On Nonparametric Estimation of the Fisher Information with Plug-in Estimators

Wei Cao*, Alex Dytso†, Michael Faußt, H. Vincent Poor†, and Gang Feng*
*National Key Lab of Science and Technology on Communications,
University of Electronic Science and Technology of China
†Department of Electrical Engineering, Princeton University
Email: wcao@std.uestc.edu.cn, {adysto, mfauss, poor}@princeton.edu, fenggang@uestc.edu.cn

***Abstract*—This paper considers a problem of estimation of the Fisher information from a random sample of size n. First, an estimator proposed by Bhattacharya is revisited and improved convergence rates are derived. Second, a new estimator, termed clipped estimator, is proposed. The new estimator is shown to have superior rates of convergence as compared to the Bhattacharya estimator, albeit with different regularity conditions. Third, both of the estimators are evaluated for the practically relevant case of a random variable contaminated by Gaussian noise. Moreover, using Brown's identity, which relates the Fisher information to the minimum mean squared error (MMSE) in Gaussian noise, a consistent estimator for the MMSE is proposed.**

***Index Terms*—Nonparametric estimation, Fisher information, MMSE, kernel estimation.**

I. Introduction

This work, based on an independent random sample $Y_1, \ldots, Y_n$ according to a common probability density function (pdf) f, considers estimation of the *Fisher information* of f, which is given by

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

where f' is the derivative of f.

Estimation of the Fisher information in (1) was first considered by Bhattacharya in [1], where a large sample regime was studied. In [1], a *plug-in estimator* was proposed based on the estimates of f and f' using the *kernel method*. Amongst other things, the work in [1] produced error bounds on the estimation of density and its derivative, and, under some regularity conditions, the proposed Fisher information estimator was shown to be consistent.

Estimation of the derivatives of pdfs is important for the plug-in methods. In particular, the kernel based methods for estimation of the derivatives of a pdf have received considerable attention. For example, the work of Schuster [2] considered estimation of higher-order derivatives of a pdf and has shown that under mild regularity conditions the estimation error for the higher-order derivatives can be controlled by the estimation error for the corresponding *cumulative distribution function* (cdf). The interested reader is also referred to [3]–[6], and [7] and references therein.

This research was supported in part by the U.S. National Science Foundation under Grants CCF-0939370 and CCF-1513915.

As previously mentioned, the estimation of Fisher information was first considered in [1]. The bounds of [1] have been revised by Dmitriev and Tarasenko in [8]. The work of [8] was also the first to consider a problem of entropy estimation. The techniques of [1] and [8] have been generalized by Nadaraya and Sokhadze in [9] to functionals that depend on the first m-th derivatives of the density. The work in [9] made an additional assumption that the density and all of the derivatives were bounded which led to a better convergence rate. In this work, we will recover the rates of [9] with the original Bhattacharya assumptions.

In [10], Donoho proposed a two-step procedure for estimating the Fisher information. In the first step, compute the empirical cdf. In the second step, compute the smallest Fisher information attained on the ball centered at the empirical cdf where the radius of the ball is defined via the Kolmogorov distance. Finally, use the computed Fisher information as an estimate of the actual Fisher information. This method is closely related to the method of Huber splines [11].

Estimation of the *parametric Fisher information*[1] has also received some attention in the literature. Particularly, Spall in [12] proposed to use a plug-in method by first performing nonparametric density estimation by perturbing each of the experiments followed by numerical gradient computation and followed by averaging. A non plug-in method was shown by Berisha and Hero in [13] where it was proposed to estimate an f-divergence and then estimate the parametric Fisher information by using the fact that f-divergences locally behave like the parametric Fisher information.

Finally, we note that estimation of the Fisher information falls under the umbrella of *estimation of nonlinear functionals*; see for example [14]. Most of the information theoretic measures, such as entropy, relative entropy, and mutual information, are nonlinear functionals and have recently received considerable attention; the interested reader is referred to [15]–[17], and [18] and references therein.

The organization of the paper is as follows. Section II revisits the Bhattacharya estimator. In particular, Theorem 2

[1]Let $\{f(x;\theta)\}, \theta \in \Theta$ denote an indexed set of pdfs, the parametric Fisher information is given by $I(\theta) = \int_{t\in\mathbb{R}} \frac{(\partial_\theta f(t;\theta))^2}{f(t;\theta)}\mathrm{d}t$. The definition of the Fisher information in (1) agrees with the parametric one for the shift family, *i.e.*, $f(x;\theta) = f(x-\theta)$.

provides improved rates to those found in [1] and [8]. Section III, to remedy the slow convergence of Bhattacharya estimator, proposes a new estimator, which is termed clipped estimator. In particular, Theorem 3 shows that the clipped estimator has better rates of convergence than the Bhattacharya estimator, albeit with different assumptions on the pdf. Section IV evaluates the convergence rates of the two estimators for the practically relevant case of a random variable contaminated by Gaussian noise. Moreover, using Brown's identity, which relates the Fisher information to the minimum mean squared error (MMSE), a consistent estimator for the MMSE is proposed and the rates of convergence are evaluated in Proposition 1. Section V concludes the paper.

Notation: Throughout the paper deterministic quantities are denoted by lowercase letters, and random variables are denoted by uppercase letters. The expected value and variance of X are denoted by $\mathbb{E}[X]$ and $\mathrm{Var}(X)$, respectively. The gamma function is denoted by $\Gamma(\cdot)$.

II. The Bhattacharya Estimator

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

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

for some $k_n \ge 0$, and where f and f' are estimated by using the kernel method as follows:

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

where $a > 0$ is the bandwidth parameter. The kernel $k(\cdot)$ is assumed to be a continuously differential pdf. For example, in Section IV we will take $k(t) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$.

A. Estimating Density and Its Derivatives

In order to analyze the Bhattacharya estimator it is necessary to obtain rates of convergence for f_n and f_n'. The following theorem, which is largely based on the proof by Schuster in [2], presents such rates. More specifically, the next theorem refines the proof of [2], which relies on the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality for the empirical cdf, by using the best possible constant for the DKW inequality shown in [19].

