

# TESTING COMPOSITE HYPOTHESES, HERMITE POLYNOMIALS, AND OPTIMAL ESTIMATION OF A NONSMOOTH FUNCTIONAL

BY T. TONY CAI \* AND MARK G. LOW

*University of Pennsylvania*

A general lower bound is developed for the minimax risk when estimating an arbitrary functional. The bound is based on testing two composite hypotheses and is shown to be effective in estimating the non-smooth functional  $\frac{1}{n} \sum |\theta_i|$  from an observation  $Y \sim N(\theta, I_n)$ . This problem exhibits some features that are significantly different from those that occur in estimating conventional smooth functionals. This is a setting where standard techniques fail to yield sharp results.

A sharp minimax lower bound is established by applying the general lower bound technique based on testing two composite hypotheses. A key step is the construction of two special priors and bounding the chi-square distance between two normal mixtures. An estimator is constructed using approximation theory and Hermite polynomials and is shown to be asymptotically sharp minimax when the means are bounded by a given value  $M$ . It is shown that the minimax risk equals  $\beta_*^2 M^2 \left(\frac{\log \log n}{\log n}\right)^2$  asymptotically, where  $\beta_*$  is the Bernstein constant.

The general techniques and results developed in the present paper can also be used to solve other related problems.

**1. Introduction.** Minimax risk is one of the most commonly used benchmarks for evaluating the performance of any estimation method. For this reason considerable effort has been made developing minimax theories in the nonparametric function estimation literature. A key step in all these developments is the derivation of minimax lower bounds. Several effective lower bound techniques based on testing have been introduced in the literature and 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).

For estimation of quadratic functionals the story is somewhat more complicated. If the parameter space is not too "large" regular parametric rate

---

\*The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.

*Keywords and phrases:* Primary 62G07; secondary 62620

*Keywords and phrases:* Best polynomial approximation,  $\ell_1$  norm, composite hypotheses, Hermite polynomial, minimax lower bound, nonsmooth functional, optimal rate of convergence.

of convergence can be attained. However Bickel and Ritov (1988) showed that when the parameter space is too large the essential difficulty of such problems cannot be captured by testing a simple null versus a simple alternative. Instead 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, for example, Cai and Low (2005), where upper and lower bounds are constructed for quadratic functionals over many different parameter spaces.

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 of the drift function under the white noise model. One of the key observations in that paper was the need to consider testing between two composite hypotheses where the  $L_r$  norm is not constant on either of these composite hypotheses and 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).

The purpose of the present paper is to advance these ideas further. We first develop a new general minimax lower bound technique for estimating any functional  $T$  based on testing two composite hypotheses. For any two priors, say  $\mu_0$  and  $\mu_1$ , on the parameter space we obtain a lower bound on the expected squared bias with respect to  $\mu_1$  under a constraint on the upper bound of the expected mean squared error with respect to  $\mu_0$ . The lower bound depends on the difference between the expected value of  $T$  over each of the priors and also on the variance of  $T$  under  $\mu_0$ . The bound also depends on the chi-square distance between the two marginal distributions of the observations, one over  $\mu_0$ , the other over  $\mu_1$ . Some of the technical tools for deriving minimax lower bounds developed earlier in the literature can be seen as special cases of the general result given in the present paper.

We then consider specifically the problem of estimating the  $\ell_1$  norm of a multivariate normal mean vector. This nonsmooth functional estimation problem exhibits some features that are significantly different from those in estimating smooth functionals in terms of the optimal rates of convergence as well as the technical tools needed for the analysis of both the minimax lower bounds and the construction of the optimal estimators.

Let  $y_1, y_2, \dots, y_n$  be independent normal random variables where  $y_i \sim N(\theta_i, 1)$ . The problem of focus in this paper is that of estimating

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

where we assume that either  $|\theta_i| \leq M$  for some constant  $M > 0$  or that

there are no constraints on the  $\theta_i$ . In the present paper we develop optimal estimators of  $T(\theta)$  along with minimax lower bounds. In particular for the bounded case we construct an asymptotically sharp minimax estimator using approximation theory and Hermite polynomials. By combining the minimax lower and upper bounds developed in later sections, the main results on the minimax estimation of the functional  $T(\theta)$  can be summarized in the following theorem.

**Theorem 1** *Let  $Y \sim N(\theta, I_n)$  and let  $T(\theta) = \frac{1}{n} \sum_{i=1}^n |\theta_i|$ . For a fixed constant  $M > 0$ , denote by  $\Theta_n(M) = \{\theta \in \mathbb{R}^n : |\theta_i| \leq M\}$ . Then the minimax risk for estimating the functional  $T(\theta)$  based on  $Y$  over  $\Theta_n(M)$  satisfies*

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

where  $\beta_* \approx 0.28017$  is the Bernstein constant, and the minimax risk for estimating the functional  $T(\theta)$  over  $\mathbb{R}^n$  satisfies

$$(3) \quad \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^n} E(\hat{T} - T(\theta))^2 \asymp \frac{1}{\log n}.$$

These rates are dramatically different from the usual parametric or algebraic rates of convergence for estimating smooth functionals. The fundamental difficulty of estimating the functional  $T(\theta)$  can be traced back to the nondifferentiability of the absolute value function at the origin. This is reflected both in the derivation of the lower bounds and the construction of the optimal estimators. Best polynomial approximation and Hermite polynomials play major roles in the derivation of the lower bounds as well as in the construction of the optimal estimators.

The minimax lower bounds are established by applying the general lower bound technique to two carefully constructed composite hypotheses. In the present context to obtain good lower bounds neither prior can be degenerate. A key step, is the construction of two mixture priors which have a large difference in the expected values of the functional while making the chi-square distance between the two mixture models small. In order to turn this heuristic idea into an effective tool it is necessary to be able to bound the chi-square distance between two normal mixture models. In previous applications such bounds have only been given in the much simpler case when one of the mixtures is degenerate. See, e.g., Cai and Low (2005) and Wang, Brown, Cai, and Levine (2008).

The construction of the optimal estimators of the nonsmooth functional  $T(\theta)$  is significantly more complicated than those for linear or quadratic

functionals. For optimal estimation of  $T(\theta)$  over the bounded set  $\Theta_n(M)$ , we first use the best polynomial approximation  $G_K^*(x) = \sum_{k=0}^K g_{2k}^* x^{2k}$  of the absolute value function  $|x|$ . Then for each  $i$  and each  $k$  we form an unbiased estimate of  $\theta_i^k$  using the Hermite polynomials. Putting these terms together for a given  $i$  yields an estimate of  $|\theta_i|$ . An effective estimate of the functional  $T$  can then be constructed by averaging these estimates of  $|\theta_i|$ . We show that by carefully selecting the cutoff  $K = K_n$  the resulting estimator is asymptotically sharp minimax. This estimator is, however, not optimal over the unbounded parameter space  $\mathbb{R}^n$ . An additional testing step is used to construct a hybrid estimator and it is shown that the estimator is rate optimal for estimating  $T(\theta)$  over  $\mathbb{R}^n$ . In addition, we also consider the estimation of  $T(\theta)$  over a parameter space where the mean  $\theta$  is a high-dimensional sparse vector with a small fraction of nonzero coordinates.

The rest of the paper is organized as follows. In Section 2 we derive the general lower bounds for estimating any functional  $T$  based on testing two composite hypotheses. In Section 3 we bound the chi-square distance between two normal mixture models and apply the general lower bound from Section 2 to derive minimax lower bounds for estimating the nonsmooth functional  $T(\theta)$  given in (1). Section 4 constructs an estimator of  $T(\theta)$  using best polynomial approximation and Hermite polynomials and shows that the estimator is sharp minimax for the bounded case. Section 5 considers the unbounded case. A hybrid estimator is constructed and is shown to attain the optimal rate of convergence. Section 6 treats the sparse case. Discussions on the connections and differences of our results with other related work is given in Section 7. Technical lemmas and some of the main results are proved in Section 8.

**2. General Lower Bound.** In this section, a constrained risk inequality is developed which immediately yields a general minimax lower bound based on testing two composite hypotheses.

Suppose we observe a random variable  $X$  which has a distribution  $P_\theta$  where  $\theta$  belongs to a given parameter space  $\Theta$ . Let  $\hat{T} = \hat{T}(X)$  be an estimator of a function  $T(\theta)$  based on  $X$  and denote the bias of  $\hat{T}$  by  $B(\theta) = E_\theta \hat{T} - T(\theta)$ . Let  $\Theta_0$  and  $\Theta_1$  be subsets of the parameter space  $\Theta$  where  $\Theta_0 \cup \Theta_1 = \Theta$ . Let  $\mu_0$  and  $\mu_1$  be two prior distributions supported on  $\Theta_0$  and  $\Theta_1$  respectively.

Let  $m_i$  and  $v_i^2$  be the means and variances of  $T(\theta)$  under the priors  $\mu_i$  for  $i = 0$  and 1. More specifically,

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

Write  $F_i$  for the marginal distribution of  $X$  when the prior is  $\mu_i$  for  $i = 0, 1$ . Let  $f_i$  be the density of  $F_i$  with respect to a common dominating measure of  $F_0$  and  $F_1$ . For any function  $g$  we shall write  $E_{f_0}g(X)$  for the expectation of  $g(X)$  with respect to the marginal distribution of  $X$  when the prior on  $\theta$  is  $\mu_0$ . We shall write  $E_{f_1}g(X)$  for the expectation of  $g(X)$  under  $F_1$ .

Finally define the chi-square distance between  $f_0$  and  $f_1$  by

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

The following theorem gives a lower bound for the average risk of an estimator  $\hat{T}$  under any mixture prior  $\lambda\mu_0 + (1 - \lambda)\mu_1$ ,  $0 \leq \lambda \leq 1$ .

