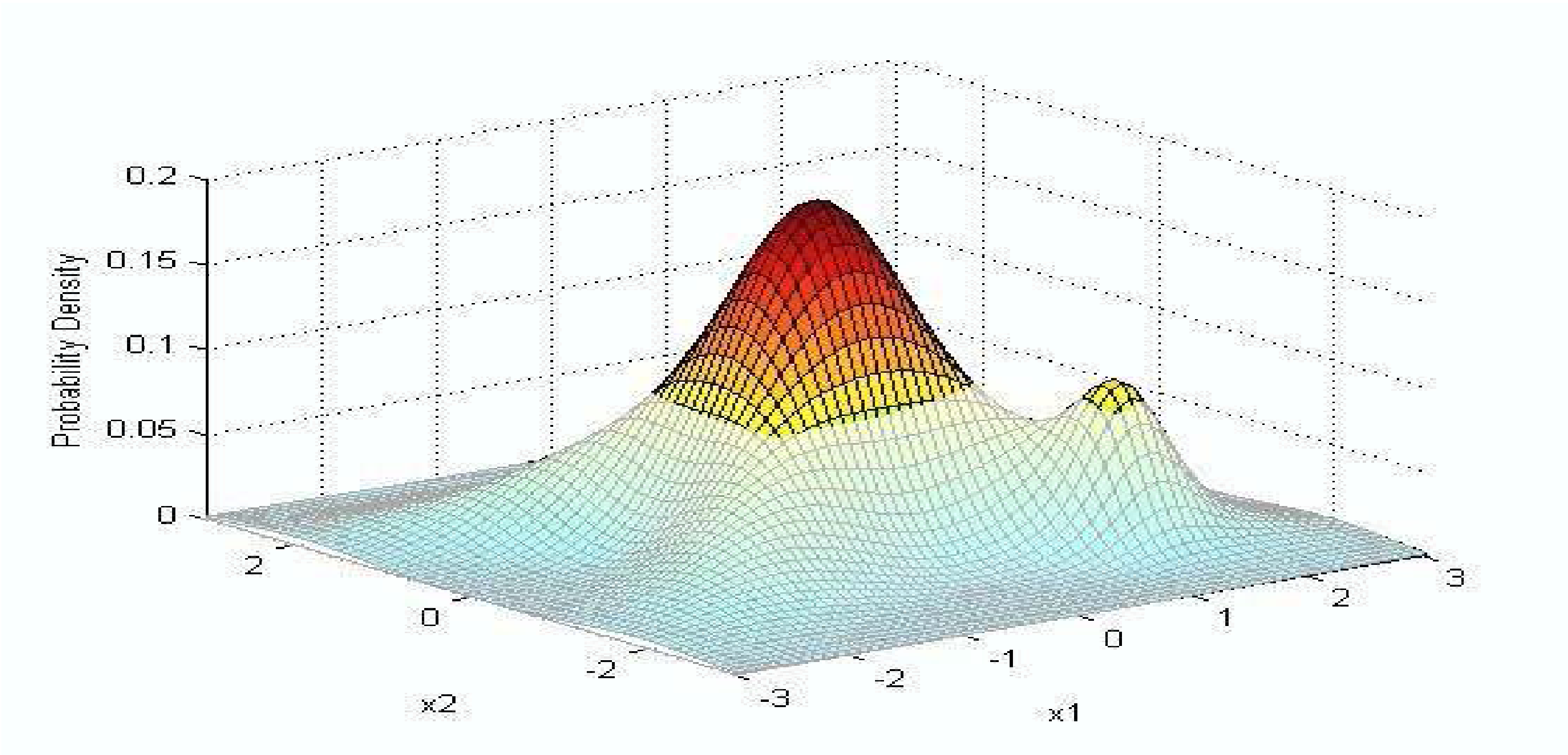
Nonsmooth Functionals

- Rényi entropy:

$$T(f) = \frac{1}{1-\alpha} \log \int f^{\alpha}(t)dt.$$

- Excess mass:

$$T(f) = \int (f(t) - \lambda)_+ dt.$$



Excess Mass

Estimating the excess mass is closely related to a wide range of applications:

- **testing multimodality** (dip test, Hartigan and Hartigan (1985), Cheng and Hall (1999), Fisher and Marron (2001))
- **estimating density level sets** (Polonik (1995), Mammen and Tsybakov (1995), Tsybakov (1997), Gayraud and Rousseau (2005), ...)
- **estimating regression contour clusters** (Polonik and Wang (2005))

Estimating the L_1 Norm

Note that $(x)_+ = \frac{1}{2}(|x| + x)$, so

$$T(f) = \int (f(t) - \lambda)_+ dt = \frac{1}{2}\int |f(t) - \lambda| dt + \frac{1}{2}\int f(t)dt - \frac{1}{2}\lambda.$$

Hence estimating the excess mass is equivalent to estimating the L_1 norm.

A key step in understanding the functional problem is the understanding of a seemingly simpler normal means problem: estimating

$$\mathbf{T}(\theta) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i}=\mathbf{1}}^{\mathbf{n}} |\theta_{\mathbf{i}}|$$

based on the sample $\boldsymbol{Y_i \overset{ind.}{\sim} N(\theta_i,\ 1),\ i = 1, \ldots, n}$.

This nonsmooth functional estimation problem exhibits some features that are significantly different from those in estimating smooth functionals.

Minimax Risk

Define

$$\Theta_n(M) = \{\theta \in \mathbb{R}^n : |\theta_i| \leq M\}.$$

Theorem 1 *The minimax risk for estimating* $T(\theta) = \frac{1}{n}\sum_{i=1}^n |\theta_i|$ *over* $\Theta_n(M)$ *satisfies*

$$\inf_{\hat{T}} \sup_{\theta \in \Theta_n(M)} E(\hat{T} - T(\theta))^2 = \beta_*^2 M^2 \left(\frac{\log\log n}{\log n}\right)^2 (1 + o(1)) \qquad (1)$$

where $\beta_* \approx 0.28017$ *is the Bernstein constant.*

The minimax risk converges to zero at a slow logarithmic rate which shows that the nonsmooth functional $T(\theta)$ is difficult to estimate.

Comparisons

In contrast the rates for estimating linear and quadratic functionals are most often algebraic. Let

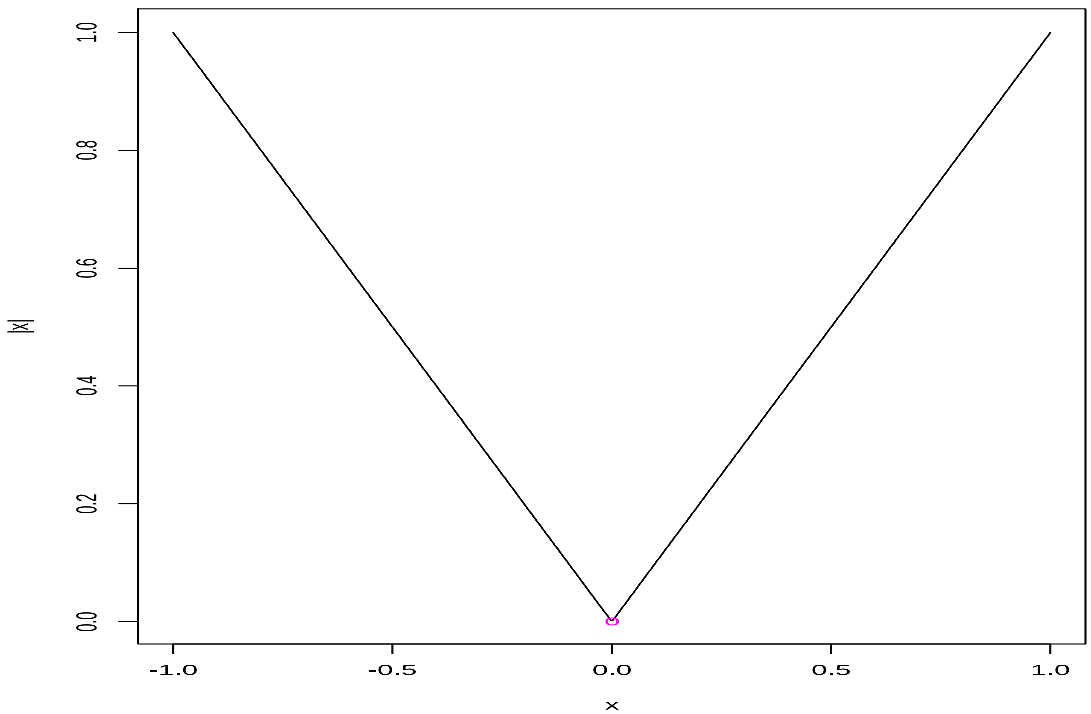
$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\theta_i \quad \text{and} \quad Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}\theta_i^2.$$