Theorem 1: Let $r \in \{0, 1\}$ and

$$v_r = \int \left|k^{(r+1)}(t)\right| \mathrm{d}t, \tag{4}$$

$$\delta_{r,a} = \sup_{t\in\mathbb{R}} \left|\mathbb{E}\left[f_n^{(r)}(t)\right] - f^{(r)}(t)\right|. \tag{5}$$

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

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}} \left|f_n^{(r)}(t) - f^{(r)}(t)\right| > \epsilon\right] \le 2\mathrm{e}^{-2n\frac{a^{2r+2}(\epsilon-\delta_{r,a})^2}{v_r^2}}. \tag{6}$$

Proof: See Appendix A. ∎

B. Analysis of the Bhattacharya Estimator

The following theorem is a non-asymptotic refinement of the result obtained by Bhattacharya in [1, Theorem 3] and Dmitriev and Tarasenko in [8, Theorem 1].

Theorem 2: Assume there exists a function ϕ such that

$$\sup_{|t|\le x} \frac{1}{f(t)} \le \phi(x) \text{ for all } x. \tag{7}$$

Then, provided that

$$\sup_{|t|\le k_n} \left|f_n^{(i)}(t) - f^{(i)}(t)\right| \le \epsilon_i,\ i \in \{0, 1\}, \tag{8}$$

and

$$\epsilon_0\phi(k_n) \le 1, \tag{9}$$

the following bound holds:

$$\begin{aligned} &|I(f) - I_n(f_n)| \\ &\le \frac{4\epsilon_1 k_n \rho_{\max}(k_n) + 2\epsilon_1^2 k_n \phi(k_n) + \epsilon_0\phi(k_n) I(f)}{1 - \epsilon_0\phi(k_n)} + c(k_n), \end{aligned} \tag{10}$$

where

$$\rho_{\max}(k_n) = \sup_{|t|\le k_n} \left|\frac{f'(t)}{f(t)}\right|, \tag{11}$$

$$c(k_n) = \int_{|t|\ge k_n} \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t. \tag{12}$$

Proof: See Appendix B. ∎

In fact, the bound in (10) is an improvement on the original bound in [1] and [8], which contains terms of the form $\epsilon_0\phi^4(k_n)$.

Note that, $\phi(k_n)$ in (7) increases with k_n (usually very fast). For example, as will be shown later, $\phi(k_n)$ increases with k_n exponentially for a random variable contaminated by Gaussian noise. This implies that while the Bhattacharya estimator converges, the rate of convergence is extremely slow, preventing the estimators from being practical.

The main problem in the convergence analysis of the estimator in (2) is that $1/f_n(x)$ is only bounded if $f(x) > \epsilon_0$. 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 rates of convergence.

III. A Clipped Estimator

In order to remedy the slow convergence of the Bhattacharya estimator, we dispense with the tail assumption in (7) but introduce a new assumption that the unknown true score function ρ is bounded (in absolute value) by a known function $\rho_{\max}$. This allows us to clip $f_n'(x)/f_n(x)$ and in turn $1/f_n(x)$ without affecting the consistency of the estimator.

Theorem 3: Assume there exists a function $\rho_{\max}$ such that

$$|\rho(t)| \le |\rho_{\max}(t)|, \tag{13}$$

for all $t \in \mathbb{R}$ and let

$$I_n^{\mathrm{c}}(f_n) = \int_{-k_n}^{k_n} \min\left\{|\rho_n(t)|, |\rho_{\max}(t)|\right\} |f_n'(t)| \, \mathrm{d}t, \quad (14)$$

where

$$\rho_n(t) = \frac{f_n'(t)}{f_n(t)}. \quad (15)$$

Under the assumptions in (8), it holds that

$$\begin{aligned} |I(f) - I_n^{\mathrm{c}}(f_n)| &\leq \max\{4\epsilon_1\Phi^1(k_n) + 2\epsilon_0\Phi^2(k_n) + c(k_n), \\ &\qquad 2\epsilon_1\Phi^1_{\max}(k_n) + 2\epsilon_0\Phi^2_{\max}(k_n)\} \quad (16) \\ &\leq 4\epsilon_1\Phi^1_{\max}(k_n) + 2\epsilon_0\Phi^2_{\max}(k_n) + c(k_n), \quad (17) \end{aligned}$$

where $c(k_n)$ is defined in (12) and

$$\Phi^m(x) = \int_{-x}^{x} |\rho^m(t)| \, \mathrm{d}t, \quad (18)$$

$$\Phi^m_{\max}(x) = \int_{-x}^{x} |\rho^m_{\max}(t)| \, \mathrm{d}t. \quad (19)$$

Proof: See Appendix C. ∎

Note that, although $\rho_{\max}(k_n)$ also increases with k_n, it usually increases much slower than $\phi(k_n)$. For example, as shown later, $\rho_{\max}(k_n)$ is linear in k_n in Gaussian noise case. As a result, a faster convergence rate can be shown for the clipped estimator.

IV. Estimation of the Fisher Information of a Random Variable Contaminated by Gaussian Noise

This section evaluates the results of Section II and Section III for an importance case of a random variable contaminated by Gaussian noise. To this end, we let f_Y denote the pdf of a random variable

$$Y = \sqrt{\mathrm{snr}}X + Z, \quad (20)$$

where X is an arbitrary random variable and 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 a very mild assumption that X has a finite second moment but otherwise is an arbitrary random variable.

The Fisher information of f_Y has several important identities that connect it to other estimation and information measures. In particular, the Fisher information can be connected to the quadratic Bayesian risk or the MMSE as follows:

$$I(f_Y) = 1 - \mathrm{snr}\,\mathrm{mmse}(X|Y), \quad (21)$$

where the MMSE is given by

$$\mathrm{mmse}(X|Y) = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right]. \quad (22)$$

In statistical literature the relationship is known as Brown's identity [20]. The Fisher information can also be connected to information measures such as mutual information, entropy, and continuous entropy via the following identities: let $Y_t = \sqrt{t}X + Z$, then

$$2I(X;Y_\gamma) = \int_0^\gamma \mathrm{mmse}(X|Y_{\mathrm{snr}})\,\mathrm{dsnr} = \int_0^\gamma \frac{1 - I(f_{Y_{\mathrm{snr}}})}{\mathrm{snr}}\,\mathrm{dsnr}, \quad (23)$$

$$2H(X) = \int_0^\infty \mathrm{mmse}(X|Y_{\mathrm{snr}})\,\mathrm{dsnr} = \int_0^\infty \frac{1 - I(f_{Y_{\mathrm{snr}}})}{\mathrm{snr}}\,\mathrm{dsnr}, \quad (24)$$

$$2h(X) = \int_0^\infty \frac{1 - I(f_{Y_{\mathrm{snr}}})}{\mathrm{snr}} - \frac{1}{2\pi\mathrm{e} + \mathrm{snr}}\,\mathrm{dsnr}. \quad (25)$$

The relationship in (23) is known as the I-MMSE identity and was shown in [21] together with the identity in (24). The identity in (25) is known as De Bruijin's identity and holds if $\lim_{\mathrm{snr}\to\infty} h\left(X + \frac{1}{\sqrt{\mathrm{snr}}}Z\right) = h(X)$, and was show in [22]; see also [21] for an alternative proof.