**Theorem 2**

(i). Suppose  $\int E_{\theta}(\hat{T}(X) - T(\theta))^2 \mu_0(d\theta) \leq \epsilon^2$ , then

$$(4) \quad \left| \int B(\theta)\mu_1(d\theta) - \int B(\theta)\mu_0(d\theta) \right| \geq |m_1 - m_0| - (\epsilon + v_0)I.$$

(ii). If  $|m_1 - m_0| > v_0I$  and  $0 \leq \lambda \leq 1$  then

$$(5) \quad \int E_{\theta}(\hat{T}(X) - T(\theta))^2 (\lambda\mu_0(d\theta) + (1-\lambda)\mu_1(d\theta)) \geq \frac{\lambda(1-\lambda)(|m_1 - m_0| - v_0I)^2}{\lambda + (1-\lambda)(I+1)^2}$$

and in particular that

$$(6) \quad \max_{i=0,1} \int E_{\theta}(\hat{T}(X) - T(\theta))^2 \mu_i(d\theta) \geq \frac{(|m_1 - m_0| - v_0I)^2}{(I+2)^2}.$$

Informally, Theorem 2 says that if the average risk of  $\hat{T}$  under  $\mu_0$  is “small”, then the change in average bias under  $\mu_0$  and under  $\mu_1$  must be “large”. In particular, this implies that the average risk under a mixture prior is “large”.

Since the maximum risk is always at least as large as the average risk, Theorem 2 yields immediately a lower bound on the minimax risk.

**Corollary 1** If  $|m_1 - m_0| > v_0I$  then

$$(7) \quad \sup_{\theta \in \Theta} E_{\theta}(\hat{T}(X) - T(\theta))^2 \geq \frac{(|m_1 - m_0| - v_0I)^2}{(I+2)^2}.$$

Simpler versions of constrained risk inequalities have been developed before, most often for studying the cost of adaptation and superefficiency. For

example, a two-point risk inequality was given in Brown and Low (1996) and used to study adaptive estimation of linear functionals. The constrained risk inequality given in the present paper allows for a richer collection of applications and is especially useful when estimating nonsmooth functionals where it is essential to test complicated composite hypotheses in order to obtain good minimax lower bounds. In particular the lower bounds given in the next section rely on Corollary 1.

**Proof of Theorem 2:** We shall also assume without loss of generality that  $m_1 \geq m_0$ . Then

$$E_{f_0} \left\{ (\hat{T}(X) - m_0) \left( \frac{f_1(X) - f_0(X)}{f_0(X)} \right) \right\} = m_1 + \int B(\theta) \mu_1(d\theta) - (m_0 + \int B(\theta) \mu_0(d\theta)).$$

Now note that

$$\begin{aligned} E_{f_0} (\hat{T}(X) - m_0)^2 &= \int E_\theta (\hat{T}(X) - m_0)^2 \mu_0(d\theta) \\ &= \int E_\theta (\hat{T}(X) - T(\theta) + T(\theta) - m_0)^2 \mu_0(d\theta) \\ &= \int E_\theta (\hat{T}(X) - T(\theta))^2 \mu_0(d\theta) + \int (T(\theta) - m_0)^2 \mu_0(d\theta) \\ &\quad + 2 \int B(\theta) (T(\theta) - m_0) \mu_0(d\theta) \\ &\leq \epsilon^2 + v_0^2 + 2v_0\epsilon = (\epsilon + v_0)^2. \end{aligned}$$

Cauchy-Schwartz Inequality now yields

$$E_{f_0} \left\{ (\hat{T}(X) - m_0) \left( \frac{f_1(X) - f_0(X)}{f_0(X)} \right) \right\} \leq \left( E_{f_0} (\hat{T}(X) - m_0)^2 \right)^{\frac{1}{2}} \cdot I \leq (\epsilon + v_0) I.$$

Hence,

$$(8) \quad m_1 + \int B(\theta) \mu_1(d\theta) - (m_0 + \int B(\theta) \mu_0(d\theta)) \leq (\epsilon + v_0) I$$

and it follows that

$$\int B(\theta) \mu_1(d\theta) - \int B(\theta) \mu_0(d\theta) \leq m_0 - m_1 + (\epsilon + v_0) I$$

which in turn yields (4).

Now consider the quadratic

$$(9) \quad J(x) = \lambda x^2 + (1 - \lambda)(a - bx)^2$$

where we assume that  $0 < \lambda < 1$ ,  $a > 0$  and  $b > 0$ . It is easy to check that  $J$  is minimized when  $x = x_{\min} = \frac{ab(1-\lambda)}{\lambda+b^2(1-\lambda)}$  and that at this value  $a - bx > 0$  and  $J(x_{\min}) = \frac{a^2\lambda(1-\lambda)}{\lambda+b^2(1-\lambda)}$ . It follows that

$$(10) \quad \lambda x^2 + (1 - \lambda)(\max(a - bx, 0))^2$$

is also minimized at this same value. Now we also have

$$\int B^2(\theta)\mu_1(d\theta) \geq (\max(m_1 - m_0 - v_0I - (I + 1)\epsilon, 0))^2.$$

It follows that for  $0 \leq \lambda \leq 1$

$$\begin{aligned} \lambda\epsilon^2 + (1 - \lambda) \int B^2(\theta)\mu_1(d\theta) &\geq \lambda\epsilon^2 + (1 - \lambda)(\max(m_1 - m_0 - v_0I - (I + 1)\epsilon, 0))^2 \\ &\geq \frac{\lambda(1 - \lambda)(|m_1 - m_0| - v_0I)^2}{\lambda + (1 - \lambda)(I + 1)^2} \end{aligned}$$

which gives (5). The final inequality (6) follows by setting  $\lambda = \frac{I+1}{I+2}$  since the minimax risk is greater than any Bayes risk. ■

### 3. Lower Bound for Estimating the $\ell_1$ Norm of Normal Means.

We now turn to the problem of optimally estimating a particular nonsmooth functional where the use of the lower bound developed in the previous section yields sharp results. Let  $y_i \stackrel{ind}{\sim} N(\theta_i, 1)$ ,  $i = 1, 2, \dots, n$ , and consider the functional  $T$  where

$$(11) \quad T(\theta) = \frac{1}{n} \sum_{i=1}^n |\theta_i|.$$

As mentioned in the introduction, there are two particularly interesting cases. One is the bounded case with  $\theta \in \Theta_n(M)$  where  $\Theta_n(M) = \{\theta \in \mathbb{R}^n : |\theta_i| \leq M\}$  with a constant  $M > 0$ . Another case is the unbounded case where  $\theta \in \mathbb{R}^n$ . It is worth noting that we need to consider the bounded case with a bound growing in  $n$  in order to solve the unbounded case. In addition, we are also interested in the sparse case where  $\theta$  is a high dimensional sparse vector with a small fraction of nonzero coordinates.

In this section the focus is on developing minimax lower bounds. The minimax upper bounds and the optimal estimation procedures will be given in the next three sections. Best polynomial approximation plays a major role in the development of the lower bound and as we shall see later also in the development of the upper bound.

3.1. *Best Polynomial Approximation of the Absolute Value Function.* Optimal polynomial approximation of the absolute value function has been well studied in approximation theory. See, for example, Bernstein (1913), Varga and Carpenter (1987), and Rivlin (1990). For a given positive integer  $k$ , let  $\mathcal{P}_k$  denote the class of all real polynomials of degree at most  $k$ . For any continuous function  $f$  on  $[-1, 1]$ , let

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

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

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

We now focus on the special case of the absolute value function  $f(x) = |x|$ . Because  $f$  is an even function, so is its best polynomial approximation. We thus only need to consider polynomials of even degrees. For any positive integer  $k$ , we shall denote by  $G_k^*$  the best polynomial approximation of degree  $2k$  to  $|x|$  and write

$$(12) \quad G_k^*(x) = \sum_{j=0}^k g_{2j}^* x^{2j}.$$

The Bernstein constant is defined as

$$\beta_* = \lim_{k \rightarrow \infty} 2k \delta_{2k}(f).$$

Bernstein (1913) showed that the limit exists and is between 0.278 and 0.286. Varga and Carpenter (1987) disproved a conjecture by Bernstein and calculated that  $\beta_* = 0.280169499$ .

The classical Chebyshev Alternation Theorem states that a polynomial  $G^* \in \mathcal{P}_k$  is the (unique) best polynomial approximation to a continuous function  $f$  if and only if the difference  $f(x) - G^*(x)$  takes consecutively its maximal value with alternating signs at least  $(k + 2)$  times. That is, there exist  $k + 2$  points  $-1 \leq x_0 < \dots < x_{k+1} \leq 1$  such that

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

In the case of the absolute value function, 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 least favorable priors used in

the derivation of the minimax lower bounds given in this section. Divide the set of the alternation points of  $G_k^*(x)$  into two subsets and denote

$$(13) \quad A_0 = \{x \in [-1, 1] : |x| - G_k^*(x) = -\delta_{2k}(|x|)\},$$

$$(14) \quad A_1 = \{x \in [-1, 1] : |x| - G_k^*(x) = \delta_{2k}(|x|)\}.$$

It follows easily from the fact that both  $|x|$  and  $G_k^*(x)$  are even functions that the set  $A_0$  contains an odd number of points and  $A_1$  has an even number of points. We shall see later that least favorable priors are necessarily supported on  $A_0$  and  $A_1$  respectively. Intuitively, this makes the priors maximally apart and yet not “testable”. It also connects the construction of the optimal estimator with the minimax lower bound.

**3.2. Minimax Lower Bounds.** We now state and prove the minimax lower bounds for estimating the nonsmooth functional  $T(\theta)$  over the bounded set  $\Theta_n(M)$  and the unbounded set  $\mathbb{R}^n$ . The derivation of the lower bounds relies heavily on the general lower bound argument given in the previous section. It also requires a careful construction of least favorable prior distributions  $\mu_0$  and  $\mu_1$  along with finding an effective upper bound for the chi-square distance between the marginal distributions.

