Electronic Journal of Statistics
Vol. 13 (2019) 3613–3653
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1617

Density estimation with contamination: minimax rates and theory of adaptation

Haoyang Liu and Chao Gao*

University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637
e-mail: hyliu@galton.uchicago.edu; chaogao@galton.uchicago.edu

Abstract: This paper studies density estimation under pointwise loss in the setting of contamination model. The goal is to estimate $f(x_0)$ at some $x_0 \in \mathbb{R}$ with i.i.d. contaminated observations:

$$X_1, \ldots, X_n \sim (1-\epsilon)f + \epsilon g$$

where g stands for a contamination distribution. We closely track the effect of contamination by the following model index: contamination proportion ϵ, smoothness of the target density β_0, smoothness of the contamination density β_1, and the local level of contamination m such that $g(x_0) \le m$. The local effect of contamination is shown to depend intricately on the interplay of these parameters. In particular, under a minimax framework, the cost

$$[\epsilon^2(1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}]$$

is shown to be the optimal cost for contamination compared with the usual minimax rate without contamination. The lower bound relies on a novel construction that involves perturbations of a density function at two different resolutions. Such a construction may be of independent interest for the study of local effect of contamination in other nonparametric estimation problems. We also study the setting without any assumption on the contamination distribution, and the minimax cost for contamination is shown to be

$$\epsilon^{\frac{2\beta_0}{\beta_0+1}}.$$

Finally, the minimax cost for adaptation is established both for smooth contamination and arbitrary contamination. Under arbitrary contamination, we show that while adaptation to either contamination proportion or smoothness only costs a logarithmic factor, adaptation to both numbers is impossible.

Keywords and phrases: Minimax rate, nonparametric functional estimation, adaptive estimation, contamination model, robust statistics, Lepski's method.

Received December 2018.

1. Introduction

Nonparametric density estimation is a well-studied classical topic [24, 8, 26]. In this paper, we consider this classical statistical task with a modern twist.

*The research is supported in part by NSF grant DMS-1712957 and NSF Career Award DMS-1847590.

Instead of assuming i.i.d. observations from a true density f, we assume

$$X_1, ..., X_n \sim (1-\epsilon)f + \epsilon g, \tag{1}$$

where g is a density not related to f, and the goal is to estimate $f(x_0)$ at some $x_0 \in \mathbb{R}$. In other words, for each observation, there is an ϵ probability that the observation is sampled from a distribution not related to the density of interest.

This problem naturally appears in both robust statistics and multiple testing literature. In robust statistics literature, g has the name "contamination", and the task is interpreted as robustly estimating a density f with contaminated data points [6]. In multiple testing literature, f and g are respectively called null density and alternative density, and the task is interpreted as estimating null density at a point [11]. In this paper, we use the name "contamination" to refer to both g and the observations generated from it.

The nature of the problem heavily depends on the assumptions put on f and g. When there is no constraint on the contamination distribution g, the data generating process (1) is also recognized as Huber's ϵ-contamination model [14, 15]. Recent work on nonparametric estimation in such a setting includes [6, 12], and the influence of contamination on minimax rates is investigated by [7, 6]. On the other hand, in the literature of multiple testing, it is more common to put parametric structural assumptions on the alternative g, and optimal rates of estimating the null density f are investigated by [16, 3]. We would also like to point out a different line of work that studies estimating f with observations generated from either $(1-\epsilon)f + \epsilon g$ [23] or $(1-\epsilon)f + \epsilon(f * g)$ [13, 28, 19], but the density g is known. In comparison, the robust estimation setting involves some unknown contamination distribution g, so that the minimax rates have very different forms.