Using the estimator of the Fisher information together with the above identities it might be possible to construct estimators for mutual information, entropy, and continuous entropy. In what follows, we will use the identity in (21) to propose an estimator for the MMSE and will evaluate the performance of that estimator. We note that an idea of using the I-MMSE identity in (23) to estimate the mutual information has been already used in [23]. Note, however, that the approach in [23] requires the existence of all moments of the distribution of X, while here we only require the existence of the second moment.

The following lemma evaluates the quantities appearing in Section II and Section III that are needed to evaluate the error bounds for the Bhattacharya and clipped estimators.

Lemma 1: Let $k(t) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}$. Then,

$$\delta_{r,a} = a \cdot \begin{cases} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\mathrm{e}}} & r = 0 \\ \frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}} & r = 1 \end{cases}, \quad (26)$$

$$v_r = \begin{cases} \sqrt{\frac{2}{\pi}} & r = 0 \\ \sqrt{\frac{2}{\mathrm{e}\pi}} & r = 1 \end{cases}, \quad (27)$$

$$\rho_{\max}(k_n) \leq \sqrt{3\,\mathrm{snr}\,\mathrm{Var}(X)} + 3k_n, \quad (28)$$

$$I(f_Y) \leq 1, \quad (29)$$

$$\phi(t) \leq \sqrt{2\pi}\mathrm{e}^{(t^2 + \mathrm{snr}\,\mathbb{E}[X^2])}, \quad (30)$$

$$c(k_n) \leq \inf_{v>0} \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v + \frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}} \left(\frac{\mathrm{snr}\,\mathbb{E}[|X|^2] + 1}{k_n^2}\right)^{\frac{v}{1+v}}. \quad (31)$$

Proof: See Appendix E. ∎

A. Convergence of the Bhattacharya Estimator

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

Theorem 4: If $a = n^{-w}$, where $w \in \left(0, \frac{1}{6}\right)$, and $k_n = \sqrt{u\log(n)}$, where $u \in (0, w)$, then

$$\mathbb{P}\left[|I_n(f_n) - I(f_Y)| \geq \varepsilon_n\right] \leq 2\mathrm{e}^{-c_1 n^{1-4w}} + 2\mathrm{e}^{-c_2 n^{1-6w}}, \quad (32)$$

where

$$\varepsilon_n \leq \frac{n^{-w}\sqrt{u\log(n)}\left(c_3 + 12\sqrt{u\log(n)} + 2c_5 n^{u-w}\right)}{1 - n^{u-w}} + \frac{c_4}{k_n} + \frac{c_5}{n^{w-u} - 1}, \quad (33)$$

and where the constants are given by

$$c_1 = \pi\left(1 - \frac{1}{\sqrt{2\pi \mathrm{e}}}\right)^2, \quad c_2 = \mathrm{e}\pi\left(1 - \frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2, \tag{34}$$

$$c_3 = 4\sqrt{3\,\mathrm{snr}\,\mathrm{Var}(X)}, \quad c_4 = \frac{2\Gamma^{\frac{1}{2}}\left(\frac{3}{2}\right)\sqrt{\mathrm{snr}\,\mathbb{E}[|X|^2]+1}}{\pi^{\frac{1}{4}}}, \tag{35}$$

$$c_5 = \sqrt{2\pi}\mathrm{e}^{\mathrm{snr}\,\mathbb{E}\left[X^2\right]}. \tag{36}$$

Proof: See Appendix F. ∎

Remark 1: If $a = n^{-w}$, where $w \in \left(0, \frac{1}{6}\right)$, and $k_n = \sqrt{u\log(n)}$, where $u \in (0, w)$, then $I_n(f_n)$ converges to $I(f_Y)$ with probability 1. In other words, the estimator is consistent.

The choice of u and w results in a trade-off between the convergence rate of the probability and the precision ε_n. Specifically, on the one hand, large values of u and w result in better precision at the cost of a lower convergence rate. On the other hand, small values of u and w improve the convergence rate but deteriorate the precision.

B. Convergence of the Clipped Estimator

From the evaluation of the Bhattacharya 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 the Bhattacharya estimator. Specifically, the precision will be shown to decay polynomially in n instead of logarithmically.

By utilizing the results in Theorem 1 and Lemma 1, we specialize the result in Theorem 3 to f_Y.

Theorem 5: If $a = n^{-w}$, where $w \in \left(0, \frac{1}{4}\right)$, and $k_n = n^u$, where $u \in (0, \frac{w}{3})$, then

$$\mathbb{P}\left[|I_n^{\mathrm{c}}(f_n) - I(f_Y)| \geq \varepsilon_n\right] \leq 2\mathrm{e}^{-c_1 n^{1-4w}} + 2\mathrm{e}^{-c_2 n^{1-6w}}, \tag{37}$$

where

$$\varepsilon_n \leq 12n^{u-w}\left(c_3 + 2n^u + n^{2u}\right) + \frac{c_4}{k_n}, \tag{38}$$

and the constants $c_i, i \in [1:4]$ are as in Theorem 4.

Proof: See Appendix G. ∎

C. Applications to the Estimation of the MMSE

Using Brown's identity in (21) we propose the following estimator for the MMSE:

$$\mathrm{mmse}_n(X|Y) = \frac{1 - I_n^{\mathrm{c}}(f_n)}{\mathrm{snr}}. \tag{39}$$

Using the above estimator, the result in Theorem 5 can now be extended to the MMSE as follows.