- It is easy to check that the usual parametric rate $\boldsymbol{n^{-1}}$ for estimating $L(\theta)$ can be easily attained by $\boldsymbol{\bar{y}}$.
- For estimating $Q(\theta)$, the parametric rate $\boldsymbol{n^{-1}}$ can be achieved over $\Theta_n(M)$ by using the unbiased estimator $\boldsymbol{\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}(y_i^2 - 1)}$.

Why Is the Problem Hard?

The fundamental difficulty of estimating $T(\theta)$ can be traced back to the **nondifferentiability of the absolute value function at the origin.**

This is reflected both in the construction of the **optimal estimators** and the derivation of the **lower bounds.**



Basic Strategy

The construction of the optimal estimator is involved. This is partly due to the nonexistence of an unbiased estimator for $|\theta_i|$.

Our strategy:

1. "smooth" the singularity at 0 by the **best polynomial approximation**;
2. construct an unbiased estimator for each term in the expansion by using the **Hermite polynomials.**

Approximation Theory

Optimal Polynomial Approximation

Optimal polynomial approximation has been well studied in approximation theory. See Bernstein (1913), Varga and Carpenter (1987), and Rivlin (1990).

Let $\mathcal{P}_m$ denote the class of all real polynomials of degree at most m. For any continuous function f on $[-1, 1]$, let

$$\delta_m(f) = \inf_{G \in \mathcal{P}_m} \max_{x \in [-1,1]} |f(x) - G(x)|.$$

A polynomial G^* is said to be a best polynomial approximation of f if

$$\delta_m(f) = \max_{x \in [-1,1]} |f(x) - G^*(x)|.$$

Chebyshev Alternation Theorem (1854)

A polynomial $G^* \in \mathcal{P}_m$ is the (unique) best polynomial approximation to a continuous function f if and only if the difference $\boldsymbol{f(x) - G^*(x)}$ **takes consecutively its maximal value with alternating signs at least** $m+2$ **times.** That is, there exist $m+2$ points $-1 \le x_0 < \cdots < x_{m+1} \le 1$ such that

$$[f(x_j) - G^*(x_j)] = \pm(-1)^j \max_{x\in[-1,1]} |f(x) - G^*(x)|, \quad j = 0, \ldots, m+1.$$

(More on the set of alternation points later.)

Absolute Value Function & Bernstein Constant

Because $|x|$ is an even function, so is its best polynomial approximation.