In this paper, we explore this problem with connections to nonparametric density estimation literature in mind. Specifically, the density function f is assumed to have a Hölder smoothness β_0. Both cases of structured and arbitrary contamination are considered and fundamental limit of this problem is studied by establishing minimax rate. In the structured contamination case, the contamination distribution g is endowed with a β_1 Hölder smoothness, and the contamination level at the point x_0 is assumed to satisfy $g(x_0) \le m$. The minimax rate of estimating $f(x_0)$ with respect to the squared error loss is shown to be of order

$$[n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2(1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}]. \tag{2}$$

The minimax rate involves three terms, and the influence of contamination on estimation is precisely characterized. The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ corresponds to the classical minimax rate of nonparametric estimation when there is no contamination. The second term $\epsilon^2(1 \wedge m)^2$ is determined by contamination on x_0. It depends on both the contamination proportion ϵ and the contamination level m. The last term $n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$ is caused by contamination on the neighborhood of x_0, which is present even if the contamination level m is zero. In the arbitrary contamination case, or equivalently under Huber's ϵ-contamination model, the

minimax rate is of order

$$[n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^{\frac{2\beta_0}{\beta_0+1}}]. \tag{3}$$

Compared with (2), the rate (3) is easier to understand in terms of the influence of the contamination. It is interesting to note that even though β_0 is the smoothness index of f, it still appears on the second term in (3). Thus, when the contamination is arbitrary, its influence on estimation is also determined by the smoothness of the target density.

We also thoroughly investigate the theory of adaptation in both settings of contamination models. Depending on specific settings, various adaptation costs are necessary. For the contamination model with structured contamination, when the contamination proportion is unknown, an optimal adaptive procedure can achieve the rate (2) with $\epsilon^2(1 \wedge m)^2$ replaced by ϵ^2. When the smoothness is unknown, an optimal adaptive procedure can achieve the rate (2) with n replaced by $n/\log n$. Similarly, for the contamination model with arbitrary contamination, the rate (3) can be achieved up to a logarithmic factor when either ϵ or β_0 is unknown. On the other hand, however, when both the contamination proportion and the smoothness are unknown, the adaptation theories are completely different for the two contamination models. For structured contamination, the adaptation cost is just the combination of the cost of unknown contamination proportion and that of unknown smoothness. In contrast, for arbitrary contamination, we show that adaptation is simply impossible when both ϵ and β_0 are unknown. In other words, it is impossible to adaptively achieve a rate of the form $n^{-r_1(\beta_0)} \vee \epsilon^{r_2(\beta_0)}$ with any two functions $r_1(\cdot)$ and $r_2(\cdot)$.

The theory of adaptation in nonparametric functional estimation without contamination is well studied in the literature. It is shown by [1, 18, 5] that a logarithmic factor must be paid for estimating a point of a density function when smoothness is not known. Adaptation costs of estimating other nonparametric functionals have been investigated in [20, 25, 17, 2, 4]. Compared with the results in the literature, the presence of contamination brings extra complication to the problem of adaptation. It is remarkable that the adaptation cost depends very sensitively on each specific setting and contamination model. The new phenomena revealed in our paper for adaptation with contamination have not been discovered before.

The rest of the paper is organized as follows. The contamination model with structured contamination is studied in Section 2 and Section 3. Results of minimax rates and costs of adaptation are given in Section 2 and Section 3, respectively. The corresponding theory of contamination model with arbitrary contamination is investigated in Section 4. In Section 5, we discuss extensions of our results to multivariate density estimation and a consistent procedure in the hardest scenario where adaptation is impossible. All proofs are given in Section 6.

We close this section by introducing notations that will be used later. For $a, b \in \mathbb{R}$, let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For an integer m, $[m]$ denotes the set $\{1, 2, ..., m\}$. For a positive real number x, $\lceil x \rceil$ is the smallest integer no smaller than x and $\lfloor x \rfloor$ is the largest integer no larger than x. For two positive

sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if $a_n \le Cb_n$ for all n with some consntant $C > 0$ independent of n. The notation $a_n \asymp b_n$ means we have both $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Given a set S, $|S|$ denotes its cardinality, and $\mathbb{1}_S$ is the associated indicator function. We use $\mathbb{P}$ and $\mathbb{E}$ to denote generic probability and expectation whose distribution is determined from the context. The notation $\mathbb{E}(X : S)$ stands for $\mathbb{E}(X\mathbb{1}_S)$. The class of infinitely differentiable functions on $\mathbb{R}$ is denoted by $\mathcal{C}^\infty(\mathbb{R})$. For two probability measures $\mathbb{P}$ and $\mathbb{Q}$, the chi-squared divergence is defined as $\chi^2(\mathbb{P}, \mathbb{Q}) = \int \frac{d\mathbb{P}^2}{d\mathbb{Q}} - 1$, and the total variation distance is defined as $\mathsf{TV}(\mathbb{P}, \mathbb{Q}) = \sup_B |\mathbb{P}(B) - \mathbb{Q}(B)|$. Throughout the paper, C, c and their variants denote generic constants that do not depend on n. Their values may change from place to place.

