entropy

MDPI

Article

Finite-Sample Bounds on the Accuracy of Plug-In Estimators of Fisher Information

Wei Cao [1], Alex Dytso [2,*], Michael Fauß [3] and H. Vincent Poor [3]

[1] National Key Lab of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China; clarissa.cao@hotmail.com
[2] Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102, USA
[3] Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544, USA; mfauss@princeton.edu (M.F.); poor@princeton.edu (H.V.P.)
* Correspondence: alex.dytso@njit.edu

Abstract: Finite-sample bounds on the accuracy of Bhattacharya's plug-in estimator for Fisher information are derived. These bounds are further improved by introducing a clipping step that allows for better control over the score function. This leads to superior upper bounds on the rates of convergence, albeit under slightly different regularity conditions. The performance bounds on both estimators are evaluated for the practically relevant case of a random variable contaminated by Gaussian noise. Moreover, using Brown's identity, two corresponding estimators of the minimum mean-square error are proposed.

Keywords: nonparametric estimation; Fisher information; MMSE; kernel estimation

Citation: Cao, W.; Dytso, A.; Fauß, M.; Poor, H.V. Finite-Sample Bounds on the Accuracy of Plug-In Estimators of Fisher Information. *Entropy* **2021**, *23*, 545. https://doi.org/10.3390/e23050545

Academic Editor: Osvaldo Anibal Rosso

Received: 15 March 2021
Accepted: 21 April 2021
Published: 28 April 2021

1. Introduction

This work considers the problem of estimating the Fisher information for the location of a univariate probability density function (PDF) f based on n random samples $Y_1, \ldots, Y_n$ independently drawn from f. To clarify, the Fisher information of a differentiable density function f is given by

$$I(f) = \int_{\{f(t)>0\}} \frac{(f'(t))^2}{f(t)} \mathrm{d}t, \tag{1}$$

where f' is the derivative of f. For the remainder of the paper, it is assumed that $\{f(t) > 0\} = \mathbb{R}$, but an extension to the general case is not difficult. The paper considers plug-in estimators based on kernel density estimates of f. That is, the Fisher information is estimated by plugging a kernel density estimate of f into the right-hand side of (1).

Estimation of the Fisher information in (1) via a plug-in estimator based on kernel density estimates was first considered by Bhattacharya in [1]. Bhattacharya showed that, under mild conditions on f, the plug-in estimator is consistent for a large class of kernels, and he provided bounds on its accuracy in the large (asymptotic) sample regime. These bounds were later revised and improved by Dmitriev and Tarasenko in [2]. However, to the best of our knowledge, no finite-sample regime bounds on the accuracy of Bhattacharya's estimator can be found in the literature. The paper aims at closing this gap.

Bounds on the accuracy of plug-in estimators rely on bounds on the accuracy of the underlying density estimators. For kernel-based density estimators, such bounds have received considerable attention in the literature. For example, Schuster [3] showed that, under mild regularity conditions, the estimation error for higher-order derivatives can be controlled by the estimation error for the corresponding cumulative distribution function (CDF). The interested reader is referred to [4–8] and the references therein. In this paper, as a preliminary result for the analysis of Bhattacharya's estimator, the bounds in [3] are further tightened by replacing some suboptimal constants with the optimal ones.

A problem that arises in the performance analysis of plug-in estimators for Fisher information is that the score function of the estimated density, that is, the ratio of the derivative of the PDF and the PDF itself, is hard to bound, especially in the tails. Bhattacharya worked around this problem by truncating the integration range in (1), thus avoiding evaluation of the estimated score function on these critical regions. However, in order for the estimator to stay consistent, this truncation has to be done rather aggressively so that the error introduced by ignoring the tails can outweigh the approximation error introduced by the density estimate. In this paper, we propose a simple remedy that allows for a much less aggressive truncation of the integration range and, in turn, for significantly tighter bounds on the approximation error. Namely, we propose the clipping of the score function whenever it exceeds a suitably chosen upper bound. In the vast majority of cases, the corresponding clipped estimates of Fisher estimation are identical to their non-clipped counterparts, meaning that the clipping has a negligible influence on the estimation accuracy. However, the knowledge that extreme values of the score would have been clipped, had they occurred, allows for much-improved performance guarantees.

It should be explicitly stated that this paper does not address the question of how best to estimate Fisher information. Although this question is highly interesting and relevant, it is far beyond the scope of this work. In addition, it is not the aim of the paper to compare the plug-in estimator to alternative estimators for the Fisher information or to claim that it provides superior result. A variety of well-motivated parametric and nonparametric Fisher information estimators have been proposed in the literature; see, for example, [9–11] and the references therein. However, comparing and contrasting these estimators in a fair manner is not straightforward and arguably constitutes a research question in its own right. Finally, the problem of obtaining estimator-independent bounds on the sample complexity of Fisher information falls under the umbrella of estimation of nonlinear functionals; see, for example, [12]. Most of the commonly used information measures, such as entropy, relative entropy, and mutual information, are nonlinear functionals, and their estimation has recently received considerable attention; the interested reader is referred to [13–17] and the references therein.

Despite its limited scope, we are convinced that the work presented in this paper is useful in a wider context. First, from a theoretical point of view, it strengthens some classic results in nonparametric estimation and, as explained above, provides bounds for the finite-sample regime, thus filling a gap in the literature. Second, from a practical perspective, the Fisher information typically provides useful bounds or limits on the estimation error (e.g., the well-known Cramér–Rao lower bound), but is not in itself the quantity of interest—an exception is the case of estimating a random signal in additive Gaussian noise, where the minimum mean square error (MMSE) and other relevant quantities can be expressed in term of the Fisher information. The problem of estimating Fisher information also arises in image processing, model selection, experimental design, and many more areas. Applications of our results include, for example, to provide the Cramér–Rao bound and, for the case of a random variable in additive Gaussian noise, to address the power allocation problem [18]. These connections will be discussed in more detail in Section 4. Most often, however, Fisher information plays the role of side information, and its estimation does not warrant investing large computational resources. This prevents the use of sophisticated estimators, which require solving non-trivial optimization problems. In contrast, kernel density estimates are relatively easy to compute and have been widely used in nonparametric statistics so that efficient implementations in software or even hardware [19] are readily available. Hence, for the foreseeable future, plug-in estimators are bound to remain a common and often the only viable option for estimating Fisher information in practice.

The paper is organized as follows: Section 2 revisits Bhattacharya's estimator. In particular, Theorem 1 provides explicit and tighter non-asymptotic bounds on its convergence rate, improving the results in [1,2]. Furthermore, Theorem 2 provides an alternative bound under the additional assumption that the density function is upper bounded within any given interval. The explicit non-asymptotic results enable us to see that the sample

complexity of Bhattacharya's estimator is considerable and that the potentially unbounded score function is a critical bottleneck for tighter bounds. Section 3 proposes a "harmless" modification of Bhattacharya's estimator, namely, a clipping of the estimated score function, which is shown to be sufficient to remedy its large sample complexity. In particular, Theorem 3 shows that the clipped estimator has significantly better bounds on rates of convergence, albeit with slightly different assumptions on the PDF. Section 4 evaluates the convergence rates of the two estimators for the practically relevant case of a random variable contaminated by additive Gaussian noise. Moreover, using Brown's identity, which relates the Fisher information and the MMSE, consistent estimators for the MMSE are proposed and their rates of convergence are evaluated in Proposition 1. Section 5 concludes the paper.

Notation

The expected value and variance of a random variable X are denoted by $\mathbb{E}[X]$ and $\mathrm{Var}(X)$, respectively. The gamma function is denoted by $\Gamma(\cdot)$. Estimators of a PDF f based on n samples are denoted by f_n. No notational distinction is made between an estimator, which is a random variable, and its realizations (estimates), which are deterministic. However, the difference will be clear from the context or will be highlighted explicitly otherwise. The nth derivative of a function $F\colon \mathbb{R} \to \mathbb{R}$ is denoted by $F^{(n)}$; the first-order derivative is also denoted by F' to improve readability.

2. Bhattacharya's Estimator

In this section, we revisit the asymptotically consistent estimator proposed by Bhattacharya in [1] and produce explicit and non-asymptotic bounds on its accuracy.

Bhattacharya's estimator is given by

$$I_n = \int_{-k_n}^{k_n} \frac{(f_n'(t))^2}{f_n(t)}\,\mathrm{d}t, \tag{2}$$

where $k_n \geq 0$ determines the integration interval as a function of the sample size n and the unknown functions f and f' are replaced by their kernel estimates, that is,

$$f_n(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_0} K\left(\frac{t - Y_i}{a_0}\right), \tag{3}$$

$$f_n'(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_1} K'\left(\frac{t - Y_i}{a_1}\right). \tag{4}$$

Here, $a_0, a_1 > 0$ are bandwidth parameters, and $K\colon \mathbb{R} \to \mathbb{R}$ denotes the kernel, which is assumed to satisfy certain regularity conditions that will be discussed later in this section.

2.1. Estimating a Density and Its Derivatives

In order to analyze plug-in estimators, it is necessary to obtain rates of convergence for f_n and f_n', that is, the kernel estimators of the density and its derivative. The following result, which is largely based on the proof by Schuster in [3], provides such rates. The proof in [3] makes use of the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality for the empirical CDF. The next lemma refines the result in [3] by using the best possible constant for the DKW inequality shown in [20].