For any positive integer K, denote by G_K^* the best polynomial approximation of degree $2K$ to $|x|$ and write

$$G_K^*(x) = \sum_{k=0}^{K} g_{2k}^* x^{2k}. \tag{2}$$

For the absolute value function $f(x) = |x|$, Bernstein (1913) proved that

$$\lim_{K\to\infty} 2K\delta_{2K}(f) = \beta_*$$

where β_* is now known as the **Bernstein constant**. Bernstein (1913) showed

$$0.278 < \beta_* < 0.286.$$

Bernstein Conjecture

Note that the average of the two bounds equals 0.282. Bernstein (1913) noted as a **"curious coincidence"** that the constant

$$\frac{1}{2\sqrt{\pi}} = 0.2820947917\cdots$$

and made a conjecture known as the **Bernstein Conjecture:**

$$\beta_* = \frac{1}{2\sqrt{\pi}}.$$

It remained as an open conjecture for 74 years!

In 1987, Varga and Karpenter proved that the Bernstein Conjecture was in fact wrong. They computed β_* to the 95th decimal places,

$$\beta_* = 0.28016\ 94990\ 23869\ 13303\ 64364\ 91230\ 67200\ 00424\ 82139\ 81236\ \cdots$$

Alternative Approximation

The best polynomial approximation G_K^* is not convenient to construct. An explicit and nearly optimal polynomial approximation G_K can be easily obtained by using the Chebyshev polynomials.

The Chebyshev polynomial of degree k is defined by $\boldsymbol{\cos(k\theta) = T_k(\cos\theta)}$ or

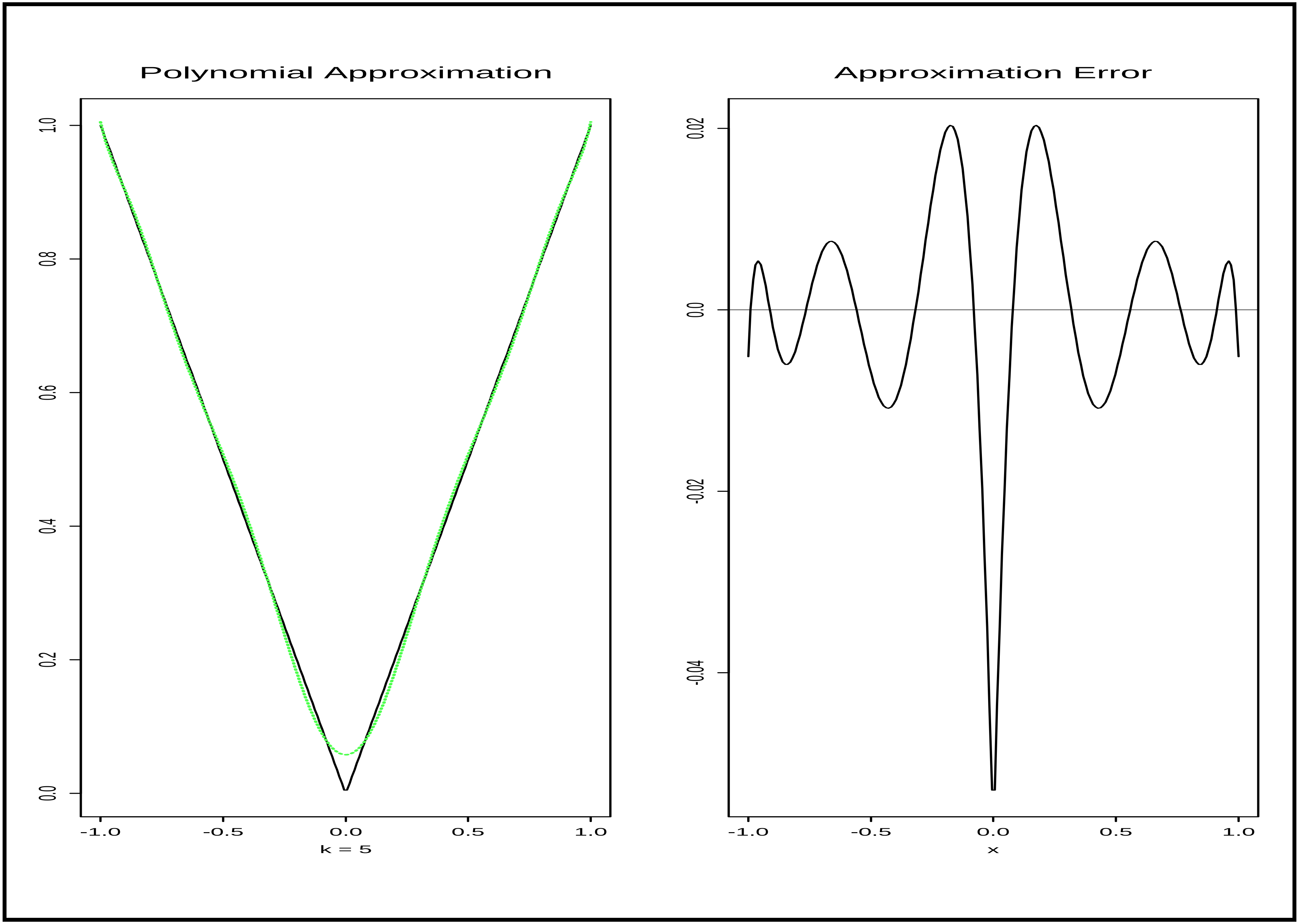
$$T_k(x) = \sum_{j=0}^{[k/2]} (-1)^j \frac{k}{k-j} \binom{k-j}{j} 2^{k-2j-1} x^{k-2j}.$$