**Theorem 3** *Let  $y_i \sim N(\theta_i, 1)$ ,  $i = 1, \dots, n$ , be independent normal random variables and let  $T(\theta) = \frac{1}{n} \sum_{i=1}^n |\theta_i|$ . For a fixed constant  $M > 0$ , denote by  $\Theta_n(M) = \{\theta \in \mathbb{R}^n : |\theta_i| \leq M\}$ . Then, the minimax risk for estimating  $T(\theta)$  over the parameter space  $\Theta_n(M)$  is bounded from below as*

$$(15) \quad \inf_{\hat{T}} \sup_{\theta \in \Theta_n(M)} E(\hat{T} - T(\theta))^2 \geq \beta_*^2 M^2 \left( \frac{\log \log n}{\log n} \right)^2 (1 + o(1))$$

where  $\beta_*$  is the Bernstein constant. Without any constraint on the parameters, the minimax risk satisfies

$$(16) \quad \inf_{\hat{T}} \sup_{\theta \in \mathbb{R}^n} E(\hat{T} - T(\theta))^2 \geq \frac{4\beta_*^2}{9e^2 \log n} (1 + o(1)).$$

The minimax lower bounds given in Theorem 3 converge to zero at a slow logarithmic rate showing that the nonsmooth functional  $T(\theta)$  is difficult to estimate. In contrast the rates for estimating linear and quadratic functionals are most often algebraic. In particular 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 of convergence over  $\mathbb{R}^n$  for estimating the linear functional  $L(\theta)$  can be attained by the sample average  $\bar{y}$ . For estimating the quadratic functional  $Q(\theta)$ , the parametric rate can be achieved over  $\Theta_n(M)$  by using the unbiased estimator  $\hat{Q} = \frac{1}{n} \sum_{i=1}^n (y_i^2 - 1)$ .

We shall show in the next section that the minimax lower bound  $\beta_*^2 M^2 \left( \frac{\log \log n}{\log n} \right)^2$  for  $\Theta_n(M)$  is in fact asymptotically sharp and the rate of convergence  $\frac{1}{\log n}$  for  $\mathbb{R}^n$  is optimal. The optimal procedures are constructed using the Hermite polynomials. These procedures are much more involved than those for estimating the linear and quadratic functionals discussed above.

A crucial tool in the proof of the lower bounds as well as in the construction of the optimal procedures is the application of properties of Hermite polynomials. Let  $H_k$  be the Hermite polynomial defined by

$$(17) \quad \frac{d^k}{dy^k} \phi(y) = (-1)^k H_k(y) \phi(y).$$

For this version of the Hermite polynomial

$$(18) \quad \int H_k^2(y) \phi(y) dy = k! \quad \text{and} \quad \int H_k(y) H_j(y) \phi(y) dy = 0$$

when  $k \neq j$ .

Another key technical tool for the proof of Theorem 3 is the construction of two priors with special properties.

**Lemma 1** *For any given even integer  $k > 0$ , there exists two probability measures  $\nu_0$  and  $\nu_1$  on  $[-1, 1]$  that satisfy the following conditions:*

- $\nu_0$  and  $\nu_1$  are symmetric around 0,
- $\int t^l \nu_1(dt) = \int t^l \nu_0(dt)$ , for  $l = 0, 1, \dots, k$ ,
- $\int |t| \nu_1(dt) - \int |t| \nu_0(dt) = 2\delta_k$ ,

where  $\delta_k$  is the distance in the uniform norm on  $[-1, 1]$  from the absolute value function  $f(x) = |x|$  to the space  $\mathcal{P}_k$  of polynomials of no more than degree  $k$ .

As discussed earlier,  $\delta_k = \beta_* k^{-1} (1 + o(1))$  as  $k \rightarrow \infty$ , where  $\beta_*$  is the Bernstein constant. See Section 7 for further discussions. The proof of Lemma 1 is given in Section 8.

**Proof of Theorem 3:** For a given even integer  $k_n$ , let  $\nu_0$  and  $\nu_1$  be two probability measures possessing the properties given in Lemma 1. Let  $g(x) = Mx$  and let  $\mu_i$  be the measures on  $[-M, M]$  defined by  $\mu_i(A) = \nu_i(g^{-1}(A))$  for  $i = 0$  and 1. It follows that

- $\mu_0$  and  $\mu_1$  are symmetric around 0
- $\int t^l \mu_1(dt) = \int t^l \mu_0(dt)$ , for  $l = 0, 1, \dots, k_n$
- $\int |t| \mu_1(dt) - \int |t| \mu_0(dt) = 2M\delta_{k_n}$ .

Let  $\mu_1^n$  and  $\mu_0^n$  be the product priors  $\mu_i^n = \prod_{j=1}^n \mu_i$ . In other words we put down  $n$  independent priors on the coordinates. We have

$$E_{\mu_1^n} T(\theta) - E_{\mu_0^n} T(\theta) = E_{\mu_1} |\theta_1| - E_{\mu_0} |\theta_1| = 2M\delta_{k_n}.$$

and

$$E_{\mu_0^n} (T(\theta) - E_{\mu_0^n} T(\theta))^2 = \frac{1}{n} E_{\mu_0} (|\theta_1| - E_{\mu_0} |\theta_1|)^2 \leq \frac{M^2}{n}.$$

Set  $f_{0,M}(y) = \int \phi(y-t)\mu_0(dt)$  and  $f_{1,M}(y) = \int \phi(y-t)\mu_1(dt)$ . Note that since  $g(x) = \exp(-x)$  is a convex function of  $x$  and  $\mu_0$  is symmetric

$$\begin{aligned} f_{0,M}(y) &\geq \frac{1}{\sqrt{2\pi}} \exp\left(-\int \frac{(y-t)^2}{2} \mu_0(dt)\right) \\ &= \phi(y) \exp\left(-\frac{1}{2} M^2 \int t^2 \nu_0(dt)\right) \\ &\geq \phi(y) \exp\left(-\frac{1}{2} M^2\right). \end{aligned}$$

Let  $H_r$  be the Hermite polynomial defined in (17). Then

$$\phi(y - \alpha t) = \sum_{k=0}^{\infty} H_k(y) \phi(y) \frac{\alpha^k t^k}{k!}$$

and it follows that

$$\int \frac{(f_{1,M}(y) - f_{0,M}(y))^2}{f_{0,M}(y)} dy \leq e^{\frac{M^2}{2}} \sum_{k=k_n+1}^{\infty} \frac{1}{k!} M^{2k}$$

Now set

$$I_n^2 = \int \frac{(\prod_{i=1}^n f_{1,M}(y_i) - \prod_{i=1}^n f_{0,M}(y_i))^2}{\prod_{i=1}^n f_{0,M}(y_i)} dy_1 dy_2 \dots dy_n.$$

Then

$$\begin{aligned} I_n^2 &= \int \frac{(\prod_{i=1}^n f_{1,M}(y_i))^2}{\prod_{i=1}^n f_{0,M}(y_i)} dy_1 dy_2 \dots dy_n - 1 \\ &= \left(\prod_{i=1}^n \int \frac{(f_{1,M}(y_i))^2}{f_{0,M}(y_i)} dy_i\right) - 1 \\ (19) \quad &\leq \left(1 + e^{\frac{M^2}{2}} \sum_{k=k_n+1}^{\infty} \frac{1}{k!} M^{2k}\right)^n - 1 \\ &\leq \left(1 + e^{\frac{3M^2}{2}} \frac{1}{k_n!} M^{2k_n}\right)^n - 1 \end{aligned}$$

Now note that  $k! > (\frac{k}{e})^k$ . Hence

$$(20) \quad I_n^2 \leq (1 + e^{\frac{3M^2}{2}} (\frac{eM^2}{k_n})^{k_n})^n - 1.$$

Now let  $k_n$  be the smallest positive integer satisfying  $k_n \geq \frac{\log n}{\log \log n} + \frac{\log n}{(\log \log n)^{3/2}}$ .

It is easy to check that  $I_n \rightarrow 0$ . Noting that  $v_0 \leq \frac{M}{\sqrt{n}}$  and applying Corollary 1 yields

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

and (15) follows.

For the proof of (16), let  $M = \sqrt{\log n}$  and take  $k_n$  to be the smallest positive integer satisfying  $k_n \geq (1.5)e \log n$ . We may bound  $I_n$  starting from (19) and then noting that for some constant  $D > 0$

$$(21) \quad I_n^2 \leq (1 + e^{\frac{M^2}{2}} D \frac{1}{k_n!} M^{2k_n})^n - 1 \leq (1 + Dn^{\frac{1}{2}} (\frac{e \log n}{(3/2)e \log n})^{k_n})^n - 1 \rightarrow 0$$

It is then easy to check that Corollary (1) now yields (16). ■

**Remark 1** In the bounded case, we shall show in Section 4 that the minimax lower bound  $\beta_*^2 M^2 \left(\frac{\log \log n}{\log n}\right)^2$  is asymptotically sharp. It can be seen from the proof of Theorem 2 that this minimax risk corresponds to the Bayes risk of the least favorable prior which is asymptotically equal to the prior  $\frac{1}{2}(\mu_0 + \mu_1)$ .

**Remark 2** The proof of (16) can be used to show that for any constant  $c > 0$ , there exists another constant  $d > 0$  such that

$$(22) \quad \inf_{\hat{T}} \sup_{\theta \in \Theta_n(\sqrt{c \log n})} E(\hat{T} - T(\theta))^2 \geq \frac{d}{\log n} (1 + o(1)).$$

#### 4. Optimal Estimation of the $\ell_1$ Norm of Bounded Normal Means.

Section 3 developed minimax lower bounds for estimating the nonsmooth functional  $T(\theta)$ . Although the minimax lower bounds converge slowly, they are also difficult to attain. The difficulty of the estimation problem stems from the fact that the absolute value function is not differentiable at 0. In this section we shall consider the bounded case and construct an estimator that relies on the best polynomial approximation to the absolute value function and the use of Hermite polynomials. The estimator is then shown to be asymptotically sharp minimax. The unbounded case and the sparse case will be treated in the next two sections.

4.1. *Polynomial Approximation.* The construction of the rate optimal estimator is involved. This is partly due to the nonexistence of an unbiased estimator for  $|\theta_i|$ . Our strategy is to “smooth” the singularity at 0 by a polynomial approximation and construct an unbiased estimator for each term in the expansion by using the Hermite polynomials.

The optimal estimator relies on the best polynomial approximation  $G_K^*$  of the absolute value function. A drawback of using  $G_K^*$  is that it is not convenient to construct. An explicit and nearly optimal polynomial approximation  $G_K$  can be easily obtained by using the Chebyshev polynomials. Note that the Chebyshev polynomial (of the first kind) of degree  $k$  is defined as

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

The following expansion can be found, for example, in Rivlin (1974):

$$(23) \quad |x| = \frac{2}{\pi} T_0(x) + \frac{4}{\pi} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{T_{2k}(x)}{4k^2 - 1}$$

where  $T_{2k}(x)$  is the Chebyshev polynomial of degree  $2k$ . Consider the truncated version of the expansion (23) and let

$$(24) \quad G_K(x) = \frac{2}{\pi} T_0(x) + \frac{4}{\pi} \sum_{k=1}^K (-1)^{k+1} \frac{T_{2k}(x)}{4k^2 - 1}.$$

We can also write  $G_K(x)$  as

$$(25) \quad G_K(x) = \sum_{k=0}^K g_{2k} x^{2k}.$$

The following lemma provides uniform error bounds of  $G_K^*$  and  $G_K$  over the interval  $[-1, 1]$  as well as bounds on the coefficients  $g_{2k}^*$  and  $g_{2k}$ . These bounds are useful in the analysis of the optimal estimators.