Lemma 1. *Let $r \in \{0, 1\}$ and*

$$v_r = \int_{-\infty}^{\infty} \left|K^{(r+1)}(t)\right| \mathrm{d}t, \tag{5}$$

$$\delta_{r,a_r} = \sup_{t\in\mathbb{R}} \left|\mathbb{E}\left[f_n^{(r)}(t)\right] - f^{(r)}(t)\right|. \tag{6}$$

Then, for any $\epsilon > \delta_{r,a_r}$ and any $n \geq 1$, the following bound holds:

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n^{(r)}(t) - f^{(r)}(t)\right| > \epsilon\right] \leq 2\mathrm{e}^{-2n\frac{a_r^{2r+2}(\epsilon-\delta_{r,a_r})^2}{v_r^2}}. \tag{7}$$

Proof. See Appendix A. □

2.2. Analysis of Bhattacharya's Estimator

The following theorem is a non-asymptotic refinement of the result obtained by Bhattacharya in Theorem 3 of [1] and Dmitriev and Tarasenko in Theorem 1 of [2].

Theorem 1. *Assume that there exists a function $\phi\colon \mathbb{R} \to \mathbb{R}$ such that*

$$\sup_{|t|\leq x}\frac{1}{f(t)} \leq \phi(x), \quad \forall x \in \mathbb{R}. \tag{8}$$

Then, provided that

$$\sup_{|t|\leq k_n}\left|f_n^{(r)}(t) - f^{(r)}(t)\right| \leq \epsilon_r, \quad r \in \{0,1\}, \tag{9}$$

and

$$\epsilon_0\phi(k_n) < 1, \tag{10}$$

the following bound holds:

$$|I(f) - I_n| \leq \frac{4\epsilon_1 k_n \rho_{\max}(k_n) + 2\epsilon_1^2 k_n \phi(k_n) + \epsilon_0\phi(k_n)I(f)}{1 - \epsilon_0\phi(k_n)} + c(k_n), \tag{11}$$

where

$$\rho_{\max}(k_n) = \sup_{|t|\leq k_n}\left|\frac{f'(t)}{f(t)}\right|, \tag{12}$$

$$c(k_n) = I(f) - \int_{-k_n}^{k_n}\frac{(f'(t))^2}{f(t)}\,\mathrm{d}t. \tag{13}$$

Proof. See Appendix B. □

The bound in (11) is an improvement of the original bound in [1,2], which contains terms of the form $\epsilon_0\phi^4(k_n)$.

Note that $\phi(k_n)$ in (8) can be rapidly increasing with k_n. For example, as will be shown later, $\phi(k_n)$ increases super-exponentially with k_n for a random variable contaminated by Gaussian noise. This implies that, while Bhattacharya's estimator converges, the rate of convergence guaranteed by the bound in (11) is extremely slow. A modified bound is proposed in the subsequent theorem.

Theorem 2. *Assume that $f(t)$ is bounded on the interval $t \in [-k_n, k_n]$, i.e.,*

$$\sup_{|t|\leq k_n} f(t) \leq f_0, \tag{14}$$

for some $f_0 \in \mathbb{R}$. If the assumptions in (8), (9), and (10) hold, then

$$|I(f) - I_n| \leq \left[\epsilon_1\Big(4 + d_f(k_n) + d_{f_n}(k_n)\Big) + \epsilon_0\Big(2 + d_{f_n}(k_n)\Big)\rho_{\max}(k_n)\right]\psi(\epsilon_0, k_n) + c(k_n), \tag{15}$$

where $\rho_{\max}$ *and* c *are given by* (12) *and* (13)*, respectively,*

$$\psi(\epsilon_0, k_n) = \max\left(\log(f_0 + \epsilon_0), \log\left(\frac{\phi(k_n)}{1 - \epsilon_0\phi(k_n)}\right)\right), \tag{16}$$

and $d_g(k_n)$ *denotes the number of zeros of the derivative of the function* g *on the interval* $[-k_n, k_n]$*, i.e.,*

$$d_g(k_n) = \left|\left\{t \in [-k_n, k_n] : g'(t) = 0\right\}\right|. \tag{17}$$

Proof. See Appendix C. □

Remark 1. *Note that* ψ *in* (15) *is on the order of* $\log(\phi(k_n))$*, which typically increases much more slowly with* k_n *than* ϕ *in* (11)*. As a result, the bound in Theorem 2 can lead to a better bound on the convergence rate than that in Theorem 1, given appropriate upper bounds on* d_f *and* d_{f_n}*. Since Gaussian blurring of a univariate density function never creates new maxima, we have that* $d_{f_Y} \leq d_{f_X}$*, which is a constant. However, to the best of our knowledge, the only known upper bound on* d_{f_n} *is given by* $d_{f_n} \leq n$ *[21] (Theorem 2), which is not useful in practice. Despite this drawback, we include Theorem 2 for the sake of completeness and in the hope that tighter bounds on* d_{f_n} *might be established in the future.*

The main problem in the convergence analysis of the estimator in (2) is that $1/f_n(t)$ is only bounded if $f(t) > \epsilon_0$. For distributions with sub-Gaussian tails, this implies that the interval $[-k_n, k_n]$, on which this is guaranteed to be the case, grows sub-logarithmically (compare Theorem 4), causing the required number of samples to grow super-exponentially. In next section, we propose an estimator that has better guaranteed rates of convergence.

3. The Clipped Bhattacharya Estimator

In order to remedy the slow guaranteed convergence rates of Bhattacharya's estimator, we dispense with the tail assumption in (8), but introduce the new assumption that the unknown true score function $\rho(t) = f'(t)/f(t)$ is bounded (in absolute value) by a known function $\overline{\rho}$. This allows us to clip $f_n'(t)/f_n(t)$ and, in turn, $1/f_n(t)$ without affecting the consistency of the estimator.

Theorem 3. *Assume that there exists a function* $\overline{\rho}\colon \mathbb{R} \to \mathbb{R}$ *such that*

$$|\rho(t)| \leq |\overline{\rho}(t)|, \quad \forall t \in \mathbb{R} \tag{18}$$

and let

$$I_n^c = \int_{-k_n}^{k_n} \min\{|\rho_n(t)|\,,\, |\overline{\rho}(t)|\}\, |f_n'(t)|\, \mathrm{d}t, \tag{19}$$

where

$$\rho_n(t) = \frac{f_n'(t)}{f_n(t)}. \tag{20}$$

Under the assumptions in (9)*, it holds that*

$$|I(f) - I_n^c| \leq \max\left\{4\epsilon_1\Phi^1(k_n) + 2\epsilon_0\Phi^2(k_n) + c(k_n),\, 3\epsilon_1\Phi^1_{max}(k_n) + \epsilon_0\Phi^2_{max}(k_n)\right\} \tag{21}$$

$$\leq 4\epsilon_1\Phi^1_{max}(k_n) + 2\epsilon_0\Phi^2_{max}(k_n) + c(k_n), \tag{22}$$

where $c(k_n)$ *is defined in* (13) *and*

$$\Phi^m(x) = \int_{-x}^{x} |\rho^m(t)|\, \mathrm{d}t, \tag{23}$$

$$\Phi^m_{max}(x) = \int_{-x}^{x} |\overline{\rho}^m(t)|\, \mathrm{d}t. \tag{24}$$

In addition, if $f(t)$ is bounded as in (14), then

$$\Phi^m(k_n) \leq \min\left\{(2+d_f)\overline{\rho}^{m-1}(k_n)\psi(0,k_n)\,,\,\Phi^m_{max}(k_n)\right\}, \tag{25}$$

where ψ and d_f are defined in (16) and (17), respectively.

Proof. See Appendix D. □

For the upper-bound function $\overline{\rho}(t)$ in assumption (18), in practice, we can set $\overline{\rho}(k_n) = \rho_{\max}(k_n)$ if the latter is available. Although $\rho_{\max}(k_n)$ also increases with k_n, it usually increases much more slowly than $\phi(k_n)$. For example, as shown later, $\rho_{\max}(k_n)$ is linear in k_n in the Gaussian noise case. As a result, better bounds on the convergence rate can be shown for the clipped estimator.

4. Estimation of the Fisher Information of a Random Variable in Gaussian Noise

This section evaluates the results of Sections 2 and 3 for the important special case of a random variable contaminated by additive Gaussian noise. To this end, we let f_Y denote the PDF of a random variable

$$Y = \sqrt{\mathsf{snr}}X + Z, \tag{26}$$

where $\mathsf{snr} > 0$ is a signal-to-noise-ratio parameter, X is an arbitrary random variable, Z is a standard Gaussian random variable, and X and Z are independent. We are interested in estimating the Fisher information of f_Y. We only make the very mild assumption that X has a finite second moment, but otherwise, it is allowed to be an arbitrary random variable. We further assume that snr is known and that Gaussian kernels are used in the density estimators, i.e.,

$$K(t) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}. \tag{27}$$

The following lemma provides explicit expressions for the quantities appearing in Sections 2 and 3 that are needed to evaluate the error bounds for the Bhattacharya and the clipped estimator.

