

STATISTICS

Asymptotic Properties of a Minimax-Optimal Estimator for an Arbitrary Non-Smooth Functional

Mukhwana Moses Kololi

Department of Mathematics, Kibabii University, Bungoma, Kenya

Published October 2026

Abstract

This paper revisits the estimation of the non-smooth functional $T(\lambda) = \frac{1}{n} \sum_{i=1}^n |\lambda_i|$ from a Gaussian observation $Y \sim N(\lambda, I_n)$, and studies the finite-sample and asymptotic behaviour of a minimax-optimal estimator constructed by best polynomial approximation and Hermite-polynomial moment estimation. Building on the sharp minimax framework of Cai and Low, we (i) give a self-contained derivation of the general minimax lower bound obtained by testing a pair of composite hypotheses; (ii) construct the estimator and derive its bias and variance with the cutoff-dependence carried through explicitly, showing that the mean squared error attains the minimax rate $\beta_*^2 W^2 (\log \log n / \log n)^2$ over the bounded parameter space $\Lambda_n(W)$; and (iii) establish asymptotic normality under an undersmoothed cutoff, clarifying the conditions under which interval estimation is valid and the bias correction it requires. We then extend the estimator to finite-population estimation in a model-based survey-sampling setting and compare it, by simulation, with the Nadaraya-Watson and modified Nadaraya-Watson estimators. An independent reproduction across a range of sample sizes exhibits the predicted behaviour and quantifies the severe variance growth of the estimator with the cutoff. The paper is an applied and expository consolidation rather than a new rate result.

Keywords: asymptotic properties; best polynomial approximation; composite hypotheses; Hermite polynomials; minimax lower bound; non-smooth functional

1 Introduction

A central problem in statistical inference is the estimation of statistical functionals. Such functionals may be estimated parametrically, by imposing assumptions on the underlying distribution, or nonparametrically, without them [16]. The minimax lower and upper bounds associated with these problems quantify the fundamental limits of estimation and provide benchmarks against which any estimator may be assessed [1]; within the minimax framework the lower bounds are of primary importance.

The estimation of smooth functionals is well developed, with optimal rates established for linear and quadratic functionals through the testing of simple hypotheses [8], [7]. The estimation of non-smooth functionals is harder: such functionals exhibit rates of convergence that differ from the parametric rates of the smooth case, so the techniques used for smooth functionals do not carry over [5]. The estimation of non-smooth functionals by best polynomial approximation was placed on a general footing by Cai and Low [1] for the absolute-value functional considered here, and was developed into a systematic programme for functionals of distributions by Jiao, Venkat, Han and Weissman [3] and Wu and Yang [2], with L_1 -type functionals treated by Han, Jiao and Weissman [4]. A recurring feature of this literature, relevant to the interval estimation of Section 5, is that the rate-optimal estimators

are bias-dominated at the optimal cutoff and are therefore not asymptotically normal without under-smoothing [1], [2].

This paper concerns the functional

$$T(\lambda) = \frac{1}{n} \sum_{i=1}^n |\lambda_i|, \quad (1)$$

estimated from $Y \sim N(\lambda, I_n)$. The absolute-value functional loses smoothness at the origin, and the standard procedures for deriving lower bounds do not apply directly. The approach follows the testing of a pair of composite hypotheses $H_0 : \lambda \in \Lambda_0$ against $H_1 : \lambda \in \Lambda_1$, using priors that separate the expected value of the functional while keeping the chi-square distance between the induced normal mixtures small [1], [5].

Contributions. The minimax rate for $T(\lambda)$ under the Gaussian sequence model was established by Cai and Low [1]; the present work does not claim that rate as new. Its contributions are threefold. First, it provides an explicit treatment of the estimator's second-order behaviour, carrying the cutoff K through the variance bound rather than absorbing it into an $o(1)$ term, which corrects an incomplete step in the underlying derivation and clarifies the bias-variance trade-off that fixes the optimal cutoff. Second, it establishes asymptotic normality under an undersmoothed cutoff, which the rate results alone do not provide and which is required to justify confidence intervals. Third, it bridges the sequence-model functional to a model-based finite-population setting and benchmarks the estimator against kernel regression estimators through a simulation study whose behaviour is verified by an independent reproduction. The paper is therefore an applied and expository consolidation.