**Lemma 2** *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 (24). Then*

$$(26) \quad \max_{x \in [-1, 1]} |G_K^*(x) - |x|| \leq \frac{\beta_*}{2K} (1 + o(1))$$

$$(27) \quad \max_{x \in [-1, 1]} |G_K(x) - |x|| \leq \frac{2}{\pi(2K + 1)}.$$

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

$$(28) \quad |g_{2k}^*| \leq 2^{3K} \quad \text{and} \quad |g_{2k}| \leq 2^{3K}.$$

The uniform error bounds (26) and (27) were proved in Bernstein (1913). The proof of the bound on the coefficients  $g_{2k}^*$  and  $g_{2k}$  is given in Section 8.

4.2. *The Construction of the Optimal Estimator.* We shall now use the best polynomial approximation  $G_K^*(x)$  and the Hermite polynomials to construct an estimator of  $T(\theta)$  that is asymptotically sharp minimax over the bounded parameter space  $\Theta_n(M)$ . We first consider the special case of  $M = 1$ . The case of a general  $M$  involves an additional rescaling step.

When  $M = 1$ , it follows from Lemma 2 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.

Let  $\phi$  be the density function of a standard normal variable. Recall that for positive integers  $k$

$$(29) \quad \frac{d^k}{dy^k} \phi(y) = (-1)^k H_k(y) \phi(y)$$

where  $H_k$  is a Hermite polynomial with respect to  $\phi$ . It is well known that for  $X \sim N(\mu, 1)$ ,  $H_k(X)$  is an unbiased estimate of  $\mu^k$  for any positive integer  $k$ , i.e.,  $EH_k(X) = \mu^k$ .

Since  $H_k(y_i)$  is an unbiased estimate of  $\theta_i^k$  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

$$(30) \quad \widehat{T}_K(\theta) = \sum_{k=0}^K g_{2k}^* \bar{B}_{2k}.$$

For estimating the functional  $T(\theta)$  over the bounded parameter space  $\Theta_n(M)$  for a general  $M > 0$ , we shall first rescale each  $\theta_i$  and then approximate  $|\theta_i|$  term by term. More specifically, let  $|\theta'_i| = M^{-1} \theta_i$ . Then  $|\theta'_i| \leq 1$  for  $i = 1, \dots, n$  and

$$||\theta'_i| - G_K^*(\theta'_i)| \leq \frac{\beta_*}{2K} (1 + o(1)), \quad \text{for all } |\theta'_i| \leq 1.$$

Hence,

$$||\theta_i| - \tilde{G}_K^*(\theta_i)| \leq \frac{\beta_* M}{2K} (1 + o(1)), \quad \text{for all } |\theta_i| \leq M$$

where  $\tilde{G}_K^*(x) = \sum_{k=0}^K \tilde{g}_{2k}^* x^{2k}$  with  $\tilde{g}_{2k}^* = g_{2k}^* M^{-2k+1}$ .

Again,  $H_k(y_i)$  is an unbiased estimate of  $\theta_i^k$ . We estimate  $b_{2k}(\theta) \equiv \frac{1}{n} \sum_{i=1}^n \theta_i^{2k}$  by

$$(31) \quad \bar{B}_{2k} = \frac{1}{n} \sum_{i=1}^n H_{2k}(y_i)$$

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

$$(32) \quad T_K(\widehat{\theta}; M) = \sum_{k=0}^K \tilde{g}_{2k}^* \bar{B}_{2k} = \sum_{k=0}^K g_{2k}^* M^{-2k+1} \bar{B}_{2k}.$$

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

$$(33) \quad K = K_* \equiv \frac{\log n}{2 \log \log n}$$

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

$$(34) \quad T_*(\widehat{\theta}) \equiv T_{K_*}(\widehat{\theta}; M) = \sum_{k=0}^{K_*} \tilde{g}_{2k}^* \bar{B}_{2k}.$$

**4.3. Optimality of the Estimator.** We now study the property of the estimator defined in (34). The following result shows that the estimator  $T_*(\widehat{\theta})$  is asymptotically sharp minimax, i.e., it achieves the exact minimax lower bound given in Theorem 3 asymptotically.

**Theorem 4** *Let  $y_i \sim N(\theta_i, 1)$  be independent normal random variables with  $|\theta_i| \leq M$ ,  $i = 1, \dots, n$ . Let  $T(\theta) = n^{-1} \sum_{i=1}^n |\theta_i|$ . The estimator  $T_*(\widehat{\theta})$  given in (34) satisfies*

$$(35) \quad \sup_{\theta \in \Theta_n(M)} E(T_*(\widehat{\theta}) - T(\theta))^2 \leq \beta_*^2 M^2 \left( \frac{\log \log n}{\log n} \right)^2 (1 + o(1)).$$

**Remark 3** If  $G_K(x)$ , instead of  $G_K^*(x)$ , is used in the construction of the estimator  $T_*(\widehat{\theta})$ , the resulting estimator  $\widehat{T}(\theta)$  satisfies

$$(36) \quad \sup_{\theta \in \Theta_n(M)} E(\widehat{T}(\theta) - T(\theta))^2 \leq 4\pi^{-2} M^2 \left( \frac{\log \log n}{\log n} \right)^2 (1 + o(1)).$$

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

We need the following variance bounds for the proof of Theorem 4 as well as other results given in the later sections.

**Lemma 3** *Let  $X \sim N(\mu, 1)$ , then*

$$E(H_k^2(X)) = k! \sum_{j=0}^k \binom{k}{j} \mu^{2j} \frac{1}{j!}.$$

*Consequently*

$$\text{Var}(H_k(X)) \leq E(H_k^2(X)) \leq e^{\mu^2} k^k.$$

*If  $|\mu| \leq M$  and  $M^2 \geq k$ , then*

$$\text{Var}(H_k(X)) \leq E(H_k^2(X)) \leq (2M^2)^k.$$

The proof of Lemma 3 is given in Section 8.

**Proof of Theorem 4:** In the proof we shall assume  $M \geq 1$ . The case of  $M < 1$  is similar. Note that  $E\bar{B}_{2k} = b_{2k}(\theta)$  for  $k \geq 0$  and hence,

$$ET_K(\widehat{\theta}; M) = \sum_{k=0}^K \tilde{g}_{2k}^* b_{2k}(\theta) = \frac{1}{n} \sum_{i=1}^n \tilde{G}_K^*(\theta_i).$$

The bias of  $\widehat{T}(\theta)$  can then be bounded easily as follows. For any  $\theta \in \Theta_n(M)$ ,

$$|ET_K(\widehat{\theta}; M) - T(\theta)| = \left| \frac{1}{n} \sum_{i=1}^n \tilde{G}_K^*(\theta_i) - \frac{1}{n} \sum_{i=1}^n |\theta_i| \right| \leq \frac{1}{n} \sum_{i=1}^n \left| \tilde{G}_K^*(\theta_i) - |\theta_i| \right| \leq \frac{\beta_* M}{2K} (1 + o(1)).$$

Now we consider the variance of  $T_K(\widehat{\theta}; M)$ . It follows from Lemma 3 that the variance of  $\bar{B}_{2k}$  satisfies

$$\text{Var}(\bar{B}_{2k}) = n^{-2} \sum_{i=1}^n \text{Var}(H_{2k}(y_i)) \leq e^{M^2} (2k)^{2k} n^{-1}.$$

To bound the variance of  $T_K(\widehat{\theta}; M)$ , first note that for any random variables  $X_i$ ,  $i = 1, \dots, n$ ,

$$(37) \quad E\left(\sum_{i=1}^n X_i\right)^2 \leq \left(\sum_{i=1}^n (EX_i^2)^{1/2}\right)^2.$$

It then follows that for all  $\theta \in \Theta_n(M)$ ,

$$\begin{aligned} \text{Var}(T_K(\widehat{\theta}; M)) &\leq \left\{ \sum_{k=1}^K |\tilde{g}_{2k}^*| \text{Var}^{\frac{1}{2}}(\bar{B}_{2k}) \right\}^2 \leq \left\{ \sum_{k=1}^K |g_{2k}^*| M^{-2k+1} e^{M^2/2} (2k)^k \right\}^2 \cdot n^{-1} \\ &\leq 2e^{M^2} 2^{8K} K^{2K} n^{-1}. \end{aligned}$$

Hence, the mean squared error of  $T_K(\widehat{\theta}; M)$  is bounded by

$$(38) \quad E(T_K(\widehat{\theta}; M) - T(\theta))^2 \leq \frac{\beta_*^2 M^2}{(2K)^2} (1 + o(1)) + 2e^{M^2} 2^{8K} K^{2K} n^{-1}.$$

Now set

$$K_* = \frac{\log n}{2 \log \log n}.$$

Then the second term in (38) is negligible relative to the first term and we have, for all  $\theta \in \Theta_n(M)$ ,

$$E(T_*(\widehat{\theta}) - T(\theta))^2 \leq \beta_*^2 M^2 \left( \frac{\log \log n}{\log n} \right)^2 (1 + o(1)). \quad \blacksquare$$

**5. Estimating the  $\ell_1$  Norm of Unbounded Normal Means.** We now turn to the unbounded case where no restriction is imposed on the values of the means  $\theta_i$ . This case is more difficult than the bounded case. We shall construct an estimator of  $T(\theta)$  that attains the optimal rate of convergence, but not the optimal constant, for the unbounded case. In the construction below, both  $G_K^*$  and  $G_K$  work. For concreteness, hereafter we shall focus on using  $G_K$  instead of the best polynomial approximation  $G_K^*$ .

It turns out that a key step towards solving this general problem is to understand the estimation problem where the means are bounded with the bound growing with the sample size  $n$ . We shall thus first treat this case and then consider rate-optimal estimation for the general case.