Proposition 1: If $a = n^{-w}$, where $w \in \left(0, \frac{1}{4}\right)$, and $k_n = n^u$, where $u \in (0, \frac{w}{3})$, then

$$\begin{aligned}&\mathbb{P}\left[|\mathrm{mmse}_n(X|Y) - \mathrm{mmse}(X|Y)| \geq \mathrm{snr}\,\varepsilon_n\right] \\ &\leq 2\mathrm{e}^{-c_1 n^{1-4w}} + 2\mathrm{e}^{-c_2 n^{1-6w}}\end{aligned} \tag{40}$$

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

V. Conclusion

This work has focused on the estimation of Fisher information of a random variable based on the plug-in estimators of the density and its derivative. The paper considered two estimators of the Fisher information. The first estimator is the estimator due to Bhattacharya. For this estimator, new sharper convergence results have been provided. The paper has also proposed a second estimator, termed clipped estimator, which provides better convergence rates than the Bhattacharya estimator.

The results of both estimators have been specialized to the practically relevant case of a Gaussian noise contaminated random variable. Moreover, using special proprieties of the Gaussian noise case, an estimator for the minimum mean square error (MMSE) has been proposed, and the convergence rates have been analyzed. This was done by using Brown's identity, which connects the Fisher information and the MMSE.

An interesting future direction would be to study the Gaussian noise case further. For example, by making further assumptions on the variable X (only a finite second-moment assumption has been made here), tighter bounds might be possible.

Appendix A
Proof of Theorem 1

Our starting place is the following bound due to [2, p.1188]:

$$\sup_{t\in\mathbb{R}}\left|\mathbb{E}[f_n^{(r)}(t)] - f_n^{(r)}(t)\right| \leq \frac{v_r}{a^{r+1}}\sup_{t\in\mathbb{R}}|F_n(t) - F(t)|, \tag{41}$$

where F is the cdf of f, F_n is an empirical cdf, and v_r is defined as (4). Now let $\delta_{r,a}$ to be (5), and consider the following sequence of bounds:

$$\begin{aligned}&\mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n^{(r)}(t) - f^{(r)}(t)\right| > \epsilon\right] \\ &\leq \mathbb{P}\left[\sup_{t\in\mathbb{R}}\left|f_n^{(r)}(t) - \mathbb{E}[f_n^{(r)}(t)]\right| > \epsilon - \delta_{r,a}\right] &(42)\\ &\leq \mathbb{P}\left[\sup_{t\in\mathbb{R}}|F_n(t) - F(t)| > \frac{a^{r+1}(\epsilon - \delta_{r,a})}{v_r}\right] &(43)\\ &\leq 2\mathrm{e}^{-2n\frac{a^{2r+2}(\epsilon-\delta_{r,a})^2}{v_r^2}}, &(44)\end{aligned}$$

where (42) follows by using the triangle inequality; (43) follows by using the bound in (41); and (44) follows by using the sharp DKW inequality [19]

$$\mathbb{P}\left[\sup_{t\in\mathbb{R}}|F_n(t) - F(t)]| > \epsilon\right] \leq 2\mathrm{e}^{-2n\epsilon^2}. \tag{45}$$

Appendix B
Proof of Theorem 2

First, using the triangle inequality we have that

$$|I(f) - I_n(f_n)| \leq \left|\int_{|t|\leq k_n}\frac{(f_n'(t))^2}{f_n(t)} - \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right| + c(k_n). \tag{46}$$

Next, we bound the first term in (46)

$$\left|\int_{|t|\le k_n} \frac{(f_n'(t))^2}{f_n(t)} - \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right| \tag{47}$$

$$= \left|\int_{|t|\le k_n} \frac{f(t)(f_n'(t))^2 - f_n(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| \tag{48}$$

$$\le \left|\int_{|t|\le k_n} \frac{f(t)(f_n'(t))^2 - f(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| \tag{49}$$

$$+ \left|\int_{|t|\le k_n} \frac{f(t)(f'(t))^2 - f_n(t)(f'(t))^2}{f_n(t)f(t)}\,\mathrm{d}t\right| \tag{50}$$

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

$$+ \left|\int_{|t|\le k_n} \frac{f_n(t)-f(t)}{f_n(t)}\frac{(f'(t))^2}{f(t)}\,\mathrm{d}t\right| \tag{51}$$

$$\le \sup_{|t|\le k_n} \frac{|f_n'(t)+f'(t)|}{f_n(t)}\,|f_n(t)-f(t)|\,2k_n$$

$$+ \sup_{|t|\le k_n} \frac{|f_n(t)-f(t)|}{f_n(t)} \int_{|t|\le k_n} \frac{(f'(t))^2}{f(t)}\,\mathrm{d}t \tag{52}$$

$$\le \sup_{|t|\le k_n} \frac{|f_n'(t)+f'(t)|}{f_n(t)}\epsilon_1 2k_n + \sup_{|t|\le k_n}\frac{1}{f_n(t)}\epsilon_0 I(f), \tag{53}$$

where the last bound follows from the assumptions in (8). Now consider the first term in (53)

$$\sup_{|t|\le k_n} \frac{|f_n'(t)+f'(t)|}{f_n(t)} \le \sup_{|t|\le k_n} \frac{2|f'(t)|+\epsilon_1}{f_n(t)} \tag{54}$$

$$\le \sup_{|t|\le k_n} \frac{2|f'(t)|+\epsilon_1}{f(t)-f(t)+f_n(t)} \tag{55}$$

$$\le \sup_{|t|\le k_n} \frac{2|f'(t)|+\epsilon_1}{f(t)-\epsilon_0} \tag{56}$$

$$= \sup_{|t|\le k_n} \frac{2|\frac{f'(t)}{f(t)}| + \frac{\epsilon_1}{f(t)}}{1-\frac{\epsilon_0}{f(t)}} \tag{57}$$

$$\le \frac{2\sup_{|t|\le k_n}|\frac{f'(t)}{f(t)}| + \epsilon_1\phi(k_n)}{1-\epsilon_0\phi(k_n)}, \tag{58}$$

where the bound in (56) follows from the assumptions in (8) and the properties of ϕ that imply

$$\epsilon_0\phi(k_n)\le 1 \Rightarrow \frac{\epsilon_0}{f(t)} \le 1, \forall |t|\le k_n \tag{59}$$