Let

$$G_K(x) = \frac{2}{\pi} T_0(x) + \frac{4}{\pi} \sum_{k=1}^{K} (-1)^{k+1} \frac{T_{2k}(x)}{4k^2 - 1}. \tag{3}$$

We can also write $G_K(x)$ as

$$G_K(x) = \sum_{k=0}^{K} g_{2k} x^{2k}. \tag{4}$$


Polynomial Approximation
1.0
0.8
0.6
0.4
0.2
0.0
-1.0
-0.5
0.0
0.5
1.0
k = 5
Approximation Error
0.02
0.0
-0.02
-0.04
-1.0
-0.5
0.0
0.5
1.0
x

Approximation Error

Lemma 1 *Let* $G_K^*(x) = \sum_{k=0}^{K} g_{2k}^* x^{2k}$ *be the best polynomial approximation of degree* $2K$ *to* $|x|$ *and let* G_K *be defined in (3). Then*

$$\max_{x \in [-1,1]} |G_K^*(x) - |x|| \leq \frac{\beta_*}{2K}(1 + o(1)) \tag{5}$$

$$\max_{x \in [-1,1]} |G_K(x) - |x|| \leq \frac{2}{\pi(2K+1)}. \tag{6}$$

The coefficients g_{2k}^* *and* g_{2k} *satisfy for all* $0 \leq k \leq K$,

$$|g_{2k}^*| \leq 2^{3K} \quad \textit{and} \quad |g_{2k}| \leq 2^{3K}. \tag{7}$$

Construction of the Optimal Procedure

Construction of the Optimal Estimator

We shall focus on the special case of $M = 1$. The case of a general M involves an additional rescaling step.

When $M = 1$, it follows from Lemma 1 that each $|\theta_i|$ can be well approximated by $G_K^*(\theta_i) = \sum_{k=0}^{K} g_{2k}^* \theta_i^{2k}$ on the interval $[-1, 1]$ and hence the functional $T(\theta) = \frac{1}{n} \sum_{i=1}^{n} |\theta_i|$ can be approximated by

$$\tilde{T}(\theta) = \frac{1}{n} \sum_{i=1}^{n} G_K^*(\theta_i) = \sum_{k=0}^{K} g_{2k}^* b_{2k}(\theta)$$

where $b_{2k}(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \theta_i^{2k}$.

Note that $\tilde{T}(\theta)$ is a smooth functional and we shall estimate $b_{2k}(\theta)$ separately for each k by using the Hermite polynomials.