5.1. *Estimating the  $\ell_1$  Norm with a Growing Bound.* Suppose  $y_i \stackrel{\text{ind}}{\sim} N(\theta_i, 1)$ ,  $i = 1, 2, \dots, n$ , where  $|\theta_i| \leq M_n$  for  $i = 1, \dots, n$ , with  $M_n = \sqrt{c \log n}$  for some  $c > 1$ . As in the last section, we estimate  $T(\theta)$  by first rescaling and define the estimator of  $T(\theta)$  by

$$(39) \quad T_K(\widehat{\theta}; M_n) = \sum_{k=0}^K \tilde{g}_{2k} \bar{B}_{2k}.$$

where  $\tilde{g}_{2k} = g_{2k} M_n^{-2k+1}$  and  $\bar{B}_{2k} = \frac{1}{n} \sum_{i=1}^n H_{2k}(y_i)$ .

**Theorem 5** *Let  $y_i \sim N(\theta_i, 1)$  be independent normal random variables with  $|\theta_i| \leq M_n$ ,  $i = 1, \dots, n$ , where  $M_n = \sqrt{c \log n}$  for some  $c > 1$ . Let  $T(\theta) = n^{-1} \sum_{i=1}^n |\theta_i|$ . The estimator  $T_K(\widehat{\theta}; M_n)$  given in (39) with  $K = \frac{1}{7} \log n - (\log n)^{\frac{1}{2}}$  satisfies*

$$(40) \quad \sup_{\theta \in \Theta_n(M_n)} E(T_K(\widehat{\theta}; M_n) - T(\theta))^2 \leq \frac{49c}{\pi^2} (\log n)^{-1} (1 + o(1)).$$

This upper bound together with the minimax lower bound (22) show that the estimator  $T_K(\widehat{\theta}; M_n)$  defined in (39) with  $K = \frac{1}{7} \log n - (\log n)^{\frac{1}{2}}$  is minimax rate optimal in this case. We shall show that the difficulty of estimating  $T(\theta)$  over  $\mathbb{R}^n$  is essentially the same as estimating over  $\Theta_n(M_n)$  with an appropriate choice of  $M_n$  of order  $\sqrt{\log n}$ . However, the construction of the rate-optimal estimator of  $T(\theta)$  over  $\mathbb{R}^n$  is much more complicated.

The proof of Theorem 5 is given in Section 8.

*5.2. Rate Optimal Estimator for the Unbounded Case.* We now turn to the unbounded case. It is helpful to provide some intuition and motivation before we formally describe the estimation procedure. Consider the one dimensional case. Suppose we observe  $X \sim N(\mu, 1)$  and wish to estimate  $|\mu|$ . Set  $M_n = 8\sqrt{\log n}$ . Let  $\mu' = M_n^{-1}\mu$ . Then  $|\mu'| \leq 1$  and

$$||\mu'| - G_K(\mu')| \leq \frac{2}{\pi(2K+1)}, \quad \text{for all } |\mu'| \leq 1.$$

Hence,

$$||\mu| - \tilde{G}_K(\mu)| \leq \frac{2M_n}{\pi(2K+1)}, \quad \text{for all } |\mu| \leq M_n$$

where  $\tilde{G}_K(\mu) = \sum_{k=0}^K \tilde{g}_{2k} \mu^{2k}$  with  $\tilde{g}_{2k} = g_{2k} M_n^{-2k+1}$ . Again,  $H_k(X)$  is an unbiased estimate of  $\mu^k$ . Set  $K = \frac{1}{12} \log n$  and

$$(41) \quad S_K(x) = \sum_{k=0}^K g_{2k} M_n^{-2k+1} H_{2k}(x).$$

We define an estimator of  $|\mu|$  by a truncated version of  $S_K(X)$ ,

$$(42) \quad \delta(X) = \min \{S_K(X), n\}.$$

It is easy to see that  $\delta(X)$  is a good estimate of  $|\mu|$  when  $|\mu|$  is small. On the other hand, when  $|\mu|$  is large,  $\delta(X)$  is no longer a good estimator of  $|\mu|$  because the variance of  $\delta(X)$  is very large. When  $|\mu|$  is large, a good estimate

of  $|\mu|$  is simply  $|X|$ . Therefore, for the unbounded case, a good strategy is to estimate  $|\mu|$  by  $\delta(X)$  when  $|X|$  is not too large and estimate  $|\mu|$  by  $|X|$  when  $|X|$  is large.

We now formally state the procedure for estimating  $T(\theta)$  as follows. We shall first use the idea of sample splitting. Note that observing  $y_i \sim N(\theta_i, 1)$  is equivalent to observing  $y_{il} \stackrel{iid}{\sim} N(\theta_i, 2)$ , for  $l = 1, 2$ . (One can generate  $y_{i1}$  and  $y_{i2}$  from  $y_i$ . Let  $z_i \sim N(0, 1)$  be independent of  $y_i$  and set  $y_{i1} = y_i + z_i$  and  $y_{i2} = y_i - z_i$ . Then  $y_{il} \stackrel{iid}{\sim} N(\theta_i, 2)$ .) Write  $x_{il} = \frac{1}{\sqrt{2}}y_{il}$  for  $l = 1, 2$  and  $i = 1, \dots, n$ . Then  $x_{il} \stackrel{iid}{\sim} N(\theta'_i, 1)$ , for  $l = 1, 2$ , with  $\theta'_i = \theta_i/\sqrt{2}$ . Estimating  $T(\theta)$  based on  $\{y_i\}$  is thus equivalent to estimating  $\sqrt{2}T(\theta')$  based on  $\{x_{il}\}$ . We shall construct an estimate  $\widehat{T(\theta')}$  for  $T(\theta')$  and estimate  $T(\theta)$  by  $\sqrt{2}\widehat{T(\theta')}$ .

We define the estimate of  $T(\theta') = n^{-1} \sum_{i=1}^n |\theta'_i|$  by

$$(43) \quad \widehat{T(\theta')} = \frac{1}{n} \sum_{i=1}^n \left\{ \delta(x_{i1})I(|x_{i2}| \leq 2\sqrt{2\log n}) + |x_{i1}|I(|x_{i2}| > 2\sqrt{2\log n}) \right\}$$

where  $\delta(\cdot)$  is defined in (41) and (42), and define the estimator of  $T(\theta) = n^{-1} \sum_{i=1}^n |\theta_i|$  by

$$(44) \quad \widehat{T(\theta)} = \sqrt{2}\widehat{T(\theta')}.$$

Here  $|x_{i2}|$  is used to test to size of  $\theta'_i$  and based on the test we use either  $\delta(x_{i1})$  or  $|x_{i1}|$  to estimate  $|\theta'_i|$ .

The following theorem shows that  $\widehat{T(\theta)}$  attains the rate of convergence  $(\log n)^{-1}$  over the whole parameter space  $\mathbb{R}^n$ .

**Theorem 6** *The estimator  $\widehat{T(\theta)}$  defined in (43) and (44) satisfies, for all  $\theta \in \mathbb{R}^n$ ,*

$$(45) \quad E(\widehat{T(\theta)} - T(\theta))^2 \leq \frac{C}{\log n}(1 + o(1))$$

for some constant  $C > 0$ .

Together with the minimax lower bound given in Theorem 3, Theorem 6 shows that the hybrid estimator is rate optimal over the parameter space  $\mathbb{R}^n$ . The proof of Theorem 6 is involved and is given in Section 8. The key is to analyze the bias and variance of a single component.

**6. Estimating the  $\ell_1$  Norm of Sparse Normal Means.** In high dimensional problems, an especially interesting case is when the mean vector is sparse, i.e., only a small proportion of the  $\theta_i$ 's are nonzero. Suppose we

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

Denote the  $\ell_0$  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

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

In this section we consider the problem of estimating the average of the absolute value of the nonzero means. For  $\theta \in \Theta_{k_n}$ ,

$$(46) \quad T(\theta) = \text{average}\{|\theta_i| : \theta_i \neq 0\} = \frac{1}{k_n} \sum_{i=1}^n |\theta_i|.$$

We calibrate the sparsity parameter  $k_n$  by  $k_n = n^\beta$  for  $0 < \beta \leq 1$ . The following result shows that for  $0 < \beta \leq \frac{1}{2}$ , it is not possible to estimate the functional  $T(\theta)$  consistently.

**Theorem 7** *Let  $k_n = n^\beta$ . Then for all  $0 < \beta \leq \frac{1}{2}$ , the minimax risk satisfies*

$$(47) \quad \inf_{\widehat{T(\theta)}} \sup_{\theta \in \Theta_{k_n}} E(\widehat{T(\theta)} - T(\theta))^2 \geq C$$

for some constant  $C > 0$ .

The proof of Theorem 7 is analogous to that of Theorem 7 in Cai and Low (2004) and we omit it here for reasons of space.

We now turn to the more interesting case where  $k_n = n^\beta$  with  $\frac{1}{2} < \beta \leq 1$ . The following result show that the minimax rate of convergence in this case is  $(\log n)^{-1}$ .

**Theorem 8** *Let  $k_n = n^\beta$  for some  $\frac{1}{2} < \beta < 1$ . Then the minimax risk for estimating the functional  $T(\theta)$  over  $\Theta_{k_n}$  satisfies*

$$(48) \quad \inf_{\widehat{T(\theta)}} \sup_{\theta \in \Theta_{k_n}} E(\widehat{T(\theta)} - T(\theta))^2 \asymp \frac{C}{\log n}$$

The proof of the lower bound in Theorem 8 is similar to that of Theorem 3. The upper bound can be attained by a modified version of the estimator  $\widehat{T(\theta)}$  defined in (43) and (44). The key in the construction is to have estimates of the individual coordinates that perform well when the coordinates are zero. This can be achieved by using the polynomial approximation  $G_K(x)$  (or  $G_K^*(x)$ ) without the constant term.