2. Minimax rates with structured contamination

2.1. Results and implications

Consider i.i.d. observations $X_1, ..., X_n \sim (1-\epsilon)f + \epsilon g$. The goal is to estimate f at a given point. Without loss of generality, we aim to estimate $f(0)$. In other words, for every $i \in [n]$, we have $X_i \sim f$ with probability $1-\epsilon$ and $X_i \sim g$ with probability ϵ. Thus, there are approximately $n\epsilon$ observations that are not related to the density function f, which are referred to as contamination.

To study the fundamental limit of estimating f with contaminated data, we need to specify appropriate regularity conditions on both f and g. We first define the Hölder class by

$$\Sigma(\beta, L) = \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, \left| f^{(\lfloor \beta \rfloor)}(x_1) - f^{(\lfloor \beta \rfloor)}(x_2) \right| \le L|x_1 - x_2|^{\beta - \lfloor \beta \rfloor} \right.$$
$$\left. \text{for any } x_1, x_2 \in \mathbb{R} \right\}.$$

Here, β stands for the smoothness parameter, and L stands for the radius of the function space. The Hölder class of density functions is defined as

$$\mathcal{P}(\beta, L) = \left\{ f : \mathbb{R} \to [0, \infty) \,\middle|\, f \in \Sigma(\beta, L), \int f = 1 \right\}.$$

Finally, we define the class of mixtures in the form of $(1-\epsilon)f + \epsilon g$ by

$$\mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$$
$$= \left\{ (1-\epsilon)f + \epsilon g \,\middle|\, f \in \mathcal{P}(\beta_0, L_0), g \in \mathcal{P}(\beta_1, L_1), g(0) \le m \right\}.$$

This class is indexed by several numbers. Throughout the paper, we refer to ϵ as contamination proportion and m as contamination level at 0. The pair (β_0, L_0) controls the smoothness of the density function f that we want to estimate,

and the pair (β_1, L_1) controls the smoothness of the contamination density g. Among the six numbers, ϵ and m are allowed to depend on the sample size n, but the numbers $\beta_0, \beta_1, L_0, L_1$ are all assumed to be constants that do not depend on n throughout the paper. It is also assumed that $\epsilon \le 1/2$.

The minimax risk of estimation is defined as (notice that we suppress the dependence on n for $\mathcal{R}$)

$$\begin{aligned}&\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)\\ &\quad = \inf_{\widehat{f}(0)} \sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{X_1,\ldots,X_n\sim p}\left(\widehat{f}(0) - f(0)\right)^2,\end{aligned}$$

where the notation $p(\epsilon, f, g)$ is used to denote the density $(1-\epsilon)f + \epsilon g$. Later in the paper, we will shorthand $\mathbb{E}_{X_1,\ldots,X_n\sim p}$ by $\mathbb{E}_{p^n}$. Obviously, the minimax risk becomes smaller if ϵ gets smaller or n gets larger. Besides the role of ϵ and n, the other model indices are also expected to affect the difficulty of the problem, as listed in the following.

- The smoothness of f: From classical density estimation theory, we know the smoother f is, the easier it is to estimate $f(0)$.
- The level of $g(0)$: Intuitively, the smaller $g(0)$ is, the smaller its influence is on $f(0)$, and thus the easier the problem is.
- The smoothness of g: Intuitively, the smoother g is, the less the contamination effect can spread, and thus the easier it is to account for the effect of g in the contamination model.

Now we present the following theorem of minimax rate, that justifies our intuition above.

Theorem 2.1. *Under the setting above, we have*

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2(1\wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}}\epsilon^{\frac{2}{2\beta_1+1}}]. \quad (4)$$

In other words, $\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ can be upper and lower bounded by the right hand side of (4) up to a constant that only depends on $\beta_0, \beta_1, L_0, L_1$.