Hermite Polynomials

Let ϕ be the density function of a standard normal variable. For positive integers k, Hermite polynomial H_k is defined by

$$\frac{d^k}{dy^k}\phi(y) = (-1)^k H_k(y)\phi(y). \tag{8}$$

The following result is well known.

Lemma 2 *Let $X \sim N(\mu, 1)$. $H_k(X)$ is an unbiased estimate of μ^k for any positive integer k, i.e.,*

$$E_\mu H_k(X) = \mu^k.$$

Also,

$$\int H_k^2(y)\phi(y)dy = k! \quad \textit{and} \quad \int H_k(y)H_j(y)\phi(y)dy = 0 \tag{9}$$

when $k \neq j$.

Optimal Estimator

Since $H_k(y_i)$ is an unbiased estimate of θ_i^k for each i, we can estimate $b_k(\theta) \equiv \frac{1}{n}\sum_{i=1}^n \theta_i^k$ by $\bar{B}_k = \frac{1}{n}\sum_{i=1}^n H_k(y_i)$ and define the estimator of $T(\theta)$ by

$$\widehat{T_K(\theta)} = \sum_{k=0}^{K} g_{2k}^* \bar{B}_{2k}. \tag{10}$$

The performance of the estimator $\widehat{T_K(\theta)}$ clearly depends on the choice of the cutoff K. We shall specifically choose

$$K = K_* \equiv \frac{\log n}{2\log\log n} \tag{11}$$

and define the final estimator of $T(\theta)$ by

$$\widehat{T_*(\theta)} \equiv \widehat{T_{K_*}(\theta)} = \sum_{k=0}^{K_*} g_{2k}^* \bar{B}_{2k}. \tag{12}$$

Optimality of the Estimator

Theorem 2 *Let* $y_i \sim N(\theta_i, 1)$ *be independent normal random variables with* $|\theta_i| \le M$, $i = 1, ...n$. *Let* $T(\theta) = n^{-1} \sum_{i=1}^n |\theta_i|$. *The estimator* $\widehat{T_*(\theta)}$ *given in (12) satisfies*

$$\sup_{\theta \in \Theta_n(M)} E(\widehat{T_*(\theta)} - T(\theta))^2 \le \beta_*^2 M^2 \left(\frac{\log \log n}{\log n} \right)^2 (1 + o(1)). \tag{13}$$

Remark: If $G_K(x)$, instead of $G_K^*(x)$, is used in the construction of the estimator $\widehat{T_*(\theta)}$, the resulting estimator $\widehat{T(\theta)}$ satisfies

$$\sup_{\theta \in \Theta_n(M)} E(\widehat{T(\theta)} - T(\theta))^2 \le 4\pi^{-2} M^2 \left(\frac{\log \log n}{\log n} \right)^2 (1 + o(1)). \tag{14}$$

The ratio of this upper bound to the minimax risk is $4\pi^{-2}/\beta_*^2 \approx 5.16$.

Minimax Lower Bound via Testing Fuzzy Hypotheses

Minimax Lower Bound

The upper bound $\beta_*^2 M^2 \left(\frac{\log\log n}{\log n}\right)^2$ is in fact **asymptotically sharp.**

The standard lower bound arguments fail to yield the desired rate of convergence. New technical tools are needed.

Standard Lower Bound Argument

Deriving minimax lower bounds is a key step in developing a minimax theory.

- **Testing a pair of simple hypotheses** $H_0 : \theta = \theta_0 \ vs. \ H_1 : \theta = \theta_1$. For estimation of linear functionals, it is often sufficient to derive the optimal rate of convergence based on testing a pair of simple hypotheses. Le Cam's method is a well known approach based on this idea. See, for example, Le Cam (1973), and Donoho and Liu (1991).

- **Testing a composite hypothesis against a simple null** $H_0 : \theta = \theta_0 \ vs. \ H_1 : \theta \in \Theta_1$. For estimation of quadratic functionals, rate optimal lower bounds can often be provided by testing a simple null versus a composite alternative where the value of the functional is constant on the composite alternative. See, e.g., C. & L. (2005).