Lemma 2. *Let K be as in (27). Then,*

$$\delta_{r,a_r} \leq a_r \cdot \begin{cases} \frac{1}{\pi\sqrt{e}}, & r=0 \\ \frac{\frac{2}{\mathrm{e}}+1}{\pi}, & r=1, \end{cases} \tag{28}$$

$$v_r = \begin{cases} \sqrt{\frac{2}{\pi}}, & r=0 \\ \sqrt{\frac{2}{\mathrm{e}\pi}}, & r=1, \end{cases} \tag{29}$$

$$\rho_{\max}(k_n) \leq \sqrt{3\mathsf{snr}\,\mathrm{Var}(X)} + 3k_n, \tag{30}$$

$$I(f_Y) \leq 1, \tag{31}$$

$$\phi(t) \leq \sqrt{2\pi}\mathrm{e}^{t^2+\mathsf{snr}\,\mathbb{E}[X^2]}. \tag{32}$$

Proof. See Appendix F. □

We now bound $c(k_n)$. To this end, we need the notion of sub-Gaussian random variables: A random variable X is said to be α-sub-Gaussian if

$$\mathbb{E}[\mathrm{e}^{tX}] \leq \mathrm{e}^{\frac{\alpha^2t^2}{2}} \quad \forall t \in \mathbb{R}. \tag{33}$$

Lemma 3. *Suppose that* $\mathbb{E}[X^2] < \infty$. *Then,*

$$c(k_n) \leq \inf_{v>0} \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}} \left(\frac{\mathsf{snr}\,\mathbb{E}[X^2]+1}{k_n^2}\right)^{\frac{v}{1+v}}. \tag{34}$$

In addition, if $|X|$ *is* α*-sub-Gaussian, then*

$$c(k_n) \leq \inf_{v>0} \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}} \left(2\mathrm{e}^{\frac{\alpha^2\mathsf{snr}-k_n^2}{2}}\right)^{\frac{v}{1+v}}. \tag{35}$$

Proof. See Appendix G. □

4.1. Convergence of Bhattacharya's Estimator

By combining the results in Lemma 1, Theorem 1, Lemma 2, and Lemma 3, we have the following theorem.

Theorem 4. *Let* K *be as in* (27). *Choose the parameters of Bhattacharya's estimator as follows:* $a_0 = n^{-w_0}$, *where* $w_0 \in \left(0, \frac{1}{4}\right)$, $a_1 = n^{-w_1}$, *where* $w_1 \in \left(0, \frac{1}{6}\right)$, *and* $k_n = \sqrt{u\log(n)}$, *where* $u \in (0, \min(w_0, w_1))$. *Then, for* $n^{w_0-u} > c_5$,

$$\mathbb{P}[|I_n - I(f_Y)| \geq \varepsilon_n] \leq 2\mathrm{e}^{-c_1 n^{1-4w_0}} + 2\mathrm{e}^{-c_2 n^{1-6w_1}}, \tag{36}$$

where

$$\varepsilon_n \leq \frac{n^{-w_1}\sqrt{u\log(n)}\left(4c_3 + 12\sqrt{u\log(n)} + 2c_5 n^{u-w_1}\right) + c_5 n^{u-w_0}}{1 - c_5 n^{u-w_0}} + \frac{c_4}{\sqrt{u\log(n)}}, \tag{37}$$

and where the constants are given by

$$c_1 = \pi\left(1 - \frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2, \tag{38}$$

$$c_2 = \mathrm{e}\pi\left(1 - \frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2, \tag{39}$$

$$c_3 = \sqrt{3\mathsf{snr}\,\mathrm{Var}(X)}, \tag{40}$$

$$c_4 = \frac{2\Gamma^{\frac{1}{2}}\left(\frac{3}{2}\right)\sqrt{\mathsf{snr}\,\mathbb{E}[X^2]+1}}{\pi^{\frac{1}{4}}}, \tag{41}$$

$$c_5 = \sqrt{2\pi}\mathrm{e}^{\mathsf{snr}\,\mathbb{E}[X^2]}. \tag{42}$$

In addition, if $|X|$ *is* α*-sub-Gaussian, then*

$$\varepsilon_n \leq \frac{n^{-w_1}\sqrt{u\log(n)}\left(4c_3 + 12\sqrt{u\log(n)} + 2c_5 n^{u-w_1}\right) + c_5 n^{u-w_0}}{1 - c_5 n^{u-w_0}} + c_6 n^{-\frac{u}{4}}, \tag{43}$$

where

$$c_6 = \frac{2^{\frac{3}{2}}\Gamma^{\frac{1}{2}}\left(\frac{3}{2}\right)\mathrm{e}^{\frac{\alpha^2\mathsf{snr}}{4}}}{\pi^{\frac{1}{4}}}. \tag{44}$$

Proof. See Appendix H. □

Note that the parameters k_n, a_0, and a_1 are chosen so as to guarantee the convergence of $I_n(f_n)$ to $I(f_Y)$ with probability 1. For the details, please refer to the proof in Appendix H.

The parameters u and w in the above theorem are auxiliary variables that couple the bandwidth of the kernel density estimators in (3) and (4) with the integration range of the Fisher information estimator in (2). Choosing them according to Theorem 4 results in a trade-off between precision, ε_n, and confidence, i.e., the probability of the estimation error exceeding ε_n. On the one hand, small values of u and large values of w result in better precision (i.e., small ε_n) at the cost of a lower confidence (i.e., large probability of exceeding ε_n). On the other hand, large values of u and small values of w improve the confidence but deteriorate the precision. In turn, this also affects the convergence rates, meaning that faster convergence of the precision can be achieved at the expense of a slower convergence of the confidence and vice versa.

4.2. Convergence of the Clipped Estimator

From the evaluation of Bhattacharya's estimator in Theorem 4, it is apparent that the bottleneck term is the truncation parameter $k_n = \sqrt{u \log(n)}$, which results in slow precision decay of the order $\varepsilon_n = O\left(\frac{1}{\sqrt{u \log(n)}}\right)$. Next, it is shown that the clipped estimator results in improved precision over Bhattacharya's estimator. Specifically, the precision will be shown to decay polynomially in n instead of logarithmically. Another benefit of the clipped estimator is that its error analysis holds for every $n \geq 1$.

By utilizing the results in Lemma 1, Lemma 2, and Lemma 3, we specialize the result in Theorem 2 to the Gaussian noise case.

Theorem 5. *Let K be as in (27). Choose the parameters of the clipped estimator as follows: $a_0 = n^{-w_0}$, where $w_0 \in \left(0, \frac{1}{4}\right)$, $a_1 = n^{-w_1}$, where $w_1 \in \left(0, \frac{1}{6}\right)$, and $k_n = n^u$, where $u \in \left(0, \min\left(\frac{w_0}{3}, \frac{w_1}{2}\right)\right)$. Then, for $n \geq 1$*

$$\mathbb{P}[|I_n^c - I(f_Y)| \geq \varepsilon_n] \leq 2\mathrm{e}^{-c_1 n^{1-4w_0}} + 2\mathrm{e}^{-c_2 n^{1-6w_1}}, \tag{45}$$

where

$$\varepsilon_n \leq 4n^{3u-w_0}\left(c_3^2 n^{-2u} + 3\right) + 4n^{2u-w_1}\left(2c_3 n^{-u} + 3\right) + c_4 n^{-u}, \tag{46}$$

and the constants c_1 to c_4 are as in Theorem 4. In addition, if $|X|$ is α-sub-Gaussian, then

$$\varepsilon_n \leq 4n^{3u-w_0}\left(c_3^2 n^{-2u} + 3\right) + 4n^{2u-w_1}\left(2c_3 n^{-u} + 3\right) + c_6\mathrm{e}^{-\frac{n^{2u}}{4}}, \tag{47}$$

where c_6 is given by (44).

Proof. See Appendix I. □

Again, the parameters k_n, a_0, and a_1 are chosen to guarantee the consistency of the estimator. For further details, please refer to Appendix I.

4.3. Applications to the Estimations of the MMSE

As discussed in the introduction, the Fisher information is often merely a proxy for the actual quantity of interest. One accuracy measure that is typically of interest is the MMSE, which is defined as

$$\mathrm{mmse}(X|Y) = \mathbb{E}\Big[(X - \mathbb{E}[X|Y])^2\Big]. \tag{48}$$

In additive Gaussian noise, the MMSE can not only be bounded by the Fisher information, but both are related via Brown's identity:

$$I(f_Y) = 1 - \mathsf{snr}\,\mathrm{mmse}(X|Y). \tag{49}$$

Based on this relation, we propose the following estimators for the MMSE:

$$\operatorname{mmse}_n(X, \mathsf{snr}) = \frac{1 - I_n}{\mathsf{snr}} \tag{50}$$

and

$$\operatorname{mmse}_n^{\mathrm{c}}(X, \mathsf{snr}) = \frac{1 - I_n^{\mathrm{c}}}{\mathsf{snr}}. \tag{51}$$

The results for the estimators of Fisher information in Theorem 4 and Theorem 5 can be immediately extended to the MMSE estimators as follows.

Proposition 1. *Let K be as in (27), and let w_0, w_1, and n be such that they satisfy the conditions in Theorem 4. It then holds that*

$$\mathbb{P}[|\operatorname{mmse}_n(X, \mathsf{snr}) - \operatorname{mmse}(X, \mathsf{snr})| \geq \mathsf{snr}\varepsilon_n] \leq 2\mathrm{e}^{-c_1 n^{1-4w_0}} + 2\mathrm{e}^{-c_2 n^{1-6w_1}}, \tag{52}$$

where ε_n, c_1, and c_2 are given in Theorem 4.