Theorem 2.1 completely characterizes the difficulty of estimating $f(0)$ with contaminated data. The three terms in the rate (4) have different but very clear meanings. The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ is the classical minimax rate of estimating a smooth function at a given point without contamination. The second term $\epsilon^2(1\wedge m)^2$ is proportional to the squared of the product of contamination level and contamination proportion. The last term $n^{-\frac{2\beta_1}{2\beta_1+1}}\epsilon^{\frac{2}{2\beta_1+1}}$ is perhaps the most interesting. Here the effect of ϵ is powered by an exponent depending on β_1, and it stands for the interaction between the contamination proportion and the contamination smoothness. The fact that it does not depend on m implies that we have to pay this price with contaminated data even if $g(0) = 0$.

To further understand the implications of Theorem 2.1, we present the following illustrative special cases of the minimax rate (4). First, when $\epsilon = 0$, we

get

$$\mathcal{R}(0, \beta_0, \beta_1, L_0, L_1, m) \asymp n^{-\frac{2\beta_0}{2\beta_0+1}}.$$

This is simply the classical minimax rate of estimating $f(0)$ without contamination.

Next, to understand the role of m, we consider two extreme cases of $m = 0$ and $m = \infty$. From (4), we have

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, 0) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}],$$

and

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, \infty) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee \epsilon^2.$$

The case of $m = 0$ is particularly interesting. It implies $g(0) = 0$, and one may expect that the contamination would have no influence on the minimax rate. This intuition is not true because of the term $n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$. Since nonparametric estimation of $f(0)$ also depends on the values of the density function at a neighborhood of 0, the contamination from g can still have an effect on the neighborhood of 0 despite that $g(0) = 0$. A smaller value of β_1 allows a greater perturbation by g on the neighborhood of 0. When $m = \infty$, the minimax rate has a simple form of $[n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee \epsilon^2$. The influence on the minimax rate from contamination is always ϵ^2, regardless of the smoothness β_1.

Finally, we consider the cases of $\beta_1 = 0$ and $\beta_1 = \infty$. In fact, the Hölder class $\Sigma(\beta, L)$ with $\beta_1 = \infty$ is not well defined, but the discussion below still holds for a sufficiently large constant β_1. From (4), we have

$$\mathcal{R}(\epsilon, \beta_0, 0, L_0, L_1, m) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee \epsilon^2,$$

and

$$\mathcal{R}(\epsilon, \beta_0, \infty, L_0, L_1, m) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2 (1 \wedge m)^2].$$

The influence of the contamination takes the forms of ϵ^2 and $\epsilon^2(1 \wedge m)^2$ for the two extreme cases. This immediately implies that for any values of $\epsilon, \beta_0, \beta_1, L_0, L_1, m$, we have

$$[n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2 (1 \wedge m)^2] \lesssim \mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \lesssim [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee \epsilon^2.$$

In other words, the influence of contamination on the minimax rate is sandwiched between $m^2\epsilon^2$ and ϵ^2.

2.2. *Upper bounds*

The minimax rate (4) can be achieved by a simple kernel density estimator that takes the form

$$\widehat{f}_h(0) = \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right). \tag{5}$$

This estimator is slightly different from the classical kernel density estimator because it is normalized by $\frac{1}{n(1-\epsilon)}$ instead of $\frac{1}{n}$. The knowledge of the contamination proportion ϵ is very critical to achieve the minimax rate (4). Later, we will show in Section 3.2 that the minimax rate (4) cannot be achieved if ϵ is not known.

We introduce the following class of kernel functions.

$$\mathcal{K}_l(L) = \left\{ K : \mathbb{R} \to \mathbb{R} \Big| \int K = 1, \int x^j K(x)dx = 0 \text{ for all } j \in [l], \right.$$
$$\left. \|K\|_\infty \vee \int K^2 \vee \int |x|^l |K(x)|dx \le L \right\}.$$

The class $\mathcal{K}_l(L)$ collects all bounded and squared integrable kernel functions of order l. The number $L > 0$ is assumed to be a constant throughout the paper. We refer to [8] for examples of kernel functions in the class $\mathcal{K}_l(L)$.