- **Other techniques:** Assouad's Lemma, Fano's Lemma, ...

General Lower Bound Argument

- Observe $X \sim P_\theta$ where $\theta \in \Theta = \Theta_0 \cup \Theta_1$, and wish to estimate a function $T(\theta)$ based on X.
- Let μ_0 and μ_1 be two priors supported on Θ_0 and Θ_1 respectively. Let

$$m_i = \int T(\theta)\mu_i(d\theta) \quad \text{and} \quad v_i^2 = \int (T(\theta) - m_i)^2 \mu_i(d\theta).$$

- Write f_i for the marginal density of X when the prior is μ_i and define the chi-square distance between f_0 and f_1 by

$$I = \left\{ E_{f_0} \left(\frac{f_1(X)}{f_0(X)} - 1 \right)^2 \right\}^{\frac{1}{2}}$$

Remark: The chi-square distance I can be hard to compute when f_0 is a mixture distribution.

General Minimax Lower Bound

Theorem 3 *If* $|m_1 - m_0| > v_0 I$, *then*

$$\sup_{\theta \in \Theta} E(\hat{T}(X) - T(\theta))^2 \geq \frac{(|m_1 - m_0| - v_0 I)^2}{(I+2)^2}. \tag{15}$$

Remark: This general minimax lower bound is obtained through testing the hypotheses:

$$H_0 : \theta \sim \mu_0 \quad vs. \quad H_1 : \theta \sim \mu_1.$$

More general results on the bias and the Bayes risks can be derived.

Minimax Lower Bound

Theorem 4 *Let* $y_i \overset{ind.}{\sim} N(\theta_i, 1)$*,* $i = 1, ..., n$*, and let* $T(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\theta_i|$*. Then, the minimax risk for estimating* $T(\theta)$ *over* $\Theta_n(M)$ *satisfies*

$$\inf_{\hat{T}} \sup_{\theta \in \Theta_n(M)} E(\hat{T} - T(\theta))^2 \geq \beta_*^2 M^2 \left(\frac{\log \log n}{\log n}\right)^2 (1 + o(1)) \qquad (16)$$

where β_* *is the Bernstein constant.*

Three major components in the derivation of the lower bounds:

- **The general lower bound argument;**
- **A careful construction of least favorable priors μ_0 and μ_1;**
- **Bounding the chi-square distance between the marginal distributions. (Moment matching & Hermite polynomials)**

Alternation Points & Least Favorable Priors

The best polynomial approximation $G_K^*(x)$ has at least $2K+2$ alternation points. The set of these alternation points is important in the construction of the fuzzy hypotheses.

Divide the set of the alternation points of $G_K^*(x)$ into two subsets and denote

$$
\begin{aligned}
A_0 &= \{x \in [-1,1] : |x| - G_K^*(x) = -\delta_{2K}(|x|)\}, \\
A_1 &= \{x \in [-1,1] : |x| - G_K^*(x) = \delta_{2K}(|x|)\}.
\end{aligned}
$$

The priors μ_0 and μ_1 used in the construction of the fuzzy hypotheses in the proof of Theorem 4 are supported on A_0 and A_1 respectively.

Intuitively, **this makes the priors μ_0 and μ_1 maximally apart and yet not "testable".**

It also connects the construction of the optimal estimator with the minimax lower bound.

Other Parameter Spaces

Theorem 5 *Let* $Y \sim N(\theta, I_n)$ *and let* $T(\theta) = \frac{1}{n}\sum_{i=1}^{n} |\theta_i|$. *The minimax risk for estimating the functional* $T(\theta)$ *over* $\mathbb{R}^n$ *satisfies*

$$\inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^n} E(\hat{T} - T(\theta))^2 \asymp \frac{1}{\log n}. \tag{17}$$

The lower bound can be derived in a similar way, the construction of the optimal estimator is much more involved.