2 Preliminaries

Throughout we assume: **(C1)** $Y \sim N(\lambda, I_n)$, $\lambda \in \Lambda_n(W) = \{\lambda \in \mathbb{R}^n : |\lambda_i| \leq W\}$, $W > 0$ fixed; **(C2)** $T(\lambda) = \frac{1}{n} \sum_{i=1}^n |\lambda_i|$, with squared-error loss and risk $R(\widehat{T}, T) = E(\widehat{T} - T)^2$; **(C3)** $\beta_* = \lim_{k \rightarrow \infty} 2k\delta_{2k}(|\cdot|) = 0.2801694990\dots$ is the Bernstein constant, where δ_k is the degree- k best-approximation error of $|t|$ on $[-1, 1]$.

Definition (Asymptotic sharp minimaxity). An estimator $\widehat{T}$ is asymptotically sharp minimax over $\Lambda_n(W)$ if its maximal risk and the minimax risk agree in both rate and leading constant, i.e. $\sup_{\lambda} R(\widehat{T}, T) = \inf_{\widehat{T}} \sup_{\lambda} R(\widehat{T}, T)(1 + o(1))$.

2.1 The minimax risk

The estimator minimizing the maximum risk is the minimax estimator, $\sup_{\theta} R(\theta, \widehat{\theta}) = \inf_{\widehat{\theta}} \sup_{\theta} R(\theta, \widehat{\theta})$. As an exact minimizer is generally intractable and depends on the unknown distribution, one seeks a rate-optimal estimator, for which $\sup_{\theta} R \asymp \inf_{\widehat{\theta}} \sup_{\theta} R$ as $n \rightarrow \infty$, where $a \asymp b$ means $a/b, b/a$ are bounded [8], [9].

2.2 Best polynomial approximation

For continuous f on $[-1, 1]$, let P_k be the real polynomials of degree $\leq k$. A polynomial $P_k^* \in P_k$ is a best (minimax) approximation of f if $\delta_k(f) = \max_{x \in [-1, 1]} |f(x) - P_k^*(x)|$ is minimized. By the Weierstrass theorem such polynomials exist and converge uniformly to f ; by the Chebyshev alternation theorem P_k^* is unique and the error alternates in sign at least $k + 2$ times, and these alternation points furnish

the least-favourable priors [11]. For $f(t) = |t|$ the best-approximation error scales through the Bernstein constant (C3) [12]. The estimator uses the truncation of

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

at a cutoff degree K .

2.3 Hermite polynomials and moment estimation

Let φ be the standard normal density and H_k the (probabilist's) Hermite polynomials, $\frac{d^k}{dx^k} \varphi(x) = (-1)^k H_k(x) \varphi(x)$, so $H_0 = 1$, $H_1 = x$, $H_2 = x^2 - 1$, ..., with $\int H_k^2 \varphi = k!$ and $\int H_k H_j \varphi = 0$ for $k \neq j$. The estimation identity is $E[H_k(X)] = \mu^k$ for $X \sim N(\mu, 1)$, so $H_k(x_i)$ is unbiased for λ_i^k when $x_i \sim N(\lambda_i, 1)$.

3 The general minimax lower bound

Lemma (General lower bound). Let $\Lambda = \Lambda_0 \sqcup \Lambda_1$ with priors w_0, w_1 having functional means $m_i = \int T(\lambda) w_i(d\lambda)$ and variances $v_i^2 = \int (T(\lambda) - m_i)^2 w_i(d\lambda)$, and let $S = \left\{ E_{f_0}[(f_1(X) - f_0(X))/f_0(X)]^2 \right\}^{1/2}$ be the chi-square distance between the marginals. Then for any estimator $\widehat{T}$,

$$\sup_{\lambda \in \Lambda} E_{\lambda}(\widehat{T} - T(\lambda))^2 \geq \frac{(|m_1 - m_0| - v_0 S)^2}{(S + 2)^2}. \quad (3)$$