Theorem 2.2. *For the estimator $\widehat{f}(0) = \widehat{f}_h(0)$ with some $K \in \mathcal{K}_{\lfloor \beta_0 \vee \beta_1 \rfloor}(L)$ and $h = n^{-\frac{1}{2\beta_0+1}} \wedge n^{-\frac{1}{2\beta_1+1}} \epsilon^{-\frac{2}{2\beta_1+1}}$, we have*

$$\sup_{p(\epsilon,f,g) \in \mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n} \left(\widehat{f}(0) - f(0) \right)^2$$
$$\lesssim [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2 (1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}].$$

Theorem 2.2 reveals an interesting choice of the bandwidth $h = n^{-\frac{1}{2\beta_0+1}} \wedge n^{-\frac{1}{2\beta_1+1}} \epsilon^{-\frac{2}{2\beta_1+1}}$. Compared with the optimal bandwidth of order $n^{-\frac{1}{2\beta_0+1}}$ in classical nonparametric function estimation, the h in the structured contamination setting is always smaller. The choice of bandwidth is a consequences of the specific bias-variance tradeoff under the structured contamination model. As an interesting contrast, in the case of arbitrary contamination, the optimal choice of bandwidth is always larger than the usual one, see Section 4.

The error bound in Theorem 2.2 can be found through a classical bias-variance tradeoff argument. We can decompose the difference $\widehat{f}(0) - f(0)$ as

$$(\widehat{f}(0) - \mathbb{E}\widehat{f}(0)) + \left(\mathbb{E}\widehat{f}(0) - f(0) - \frac{\epsilon}{1-\epsilon} g(0) \right) + \frac{\epsilon}{1-\epsilon} g(0). \tag{6}$$

Here, the first term is the stochastic error. The second term gives the approximation error of the kernel convolution. The last term is caused by the contamination at 0. Direct analysis of the three terms gives the bound

$$\mathbb{E} \left(\widehat{f}(0) - f(0) \right)^2 \lesssim \frac{1}{nh} + (h^{2\beta_0} + \epsilon^2 h^{2\beta_1}) + \epsilon^2 (m \wedge 1)^2. \tag{7}$$

Now with the choice $h = n^{-\frac{1}{2\beta_0+1}} \wedge n^{-\frac{1}{2\beta_1+1}} \epsilon^{-\frac{2}{2\beta_1+1}}$, we obtain the error bound in Theorem 2.2. For detailed derivation, see the proof of Theorem 2.2 in Section 6.1.

2.3. Lower bounds

In this section, we study the lower bound part of the minimax rate (4). We first state a theorem.

Theorem 2.3. *We have*

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^2(1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}].$$

The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ is the classical minimax lower bound for nonparametric estimation. Thus, we will only give here a overview of how to derive the second and the third terms. Two specific functions are used as building blocks for our construction, and their definitions and properties are summarized in the following two lemmas.

Lemma 2.1. *Let* $l(x) = e^{-\frac{1}{1-x^2}} \mathbb{1}_{\{|x|\le 1\}}$. *Define*

$$a(x) = \begin{cases} c_0 l(x+1), & -2 \le x \le 0, \\ c_0 l(x-1), & 0 \le x \le 2. \end{cases}$$

The constant c_0 *is chosen so that* $\int a = 1$. *It satisfies the following properties:*

1. *a is an even density function compactly supported on* $[-2, 2]$.
2. $a(0) = 0$.
3. *For any constants* $\beta, L > 0$, *there exists a constant* $c > 0$, *such that* $ca(cx) \in \mathcal{P}(\beta, L) \cap \mathcal{C}^\infty(\mathbb{R})$.
4. *For any small constant* $c > 0$, *a is uniformly lower bounded by a positive constant on* $[-1, -c] \cup [c, 1]$, *and it is uniformly upper bounded by a positive constant on* $\mathbb{R}$.