Proposition 2. *Let K be as in (27), and let w_0 and w_1 be such that they satisfy the conditions in Theorem 5. It then holds that*

$$\mathbb{P}[|\operatorname{mmse}_n^c(X|Y) - \operatorname{mmse}(X|Y)| \geq \mathsf{snr}\varepsilon_n] \leq 2\mathrm{e}^{-c_1 n^{1-4w_0}} + 2\mathrm{e}^{-c_2 n^{1-6w_1}}, \tag{53}$$

where ε_n, c_1, and c_2 are given in Theorem 5.

4.4. Sample Complexity

Finally, we demonstrate the difference in the bounds on the convergence rates between Bhattacharya's estimator and its clipped version by comparing the sample complexity of the two estimators, that is, the required number of samples to guarantee a given accuracy with a given confidence. MATLAB implementations of both estimators, as well as the code used to generate the figures below, can be found in [22].

To this end, we consider the simple example of estimating the density of a Gaussian random variable in additive Gaussian noise. More precisely, we assume that X and Z in (26) are independent and identically distributed according to the standard normal distribution $\mathcal{N}(0,1)$, and that $\mathsf{snr} = 1$. This trivially implies that X is α-sub-Gaussian with $\alpha = 1$. In order to make the comparison as fair as possible, the parameters of the kernel estimators, a_0, a_1, and k_n, are not chosen according to Theorem 4 or Theorem 5, but are calculated by numerically minimizing the required number of samples; see [22] for details.

Let $P_{\mathrm{err}} = \mathbb{P}[|I_n - I(f_Y)| \geq \varepsilon_n]$. The left-hand plot in Figure 1 shows the corresponding bounds on the sample complexities of the two estimators with $P_{\mathrm{err}} = 0.2$ and ε_n varying from 0.1 to 0.9. Note that the results with larger ε_n are not shown because $I(f_Y) \leq 1$, as shown in Lemma 2. Moreover, the right-hand plot in Figure 1 shows the sample complexities for $\varepsilon_n = 0.5$ with P_{err} varying from 0.1 to 0.9. By inspection, the clipped estimator reduces the sample complexity by several orders of magnitude; note that the y-axes scale logarithmically. As discussed before, this does not imply that the clipped estimator is more accurate in general. However, it does imply that the clipped estimator provides significantly better worst-case performance, i.e., it requires significantly fewer samples to guarantee a certain precision or confidence. Finally, note that this improvement comes at a low cost in terms of complexity and regularity assumptions. The complexity of both algorithms is almost identical, with the clipped estimator only requiring an additional evaluation of $\bar{\rho}$. The regularity conditions are identical for bounded density functions, and slightly stronger for the clipped estimator for unbounded density functions.

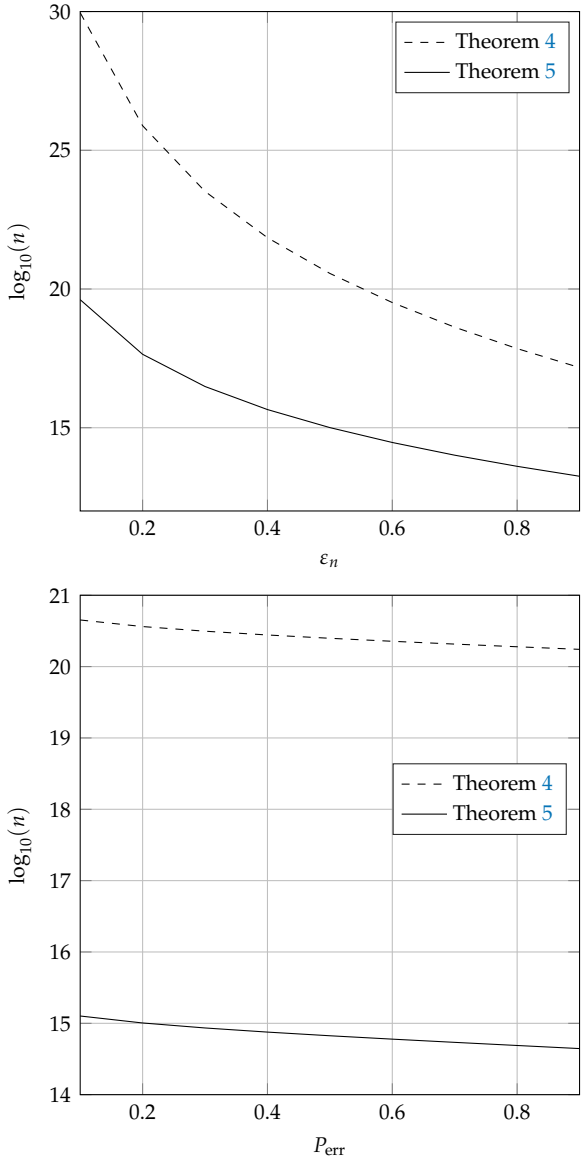

Figure 1. Sample complexity with Gaussian input. Left: number of samples required versus error of the estimators I_n and I_n^c given $P_{\text{err}} = 0.2$. Right: number of samples required versus confidence of the estimators with given $\varepsilon_n = 0.5$.

5. Conclusions

This work focused on the estimation of the Fisher information for the location of a univariate random variable using plug-in estimators based on estimators of the PDF and its derivative. Two estimators of the Fisher information were considered. The first estimator is the estimator due to Bhattacharya, for which new, sharper convergence results were shown. The paper also proposed a second estimator, termed a clipped estimator, which provides better bounds on the convergence rates. The accuracy bounds on both estimators were specialized to the practically relevant case of a random variable contaminated by additive Gaussian noise. Moreover, using special proprieties of the Gaussian noise case, two estimators for the MMSE were proposed, and their convergence rates were analyzed. This was done by using Brown's identity, which connects the Fisher information and the MMSE. Finally, using a numerical example, it was demonstrated that the proposed clipped estimator can achieve a significantly lower sample complexity at little or no additional cost.

Author Contributions: Investigation, Writing—original draft: W.C.; Investigation, Writing - original draft: A.D.; Investigation, Writing—original draft: M.F.; Supervision, Resources, Funding acquisition, Writing—review & editing: H.V.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the U.S. National Science Foundation under Grants CCF-0939370 and CCF-1908308, and in part by the German Research Foundation (DFG) under Grant 424522268.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. A Proof of Lemma 1

Our starting point is the following bound due to [3] (p. 1188):

$$\sup_{t\in\mathbb{R}}\left|\mathbb{E}\left[f_n^{(r)}(t)\right]-f_n^{(r)}(t)\right| \leq \frac{v_r}{a_r^{r+1}}\sup_{t\in\mathbb{R}}|F_n(t)-F_Y(t)|, \tag{A1}$$

where F is the CDF of f, F_n is the empirical CDF, and v_r is defined in (5). Now, let δ_{r,a_r} be as in (6), and consider the following sequence of bounds:

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n^{(r)}(t)-f^{(r)}(t)\right|>\epsilon\right] \leq \mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n^{(r)}(t)-\mathbb{E}[f_n^{(r)}(t)]\right|>\epsilon-\delta_{r,a_r}\right] \tag{A2}$$

$$\leq \mathbb{P}\left[\sup_{t\in\mathbb{R}}|F_n(t)-F(t)|>\frac{a_r^{r+1}(\epsilon-\delta_{r,a_r})}{v_r}\right] \tag{A3}$$

$$\leq 2\mathrm{e}^{-2n\frac{a_r^{2r+2}(\epsilon-\delta_{r,a_r})^2}{v_r^2}}, \tag{A4}$$

where (A2) follows by using the triangle inequality; (A3) follows by using the bound in (A1); and (A4) follows by using the sharp DKW inequality [20]:

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}}|F_n(t)-F(t)|>\epsilon\right] \leq 2\mathrm{e}^{-2n\epsilon^2}. \tag{A5}$$

This concludes the proof.

Appendix B. A Proof of Theorem 1

First, using the triangle inequality, we have that

$$|I(f)-I_n| \leq \left|\int_{-k_n}^{k_n}\frac{(f_n'(t))^2}{f_n(t)}-\frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right|+c(k_n). \tag{A6}$$