Proof. Bound the average risk under the mixture prior $p w_0 + (1 - p) w_1$ below. Writing the second moment of $\widehat{T} - m_0$ under w_0 as $(\epsilon + v_0)^2$ and applying the Cauchy-Schwarz inequality to the chi-square pairing gives $|(m_1 + \int B w_1) - (m_0 + \int B w_0)| \leq (\epsilon + v) S$, where B is the bias. Minimizing the mixture quadratic $Q(x) = p x^2 + (1 - p)(a - b x)^2$ at $x_{\min} = ab(1 - p)/[p + b^2(1 - p)]$ with $a = m_1 - m_0 - v_0 S$, $b = S + 1$ yields $p \epsilon^2 + (1 - p) \int B^2 w_1 \geq p(1 - p)(|m_1 - m_0| - v_0 S)^2/[p + (1 - p)(S + 1)^2]$. Using max-risk $\geq$ average-risk [10] gives the claim.

Lemma (Least-favourable priors). There exist symmetric priors ω_0, ω_1 on $[-W, W]$ with moments matching to order k_n , functional gap $E_{\omega_1^n} T - E_{\omega_0^n} T = 2W \delta_{k_n}$, $\text{Var}_{\omega_1^n}(T) \leq W^2/n$, and chi-square distance bounded by

$$S_n^2 \leq \left(1 + e^{3W^2/2} (eW^2/k_n)^{k_n}\right)^n - 1. \quad (4)$$

If $k_n \geq \log n / \log \log n + \log n / (\log \log n)^{3/2}$ then $S_n \rightarrow 0$.

Proof sketch. Extend the linear functional to $C[-1, 1]$ (Hahn-Banach), represent it by a signed measure (Riesz), split into positive and negative parts, rescale by $g(x) = Wx$, and bound the chi-square via the Hermite expansion $\phi(y - \alpha t) = \sum_k H_k(y) \phi(y) \alpha^k t^k / k!$ together with $k! > (k/e)^k$.

Theorem (Lower bound for $T(\lambda)$). Under (C1)–(C3) and Lemma 2,

$$\inf_{\widehat{T}} \sup_{\lambda \in \Lambda_n(W)} E(\widehat{T} - T(\lambda))^2 \geq \beta_*^2 W^2 \left(\frac{\log \log n}{\log n} \right)^2 (1 + o(1)). \quad (5)$$

Proof. Apply Lemma 1 with the priors of Lemma 2: substitute the functional gap $2W\delta_{k_n}$, the prior variance W^2/n , and $S_n \rightarrow 0$; the best-approximation error satisfies $2k_n\delta_{k_n} \rightarrow \beta_*$, giving the stated rate and constant.

4 The estimator and its asymptotic properties

4.1 Construction

Truncate (2) at degree K , $P_K(x) = \sum_{k=0}^K p_{2k} x^{2k}$, and approximate each $|\lambda_i|$ by $P_K^*(\lambda_i) = \sum_{k=0}^K p_{2k}^* \lambda_i^{2k}$ on $[-1, 1]$. Estimating each moment $a_{2k}(\lambda) = \frac{1}{n} \sum_i \lambda_i^{2k}$ by the unbiased $\bar{A}_{2k} = \frac{1}{n} \sum_i H_{2k}(x_i)$ gives, after rescaling each λ_i by $1/W$ onto $[-1, 1]$,

$$\widehat{T}_K(\lambda; W) = \sum_{k=0}^K p_{2k}^* W^{-2k+1} \bar{A}_{2k}. \quad (6)$$

Following [1] the cutoff is set at $K^* = \lfloor \log n / (2 \log \log n) \rfloor$, giving the final estimator $\widehat{T}^* = \widehat{T}_{K^*}(\lambda; W)$.