The Sparse Case

Suppose we observe $y_i \overset{ind}{\sim} N(\theta_i,\ 1)$, $i = 1, 2, ..., n$ where the mean vector θ is sparse : only a small fraction of components are nonzero, and the locations of the nonzero components are unknown.

Denote the ℓ_0 quasi-norm by $\|\theta\|_0 = \text{Card}(\{i :\ \theta_i \neq 0\})$. Fix k_n, the collection of vectors with exactly k_n nonzero entries is

$$\Theta_{k_n} = \ell_0(k_n) = \{\theta \in \mathbb{R}^n :\ \|\theta\|_0 = k_n\}.$$

Suppose we wish to estimate the average of the absolute value of the nonzero means,

$$T(\theta) = \text{average}\{|\theta_i| :\ \theta_i \neq 0\} = \frac{1}{\|\theta\|_0} \sum_{i=1}^{n} |\theta_i|. \tag{18}$$

The Sparse Case

We calibrate the sparsity parameter k_n by $k_n = n^\beta$ for $0 < \beta \leq 1$.

When $0 < \beta \leq \frac{1}{2}$, it is not possible to estimate the functional $T(\theta)$ consistently.

Theorem 6 *Let $k_n = n^\beta$. Then for all $0 < \beta \leq \frac{1}{2}$, the minimax risk satisfies*

$$\inf_{\widehat{T(\theta)}} \sup_{\theta \in \Theta_{k_n}} E(\widehat{T(\theta)} - T(\theta))^2 \geq C \tag{19}$$

for some constant $C > 0$.

The Sparse Case

Theorem 7 *Let $k_n = n^\beta$ for some $\frac{1}{2} < \beta < 1$. Then the minimax risk for estimating the functional $T(\theta)$ over Θ_{k_n} satisfies*

$$\inf_{\widehat{T(\theta)}} \sup_{\theta \in \Theta_{k_n}} E(\widehat{T(\theta)} - T(\theta))^2 \asymp \frac{C}{\log n}. \tag{20}$$

Discussions

Discussions

Lepski, Nemirovski and Spokoiny (1999) used a **Fourier series approximation** of $|x|$ and the estimate is based on unbiased estimates of individual terms in the approximation.

- The maximum error of the best K-term Fourier series approximation is of order K^{-1}.
- The variance bound of the estimator based on the K-term Fourier series approximation is of order $\boldsymbol{e^{CK^2}}$, whereas the variance of our estimator based on the polynomial approximation of degree K grows at the rate of $\boldsymbol{K^K = e^{K \log K}}$.
- So the variance of the polynomial-based estimator is much smaller than that of the corresponding estimator using Fourier series.
- This allows for more terms to be used in the polynomial approximation thus reducing the bias of the estimate.

Discussions

- In the bounded case, the best rate of convergence for estimators using Fourier series approximation can be shown to be $(\log n)^{-1}$, which is sub-optimal relative to the minimax rate $(\frac{\log \log n}{\log n})^2$.
- Another drawback of the Fourier series method is that it cannot be used for the unbounded case.

Concluding Remarks

- Nonsmooth functional estimation problems exhibit some features that are significantly different from those in estimating smooth functionals.
- We showed that the asymptotic risk for estimating $T(\theta) = \frac{1}{n}\sum|\theta_i|$ is

$$\beta_*^2 M^2 \left(\frac{\log\log n}{\log n}\right)^2.$$

- The general techniques and results developed here can be used to solve other related problems.
  - When the approach taken in this paper is used for estimating the L_1 norm of a regression function, both the upper and lower bounds given in Lepski, Nemirovski and Spokoiny (1999) are improved.
  - The techniques can also be used for estimating other nonsmooth functionals such as excess mass. See C. & L. (2011).

Paper

Cai, T., & Low, M. (2011). **Testing composite hypotheses, Hermite polynomials, and optimal estimation of a nonsmooth functional.** *The Annals of Statistics*, to appear.

Available at: http://stat.wharton.upenn.edu/∼tcai