$$\Rightarrow \epsilon_0 \le f(t), \forall |t|\le k_n; \tag{60}$$

and the bound in (58) follows form the definition of ϕ in (7). Now consider the second term in (53)

$$\sup_{|t|\le k_n}\frac{1}{f_n(t)} = \sup_{|t|\le k_n}\frac{1}{f_n(t)-f(t)+f(t)} \tag{61}$$

$$\le \sup_{|t|\le k_n}\frac{1}{f(t)-\epsilon_0} \tag{62}$$

$$= \sup_{|t|\le k_n}\frac{1}{1-\frac{\epsilon_0}{f(t)}}\frac{1}{f(t)} \tag{63}$$

$$\le \frac{1}{1-\epsilon_0\phi(k_n)}\phi(k_n), \tag{64}$$

where (63) follows by using similar steps leading to the bound in (56); and (64) follows from the definition of ϕ.

Combining the bounds in (46), (53), (58), and (64) concludes the proof.

Appendix C
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 a gross overestimation, yet permissive enough to avoid a 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

$$\underline{I}_n(f_n) = \int_{-k_n}^{k_n} \frac{\lceil f_n'(t)-\epsilon_1\rfloor^2}{f_n(t)+\epsilon_0}\,\mathrm{d}t, \tag{65}$$

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

$$\lceil f(t)-\epsilon\rfloor = \begin{cases} f(t)-\epsilon, & f(t)>\epsilon \\ 0, & -\epsilon\le f(t)\le\epsilon \\ f(t)+\epsilon, & f(t)<-\epsilon. \end{cases} \tag{66}$$

Next, consider the estimator

$$\overline{I}_n(f_n) = \int_{-k_n}^{k_n} \frac{\lceil f_n'(t)-\gamma_{1,n}(t)\rfloor^2}{f_n(t)+\gamma_{0,n}(t)}\,\mathrm{d}t, \tag{67}$$

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

$$|\rho_n(t)| \le |\rho_{\max}(t)|, \tag{68}$$

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

$$|\rho_n(t)| > |\rho_{\max}(t)|, \tag{69}$$

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

$$\frac{\lceil f_n'(t)-\gamma_{1,n}(t)\rfloor}{f_n(t)+\gamma_{0,n}(t)} = \rho_{\max}(t). \tag{70}$$

Note that since

$$\left|\frac{\lceil f_n'(t)-\epsilon_1\rfloor}{f_n(t)+\epsilon_0}\right| \le |\rho(t)| \le |\rho_{\max}(t)|, \tag{71}$$

this is always possible.

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

$$\underline{I}_n(f_n) \le I(f), \tag{72}$$

$$\underline{I}_n(f_n) \le I_n^{\mathrm{c}}(f), \tag{73}$$

$$I_n^{\mathrm{c}}(f) \le \overline{I}_n(f_n) + \epsilon_0\Phi^2_{\max}(k_n), \tag{74}$$

$$I(f) - \underline{I}_n(f_n) \le 4\epsilon_1\Phi^1(k_n) + 2\epsilon_0\Phi^2(k_n) + c(k_n), \tag{75}$$

$$\overline{I}_n(f_n) - \underline{I}_n(f_n) \le 2\epsilon_1\Phi^1_{\max}(k_n) + \epsilon_0\Phi^2_{\max}(k_n). \tag{76}$$

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

$$I(f) - I_n^{\mathrm{c}}(f) \le I(f) - \underline{I}_n(f) \tag{77}$$

$$\leq 4\epsilon_1\Phi^1(k_n) + 2\epsilon_0\Phi^2(k_n) + c(k_n). \quad (78)$$

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

$$I_n^{\mathrm{c}}(f) - I(f) \leq \overline{I}_n(f_n) - \underline{I}_n(f) + \epsilon_0\Phi^2_{\max}(k_n) \quad (79)$$
$$\leq 2\epsilon_1\Phi^1_{\max}(k_n) + 2\epsilon_0\Phi^2_{\max}(k_n). \quad (80)$$

The statement in Theorem 3 follows.

Appendix D
Proof of Estimator Relations in Theorem 3

The bound in (72) follows directly from the fact that under the assumptions in (8)

$$\frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \leq \frac{(f'(t))^2}{f(t)}. \quad (81)$$

Analogously, (73) follows from

$$\frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0} \leq |\rho(t)||\lceil f_n'(t) - \epsilon_1 \rfloor| \leq |\rho(t)||f_n'(t)| \quad (82)$$

In order to show (75), note that under the assumptions in (8) it holds that

$$f_n(t) + \epsilon_0 \geq f(t), \quad (83)$$
$$|\lceil f_n'(t) - \epsilon_1 \rfloor| \leq |f'(t)|, \quad (84)$$
$$(f_n(t) + \epsilon_0) - f(t) \leq 2\epsilon_0, \quad (85)$$
$$|\lceil f_n'(t) - \epsilon_1 \rfloor - f'(t) \leq 2\epsilon_1. \quad (86)$$

Hence, in analogy to Theorem 2, the estimation error of $\underline{I}_n(f_n)$ can be written as

$$I(f) - \underline{I}_n(f_n) = \int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)} - \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0}\, \mathrm{d}t + c(k_n). \quad (87)$$

Using the same arguments as in the proof of Theorem 2, the integral term on the right hand side of (87) can be bounded by