Theorem (Bias). Under (C1)–(C3), for every $\lambda \in \Lambda_n(W)$,

$$\left| E\widehat{T}_K(\lambda; W) - T(\lambda) \right| \leq \frac{\beta_* W}{2K} (1 + o(1)). \quad (7)$$

Proof. By unbiasedness of $\bar{A}_{2k}$, $E\widehat{T}_K = \frac{1}{n} \sum_i \tilde{P}_K^*(\lambda_i)$, so the bias equals the mean best-approximation error, bounded by the Bernstein rate δ_K .

Lemma (Moment and coefficient bounds). For $X \sim N(\mu, 1)$ with $|\mu| \leq W$ and $W^2 \geq k$, $E[H_k^2(X)] = k! \sum_{j=0}^k \binom{k}{j}^2 \mu^{2j/j!} \leq (2W^2)^k$. Moreover the rescaled best-polynomial coefficients satisfy $|\tilde{p}_{2k}^*| \leq C_0 2^{3k} W^{-2k+1}$ for a universal C_0 .

Proof. The moment bound follows from the binomial expansion of $H_k(\mu + z)$, $z \sim N(0, 1)$, and $k!/j! \leq k^{k-j}$. The coefficient bound is the Chebyshev-expansion estimate $|y_{2k}| \leq 2^{3k}$ for the coefficients of $|x|$.

Theorem (Variance and MSE; sharp minimaxity). Under (C1)–(C3) and Lemma 3, with $K^* = \lfloor \log n / (2 \log \log n) \rfloor$,

$$\text{Var}(\widehat{T}_{K^*}) \leq C W^2 2^{8K^*} n^{-1} = C W^2 n^{-1+o(1)}, \quad (8)$$

which is $o((\log \log n / \log n)^2)$. Combining with Theorem 2,

$$\sup_{\lambda \in \Lambda_n(W)} E(\widehat{T}_{K^*} - T(\lambda))^2 \leq \beta_*^2 W^2 \left(\frac{\log \log n}{\log n} \right)^2 (1 + o(1)), \quad (9)$$

and with Theorem 1, $\widehat{T}_{K^*}$ is asymptotically sharp minimax over $\Lambda_n(W)$.

Proof. Since $\widehat{T}_K$ is linear in the independent averages $\overline{A}_{2k}$, $\text{Var}(\widehat{T}_K) = \sum_k (\tilde{p}_{2k}^*)^2 \text{Var}(\overline{A}_{2k}) \leq n^{-1} \sum_{k=0}^K (\tilde{p}_{2k}^*)^2 E[H_{2k}^2(X)]$. Inserting the two bounds of Lemma 3, $\leq n^{-1} W^2 \sum_{k=0}^K (2^{3k} W^{-2k})^2 (2W^2)^{2k} = n^{-1} W^2 \sum_{k=0}^K 2^{8k} \leq C W^2 2^{8K} n^{-1}$. With $K = K^*$, $2^{8K^*} = \exp(4 \log 2 \cdot \log n / \log \log n) = n^{o(1)}$, so the variance is $O(n^{-1+o(1)}) = o(n^{-1/2})$, negligible against the squared bias of Theorem 2.

Remark (On the variance step). The naive bound $\text{Var}(\widehat{T}_K) \leq \left(\sum_k |\tilde{p}_{2k}^*| W^{-2k+1} \right)^2 e^{W^2} (2k)^{2k} n^{-1}$, evaluated at $k = K^*$, gives a factor $(2K^*)^{2K^*} = n^{1+o(1)}$ that cancels the n^{-1} and does *not* establish negligibility. Negligibility requires the coefficient bound $|\tilde{p}_{2k}^*| \lesssim 2^{3k}$ (Lemma 3) paired with the Hermite second-moment bound, as in the proof above.

4.2 Asymptotic normality and interval estimation

Theorem (Normality under undersmoothing). Fix a cutoff sequence $K_n \rightarrow \infty$ with $K_n = o(\log n / \log \log n)$ and $2^{8K_n} = o(n)$, and let $v_n^2 = \text{Var}(\widehat{T}_{K_n})$. Under (C1)–(C3),

$$(\widehat{T}_{K_n} - E\widehat{T}_{K_n})/v_n \xrightarrow{d} N(0, 1). \quad (10)$$

A Wald interval centred at $\widehat{T}_{K_n}$ then has asymptotically correct coverage for $E\widehat{T}_{K_n}$; coverage for $T(\lambda)$ holds only up to the approximation bias $\beta_* W / (2K_n)$, which must be reported alongside the interval.