Next, we bound the first term in (A6) :

$$\left|\int_{-k_n}^{k_n} \frac{(f_n'(t))^2}{f_n(t)} - \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right|$$

$$= \left|\int_{-k_n}^{k_n} \frac{f(t)(f_n'(t))^2 - f_n(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| \tag{A7}$$

$$\leq \left|\int_{-k_n}^{k_n} \frac{f(t)(f_n'(t))^2 - f(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| + \left|\int_{-k_n}^{k_n} \frac{f(t)(f'(t))^2 - f_n(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| \tag{A8}$$

$$= \left|\int_{-k_n}^{k_n} \frac{(f_n'(t))^2 - (f'(t))^2}{f_n(t)}\,\mathrm{d}t\right| + \left|\int_{-k_n}^{k_n} \frac{f_n(t) - f(t)}{f_n(t)} \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right| \tag{A9}$$

$$\leq \sup_{|t|\leq k_n} \frac{|f_n'(t) + f'(t)|}{f_n(t)} |f_n(t) - f(t)| 2k_n + \sup_{|t|\leq k_n} \frac{|f_n(t) - f(t)|}{f_n(t)} \int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t \tag{A10}$$

$$\leq \sup_{|t|\leq k_n} \frac{|f_n'(t) + f'(t)|}{f_n(t)} \epsilon_1 2k_n + \sup_{|t|\leq k_n} \frac{1}{f_n(t)} \epsilon_0 I(f), \tag{A11}$$

where the last bound follows from the assumptions in (9). Now, consider the first term in (A11):

$$\sup_{|t|\leq k_n} \frac{|f_n'(t) + f'(t)|}{f_n(t)} \leq \sup_{|t|\leq k_n} \frac{2|f'(t)| + \epsilon_1}{f_n(t)} \tag{A12}$$

$$\leq \sup_{|t|\leq k_n} \frac{2|f'(t)| + \epsilon_1}{f(t) - f(t) + f_n(t)} \tag{A13}$$

$$\leq \sup_{|t|\leq k_n} \frac{2|f'(t)| + \epsilon_1}{f(t) - \epsilon_0} \tag{A14}$$

$$= \sup_{|t|\leq k_n} \frac{2\left|\frac{f'(t)}{f(t)}\right| + \frac{\epsilon_1}{f(t)}}{1 - \frac{\epsilon_0}{f(t)}} \tag{A15}$$

$$\leq \frac{2\sup_{|t|\leq k_n}\left|\frac{f'(t)}{f(t)}\right| + \epsilon_1 \phi(k_n)}{1 - \epsilon_0 \phi(k_n)}, \tag{A16}$$

where the bound in (A14) follows from the assumptions in (9) and the properties of ϕ, which imply

$$\epsilon_0 \phi(k_n) < 1 \quad \Rightarrow \quad \epsilon_0 < f(t), \quad \forall |t| \leq k_n; \tag{A17}$$

and the bound in (A16) follows from the definition of ϕ in (8). Now, consider the second term in (A11):

$$\sup_{|t|\leq k_n} \frac{1}{f_n(t)} = \sup_{|t|\leq k_n} \frac{1}{f(t) - f(t) + f_n(t)} \tag{A18}$$

$$\leq \sup_{|t|\leq k_n} \frac{1}{f(t) - \epsilon_0} \tag{A19}$$

$$= \sup_{|t|\leq k_n} \frac{1}{1 - \frac{\epsilon_0}{f(t)}} \frac{1}{f(t)} \tag{A20}$$

$$\leq \frac{1}{1 - \epsilon_0 \phi(k_n)} \phi(k_n), \tag{A21}$$

where (A20) follows by using similar steps, leading to the bound in (A14), and (A21) follows from the definition of ϕ.

Combining the bounds in (A6), (A11), (A16), and (A21) concludes the proof.

Appendix C. A Proof of Theorem 2

Using (9), (14), and (A21), it holds that

$$|\log(f_n)| \leq \max\left(\log(f_n), \log\left(\frac{1}{f_n}\right)\right) \tag{A22}$$

$$\leq \max\left(\log(f_0+\epsilon_0), \log\left(\frac{\phi(k_n)}{1-\epsilon_0\phi(k_n)}\right)\right) \tag{A23}$$

$$= \psi(\epsilon_0, k_n). \tag{A24}$$

Next, we bound the first term on the right-hand side of (A6). Starting with (A9) above, it holds that

$$\left|\int_{-k_n}^{k_n} \frac{(f_n'(t))^2}{f_n(t)} - \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right|$$

$$\leq \epsilon_1 \int_{-k_n}^{k_n} \left|\frac{f_n'(t)+f'(t)}{f_n(t)}\right| \mathrm{d}t + \epsilon_0 \int_{-k_n}^{k_n} \left|\frac{(f'(t))^2}{f_n(t)f(t)}\right| \mathrm{d}t \tag{A25}$$

$$\leq \epsilon_1 \int_{-k_n}^{k_n} \left|\frac{f_n'(t)}{f_n(t)}\right| \mathrm{d}t + \epsilon_1 \int_{-k_n}^{k_n} \left|\frac{f'(t)}{f_n(t)}\right| \mathrm{d}t + \epsilon_0\rho_{\max}(k_n) \int_{-k_n}^{k_n} \left|\frac{f'(t)}{f_n(t)}\right| \mathrm{d}t, \tag{A26}$$

where the inequality in (A25) follows again from (9), and the last bound follows from the triangle inequality together with the definition of $\rho_{\max}$.

Consider the integral in the first term in (A26):

$$\int_{-k_n}^{k_n} \left|\frac{f_n'(t)}{f_n(t)}\right| \mathrm{d}t$$

$$= \int_{-k_n}^{k_n} |\nabla \log(f_n(t))|\,\mathrm{d}t \tag{A27}$$

$$= \int_{-k_n}^{k_n} \mathsf{sign}(\nabla \log(f_n(t))) \cdot \nabla \log(f_n(t))\,\mathrm{d}t \tag{A28}$$

$$= \mathsf{sign}(\nabla \log(f_n(t))) \cdot \log(f_n(t))\Big|_{-k_n}^{k_n} - \int_{-k_n}^{k_n} \log(f_n(t)) \frac{\mathrm{d}}{\mathrm{d}t}\mathsf{sign}(\nabla \log(f_n(t)))\,\mathrm{d}t, \tag{A29}$$

where the last equality follows from integration by parts. The first term in (A29) can be upper bounded as

$$\mathsf{sign}(\nabla \log(f_n(t))) \cdot \log(f_n(t))\Big|_{-k_n}^{k_n} \leq 2\psi(\epsilon_0, k_n), \tag{A30}$$

where the inequality in (A30) follows from (A24). The second term in (A29) is given by

$$-\int_{-k_n}^{k_n} \log(f_n(t)) \frac{\mathrm{d}}{\mathrm{d}t}\mathsf{sign}(\nabla \log(f_n(t)))\,\mathrm{d}t = -\sum_{t\in[-k_n,k_n]:f_n'(t)=0} \log(f_n(t)) \tag{A31}$$

$$\leq d_{f_n}(k_n)\psi(\epsilon_0, k_n). \tag{A32}$$

By substituting (A30) and (A32) into (A29), one obtains

$$\int_{-k_n}^{k_n} \left|\frac{f_n'(t)}{f_n(t)}\right| \mathrm{d}t \leq (2 + d_{f_n})\psi(\epsilon_0, k_n). \tag{A33}$$

Next, we consider the integral in the second and third terms in (A26):

$$\int_{-k_n}^{k_n} \left| \frac{f'(t)}{f_n(t)} \right| \mathrm{d}t \leq \int_{-k_n}^{k_n} \left| \frac{f'(t)}{f(t) - \epsilon_0} \right| \mathrm{d}t \tag{A34}$$

$$= \int_{-k_n}^{k_n} |\nabla \log(f(t) - \epsilon_0)| \, \mathrm{d}t \tag{A35}$$

$$= \int_{-k_n}^{k_n} \mathsf{sign}(\nabla \log(f(t) - \epsilon_0)) \cdot \nabla \log(f(t) - \epsilon_0) \, \mathrm{d}t \tag{A36}$$

$$= \mathsf{sign}(\nabla \log(f(t) - \epsilon_0)) \cdot \log(f(t) - \epsilon_0) \Big|_{-k_n}^{k_n}$$

$$- \int_{-k_n}^{k_n} \log(f(t) - \epsilon_0) \frac{\mathrm{d}}{\mathrm{d}t} \mathsf{sign}(\nabla \log(f(t) - \epsilon_0)) \, \mathrm{d}t \tag{A37}$$

$$\leq (2 + d_f(k_n)) \max\left(\log(f_0 - \epsilon_0), \log\left(\frac{\phi(k_n)}{1 - \epsilon_0 \phi(k_n)} \right) \right) \tag{A38}$$

$$\leq (2 + d_f(k_n)) \psi(\epsilon_0, k_n), \tag{A39}$$

where the inequalities in (A34) follows from the assumptions in (9) and (A17), and the bound in (A38) follows by using steps similar to those leading to the bound in (A33).

Combining the bounds in (A6), (A26), (A33), and (A38) concludes the proof.

Appendix D. A Proof of Theorem 3

The difficulty in bounding the error of a clipped estimator is in showing that the clipping is strict enough to avoid gross overestimation, yet permissive enough to avoid gross underestimation. The proof presented here is based on two auxiliary estimators that are constructed to under- and overestimate $I_n^{\mathrm{c}}(f_n)$ in a controlled manner.

Let

$$\underline{I}_n = \int_{-k_n}^{k_n} \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \, \mathrm{d}t, \tag{A40}$$

where $\lceil \bullet - \epsilon \rfloor$ denotes an "ϵ-compression" operator, i.e.,

$$\lceil f(t) - \epsilon \rfloor = \begin{cases} f(t) - \epsilon, & f(t) > \epsilon \\ 0, & -\epsilon \leq f(t) \leq \epsilon \\ f(t) + \epsilon, & f(t) < -\epsilon. \end{cases} \tag{A41}$$

Next, consider the estimator

$$\overline{I}_n = \int_{-k_n}^{k_n} \frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor^2}{f_n(t) + \gamma_{0,n}(t)} \, \mathrm{d}t, \tag{A42}$$

where the functions $\gamma_{i,n} \colon \mathbb{R} \to [0, \epsilon_i]$, $i = 0, 1$ are chosen as follows: If it holds that

$$|\rho_n(t)| \leq |\overline{\rho}(t)|, \tag{A43}$$

then $\gamma_{0,n}(t) = \gamma_{1,n}(t) = 0$. If, on the other hand,

$$|\rho_n(t)| > |\overline{\rho}(t)|, \tag{A44}$$

then $\gamma_{0,n}(t)$ and $\gamma_{1,n}(t)$ are chosen such that

$$\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor}{f_n(t) + \gamma_{0,n}(t)} = \overline{\rho}(t). \tag{A45}$$

Note that because

$$\left| \frac{\lceil f_n'(t) - \epsilon_1 \rfloor}{f_n(t) + \epsilon_0} \right| \leq |\rho(t)| \leq |\overline{\rho}(t)|, \tag{A46}$$

this is always possible.