Lemma 2.2. *Let* $l(x) = e^{-\frac{1}{1-x^2}} \mathbb{1}_{\{|x|\le 1\}}$. *Define*

$$b(x) = \begin{cases} -l\left(4x+3\right), & -1 \le x \le -\frac{1}{2}, \\ l(2x), & |x| \le \frac{1}{2}, \\ -l\left(4x-3\right), & \frac{1}{2} \le x \le 1. \end{cases}$$

It satisfies the following properties:

1. *b is an even function compactly supported on* $[-1, 1]$.
2. *For any* $\beta, L > 0$, *there exists a constant* $c > 0$ *such that* $cb \in \Sigma(\beta, L) \cap \mathcal{C}^\infty(\mathbb{R})$.
3. *b is uniformly lower bounded by a positive constant on* $[-\frac{1}{4}, \frac{1}{4}]$, *and* $|b|$ *is uniformly upper bounded by a positive constant on* $\mathbb{R}$.
4. $\int b = 0$.

Both the proofs of the second and the third terms in the lower bound involve careful constructions of two pairs of densities (f, g) and $(\widetilde{f}, \widetilde{g})$. In order to show $\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim \epsilon^2(1 \wedge m)^2$, we consider the following constructions,

$$
\begin{aligned}
f(x) &= f_0(x),\\
\widetilde{f}(x) &= f_0(x) + c_1 \frac{\epsilon}{1-\epsilon}(m \wedge 1) b(x),\\
g(x) &= c_2 a(c_2 x) + c_1 (m \wedge 1) b(x),\\
\widetilde{g}(x) &= c_2 a(c_2 x).
\end{aligned}
$$

Here, the constants c_1, c_2 are chosen so that the constructed functions $f, \widetilde{f}, g, \widetilde{g}$ are well-defined densities in the desired parameter spaces. It is easy to check that with the above construction,

$$(1-\epsilon) f + \epsilon g = (1-\epsilon)\widetilde{f} + \epsilon \widetilde{g}.$$

This implies that with the presence of contamination, an estimator $\widehat{f}(0)$ cannot distinguish between the two data generating processes $(1-\epsilon)f + \epsilon g$ and $(1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$. As a consequence, an error of order $|f(0) - \widetilde{f}(0)|^2 \asymp \epsilon^2 (1 \wedge m)^2$ cannot be avoided.

The derivation of the lower bound $\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$ is more intricate. Consider the following four functions,

$$
\begin{aligned}
f(x) &= f_0(x),\\
\widetilde{f}(x) &= f_0(x) + \frac{\epsilon}{1-\epsilon} c_2 \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right)\right],\\
g(x) &= c_1 a(c_1 x) + c_2 \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right)\right]\\
&\quad - c_3 \widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right),\\
\widetilde{g}(x) &= c_1 a(c_1 x),
\end{aligned}
$$

where the definitions of the functions l, a, b are given in Lemma 2.1 and Lemma 2.2. Again, the constants c_1, c_2, c_3, c_4 are chosen properly so that the constructed functions are well-defined densities in the desired function classes.

A dominant feature of this constructions is that g is a perturbation of $\widetilde{g}$ with two levels of perturbation, respectively with bandwidth h and $\widetilde{h}$, while usual lower bound proof in nonparametric estimation involves perturbing a function at a single bandwidth level. The first level of perturbation $h^{\beta_0} l\left(\frac{x}{h}\right)$ serves to cancel the effect of the corresponding perturbation on f, while the second perturbation $-\widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right)$ serves to ensure the constraint of contamination level. Indeed, if we relate h and $\widetilde{h}$ through the equation $h^{\beta_0} \asymp \widetilde{h}^{\beta_1}$, then it is direct that $\widetilde{g}(0) = g(0) = 0$. In other words, the constructed contamination density functions g and $\widetilde{g}$ both have contamination level 0. An illustration of this construction with a two-level perturbation is given by Figure 1. The colors of the plot correspond to those in the formulas.