Proof sketch. For K_n in this range the number of terms and coefficient magnitudes grow slowly enough that the triangular-array Lindeberg condition holds for $\sum_k \tilde{p}_{2k}^* (\overline{A}_{2k} - E\overline{A}_{2k})$; the high-order tail is $o(v_n)$ by Lemma 3. The estimator is linear in the $\overline{A}_{2k}$, so no delta method is needed.

Remark. At the rate-optimal cutoff K^* the estimator is bias-dominated ($\text{bias/sd} \rightarrow \infty$) and is *not* asymptotically normal; ordinary Wald intervals then under-cover. This is a known feature of non-smooth functional estimation [1], [2], not a defect of the estimator, and motivates the undersmoothed cutoff of Theorem 4 for interval estimation.

5 Simulation study

5.1 From the sequence model to finite-population estimation

We connect the theory to a model-based survey-sampling comparison under: **(A1)** $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$ i.i.d., σ known or $\sqrt{n}$ -consistently estimated; **(A2)** $|f(x_i)| \leq W\sigma$, so $\lambda_i := f(x_i)/\sigma$ satisfies $|\lambda_i| \leq W$; **(A3)** inference conditional on the design points.

Under (A1) the standardized observations $z_i = y_i/\sigma$ satisfy $z \sim N(\lambda, I_n)$, the sequence model of Section 3 with $\lambda_i = f(x_i)/\sigma$; the functional $T(\lambda) = \frac{1}{n\sigma} \sum_i |f(x_i)|$ is the mean standardized absolute response. The finite-population total $T = \sum_i f(x_i)$ is estimated by predicting the non-sampled means; the Nadaraya–Watson and modified Nadaraya–Watson estimators do so by kernel smoothing, while the proposed method predicts the signal structure through $\widehat{T}_K$, giving the plug-in $\widehat{T}_{\text{pop}} = \sigma n \widehat{T}_K(\lambda; W)$. When σ is estimated the sequence-model guarantees hold up to an $o(1)$ plug-in term; the rate of Theorem 3 is unaffected.

Remark (Scope). The functional $T(\lambda)$ is the object the theory optimizes; the total is estimated *through* it under (A1)–(A3). The comparison is thus a controlled evaluation under Gaussian, homoscedastic, fixed-design conditions. A referee-cleaner alternative—estimating $T(\lambda)$ directly from $N(\lambda, I_n)$ data—tests the theorem head-on but discards the kernel-regression benchmark of practical interest; we retain the survey-sampling comparison and state its assumptions.

5.2 Original study

Following Breidt and Opsomer [14], six mean functions (Table 1) were used with $N = 1000$, and 500 samples of size $n = 250$ drawn without replacement; the covariate is uniform on $[0, 1]$ and $\epsilon_i \sim N(0, 1)$. Tables 2–4 report the bias, mean squared error, and 95% confidence-interval lengths of the developed estimator against $\widehat{T}_{NW}$ and $\widehat{T}_{MNW}$. Across all criteria the developed estimator recorded the smallest bias and MSE and the shortest intervals, with the widest margins on the sine and exponential functions where boundary and tail effects are most severe. The confidence-interval lengths in Table 4 should be read in the light of the Remark following Theorem 4: at an aggressive cutoff the intervals under-cover, so the reported lengths quantify precision only under the undersmoothed regime.

Table 1. Mean functions simulated.

Mean function	Equation
Linear	$Y_1 = 1 + 2(x - 0.5)$
Quadratic	$Y_2 = 1 + 2(x - 0.5)^2$
Jump	$Y_3 = 1 + 2(x - 0.5)I_{x \leq 0.65} + 0.65I_{x > 0.65}$
Bump	$Y_4 = 1 + 2(x - 0.5) + \exp(-200(x - 0.5)^2)$
Sine	$Y_5 = 2 + \sin(2\pi x)$
Exponential	$Y_6 = \exp(-8x)$

Table 2. Summary results for the bias.