As in Section 5.2 set  $K = \frac{1}{12} \log n$ , and define

$$(49) \quad \tilde{S}_K(x) = \sum_{k=1}^K g_{2k} M_n^{-2k+1} H_{2k}(x).$$

Note that here the constant term  $g_0$  is excluded. We then define an estimator of  $|\mu|$  by truncating  $\tilde{S}_K(X)$ ,

$$(50) \quad \tilde{\delta}(X) = \min \left\{ \tilde{S}_K(X), n^2 \right\}.$$

Note that the bias of the estimator  $\tilde{\delta}(X)$  is much smaller than the bias of  $\delta(X)$  when the mean of  $X$  is zero. As in Section 5.2 we split the sample into two parts and use one for testing and the other for estimation. Let  $x_{il}$  be defined as in Section 5.2. That is,  $x_{il} \stackrel{iid}{\sim} N(\theta'_i, 1)$ , for  $l = 1, 2$ , with  $\theta'_i = \theta_i / \sqrt{2}$ . We define the estimate of  $T(\theta') = k_n^{-1} \sum_{i=1}^n |\theta'_i|$  by

$$(51) \quad \widehat{T(\theta')} = \frac{1}{k_n} \sum_{i=1}^n \left\{ \tilde{\delta}(x_{i1}) I(|x_{i2}| \leq 2\sqrt{2 \log n}) + |x_{i1}| I(|x_{i2}| > 2\sqrt{2 \log n}) \right\}$$

where  $\tilde{\delta}(\cdot)$  is defined in (49) and (50), and set the estimator of  $T(\theta) = k_n^{-1} \sum_{i=1}^n |\theta_i|$  as

$$(52) \quad \widehat{T(\theta)} = \sqrt{2} \widehat{T(\theta')}.$$

It can be shown that the estimator  $\widehat{T(\theta)}$  is rate optimal for estimating  $T(\theta)$  over  $\Theta_{k_n}$ . The proof is similar to that of Theorem 6 and we omit the details here.

**7. Discussions.** The present paper was partly inspired by the general theory of estimating functionals based on i.i.d. observations given in Donoho and Liu (1991) which showed that bounds on minimax estimation can be based on testing two composite hypotheses. The difficulty of the composite testing problem was shown in Le Cam (1973 and 1986) to depend on the total variation distance between the convex hulls of the two composite hypotheses. In the present context the priors  $\mu_0$  and  $\mu_1$  used in the general lower bound of Section 2 can be viewed as picking two points in the convex hull of two subsets of the parameter space and Theorem 2 gives bounds on the risk over these two points. Sections 3 and 4 show that a careful choice of these priors yields sharp minimax lower bounds for estimating the  $\ell_1$  norm of the means of normal random variables.

Best polynomial approximation played a major role in the development of our results, both for the upper and lower bounds. Note that the last two conditions in Lemma 1 yield

$$\int (|t| - G_k^*(t))\nu_1(dt) - \int (|t| - G_k^*(t))\nu_0(dt) = 2\delta_k.$$

From the definition of  $G_k^*$  we have  $-\delta_k \leq |t| - G_k^*(t) \leq \delta_k$  for all  $-1 \leq t \leq 1$ . Since  $\nu_0$  and  $\nu_1$  are probability measures it follows that they are supported on the subsets  $A_0$  and  $A_1$  of the alternation points defined in (13) and (14) respectively.

We should also emphasize that the values of the functional  $T$  on the two set of support points are not well separated. In fact the values alternate. This is quite different from the more standard cases of estimating a linear or quadratic functional. In the case of quadratic functionals, even though the alternative hypothesis may need to be composite, the functional only takes on two values, one on the null and the other on the alternative. See for example Cai and Low (2005).

The techniques given here can also be compared to those found in Lepski, Nemirovski and Spokoiny (1999) where attention was focused on estimating the  $L_1$  norm of a regression function. In that paper lower bounds were constructed by mixing in a way similar to that used in the present paper. However, instead of bounding a chi-square distance a bound was given for the Kullback-Leibler distance. It is however not easy to provide good bounds directly for the Kullback-Leibler distance. This is particularly true in cases which correspond to parameter spaces with growing bounds. The lower bounds provided there only work in the case where the parameter space has a fixed bound.

For upper bounds, 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 can be shown easily to be of order  $K^{-1}$ , which is comparable to the best polynomial approximation of degree  $K$ . However, the variance bound of the estimator based on the  $K$ -term Fourier series approximation is of order  $e^{CK^2}$  for some constant  $C > 0$ , whereas the variance of our estimator based on the polynomial approximation of degree  $K$  grows at the rate of  $K^K = e^{K \log K}$ . So the variance of the polynomial-based estimator is much smaller than that of the corresponding estimator using Fourier series, even though the biases of the two estimators are very similar. This allows for more terms to be used in the polynomial approximation with the same variance level thus reducing the bias of the estimate. 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.

The techniques and results developed in the present paper can be used to solve other related problems. For example, 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. For reasons of space, we shall report the results elsewhere. The techniques can also be used for estimating other nonsmooth functionals such as excess mass. See Cai and Low (2010).

**8. Proofs.** In this section we first prove the technical lemmas given in the earlier sections. We then prove Theorem 5 in Section 8.2. The proof of Theorem 6 is involved and will be given in Section 8.3.

8.1. *Proof of Technical Lemmas.* **Proof of Lemma 1:** The proof of this lemma relies on the Hahn-Banach Theorem and the Riesz Representation Theorem. The argument is essentially the same as the one given in Lepski, Nemirovski and Spokoiny (1999). We include it here for completeness.

Consider the space  $C(-1, 1)$  of continuous real-valued functions on the interval  $[-1, 1]$  with uniform norm  $\|\cdot\|_\infty$ . Clearly  $f(t) = |t|$  defined on this interval  $[-1, 1]$  belongs to  $C(-1, 1)$ . Let  $\delta_k$  be the distance in uniform norm on  $[-1, 1]$  from the function  $f$  to the space of polynomials of order  $k$ . Let  $\mathcal{P}_k$  be the linear space spanned by the collection of polynomial of order  $k$  and in addition let  $\mathcal{F}_k$  be the linear space spanned by  $\mathcal{P}_k$  and  $f$ . Note that every element  $g \in \mathcal{F}_k$  can be written uniquely as  $g = cf + p_k$  where  $p_k \in \mathcal{P}_k$  and  $c \in \mathbb{R}$ . Let  $T$  be the linear functional defined by  $T(g) = T(cf + p_k) = c\delta_k$ . It is then clear that  $T = 0$  on  $\mathcal{P}_k$  and  $T(f) = \delta_k$ . Now the norm of the functional  $T$  is given by

$$\|T\| \equiv \sup\{T(g) : g \in \mathcal{F}_k, \|g\|_\infty \leq 1\}$$

It can be checked directly that the norm of this functional is equal to 1. Let  $G_k^*$  be the closest polynomial in  $\mathcal{P}_k$  to  $f$ . Then  $\|f - G_k^*\|_\infty = \delta_k$  and it follows that  $\frac{1}{\delta_k}(f - G_k^*)$  has a norm of 1. Since  $T(\frac{1}{\delta_k}(f - G_k^*)) = 1$  it follows that  $\|T\| \geq 1$ . Now suppose that  $\|T\| > 1$ . Then there exists an element  $g = cf + p_k$  with  $p_k \in \mathcal{P}_k$  such that  $\|g\|_\infty = 1$  and  $T(g) > 1$ . This implies that  $c > \frac{1}{\delta_k}$  and

$$\|f - (-\frac{1}{c}p_k)\|_\infty = \frac{1}{c} < \delta_k.$$

Since  $-\frac{1}{c}p_k \in \mathcal{P}_k$ , this is a contradiction to the definition of  $\delta_k$  which is the distance between  $f$  and  $\mathcal{P}_k$ .

Now by the Hahn-Banach Theorem the linear functional  $T$  can be extended to  $C(-1, 1)$  without increasing the norm of the functional. For simplicity we shall also call this linear functional  $T$ . It then follows from the Riesz Representation Theorem that for each  $g \in C(-1, 1)$

$$T(g) = \int_{-1}^1 g(t)\tau(dt)$$

where  $\tau$  is a Borel signed measure with total variation equal to 1.

It follows from Hahn-Jordan decomposition that there exists two positive measures  $\tau_+$  and  $\tau_-$  such that  $\tau = \tau_+ - \tau_-$ . It then follows that

$$(53) \quad \int_{-1}^1 |t|[\tau_+(dt) - \tau_-(dt)] = \delta_k \quad \text{and} \quad \int_{-1}^1 t^l \tau_+(dt) = \int_{-1}^1 t^l \tau_-(dt) \quad \text{for } l = 0, 1, \dots, k.$$

Define the measures  $\tau_-^*$  and  $\tau_+^*$  by  $\tau_-^*(S) = \tau_-(-S)$  and  $\tau_+^*(S) = \tau_+(-S)$  for all measurable sets  $S$ . Then (53) holds with  $\tau_-$  and  $\tau_+$  replaced by  $\tau_-^*$  and  $\tau_+^*$  respectively. Hence (53) is also true with  $\tau_-$  and  $\tau_+$  replaced by  $(\tau_- + \tau_-^*)/2$  and  $(\tau_+ + \tau_+^*)/2$  respectively. We can thus assume that  $\tau$  is symmetric.

Now take  $\nu = 2\tau$ . Then  $\nu$  is symmetric and

$$(54) \quad \int_{-1}^1 |t|\nu(dt) = 2\delta_k \quad \text{and} \quad \int_{-1}^1 t^l \nu(dt) = 0 \quad \text{for } l = 0, 1, \dots, k.$$

Now let  $\nu_1$  and  $\nu_0$  be the positive and the negative components of  $\nu$ . Then both  $\nu_1$  and  $\nu_0$  are symmetric. Since  $\nu$  has variation equal to 2 and  $\int_{-1}^1 \nu(dt) = 0$  it follows that  $\nu_1$  and  $\nu_0$  are both probability measures.

These measures also clearly satisfy by construction

$$\int_{-1}^1 t^l \nu_1(dt) = \int_{-1}^1 t^l \nu_0(dt)$$

for  $l = 0, 1, \dots, k$  and also

$$\int_{-1}^1 |t|\nu_1(dt) - \int_{-1}^1 |t|\nu_0(dt) = 2\delta_k. \quad \blacksquare$$

**Proof of Lemma 2:** The Chebyshev polynomial  $T_{2m}$  can be alternatively written as

$$(55) \quad T_{2m}(x) = \sum_{l=0}^m \left[ (-1)^{m-l} \sum_{j=m-l}^m \binom{2m}{2j} \binom{j}{m-l} \right] x^{2l}.$$

Write  $T_{2m}(x) = \sum_{l=0}^m t_{2l}x^{2l}$ . Then

$$(56) \quad |t_{2l}| = \sum_{j=m-l}^m \binom{2m}{2j} \binom{j}{m-l} \leq \sum_{j=m-l}^m \binom{2m}{2j} \binom{m}{m-l} \leq 2^{2m}2^m = 2^{3m}.$$

It is now easy to see that the coefficient for  $x^{2k}$  in the polynomial  $G_K(x)$  is bounded from above by

$$|g_{2k}| \leq \frac{4}{\pi} \sum_{j=k}^K \frac{2^{3j}}{4j^2 - 1} \leq 2^{3K}.$$

The bound on the coefficients  $g_{2k}^*$  of the best polynomial approximation  $G_K^*$  follows from Theorem E in Qazi and Rahman (2007) and the bound (56).