$$\int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)} - \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2}{f_n(t) + \epsilon_0}\, \mathrm{d}t \quad (88)$$
$$= \left| \int_{-k_n}^{k_n} \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2 f(t) - (f'(t))^2(f_n(t) + \epsilon_0)}{f(t)(f_n(t) + \epsilon_0)}\, \mathrm{d}t \right|$$
$$= \left| \int_{-k_n}^{k_n} \frac{\lceil f_n'(t) - \epsilon_1 \rfloor^2 f(t) - (f'(t))^2(f_n(t) + \epsilon_0)}{f(t)(f_n(t) + \epsilon_0)}\, \mathrm{d}t \right|$$
$$\leq \left| \int_{-k_n}^{k_n} |\lceil f_n'(t) - \epsilon_1 \rfloor - f'(t)| \frac{|\lceil f_n'(t) - \epsilon_1 \rfloor + f'(t)|}{f_n(t) + \epsilon_0}\, \mathrm{d}t \right|$$
$$+ \int_{-k_n}^{k_n} |f(t) - (f_n(t) + \varepsilon_0)| \frac{(f'(t))^2}{f(t)(f_n(t) + \varepsilon_0)}\, \mathrm{d}t$$
$$\leq 2\epsilon_1 \int_{-k_n}^{k_n} \frac{|\lceil f_n'(t) - \epsilon_1 \rfloor| + |f'(t)|}{f_n(t) + \epsilon_0}\, \mathrm{d}t \quad (89)$$
$$+ 2\epsilon_0 \int_{-k_n}^{k_n} \frac{(f'(t))^2}{f(t)(f_n(t) + \epsilon_0)}\, \mathrm{d}t$$
$$\leq 2\epsilon_1 \int_{-k_n}^{k_n} 2\left|\frac{f'(t)}{f(t)}\right| \mathrm{d}t + 2\epsilon_0 \int_{-k_n}^{k_n} \left|\frac{f'(t)}{f(t)}\right|^2 \mathrm{d}t \quad (90)$$
$$\leq 4\epsilon_1 \int_{-k_n}^{k_n} |\rho(t)| + 2\epsilon_0 \int_{-k_n}^{k_n} \rho^2(t)\, \mathrm{d}t \quad (91)$$
$$= 4\epsilon_1\, \Phi^1(k_n) + 2\varepsilon_0\, \Phi^2(k_n). \quad (92)$$

Using the same steps, it is not difficult to show (76), *i.e.*,

$$\overline{I}_n(f_n) - \underline{I}_n(f_n) \leq 2\epsilon_1\Phi^1_{\max}(k_n) + \epsilon_0\Phi^2_{\max}(k_n), \quad (93)$$

where the factor 2 does not arise since, in contrast to (85) and (86),

$$\lceil f_n(t) + \epsilon_0 \rfloor - \lceil f_n(t) + \gamma_{0,n}(t) \rfloor \leq \epsilon_0, \quad (94)$$
$$\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor - \lceil f_n'(t) - \epsilon_1 \rfloor \leq \epsilon_1, \quad (95)$$

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

In order to show (74), first note that for $|\rho_n(t)| \leq |\rho_{\max}(t)|$ it holds that

$$\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor^2}{f_n(t) + \gamma_{0,n}(t)} = \frac{(f_n'(t))^2}{f_n(t)} = |\rho_n(t)||f_n(t)|, \quad (96)$$

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 $|\rho_n(t)| \geq |\rho_{\max}|$ on some region of $[-k_n, k_n]$. On this region it holds that

$$\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor^2}{f_n(t) + \gamma_{0,n}(t)} \quad (97)$$
$$= \left(\frac{\lceil f_n'(t) - \gamma_{1,n}(t) \rfloor}{f_n(t) + \gamma_{0,n}(t)} \right)^2 (f_n(t) + \gamma_{0,n}(t)) \quad (98)$$
$$\leq \rho^2_{\max}(t)(f_n(t) + \varepsilon_0). \quad (99)$$

Consequently,

$$I_n^{\mathrm{c}}(f_n) - \overline{I}(f_n) \quad (100)$$
$$\leq \int_{-k_n}^{k_n} \rho^2_{\max}(t)(f_n(t) + \varepsilon_0) - \rho^2_{\max}(t) f_n(t)\, \mathrm{d}t \quad (101)$$
$$= \epsilon_0\Phi^2_{\max}(k_n). \quad (102)$$

Appendix E
Proof of Lemma 1

We begin by bounding v_r and $\delta_{r,a}$ terms needed for the estimation of density and it's derivative. First,

$$v_0 = \int |t| k(t)\, \mathrm{d}t = \sqrt{\frac{2}{\pi}}, \quad (103)$$
$$v_1 = \int \left|t^2 - 1\right| k(t)\, \mathrm{d}t = 2\sqrt{\frac{2}{\mathrm{e}\pi}}. \quad (104)$$

Second,

$$\delta_{r,a} = \left| \mathbb{E}[f_n^{(r)}(t)] - f_Y^{(r)}(t) \right|$$
$$= \left| \int \frac{1}{a} k\left(\frac{t-y}{a}\right) \left(f_Y^{(r)}(y) - f_Y^{(r)}(t) \right) \mathrm{d}y \right| \quad (105)$$
$$= \left| \int k(y) \left(f_Y^{(r)}(t + ay) - f_Y^{(r)}(t) \right) \mathrm{d}y \right| \quad (106)$$
$$\leq \sup_{t \in \mathbb{R}} \left| f_Y^{(r+1)}(t) \right| \int k(y)\, a|y|\, \mathrm{d}y \quad (107)$$

$$= a\sqrt{\frac{2}{\pi}} \sup_{t\in\mathbb{R}} \left| f_Y^{(r+1)}(t) \right|. \tag{108}$$

Now, for $r = 0$,

$$\left| f_Y^{(1)}(t) \right| = \left| \mathbb{E}\left[(t - \sqrt{\mathrm{snr}}X) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(t-\sqrt{\mathrm{snr}}X)^2}{2}} \right] \right| \tag{109}$$

$$\leq \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\mathrm{e}}}, \tag{110}$$

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

$$\left| f_Y^{(2)}(t) \right| = \left| \mathbb{E}\left[\left((t - \sqrt{\mathrm{snr}}X)^2 - 1 \right) \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(t-\sqrt{\mathrm{snr}}X)^2}{2}} \right] \right| \tag{111}$$

$$\leq \frac{1}{\sqrt{2\pi}} \frac{2}{\mathrm{e}} + \frac{1}{\sqrt{2\pi}}, \tag{112}$$