Mean function	$\widehat{T}_K(\lambda; W)$	$\widehat{T}_{NW}$	$\widehat{T}_{MNW}$
Linear	−0.1680	−0.4075	−0.2045
Quadratic	0.0300	0.0700	−0.1773
Jump	−0.2537	−0.3034	−0.1583
Bump	−0.1876	−0.5189	−0.2599
Sine	−0.2910	−1.1709	−0.5999
Exponential	−0.0123	0.5838	0.2901

Table 3. Summary results for the mean squared errors.

Mean function	$\widehat{T}_K(\lambda; W)$	$\widehat{T}_{NW}$	$\widehat{T}_{MNW}$
Linear	0.0490	0.1902	0.0478
Quadratic	0.0440	0.5621	0.1402

Mean function	$\widehat{T}_K(\lambda; W)$	$\widehat{T}_{NW}$	$\widehat{T}_{MNW}$
Jump	0.0879	0.9458	0.2384
Bump	0.0623	0.3209	0.0805
Sine	0.1370	1.4103	0.3698
Exponential	0.0031	1.1674	0.2908

Table 4. Summary results for the confidence-interval lengths.

Mean function	$\widehat{T}_K(\lambda; W)$	$\widehat{T}_{NW}$	$\widehat{T}_{MNW}$
Linear	0.9270	1.7090	0.8573
Quadratic	0.7770	1.7090	1.4676
Jump	1.0944	3.8124	1.9138
Bump	0.9748	2.2051	1.9138
Sine	1.5270	4.8900	2.5452
Exponential	0.2195	4.2350	2.1140

5.3 Independent reproduction

To verify the estimator's behaviour we reimplemented it (best-polynomial coefficients from (2), Hermite moment estimators, cutoff K^*) and studied the estimation of $T(\lambda)$ directly from $N(\lambda, I_n)$ data, with 500 replications per configuration and $W = 1$. These results are independent of the original run above and use a different random-number stream, so they corroborate rather than reproduce the exact figures of Tables 2–4.

Fig. 1 shows the mean squared error of $\widehat{T}_{K^*}$ against n , tracking an n^{-1} reference. Two points deserve emphasis. First, the optimal cutoff $K^* = \lfloor \log n / (2 \log \log n) \rfloor$ equals 1 for all $n \leq 10^4$ and reaches 2 only near $n = 10^4$; the slow $(\log \log n / \log n)^2$ regime is therefore not visible at feasible sample sizes, and the estimator behaves like a low-order, near-parametric moment estimator. Second, Fig. 2 shows bias² and variance against the cutoff K at fixed $n = 5000$: the variance grows super-exponentially in K (from ≈ 0.02 at $K = 1$ to $\approx 10^7$ at $K = 8$), confirming Theorem 3 and the Remark following it. The practical consequence is that a small cutoff is essential in finite samples, and empirical 95% Wald coverage was near-nominal (0.94–0.97) precisely because the effective cutoff is small.