■

**Proof of Lemma 3:** Write  $X = \mu + z$  with  $z \sim N(0, 1)$ . It is well known that  $E(H_k^2(z)) = k!$ ,  $E(H_i(z)H_j(z)) = 0$  for  $i \neq j$ , and

$$H_k(\mu + z) = \sum_{j=0}^k \binom{k}{j} \mu^j H_{k-j}(z).$$

Hence,

$$\begin{aligned} EH_k^2(X) &= EH^2(\mu + z) = \sum_{i=0}^k \sum_{j=0}^k \binom{k}{i} \binom{k}{j} \mu^{i+j} E(H_{k-i}(z)H_{k-j}(z)) \\ &= \sum_{j=0}^k \binom{k}{j}^2 \mu^{2j} (k-j)! \\ &= k! \sum_{j=0}^k \binom{k}{j} \mu^{2j} \frac{1}{j!}. \end{aligned}$$

Note that  $k!/j! \leq k^{k-j}$  and hence,

$$EH_k^2(X) = k! \sum_{j=0}^k \binom{k}{j} \mu^{2j} \frac{1}{j!} \leq k^k \sum_{j=0}^k \binom{k}{j} \left(\frac{\mu^2}{k}\right)^j = k^k \left(1 + \frac{\mu^2}{k}\right)^k \leq e^{\mu^2} k^k.$$

If  $|\mu| \leq M$  and  $M^2 \geq k$ , for all  $0 \leq j \leq k$ ,  $\mu^{2j} \frac{1}{j!} \leq M^{2j} \frac{1}{j!} \leq M^{2k} \frac{1}{k!}$ . Hence,

$$EH_k^2(X) = k! \sum_{j=0}^k \binom{k}{j} \mu^{2j} \frac{1}{j!} \leq k! \sum_{j=0}^k \binom{k}{j} M^{2k} \frac{1}{k!} = (2M^2)^k. \quad \blacksquare$$

8.2. *Proof of Theorem 5.* For  $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$ , denote  $b_k(\theta) \equiv \frac{1}{n} \sum_{i=1}^n \theta_i^k$ . Note that  $E\bar{B}_k = b_k(\theta)$  for  $k \geq 0$  and hence,

$$ET(\widehat{\theta}) = \sum_{k=0}^K \tilde{g}_{2k} b_{2k}(\theta) = \frac{1}{n} \sum_{i=1}^n \tilde{G}_K(\theta_i).$$

The bias of  $T(\widehat{\theta})$  can then be bounded easily as follows:

$$|ET(\widehat{\theta}) - T(\theta)| = \left| \frac{1}{n} \sum_{i=1}^n \tilde{G}_K(\theta_i) - \frac{1}{n} \sum_{i=1}^n |\theta_i| \right| \leq \frac{1}{n} \sum_{i=1}^n |G_K(\theta_i) - |\theta_i|| \leq \frac{2M_n}{\pi(2K+1)}.$$

Now we consider the variance of  $T(\widehat{\theta})$ . Note that  $M_n^2 \geq K$ . In this case, the variance of  $\bar{B}_k$  can be bounded by

$$\text{Var}(\bar{B}_{2k}) = n^{-2} \sum_{i=1}^n \text{Var}(H_{2k}(y_i)) \leq n^{-1} (2M_n^2)^{2k}.$$

Hence,

$$\begin{aligned} \text{Var}(T(\widehat{\theta})) &\leq \left\{ \sum_{k=1}^K |\tilde{g}_{2k}| \text{Var}^{\frac{1}{2}}(\bar{B}_{2k}) \right\}^2 \leq \left\{ \sum_{k=1}^K |g_{2k}| M_n^{-2k+1} 2^k M_n^{2k} \right\}^2 \cdot n^{-1} \\ &\leq 4M_n^2 2^{7K} \cdot n^{-1}. \end{aligned}$$

With  $K = \frac{1}{7} \log_2 n - (\log n)^{\frac{1}{2}}$ , the mean squared error is then bounded by

$$E(T(\widehat{\theta}) - T(\theta))^2 \leq \frac{4M_n^2}{\pi^2(2K+1)^2} + 4M_n^2 2^{7K} \cdot n^{-1} = \frac{49c}{\pi^2} (\log n)^{-1} (1 + o(1)). \quad \blacksquare$$

8.3. *Proof of Theorem 6.* We now analyze the properties of the hybrid estimator defined in (44). The key is to study the bias and variance of a single component. Let  $x_1, x_2 \stackrel{iid}{\sim} N(\mu, 1)$  and let

$$(57) \quad \xi = \xi(x_1, x_2) = \delta(x_1) I(|x_2| \leq 2\sqrt{2 \log n}) + |x_1| I(|x_2| > 2\sqrt{2 \log n}).$$

Note that

$$E(\xi) = E\delta(x_1)P(|x_2| \leq 2\sqrt{2 \log n}) + E|x_1|P(|x_2| > 2\sqrt{2 \log n}).$$

**Lemma 4** *Suppose  $I(A)$  is an indicator random variable independent of  $X$  and  $Y$ , then*

$$(58) \quad \text{Var}(XI(A) + YI(A^c)) = \text{Var}(X)P(A) + \text{Var}(Y)P(A^c) + (EX - EY)^2 P(A)P(A^c).$$

Applying Lemma 4, we have

$$\begin{aligned} \text{Var}(\xi) &= \text{Var}(\delta(x_1))P(|x_2| \leq 2\sqrt{2\log n}) + \text{Var}(|x_1|)P(|x_2| > 2\sqrt{2\log n}) \\ (59) \quad &+ (E\delta(x_1) - E|x_1|)^2P(|x_2| \leq 2\sqrt{2\log n})P(|x_2| > 2\sqrt{2\log n}). \end{aligned}$$

We also need the following lemma for the variance of  $\delta$ . (The proof is similar to Lemma 2 in Cai and Low (2005).)

**Lemma 5** *For any two random variables  $X$  and  $Y$*

$$(60) \quad \text{Var}(\min\{X, Y\}) \leq \text{Var}X + \text{Var}Y.$$

*In particular, for any random variable  $X$  and any constant  $C$*

$$(61) \quad \text{Var}(\min(X, C)) \leq \text{Var}X.$$

*Proof of Lemma 5:* Without loss of generality we can assume  $E(X) = 0$  and  $E(Y) \leq 0$ . Let  $Z = \min\{X, Y\}$ . Then

$$(62) \quad EZ^2 \leq EX^2 + EY^2$$

and

$$(63) \quad EZ \leq E(Y).$$

Hence  $(EZ)^2 \geq (EY)^2$  and consequently

$$(64) \quad \text{Var}Z = EZ^2 - (EZ)^2 \leq EX^2 + EY^2 - (EY)^2 = \text{Var}X + \text{Var}Y. \quad \blacksquare$$

**Lemma 6** *Let  $X \sim N(\mu, 1)$  and  $S_K(x) = \sum_{k=0}^K g_{2k} M_n^{-2k+1} H_{2k}(x)$  with  $M_n = 8\sqrt{\log n}$  and  $K = \frac{1}{12} \log n$ . Then for all  $|\mu| \leq 4\sqrt{2\log n}$ ,*

$$(65) \quad |ES_K(X) - |\mu|| \leq \frac{2M_n}{\pi(2K+1)}$$

$$(66) \quad ES_K^2(X) \leq n^{\frac{1}{2}} \log^5 n.$$

**Proof:** The first part follows from Lemmas 2 and 3 and the discussions in Section 5.1. To bound  $ES_K^2(X)$ , it follows from inequality (37) and Lemmas 2 and 3 that

$$\begin{aligned} ES_K^2(X) &\leq \left( \sum_{k=1}^K |g_{2k}| M_n^{-2k+1} (EH_{2k}^2(X))^{\frac{1}{2}} \right)^2 \leq 2^{6K} \left( \sum_{k=1}^K (8\sqrt{\log n})^{-2k+1} (64 \log n)^k \right)^2 \\ &\leq n^{\frac{1}{2}} \log^5 n. \quad \blacksquare \end{aligned}$$

Write  $B(\xi) = E(\xi) - |\mu|$  for the bias of  $\xi$ . We divide into three cases according to the value of  $|\mu|$ . In the first case when  $|\mu| \leq \sqrt{2\log n}$ , we shall show that the estimator behaves essentially like  $\delta(x_1)$  which is a good estimator when  $|\mu|$  is small. In the second case when  $\sqrt{2\log n} \leq |\mu| \leq 4\sqrt{2\log n}$ , we show that the hybrid estimator uses either  $\delta(x_1)$  or  $|x_1|$  and in this case both are good estimators of  $|\mu|$ . In the third case when  $|\mu|$  is large, the hybrid estimator is essentially the same as  $|x_1|$ .