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

Next we bound the score function ρ_Y

$$|\rho_Y(t)| = \left| \frac{f_Y'(t)}{f_Y(t)} \right| = \left| \sqrt{\mathrm{snr}}\, \mathbb{E}[X|Y = t] - t \right| \tag{113}$$

$$\leq \sqrt{\mathrm{snr}}\, \mathbb{E}\left[|X| \, |Y = t \right] + |t| \tag{114}$$

$$\leq \sqrt{\mathrm{snr}} \sqrt{\mathbb{E}\left[X^2 | Y = t \right]} + |t| \tag{115}$$

$$\leq \sqrt{3\,\mathrm{snr}\,\mathrm{Var}(X) + 4t^2} + |t| \tag{116}$$

$$\leq \sqrt{3\,\mathrm{snr}\,\mathrm{Var}(X)} + 3|t|, \tag{117}$$

where the equality in (113) follows by using identify $\frac{f_Y'(t)}{f_Y(t)} = \sqrt{\mathrm{snr}}\, \mathbb{E}[X|Y = t] - t$ [24]; the inequality in (115) follows from Jensen's inequality; and the inequality in (116) follows from the bound in [25, Proposition 1.2]. Using the bound in (117) it follows that

$$\rho_{\max}(k_n) = \max_{|t|\leq k_n} |\rho(t)| \leq \sqrt{3\,\mathrm{snr}\,\mathrm{Var}(X)} + 3k_n. \tag{118}$$

Now we bound the Fisher information by using the identify between the Fisher information and the MMSE, we have that

$$I(f_Y) = 1 - \mathrm{snr}\,\mathrm{mmse}(X, \mathrm{snr}) \leq 1. \tag{119}$$

Then, we provide a function ϕ

$$f_Y(t) = \mathbb{E}\left[\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(t-\sqrt{\mathrm{snr}}X)^2}{2}} \right] \tag{120}$$

$$\geq \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{\mathbb{E}\left[(t-\sqrt{\mathrm{snr}}X)^2\right]}{2}} \tag{121}$$

$$\geq \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left(t^2 + \mathrm{snr}\,\mathbb{E}[X^2]\right)} \tag{122}$$

where the inequalities follow by using Jensen's inequality and the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

Finally, we bound $c(k_n)$. Choose some $v > 0$

$$c(k_n) = \mathbb{E}\left[\rho_Y^2(Y) 1_{\{|Y|\geq k_n\}} \right] \tag{123}$$

$$\leq \mathbb{E}^{\frac{1}{1+v}} \left[|\rho_Y(Y)|^{2(1+v)} \right] \mathbb{P}^{\frac{v}{1+v}} \left[|Y| \geq k_n \right] \tag{124}$$

$$= \mathbb{E}^{\frac{1}{1+v}} \left[|\,\mathbb{E}[Z|Y]|^{2(1+v)} \right] \mathbb{P}^{\frac{v}{1+v}} \left[|Y| \geq k_n \right] \tag{125}$$

$$\leq \mathbb{E}^{\frac{1}{1+v}} \left[|Z|^{2(1+v)} \right] \mathbb{P}^{\frac{v}{1+v}} \left[|Y| \geq k_n \right] \tag{126}$$

$$= \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v + \frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}} \mathbb{P}^{\frac{v}{1+v}} \left[|Y| \geq k_n \right] \tag{127}$$

$$= \frac{2\Gamma^{\frac{1}{(1+v)}}\left(v + \frac{1}{2}\right)}{\pi^{\frac{1}{2(1+v)}}} \left(\frac{\mathrm{snr}\,\mathbb{E}[|X|^2] + 1}{k_n^2} \right)^{\frac{v}{1+v}} \tag{128}$$

where (124) follows from Holder's inequality; (125) follows by using the identity

$$\rho_Y(t) = \sqrt{\mathrm{snr}}\,\mathbb{E}[X|Y = t] - t = -\,\mathbb{E}[Z|Y = t]; \tag{129}$$

and (128) follows from Markov's inequality. This concludes the proof.

Appendix F
Proof of Theorem 4

Denote

$$\varepsilon_n = \frac{4\epsilon k_n \rho_{\max}(k_n) + \epsilon\phi(k_n) + 2\epsilon^2 k_n \phi(k_n)}{1 - \epsilon\phi(k_n)} + c(k_n). \tag{130}$$

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

$$\epsilon > \delta_{r,a}, \tag{131a}$$

$$a^{2r+2}(\epsilon - \delta_{r,a})^2 \gg \frac{1}{n}, \tag{131b}$$

$$\lim_{n\to\infty} \epsilon^2 k_n \phi(k_n) = 0, \tag{131c}$$

$$\lim_{n\to\infty} \epsilon\phi(k_n) = 0, \tag{131d}$$

$$\lim_{n\to\infty} c(k_n) = 0. \tag{131e}$$

To satisfy (131), we choose

$$a = n^{-w}, w \in \left(0, \frac{1}{6}\right), \tag{132}$$

$$k_n = \sqrt{u\log(n)}, u \in (0, w), \tag{133}$$

$$\epsilon = a. \tag{134}$$

Denote

$$\beta_r = \begin{cases} \left(1 - \frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2, & r = 0 \\ \left(1 - \frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2, & r = 1. \end{cases} \tag{135}$$

Then, together with the bounds in Lemma 1, several quantities in (131) of our concern are as follows:

$$a^{2r+2}(\epsilon - \delta_{r,a})^2 = \beta_r n^{(2r+r)w}, \tag{136a}$$

$$\epsilon^2 k_n \phi(k_n) \leq c_5 n^{u-2w} \sqrt{u\log(n)}, \tag{136b}$$

$$\epsilon\phi(k_n) = c_5 n^{u-w}, \tag{136c}$$

$$c(k_n) \leq \frac{c_4}{k_n}, \tag{136d}$$

which yield (33).

Since (8) leads to (10), one obtains

$$\mathbb{P}\left[|I_n(f_n) - I(f_Y)| \geq \varepsilon_n \right]$$

$$\leq \mathbb{P}\left[\sup_{|t|\leq k_n} |f_n(t) - f_Y(t)| \geq \epsilon\right] + \mathbb{P}\left[\sup_{|t|\leq k_n} |f_n'(t) - f_Y'(t)| \geq \epsilon\right] \tag{137}$$

$$\leq \mathbb{P}\left[\sup_{t\in\mathbb{R}} |f_n(t) - f_Y(t)| > \epsilon\right] + \mathbb{P}\left[\sup_{t\in\mathbb{R}} |f_n'(t) - f_Y'(t)| > \epsilon\right] \tag{138}$$

$$\leq 2\mathrm{e}^{-n\pi a^2\left(\epsilon - a\frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2} + 2\mathrm{e}^{-n\mathrm{e}\pi a^4\left(\epsilon - a\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2} \tag{139}$$

$$= 2\mathrm{e}^{-\pi\left(1-\frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2 n^{1-4w}} + 2\mathrm{e}^{-\mathrm{e}\pi\left(1-\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2 n^{1-6w}}, \tag{140}$$

where the inequality in (139) follows from Theorem 1, and the last step follows from (133), (132), and (134). This concludes the proof.