With the above construction, it is not hard to check that

$$p(\epsilon, f, g) - p(\epsilon, \widetilde{f}, \widetilde{g}) = -c_3 \epsilon \widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right).$$

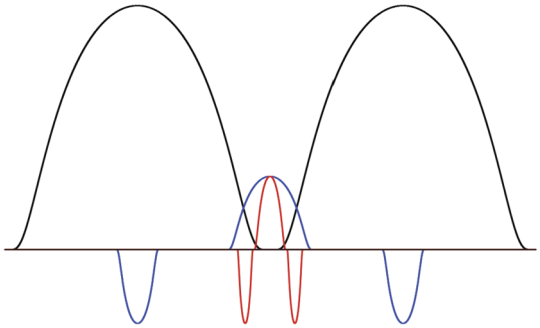

FIG 1. *An illustration of the construction of g.*

In order that an estimator cannot distinguish between the two densities $p(\epsilon, f, g) = (1-\epsilon)f + \epsilon g$ and $p(\epsilon, \widetilde{f}, \widetilde{g}) = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$, a sufficient condition is $\chi^2\left(p(\epsilon, \widetilde{f}, \widetilde{g}), p(\epsilon, f, g)\right) \lesssim n^{-1}$ (see Lemma 6.1), which leads to the choice of $\widetilde{h}$ at the order $\widetilde{h} \asymp \left(n\epsilon^2\right)^{-\frac{1}{2\beta_1+1}}$. As a consequence, an error of order

$$|f(0) - \widetilde{f}(0)|^2 \asymp \epsilon^2 h^{2\beta_0} \asymp \epsilon^2 \widetilde{h}^{2\beta_1} \asymp \epsilon^{\frac{2}{2\beta_1+1}} n^{-\frac{2\beta_1}{2\beta_1+1}}$$

cannot be avoided. A rigorous proof of Theorem 2.3 will be given in Section 6.2.

3. Adaptation theory with structured contamination

3.1. Summary of results

To achieve the minimax rate in Theorem 2.1, the kernel density estimator (5) requires the knowledge of contamination proportion ϵ and smoothness (β_0, β_1). In this section, we discuss adaptive procedures to estimate $f(0)$ without the knowledge of these parameters. However, adaptation to ϵ or to (β_0, β_1) is not free, and one can only achieve slower rates than the minimax rate (4). The adaptation cost varies for each different scenario. A summary of our results is listed below.

- When the contamination proportion is unknown, the best possible rate is

$$n^{-\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^2.$$

- When the smoothness parameters are unknown, the best possible rate is

$$\left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}\right] \vee \left[\epsilon^2(1 \wedge m)^2\right] \vee \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}\right].$$

- When both the contamination proportion and the smoothness are unknown, the best possible rate becomes

$$\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^2.$$

Compared with the minimax rate (4), the ignorance of the contamination proportion implies that m is replaced by 1 in the rate, while the ignorance of the smoothness implies that n is replaced by $n/\log n$ in the rate.

3.2. Unknown contamination proportion

The kernel density estimator (5) depends on ϵ in two ways: the normalization through $\frac{1}{n(1-\epsilon)}$ and the optimal choice of bandwidth h. Without the knowledge of ϵ, we consider the following estimator

$$\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right). \tag{8}$$

The first difference between (8) and (5) is the normalization. When ϵ is not given, we can only use $\frac{1}{n}$ in (8). Moreover, the choice of h in (8) cannot depend on ϵ.

Theorem 3.1. *For the estimator $\widehat{f}(0) = \widehat{f}_h(0)$ with some $K \in \mathcal{K}_{\lfloor \beta_0 \vee \beta_1 \rfloor}(L)$ and $h = n^{-\frac{1}{2\beta_0+1}}$, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim n^{-\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^2.$$

With the choice $h = n^{-\frac{1}{2\beta_0+1}}$, $\widehat{f}_h$ becomes the classical nonparametric density estimator. The contamination results in an extra ϵ^2 in the rate compared with the classical nonparametric minimax rate, regardless of the values of m and β_1. Note that in the current setting, the error $\widehat{f}_h(0) - f(0)$ has the following decomposition,

$$(\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0)) + (\mathbb{E}\widehat{f}_h(0) - (1-\epsilon)f(0) - \epsilon g(0)) + \epsilon(g(0) - f(0)). \tag{9}$$

The difference between (6) and (9) is resulted from different normalizations in (5) and (8). Some standard calculation gives the bound

$$\mathbb{E}(\widehat{f}_h(0) - f(0))^2 \lesssim \frac{1}{nh} \vee h^{2\beta_0} \vee \epsilon^2,$$

which implies the optimal choice of bandwidth $h = n^{-\frac{1}{2\beta_0+1}}$, and thus the rate in Theorem 3.1. A detailed proof is given in Section 6.1.