In Appendix E, it is shown that the following relations hold between the estimators defined above:

$$\underline{I}_n \leq I(f), \tag{A47}$$
$$\underline{I}_n \leq I_n^{\mathrm{c}}, \tag{A48}$$
$$I_n^{\mathrm{c}} \leq \overline{I}_n + \epsilon_1 \Phi^1_{\max}(k_n), \tag{A49}$$
$$I(f) - \underline{I}_n \leq 4\epsilon_1 \Phi^1(k_n) + 2\epsilon_0 \Phi^2(k_n) + c(k_n), \tag{A50}$$
$$\overline{I}_n - \underline{I}_n \leq 2\epsilon_1 \Phi^1_{\max}(k_n) + \epsilon_0 \Phi^2_{\max}(k_n). \tag{A51}$$

The bound in Theorem 3 can now be obtained by bounding the under- and overestimation errors separately. For $I_n^{\mathrm{c}} \leq I(f)$, it holds that

$$I(f) - I_n^{\mathrm{c}} \leq I(f) - \underline{I}_n \tag{A52}$$
$$\leq 4\epsilon_1 \Phi^1(k_n) + 2\epsilon_0 \Phi^2(k_n) + c(k_n). \tag{A53}$$

For $I_n^{\mathrm{c}} > I(f)$, it hold that

$$I_n^{\mathrm{c}} - I(f) \leq \overline{I}_n - \underline{I}_n + \epsilon_1 \Phi^1_{\max}(k_n) \tag{A54}$$
$$\leq 3\epsilon_1 \Phi^1_{\max}(k_n) + \epsilon_0 \Phi^2_{\max}(k_n). \tag{A55}$$

The bound in (22) follows. Furthermore, following the same steps as those leading to the bound in (A33), the bound in (25) follows.

Appendix E. A Proof of the Estimator Relations in Theorem 3

The bound in (A47) follows directly from the fact that under the assumptions in (9)

$$\frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \leq \frac{(f'(t))^2}{f(t)}. \tag{A56}$$

Analogously, (A48) follows from

$$\frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \leq |\rho(t)||\lceil f_n'(t) - \epsilon_1 \rfloor| \leq |\rho(t)||f_n'(t)|. \tag{A57}$$

In order to show (A50), note that under the assumptions in (9), it holds that

$$|\lceil f_n'(t) - \epsilon_1 \rfloor| \leq |f'(t)|, \tag{A58}$$
$$(f_n(t) + \epsilon_0) - f(t) \leq 2\epsilon_0, \tag{A59}$$
$$|\lceil f_n'(t) - \epsilon_1 \rfloor - f'(t)| \leq 2\epsilon_1. \tag{A60}$$

Hence, in analogy to Theorem 1, the estimation error of $\underline{I}_n$ can be written as

$$I(f) - \underline{I}_n = \int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)} - \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \, \mathrm{d}t + c(k_n). \tag{A61}$$

Using the same arguments as in the proof of Theorem 1, the integral term on the right-hand side of (A61) can be bounded by

$$\begin{aligned}
&\int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)} - \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \, \mathrm{d}t \\
&= \left| \int_{-k_n}^{k_n} \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2 f(t) - (f'(t))^2 (f_n(t) + \epsilon_0)}{f(t)(f_n(t) + \epsilon_0)} \, \mathrm{d}t \right| && \text{(A62)} \\
&\leq \left| \int_{-k_n}^{k_n} \left| \lceil f_n'(t) - \epsilon_1 \rfloor - f'(t) \right| \frac{|\lceil f_n'(t) - \epsilon_1 \rfloor + f'(t)|}{f_n(t) + \epsilon_0} \, \mathrm{d}t \right| \\
&\quad + \int_{-k_n}^{k_n} |f(t) - (f_n(t) + \varepsilon_0)| \frac{(f'(t))^2}{f(t)(f_n(t) + \varepsilon_0)} \, \mathrm{d}t && \text{(A63)} \\
&\leq 2\epsilon_1 \int_{-k_n}^{k_n} \frac{|\lceil f_n'(t) - \epsilon_1 \rfloor| + |f'(t)|}{f_n(t) + \epsilon_0} \, \mathrm{d}t + 2\epsilon_0 \int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)(f_n(t) + \epsilon_0)} \, \mathrm{d}t && \text{(A64)} \\
&\leq 2\epsilon_1 \int_{-k_n}^{k_n} 2 \left| \frac{f'(t)}{f(t)} \right| \mathrm{d}t + 2\epsilon_0 \int_{-k_n}^{k_n} \left| \frac{f'(t)}{f(t)} \right|^2 \mathrm{d}t && \text{(A65)} \\
&\leq 4\epsilon_1 \int_{-k_n}^{k_n} |\rho(t)| + 2\epsilon_0 \int_{-k_n}^{k_n} \rho^2(t) \, \mathrm{d}t && \text{(A66)} \\
&= 4\epsilon_1 \, \Phi^1(k_n) + 2\varepsilon_0 \, \Phi^2(k_n). && \text{(A67)}
\end{aligned}$$

Using the same steps, it is not difficult to show (A51), where the factor 2 does not arise because, in contrast to (A59) and (A60),

$$\lceil f_n(t) + \epsilon_0 \rfloor - \lceil f_n(t) + \gamma_{0,n}(t) \rfloor \leq \epsilon_0, \tag{A68}$$

$$\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor - \lceil f_n'(t) - \epsilon_1 \rfloor \leq \epsilon_1, \tag{A69}$$

and $c(k_n)$ does not arise because both estimators are defined on $[-k_n, k_n]$.

In order to show (A49), first note that for $|\rho_n(t)| \leq |\overline{\rho}(t)|$, it holds that

$$\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor^2}{f_n(t) + \gamma_{0,n}(t)} = \frac{(f_n'(t))^2}{f_n(t)} = |\rho_n(t)||f_n'(t)|, \tag{A70}$$

i.e., $\overline{I}_n(f_n) = I_n^{\mathrm{c}}(f_n)$. Hence, $I_n^{\mathrm{c}}(f_n) > \overline{I}(f_n)$ implies $|\rho_n(t)| \geq |\overline{\rho}(t)|$ on some region of $[-k_n, k_n]$. On this region, it holds that

$$\begin{aligned}
\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor^2}{f_n(t) + \gamma_{0,n}(t)} &= \frac{|\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor|}{f_n(t) + \gamma_{0,n}(t)} |\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor| && \text{(A71)} \\
&= |\overline{\rho}(t)| \, |\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor|. && \text{(A72)}
\end{aligned}$$

Because

$$|f_n'(t)| - |\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor| \leq \gamma_{1,n} \leq \epsilon_1, \tag{A73}$$

it follows that

$$\begin{aligned}
I_n^{\mathrm{c}}(f_n) - \overline{I}_n(f_n) &\leq \int_{-k_n}^{k_n} |\overline{\rho}(t)| \epsilon_1 \, \mathrm{d}t && \text{(A74)} \\
&\leq \epsilon_1 \Phi_{\max}^1(k_n). && \text{(A75)}
\end{aligned}$$

Appendix F. A Proof of Lemma 2

We begin by bounding v_r and δ_{r,a_r}. First,

$$v_0 = \int_{-\infty}^{\infty} |t| K(t)\,\mathrm{d}t = \sqrt{\frac{2}{\pi}}, \tag{A76}$$

$$v_1 = \int_{-\infty}^{\infty} \left|t^2 - 1\right| K(t)\,\mathrm{d}t = 2\sqrt{\frac{2}{\mathrm{e}\pi}}. \tag{A77}$$

Second,

$$\begin{aligned}
\delta_{r,a_r} &= \left|\mathbb{E}[f_n^{(r)}(t)] - f_Y^{(r)}(t)\right| && \text{(A78)}\\
&= \left|\int_{-\infty}^{\infty} \frac{1}{a_r} K\left(\frac{t-y}{a_r}\right)\left(f_Y^{(r)}(y) - f_Y^{(r)}(t)\right)\mathrm{d}y\right| && \text{(A79)}\\
&= \left|\int_{-\infty}^{\infty} K(y)\left(f_Y^{(r)}(t + a_r y) - f_Y^{(r)}(t)\right)\mathrm{d}y\right| && \text{(A80)}\\
&\le \sup_{t\in\mathbb{R}}\left|f_Y^{(r+1)}(t)\right| \int_{-\infty}^{\infty} K(y) a_r |y|\,\mathrm{d}y && \text{(A81)}\\
&= a_r\sqrt{\frac{2}{\pi}} \sup_{t\in\mathbb{R}}\left|f_Y^{(r+1)}(t)\right|. && \text{(A82)}
\end{aligned}$$

Now, for $r = 0$,

$$\begin{aligned}
\left|f_Y^{(1)}(t)\right| &= \left|\mathbb{E}\left[(t - \sqrt{\mathsf{snr}}X)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-\sqrt{\mathsf{snr}}X)^2}{2}}\right]\right| && \text{(A83)}\\
&\le \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\mathrm{e}}}, && \text{(A84)}
\end{aligned}$$

where we have used the bound $t\mathrm{e}^{-\frac{t^2}{2}} \le \frac{1}{\sqrt{\mathrm{e}}}$. For $r = 1$,

$$\begin{aligned}
\left|f_Y^{(2)}(t)\right| &= \left|\mathbb{E}\left[\left((t - \sqrt{\mathsf{snr}}X)^2 - 1\right)\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-\sqrt{\mathsf{snr}}X)^2}{2}}\right]\right| && \text{(A85)}\\
&\le \frac{1}{\sqrt{2\pi}}\frac{2}{\mathrm{e}} + \frac{1}{\sqrt{2\pi}}, && \text{(A86)}
\end{aligned}$$

where we have used the bound $t^2\mathrm{e}^{-\frac{t^2}{2}} \le \frac{2}{\mathrm{e}}$.