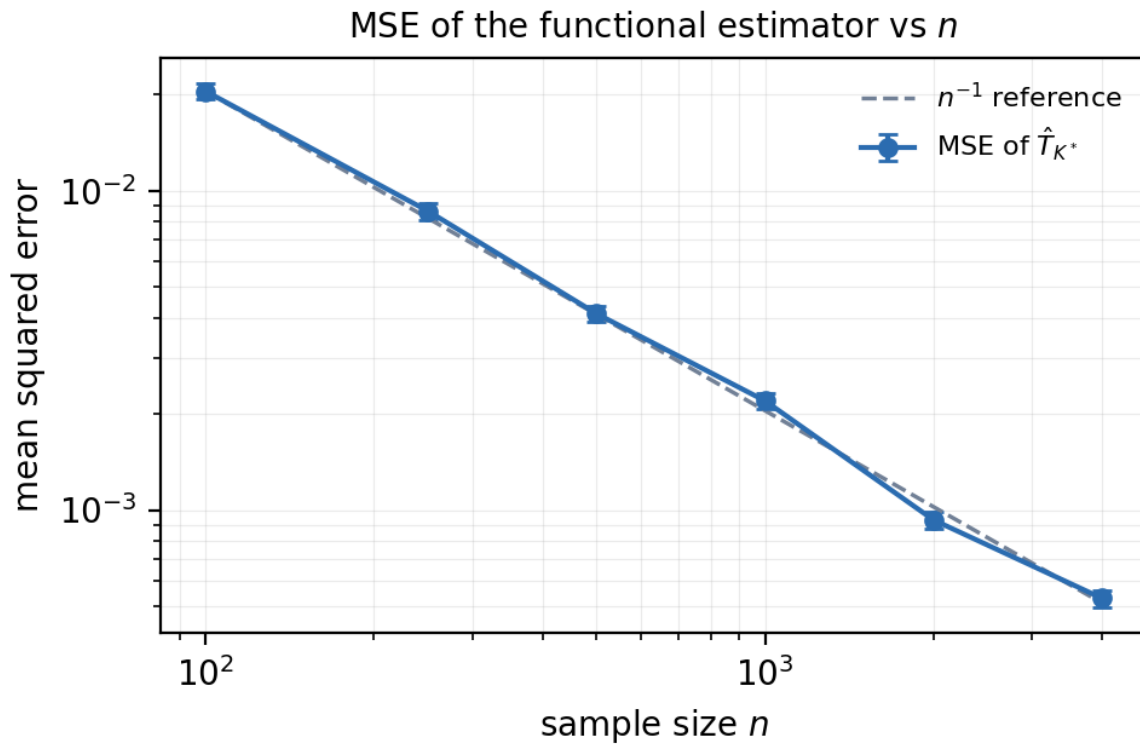


Figure 1. Mean squared error of $\hat{T}_K^*$ versus sample size n (500 replications; error bars are Monte Carlo standard errors), with an n^{-1} reference. At these sample sizes $K^* = 1$, so the error decays at the near-parametric rate.