**Case 1:**  $|\mu| \leq \sqrt{2\log n}$ . Note that  $\delta(x_1)$  can be written as  $\delta(x_1) = S_K(x_1) - (S_K(x_1) - n)I(S_K(x_1) \geq n)$  and consequently

$$\begin{aligned} |B(\xi)| &= |E\delta(x_1)P(|x_2| \leq 2\sqrt{2\log n}) + E|x_1|P(|x_2| > 2\sqrt{2\log n}) - |\mu| \\ &\leq |ES_K(x_1) - |\mu|| + E\{(S_K(x_1) - n)I(S_K(x_1) \geq n)\} \\ &\quad + (|ES_K(x_1)| + E|x_1|)P(|x_2| > 2\sqrt{2\log n}) \\ (67) \quad &\equiv B_1 + B_2 + B_3. \end{aligned}$$

Lemma 6 yields that

$$B_1 = |ES_K(x_1) - |\mu|| \leq \frac{2M_n}{\pi(2K+1)}.$$

It follows from the fact  $|\mu| \leq \sqrt{2\log n}$  and the standard bound for normal tail probability  $\Phi(-z) \leq z^{-1}\phi(z)$  for  $z > 0$  that

$$(68) \quad P(|x_2| > 2\sqrt{2\log n}) \leq 2\Phi(-\sqrt{2\log n}) \leq \frac{1}{\sqrt{\pi \log n}} n^{-1}.$$

Note that in this case

$$(69) \quad |ES_K(x_1)| = |\tilde{G}_K(\mu)| \leq |\mu| + \frac{2M_n}{\pi(2K+1)}$$

$$(70) \quad E|x_1| = |\mu| + 2\phi(\mu) - 2|\mu|\Phi(-|\mu|) \leq |\mu| + 1 \leq \sqrt{2\log n} + 1.$$

It then follows from (68)-(70) that

$$B_3 \leq (2\sqrt{2\log n} + \frac{2M_n}{\pi(2K+1)} + 1) \cdot \frac{1}{\sqrt{\pi \log n}} n^{-1} \leq 3n^{-1}.$$

Now consider  $B_2$ . Note that for any random variable  $X$  and any constant  $\lambda > 0$ ,

$$(71) \quad E(XI(X \geq \lambda)) \leq \lambda^{-1}E(X^2I(X \geq \lambda)) \leq \lambda^{-1}EX^2.$$

This together with Lemma 6 yields that

$$(72) \quad B_2 \leq E\{S_K(x_1)I(S_K(x_1) \geq n)\} \leq n^{-1}ES_K^2(x_1) \leq n^{-\frac{1}{2}} \log^5 n.$$

Combining the three terms together shows that in this case the bias is bounded by

$$|B(\xi)| \leq B_1 + B_2 + B_3 \leq \frac{M_n}{\pi K}(1 + o(1)).$$

We now consider the variance. It follows from (59) and Lemma 5 that

$$\begin{aligned} \text{Var}(\xi) &\leq \text{Var}(S_K(x_1)) + \text{Var}(|x_1|)P(|x_2| > 2\sqrt{2\log n}) + (E\delta(x_1) - E|x_1|)^2P(|x_2| > 2\sqrt{2\log n}) \\ &\leq ES_K^2(x_1) + Ex_1^2P(|x_2| > 2\sqrt{2\log n}). \end{aligned}$$

Lemma 6 and Equation (68) together yield that

$$\text{Var}(\xi) \leq n^{\frac{1}{2}} \log^5 n(1 + o(1)).$$

**Case 2.**  $\sqrt{2\log n} \leq |\mu| \leq 4\sqrt{2\log n}$ . In this case,

$$\begin{aligned} |B(\xi)| &= |E\delta(x_1)P(|x_2| \leq 2\sqrt{2\log n}) + E|x_1|P(|x_2| > 2\sqrt{2\log n}) - |\mu| \\ &\leq |E\delta(x_1) - |\mu|| + |E|x_1| - |\mu|| \\ &\leq |ES_K(x_1) - |\mu|| + E\{(S_K(x_1) - n)I(S_K(x_1) \geq n)\} + 2\phi(\mu). \end{aligned}$$

Note that  $|ES_K(x_1) - |\mu|| \leq \frac{2M_n}{\pi(2K+1)}$  and as in (72)

$$E\{(S_K(x_1) - n)I(S_K(x_1) \geq n)\} \leq n^{-\frac{1}{2}} \log^5 n.$$

Note that  $\phi(\mu) \leq \phi(\sqrt{2\log n}) \leq n^{-1}$ . Hence, again the bias is bounded by

$$|B(\xi)| \leq \frac{M_n}{\pi K}(1 + o(1)).$$

For the variance, Equation (59) and Lemma 5 yield that

$$\text{Var}(\xi) \leq \text{Var}(S_K(x_1)) + \text{Var}(x_1) + (E\delta(x_1) - E|x_1|)^2.$$

Note that

$$\begin{aligned} (E\delta(x_1) - E|x_1|)^2 &\leq [ES_K(x_1) - |\mu| + E\{(S_K(x_1) - n)I(S_K(x_1) \geq n)\} - 2\phi(\mu) + 2|\mu|\Phi(-|\mu|)]^2 \\ &\leq \frac{M_n^2}{\pi^2 K^2}(1 + o(1)). \end{aligned}$$

Hence, it follows from Lemma 5 that

$$\begin{aligned}\text{Var}(\xi) &\leq ES_K^2(x_1) + \text{Var}(x_1) + (E\delta(x_1) - E|x_1|)^2 \\ &\leq n^{\frac{1}{2}} \log^5 n (1 + o(1)).\end{aligned}$$

**Case 3.**  $|\mu| > 4\sqrt{2\log n}$ . In this case the standard bound for normal tail probability yields that

$$P(|x_2| \leq 2\sqrt{2\log n}) \leq 2\Phi(-(|\mu| - 2\sqrt{2\log n})) \leq 2\Phi\left(-\frac{|\mu|}{2}\right) \leq \frac{4}{|\mu|} \phi\left(\frac{|\mu|}{2}\right).$$

In particular,

$$P(|x_2| \leq 2\sqrt{2\log n}) \leq 2\Phi(-2\sqrt{2\log n}) \leq \frac{1}{2\sqrt{\pi \log n}} n^{-4}.$$

Hence,

$$\begin{aligned}|B(\xi)| &\leq |E|x_1| - |\mu|| + (|E\delta(x_1)| + E|x_1|)P(|x_2| \leq 2\sqrt{2\log n}) \\ &\leq 2\phi(\mu) + (n + |\mu| + 1)P(|x_2| \leq 2\sqrt{2\log n}) \\ &\leq 2\phi(\mu) + 4\phi\left(\frac{|\mu|}{2}\right) + \frac{1}{2}n^{-3} \leq 6\phi\left(\frac{|\mu|}{2}\right) + \frac{1}{2}n^{-3} \leq n^{-3}.\end{aligned}$$

For the variance, Equation (59) and Lemma 5 yield that

$$\begin{aligned}\text{Var}(\xi) &\leq \text{Var}(|x_1|) + (\text{Var}(\delta(x_1)) + (E\delta(x_1) - E|x_1|)^2)P(|x_2| \leq 2\sqrt{2\log n}) \\ &\leq 1 + (3n^2 + 2(\mu^2 + 1))P(|x_2| \leq 2\sqrt{2\log n}) = 1 + o(1).\end{aligned}$$

Putting the three cases together, we have the following.

**Proposition 1** *For all  $\mu \in \mathbb{R}$ , the bias and the variance of the estimator  $\xi$  defined in (57) satisfy*

$$(73) \quad |B(\xi)| \leq \frac{M_n}{\pi K} (1 + o(1)), \quad \text{and} \quad \text{Var}(\xi) \leq n^{\frac{1}{2}} \log^5 n (1 + o(1)).$$

**Proof of Theorem 6:** With the detailed analysis of the one-dimensional case, we are now ready to give a short proof Theorem 6. It suffices to focus on the estimator  $\widehat{T}(\theta')$  given in (43). Note that

$$\widehat{T}(\theta') = \frac{1}{n} \sum_{i=1}^n \xi(x_{i1}, x_{i2})$$

where  $\xi$  is defined in (57). It follows from Proposition 1 that the bias  $B(\widehat{T(\theta')})$  of the estimator  $\widehat{T(\theta')}$  is bounded by

$$|B(\widehat{T(\theta')})| \leq \frac{1}{n} \sum_{i=1}^n |B(\xi(x_{i1}, x_{i2}))| \leq \frac{M_n}{\pi K} (1 + o(1)).$$

and the variance of  $\widehat{T(\theta')}$  is bounded by

$$\text{Var}(\widehat{T(\theta')}) \leq \frac{1}{n^2} \sum_{i=1}^n \text{Var}(\xi(x_{i1}, x_{i2})) \leq \frac{1}{n^2} \sum_{i=1}^n n^{\frac{1}{2}} \log^5 n (1 + o(1)) \leq 64n^{-\frac{1}{2}} \log n (1 + o(1)).$$

Hence the mean squared error of  $\widehat{T(\theta')}$  satisfies

$$E(\widehat{T(\theta')} - T(\theta'))^2 \leq B^2(\widehat{T(\theta')}) + \text{Var}(\widehat{T(\theta')}) \leq \frac{M_n^2}{\pi^2 K^2} (1 + o(1)) \leq \frac{C}{\log n} (1 + o(1)). \quad \blacksquare$$

*Acknowledgment.* We thank three referees for very constructive comments which have helped significantly to improve the presentation of the paper.

## REFERENCES

- [1] Bernstein, S. N. (1913). Sur la meilleure approximation de  $|x|$  par les polynomes de degrés donnés. *Acta Math.* **37**, 1-57.
- [2] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhya Ser. A* **50**, 381-393.
- [3] Brown, L.D. and Low, M.G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24**, 2524-2535.
- [4] Cai, T. and Low, M. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.* **32**, 552 - 576.
- [5] Cai, T. and Low, M. (2005). Non-quadratic estimators of a quadratic functional. *Ann. Statist.* **33**, 2930-2956.
- [6] Cai, T. and Low, M. (2010). Estimation of excess mass and other related nonsmooth functionals. Manuscript.
- [7] Donoho, D.L. and Liu, R.C. (1991) Geometrizing Rates of Convergence II. *Ann. Statist.* **19**, 633-667.
- [8] Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the  $L_r$  norm of a regression function. *Probab. Theory Relat. Fields* **113**, 221-253.
- [9] Le Cam, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1**, 38-53.
- [10] Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
- [11] Qazi, M. A. and Rahman, Q. I. (2007). Some coefficient estimates for polynomials on the unit interval. *Serdica Mathematical Journal* **33**, 449-474.

- [12] Rivlin, T. J. (1990). *Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory*. Second edition. John Wiley & Sons, New York.
- [13] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag, New York.
- [14] Varga, R. S. and Carpenter, A. J. (1987). On a conjecture of S. Bernstein in approximation theory. *Math. USSR Sbornik* **57**, 547-560.
- [15] Wang, L., Brown, L. D., Cai, T. and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *Ann. Statist.* **36**, 646-664.