Next, we bound the score function ρ_Y as follows:

$$\begin{aligned}
|\rho_Y(t)| &= \left|\frac{f_Y'(t)}{f_Y(t)}\right| && \text{(A87)}\\
&= \left|\sqrt{\mathsf{snr}}\,\mathbb{E}[X|Y = t] - t\right| && \text{(A88)}\\
&\le \sqrt{\mathsf{snr}}\,\mathbb{E}[|X||Y = t] + |t| && \text{(A89)}\\
&\le \sqrt{\mathsf{snr}}\sqrt{\mathbb{E}[X^2|Y = t]} + |t| && \text{(A90)}\\
&\le \sqrt{3\mathsf{snr}\,\mathrm{Var}(X) + 4t^2} + |t| && \text{(A91)}\\
&\le \sqrt{3\mathsf{snr}\,\mathrm{Var}(X)} + 3|t|, && \text{(A92)}
\end{aligned}$$

where the equality in (A88) follows by using the identify $\frac{f_Y'(t)}{f_Y(t)} = \sqrt{\mathsf{snr}}\,\mathbb{E}[X|Y=t] - t$ [23], the inequality in (A90) follows from Jensen's inequality, and the inequality in (A91) follows from the bound in [24] (Proposition 1.2). Using the bound in (A92), it follows that

$$\rho_{\max}(k_n) = \max_{|t|\le k_n} |\rho(t)| \le \sqrt{3\mathsf{snr}\,\mathrm{Var}(X)} + 3k_n. \tag{A93}$$

Using the relation between the Fisher information and the MMSE, we have that

$$I(f_Y) = 1 - \mathsf{snr}\,\mathrm{mmse}(X, \mathsf{snr}) \le 1. \tag{A94}$$

Finally, the function ϕ is obtained by observing that

$$f_Y(t) = \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{(t-\sqrt{\mathsf{snr}}X)^2}{2}}\right] \tag{A95}$$

$$\ge \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{\mathbb{E}\left[(t-\sqrt{\mathsf{snr}}X)^2\right]}{2}} \tag{A96}$$

$$\ge \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\left(t^2+\mathsf{snr}\,\mathbb{E}[X^2]\right)}, \tag{A97}$$

where we used Jensen's inequality and the fact that $(a+b)^2 \le 2(a^2+b^2)$. This concludes the proof.

Appendix G. A Proof of Lemma 3

Choose some $v > 0$. Then,

$$c(k_n) = \mathbb{E}\left[\rho_Y^2(Y)1_{\{|Y|\ge k_n\}}\right] \tag{A98}$$

$$\le \mathbb{E}^{\frac{1}{1+v}}\left[\left(\rho_Y^2(Y)\right)^{1+v}\right]\mathbb{E}^{\frac{v}{1+v}}\left[\left(1_{\{|Y|\ge k_n\}}\right)^{\frac{1+v}{v}}\right] \tag{A99}$$

$$= \mathbb{E}^{\frac{1}{1+v}}\left[\left(\rho_Y^2(Y)\right)^{1+v}\right]\mathbb{E}^{\frac{v}{1+v}}\left[1_{\{|Y|\ge k_n\}}\right] \tag{A100}$$

$$= \mathbb{E}^{\frac{1}{1+v}}\left[|\rho_Y(Y)|^{2(1+v)}\right]\mathbb{P}^{\frac{v}{1+v}}[|Y| \ge k_n] \tag{A101}$$

$$= \mathbb{E}^{\frac{1}{1+v}}\left[|\mathbb{E}[Z|Y]|^{2(1+v)}\right]\mathbb{P}^{\frac{v}{1+v}}[|Y| \ge k_n] \tag{A102}$$

$$\le \mathbb{E}^{\frac{1}{1+v}}\left[|Z|^{2(1+v)}\right]\mathbb{P}^{\frac{v}{1+v}}[|Y| \ge k_n] \tag{A103}$$

$$= \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}}\mathbb{P}^{\frac{v}{1+v}}[|Y| \ge k_n] \tag{A104}$$

$$= \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}}\left(\frac{\mathsf{snr}\,\mathbb{E}[|X|^2]+1}{k_n^2}\right)^{\frac{v}{1+v}}, \tag{A105}$$

where (A99) follows from Hölder's inequality, (A102) follows by using the identity

$$\rho_Y(t) = \sqrt{\mathsf{snr}}\mathbb{E}[X|Y=t] - t = -\mathbb{E}[Z|Y=t], \tag{A106}$$

and (A105) follows from Markov's inequality.

Now, if $\mathbb{E}[X^2] < \infty$, then using Markov's inequality,

$$\mathbb{P}[|Y| \ge k_n] \le \frac{\mathbb{E}[Y^2]}{k_n^2} = \frac{\mathsf{snr}\,\mathbb{E}[|X|^2]+1}{k_n^2}. \tag{A107}$$

Moreover, using the Chernoff bound,

$$\mathbb{P}[|Y| \geq k_n] \leq \mathrm{e}^{-k_n t}\mathbb{E}\left[\mathrm{e}^{t|Y|}\right] \tag{A108}$$

$$\leq 2\mathrm{e}^{-k_n t+\frac{t^2}{2}}\mathbb{E}\left[\mathrm{e}^{t\sqrt{\mathsf{snr}}|X|}\right] \tag{A109}$$

$$= 2\mathrm{e}^{-k_n t+\frac{t^2}{2}}\mathrm{e}^{\frac{\alpha^2\mathsf{snr}}{2}}. \tag{A110}$$

Therefore,

$$c(k_n) \leq \inf_{t>0}\inf_{v>0}\frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}}2^{\frac{v}{1+v}}\mathrm{e}^{\frac{v}{1+v}\left(-k_n t+\frac{t^2}{2}+\frac{\alpha^2\mathsf{snr}}{2}\right)} \tag{A111}$$

$$\leq \inf_{v>0}\frac{2\Gamma^{\frac{1}{(1+v)}}\left(v+\frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}}2^{\frac{v}{1+v}}\mathrm{e}^{\frac{v}{1+v}\frac{\alpha^2\mathsf{snr}-k_n^2}{2}}. \tag{A112}$$

This concludes the proof.

Appendix H. A Proof of Theorem 4

Let

$$\varepsilon_n = \frac{4\epsilon_1 k_n \rho_{\max}(k_n) + 2\epsilon_1^2 k_n \phi(k_n) + \epsilon_0 \phi(k_n)}{1-\epsilon_0\phi(k_n)} + c(k_n), \tag{A113}$$

which is obtained from (11) by bounding $I(f)$ by 1 according to (31). In order to apply the bounds in Lemma 1 and Theorem 1, the following equalities/inequalities must hold for $r \in \{0,1\}$:

$$\epsilon_r > \delta_{r,a_r}, \tag{A114a}$$

$$\frac{a_r^{2r+2}(\epsilon_r-\delta_{r,a_r})^2}{v_r^2} \gg \frac{1}{n}, \tag{A114b}$$

$$\lim_{n\to\infty}\epsilon_1\rho_{\max}(k_n) = 0, \tag{A114c}$$

$$\lim_{n\to\infty}\epsilon_1^2 k_n\phi(k_n) = 0, \tag{A114d}$$

$$\lim_{n\to\infty}\epsilon_0\phi(k_n) = 0, \tag{A114e}$$

$$\lim_{n\to\infty}c(k_n) = 0. \tag{A114f}$$

To satisfy (A114), we choose

$$a_0 = \epsilon_0 = n^{-w_0}, \quad w_0 \in \left(0,\tfrac{1}{4}\right), \tag{A115}$$

$$a_1 = \epsilon_1 = n^{-w_1}, \quad w_1 \in \left(0,\tfrac{1}{6}\right), \tag{A116}$$

$$k_n = n^u, \quad u \in (0,\min(w_0,w_1)). \tag{A117}$$

Then, together with the bounds in Lemma 2, the relevant quantities in (A114) are as follows:

$$\frac{a_0^2(\epsilon_0-\delta_{0,a_0})^2}{v_0^2}=\frac{c_1}{2}n^{-4w_0}, \tag{A118a}$$

$$\frac{a_1^4(\epsilon_1-\delta_{1,a_1})^2}{v_1^2}=\frac{c_2}{2}n^{-6w_1}, \tag{A118b}$$

$$\epsilon_1 k_n \rho_{\max}(k_n)\le\left(c_3\sqrt{u\log(n)}+3u\log(n)\right)n^{-w_1}, \tag{A118c}$$

$$\epsilon_1^2 k_n \phi(k_n)\le c_5\sqrt{u\log(n)}n^{u-2w_1}, \tag{A118d}$$

$$\epsilon_0 \phi(k_n)\le c_5 n^{u-w_0}, \tag{A118e}$$

$$c(k_n)\le\frac{c_4}{\sqrt{u\log(n)}}, \tag{A118f}$$

which yields (37). Now, if $|X|$ is α-sub-Gaussian, the bound in (43) can be obtained from Lemma 3 with $v=1$.