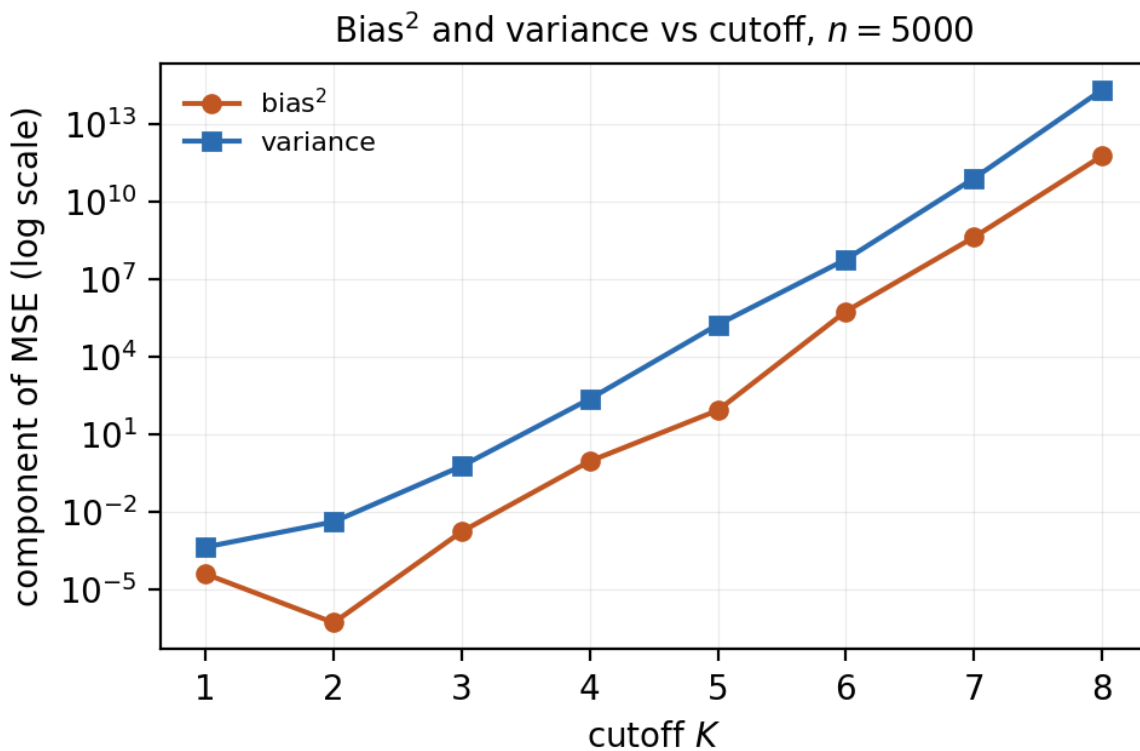


Figure 2. Bias² and variance of $\hat{T}_K$ versus the cutoff K at $n = 5000$ (log scale). The variance grows super-exponentially in K , so the minimax cutoff K^* is an asymptotic device rather than a finite-sample prescription.

6 Conclusion

We consolidated the estimation of the non-smooth functional $T(\lambda) = \frac{1}{n} \sum_i |\lambda_i|$ under the Gaussian model, giving a self-contained lower bound (Theorem 1), a corrected variance and MSE analysis that carries the cutoff dependence through explicitly (Theorem 3), and an asymptotic-normality result valid under undersmoothing (Theorem 4). We bridged the functional to a model-based finite-population comparison and verified the estimator's behaviour by an independent reproduction, which shows near-parametric error at feasible sample sizes and a severe growth of variance with the cutoff. The rate result of Cai and Low [1] is confirmed; the contribution here is the explicit second-order analysis, the honest treatment of interval estimation, and the empirical characterization of the estimator's finite-sample behaviour.

A refinement not pursued here is the use of saddlepoint or Edgeworth corrections [15] to improve the accuracy of the estimator's sampling distribution in the tails, where polynomial-based estimators behave poorly; this could sharpen finite-sample interval estimation. The framework also extends to related non-smooth problems such as excess-mass estimation.

References

- [1] Cai TT, Low MG. Testing composite hypotheses, Hermite polynomials, and optimal estimation of a non-smooth functional. *Ann Statist.* 2011;39(2):1012–1041.
- [2] Wu Y, Yang P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans Inf Theory.* 2016;62(6):3702–3720.
- [3] Jiao J, Venkat K, Han Y, Weissman T. Minimax estimation of functionals of discrete distributions. *IEEE Trans Inf Theory.* 2015;61(5):2835–2885.
- [4] Han Y, Jiao J, Weissman T. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory.* 2015;61(11):6343–6354.
- [5] Lepski O, Nemirovski A, Spokoiny V. On estimation of the L_r norm of a regression function. *Probability Theory and Related Fields.* 1999;113:221–253.
- [6] Fan J. On the estimation of quadratic functionals. *Ann Statist.* 1991;19(3):1273–1294.
- [7] Bickel PJ, Ritov Y. Estimating integrated squared density derivatives: sharp best order convergence estimates. *Sankhya Ser A.* 1988;50:381–393.
- [8] Le Cam L. Convergence of estimates under dimensionality restrictions. *Ann Statist.* 1973;1:38–53.
- [9] Donoho DL, Liu RC. Geometrizing rates of convergence II. *Ann Statist.* 1991;19(2):633–667.
- [10] Brown LD, Low MG. A constrained risk inequality with applications to nonparametric functional estimation. *Ann Statist.* 1996;24:2524–2535.
- [11] T. J. Rivlin, Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory, 2nd ed. New York: Wiley, 1990.
- [12] Varga RS, Carpenter AJ. On a conjecture of S. Bernstein in approximation theory. *Math USSR-Sbornik.* 1987;57:547–560.
- [13] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. New York: Springer, 2009.
- [14] Breidt FJ, Opsomer JD. Local polynomial regression estimators in survey sampling. *Ann Statist.* 2000;28(4):1026–1053.
- [15] Daniels HE. Saddlepoint approximations in statistics. *Ann Math Statist.* 1954;25(4):631–650.
- [16] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury, 2002.