Appendix G
Proof of Theorem 5

Denote

$$\varepsilon_n = 8\epsilon\Phi^1_{\max}(k_n) + 4\epsilon\Phi^2_{\max}(k_n) + c(k_n). \tag{141}$$

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

$$\epsilon > \delta_{r,a}, \tag{142a}$$

$$a^{2r+2}(\epsilon - \delta_{r,a})^2 \gg \frac{1}{n}, \tag{142b}$$

$$\lim_{n\to\infty} \epsilon\Phi^1_{\max}(k_n) = 0, \tag{142c}$$

$$\lim_{n\to\infty} \epsilon\Phi^2_{\max}(k_n) = 0, \tag{142d}$$

$$\lim_{n\to\infty} c(k_n) = 0. \tag{142e}$$

To satisfy (142), we choose

$$a = n^{-w}, w \in \left(0, \frac{1}{4}\right), \tag{143}$$

$$k_n = n^u, u \in \left(0, \frac{w}{3}\right), \tag{144}$$

$$\epsilon = a. \tag{145}$$

Then, together with the bounds in Lemma 1, several quantities in (142) of our concern are as follows:

$$a_r^{2r+2}(\epsilon_r - \delta_{r,a})^2 = \beta_r n^{(2r+r)w_r}, r = 0, 1, \tag{146a}$$

$$\epsilon\Phi^1_{\max}(k_n) \leq n^{u-w}\left(2c_3 + 3n^u\right), \tag{146b}$$

$$\epsilon\Phi^2_{\max}(k_n) \leq 2n^{u-w}\left(c_3 + 3n^u + 3n^{2u}\right), \tag{146c}$$

$$c(k_n) \leq \frac{c_4}{k_n}, \tag{146d}$$

which yield (38).

By using similar steps leading to (140), we have that

$$\mathbb{P}\left[|I_n^{\mathrm{c}}(f_n) - I(f_Y)| \geq \varepsilon_n\right] \leq 2\mathrm{e}^{-\pi\left(1-\frac{1}{\sqrt{2\pi\mathrm{e}}}\right)^2 n^{1-4w}} + 2\mathrm{e}^{-\mathrm{e}\pi\left(1-\frac{\frac{2}{\mathrm{e}}+1}{\sqrt{2\pi}}\right)^2 n^{1-6w}}. \tag{147}$$

This concludes the proof.

References

[1] P. Bhattacharya, "Estimation of a probability density function and its derivatives," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 373–382, 1967.

[2] E. F. Schuster, "Estimation of a probability density function and its derivatives," *The Annals of Mathematical Statistics*, vol. 40, no. 4, pp. 1187–1195, 1969.

[3] L. Rüschendorf, "Consistency of estimators for multivariate density functions and for the mode," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 243–250, 1977.

[4] B. W. Silverman, "Weak and strong uniform consistency of the kernel estimate of a density and its derivatives," *The Annals of Statistics*, pp. 177–184, 1978.

[5] G. G. Roussas, "Kernel estimates under association: Strong uniform consistency," *Statistics & Probability Letters*, vol. 12, no. 5, pp. 393–403, 1991.

[6] W. Wertz and B. Schneider, "Statistical density estimation: A bibliography," *International Statistical Review/Revue Internationale de Statistique*, pp. 155–175, 1979.

[7] A. B. Tsybakov, *Introduction to Nonparametric Estimation*. Springer, 2009.

[8] Y. G. Dmitriev and F. Tarasenko, "On the estimation of functionals of the probability density and its derivatives," *Theory of Probability & Its Applications*, vol. 18, no. 3, pp. 628–633, 1974.

[9] E. Nadaraya and G. Sokhadze, "On integral functionals of a density," *Communications in Statistics-Theory and Methods*, vol. 45, no. 23, pp. 7086–7102, 2016.

[10] D. L. Donoho, "One-sided inference about functionals of a density," *The Annals of Statistics*, vol. 16, no. 4, pp. 1390–1420, 1988.

[11] P. J. Huber, "Fisher information and spline interpolation," *The Annals of Statistics*, pp. 1029–1033, 1974.

[12] J. C. Spall, "Monte Carlo computation of the Fisher information matrix in nonstandard settings," *Journal of Computational and Graphical Statistics*, vol. 14, no. 4, pp. 889–909, 2005.

[13] V. Berisha and A. O. Hero, "Empirical non-parametric estimation of the Fisher information," *IEEE Signal Processing Letters*, vol. 22, no. 7, pp. 988–992, 2014.

[14] L. Birgé, P. Massart *et al.*, "Estimation of integral functionals of a density," *The Annals of Statistics*, vol. 23, no. 1, pp. 11–29, 1995.

[15] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4135–4159, 2012.

[16] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[17] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *arXiv preprint arXiv:1711.02141*, 2017.

[18] S. Verdú, "Empirical estimation of information measures: A literature guide," *Entropy*, vol. 21, no. 8, p. 720, 2019.

[19] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *The Annals of Probability*, pp. 1269–1283, 1990.

[20] L. D. Brown, "Admissible estimators, recurrent diffusions, and insoluble boundary value problems," *The Annals of Mathematical Statistics*, vol. 42, no. 3, pp. 855–903, 1971.

[21] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, 2005.

[22] A. J. Stam, "Some inequalities satisfied by the quantities of information of Fisher and Shannon," *Information and Control*, vol. 2, no. 2, pp. 101–112, 1959.

[23] W. Alghamdi and F. P. Calmon, "Mutual information as a function of moments," in *Proc. IEEE International Symposium on Information Theory*, 2019, pp. 3122–3126.

[24] R. Esposito, "On a relation between detection and estimation in decision theory," *Inf. Control*, vol. 12, no. 2, pp. 116–120, February 1968.

[25] M. Fozunbal, "On regret of parametric mismatch in minimum mean square error estimation," in *Proc. IEEE International Symposium on Information Theory*, 2010, pp. 1408–1412.