Because (9) leads to (11), one obtains

$$\begin{aligned}
&\mathbb{P}[|I_n(f_n)-I(f_Y)|\ge\varepsilon_n]\\
&\le\mathbb{P}\left[\sup_{|t|\le k_n}|f_n(t)-f_Y(t)|\ge\epsilon_0\right]+\mathbb{P}\left[\sup_{|t|\le k_n}\left|f_n'(t)-f_Y'(t)\right|\ge\epsilon_1\right] &\text{(A119)}\\
&\le\mathbb{P}\left[\sup_{t\in\mathbb{R}}|f_n(t)-f_Y(t)|>\epsilon_0\right]+\mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n'(t)-f_Y'(t)\right|>\epsilon_1\right] &\text{(A120)}\\
&\le 2\mathrm{e}^{-n\pi a_0^2\left(\epsilon_0-a_0\frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2}+2\mathrm{e}^{-n\mathrm{e}\pi a_1^4\left(\epsilon_1-a_1\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2} &\text{(A121)}\\
&=2\mathrm{e}^{-\pi\left(1-\frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2 n^{1-4w_0}}+2\mathrm{e}^{-\mathrm{e}\pi\left(1-\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2 n^{1-6w_1}} &\text{(A122)}\\
&=2\mathrm{e}^{-c_1 n^{1-4w_0}}+2\mathrm{e}^{-c_2 n^{1-6w_1}}, &\text{(A123)}
\end{aligned}$$

where the inequality in (A121) follows from Lemma 1, and the last step follows from (A115), (A116), and (A117). This concludes the proof.

Appendix I. A Proof of Theorem 5

Let

$$\varepsilon_n=4\epsilon_1\Phi^1_{\max}(k_n)+2\epsilon_0\Phi^2_{\max}(k_n)+c(k_n). \tag{A124}$$

To apply the bounds in Theorem 3 and Lemma 2, the following equalities/inequalities must hold for $r\in\{0,1\}$:

$$\epsilon_r>\delta_{r,a_r}, \tag{A125a}$$

$$\frac{a_r^{2r+2}(\epsilon_r-\delta_{r,a_r})^2}{v_r^2}\gg\frac{1}{n}, \tag{A125b}$$

$$\lim_{n\to\infty}\epsilon_1\Phi^1_{\max}(k_n)=0, \tag{A125c}$$

$$\lim_{n\to\infty}\epsilon_0\Phi^0_{\max}(k_n)=0, \tag{A125d}$$

$$\lim_{n\to\infty}c(k_n)=0. \tag{A125e}$$

To satisfy (A125), we choose

$$a_0 = \epsilon_0 = n^{-w_0}, \quad w_0 \in \left(0, \tfrac{1}{4}\right), \tag{A126}$$

$$a_1 = \epsilon_1 = n^{-w_1}, \quad w_1 \in \left(0, \tfrac{1}{6}\right), \tag{A127}$$

$$k_n = n^u, \quad u \in \left(0, \min\left(\tfrac{w_0}{3}, \tfrac{w_1}{2}\right)\right). \tag{A128}$$

Then, together with the bounds in Lemma 2, the relevant quantities in (A125) are as follows:

$$\frac{a_0^2(\epsilon_0 - \delta_{0,a_0})^2}{v_0^2} = \frac{c_1}{2} n^{-4w_0}, \tag{A129a}$$

$$\frac{a_1^4(\epsilon_1 - \delta_{1,a_1})^2}{v_1^2} = \frac{c_2}{2} n^{-6w_1}, \tag{A129b}$$

$$\epsilon_1 \Phi^1_{\max}(k_n) = n^{u-w_1}(2c_3 + 3n^u), \tag{A129c}$$

$$\epsilon_0 \Phi^1_{\max}(k_n) = n^{u-w_0}\left(2c_3^2 + 6n^{2u}\right), \tag{A129d}$$

$$c(k_n) \leq c_4 n^{-u}, \tag{A129e}$$

which yields (46). Moreover, if $|X|$ is α-sub-Gaussian, the bound in (47) can be obtained from Lemma 3.

Following the same steps leading to (A123), we have that

$$\mathbb{P}[|I_n^c(f_n) - I(f_Y)| \geq \varepsilon_n] \leq 2\mathrm{e}^{-\pi\left(1-\frac{1}{\sqrt{2\pi \mathrm{e}}}\right)^2 n^{1-4w}} + 2\mathrm{e}^{-\mathrm{e}\pi\left(1-\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2 n^{1-6w}} \tag{A130}$$

$$= 2\mathrm{e}^{-c_1 n^{1-4w_0}} + 2\mathrm{e}^{-c_2 n^{1-6w_1}}. \tag{A131}$$

This concludes the proof.

References

1. Bhattacharya, P. Estimation of a probability density function and its derivatives. *Sankhyā: Indian J. Stat. Ser. A* **1967**, *29*, 373–382.
2. Dmitriev, Y.G.; Tarasenko, F. On the estimation of functionals of the probability density and its derivatives. *Theory Probab. Appl.* **1974**, *18*, 628–633. [CrossRef]
3. Schuster, E.F. Estimation of a probability density function and its derivatives. *Ann. Math. Stat.* **1969**, *40*, 1187–1195. [CrossRef]
4. Rüschendorf, L. Consistency of estimators for multivariate density functions and for the mode. *Sankhyā: Indian J. Stat. Ser. A* **1977**, *39*, 243–250.
5. Silverman, B.W. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Stat.* **1978**, *6*, 177–184. [CrossRef]
6. Roussas, G.G. Kernel estimates under association: Strong uniform consistency. *Stat. Probab. Lett.* **1991**, *12*, 393–403. [CrossRef]
7. Wertz, W.; Schneider, B. Statistical density estimation: A bibliography. *Int. Stat. Rev. Int. Stat.* **1979**, *47*, 155–175.
8. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: Paris, France, 2009.
9. Donoho, D.L. One-sided inference about functionals of a density. *Ann. Stat.* **1988**, *16*, 1390–1420. [CrossRef]
10. Berisha, V.; Hero, A.O. Empirical non-parametric estimation of the Fisher information. *IEEE Signal Process. Lett.* **2014**, *22*, 988–992. [CrossRef]
11. Spall, J.C. Monte Carlo computation of the Fisher information matrix in nonstandard settings. *J. Comput. Graph. Stat.* **2005**, *14*, 889–909. [CrossRef]
12. Birgé, L.; Massart, P. Estimation of integral functionals of a density. *Ann. Stat.* **1995**, *23*, 11–29. [CrossRef]
13. Cao, W.; Dytso, A.; Fauß, M.; Poor, H.V.; Feng, G. On Nonparametric Estimation of the Fisher Information. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020.
14. Sricharan, K.; Raich, R.; Hero, A.O. Estimation of nonlinear functionals of densities with confidence. *IEEE Trans. Inf. Theory* **2012**, *58*, 4135–4159. [CrossRef]
15. Wu, Y.; Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* **2016**, *62*, 3702–3720. [CrossRef]
16. Han, Y.; Jiao, J.; Weissman, T.; Wu, Y. Optimal rates of entropy estimation over Lipschitz balls. *arXiv* **2017**, arXiv:1711.02141.
17. Verdú, S. Empirical Estimation of Information Measures: A Literature Guide. *Entropy* **2019**, *21*, 720. [CrossRef] [PubMed]

18. Lozano, A.; Tulino, A.M.; Verdú, S. Optimum power allocation for parallel Gaussian channels with arbitrary input distributions. *IEEE Trans. Inf. Theory* **2006**, *52*, 3033–3051. [CrossRef]
19. Gramacki, A.; Sawerwain, M.; Gramacki, J. FPGA-based bandwidth selection for kernel density estimation using high level synthesis approach. *Bull. Pol. Acad. Sci. Tech. Sci.* **2016**, *64*, 821–829. [CrossRef]
20. Massart, P. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **1990**, *18*, 1269–1283. [CrossRef]
21. Carreira-Perpiñán, M.Á.; Williams, C.K. On the Number of Modes of a Gaussian Mixture. Available online: http://www.cs.utoronto.ca/~miguel/papers/ps/sst03.pdf (accessed on 24 April 2021).
22. Cao, W.; Dytso, A.; Fauß, M.; Poor, H.V. MATLAB Codes for Nonparametric Estimation of the Fisher Information. Available online: https://github.com/mifauss/Fisher_Information_Estimation (accessed on 24 April 2021).
23. Esposito, R. On a relation between detection and estimation in decision theory. *Inf. Control* **1968**, *12*, 116–120. [CrossRef]
24. Fozunbal, M. On regret of parametric mismatch in minimum mean square error estimation. In Proceedings of the 2010 IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010.