In view of the form of the minimax rate (4), the rate given by Theorem 3.1 can be obtained by replacing the $\epsilon^2(1 \wedge m)^2$ in (4) with ϵ^2. A matching lower bound for adaptivity to ϵ is given by the following theorem.

Theorem 3.2. *Consider two models* $\mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ *and* $\mathcal{M}(\widetilde{\epsilon}, \beta_0, \beta_1, L_0, L_1, m)$ *with different contamination proportions. For any estimator* $\widehat{f}(0)$ *that satisfies*

$$\sup_{p(\widetilde{\epsilon},f,g)\in\mathcal{M}(\widetilde{\epsilon},\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \le C\widetilde{\epsilon}^2,$$

for some constant $C > 0$*, there exists another constant* $C' > 0$*, such that if* $C'\widetilde{\epsilon} \le \epsilon \le 1/2$*, then we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \gtrsim \epsilon^2.$$

As a consequence, we have

$$\inf_{\widehat{f}(0)} \sup_{\epsilon\in(0,1/2)} \sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 /\epsilon^2 \gtrsim 1.$$

Theorem 3.2 shows that it is impossible to achieve a rate that is faster than ϵ^2 even over only two different contamination proportions. The proof of Theorem 3.2 relies on the following construction,

$$\begin{aligned} f &= f_0, \\ g &= c_1 a(c_1 x), \\ \widetilde{f} &= \frac{1-\epsilon}{1-\widetilde{\epsilon}} f_0 + \frac{\epsilon - \widetilde{\epsilon}}{1-\widetilde{\epsilon}} c_1 a(c_1 x), \\ \widetilde{g} &= c_1 a(c_1 x). \end{aligned}$$

With an appropriate choice of the constant $c_1 > 0$, we have $(1-\epsilon)f + \epsilon g \in \mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ and $(1-\widetilde{\epsilon})\widetilde{f} + \widetilde{\epsilon}\widetilde{g} \in \mathcal{M}(\widetilde{\epsilon}, \beta_0, \beta_1, L_0, L_1, m)$. Moreover, it is easy to check that

$$(1-\epsilon)f + \epsilon g = (1-\widetilde{\epsilon})\widetilde{f} + \widetilde{\epsilon}\widetilde{g}.$$

In other words, a model with contamination proportion ϵ can also be written as a mixture that uses a different $\widetilde{\epsilon}$. Unless the contamination proportion is specified, one cannot tell the difference between $(1-\epsilon)f + \epsilon g$ and $(1-\widetilde{\epsilon})\widetilde{f} + \widetilde{\epsilon}\widetilde{g}$. This leads to a lower bound of the error, which is of order $|f(0) - \widetilde{f}(0)|^2 \asymp \epsilon^2$. A rigorous proof of Theorem 3.2 that uses a constrained risk inequality in [1] is given in Section 6.3.

3.3. Unknown smoothness

In this section, we consider the case that the smoothness numbers are unknown, but the contamination proportion is given. In view of the kernel density estimator (5) that achieves the minimax rate, we can still use the normalization by

$\frac{1}{n(1-\epsilon)}$ because of the knowledge of ϵ, but the bandwidth h needs to be picked in a data-driven way. For a given h, define

$$\widehat{f}_h(0) = \frac{1}{n(1-\epsilon)} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right).$$

With a discrete set $\mathcal{H}$ and some constant $c_1 > 0$, Lepski's method [20, 21, 22] selects a data-driven bandwidth through the following procedure,

$$\widehat{h} = \max\left\{ h \in \mathcal{H} : |\widehat{f}_h(0) - \widehat{f}_l(0)| \le c_1 \sqrt{\frac{\log n}{nl}}, \forall l \le h, l \in \mathcal{H} \right\}. \tag{10}$$

In words, we choose the largest bandwidth below which the variance dominates. If the set that is maximized over is empty, we will use the convention $\widehat{h} = \frac{1}{n}$. The estimator $\widehat{f}_{\widehat{h}}(0)$ that uses a data-driven bandwidth enjoys the following guarantee.

Theorem 3.3. *Consider the adaptive kernel density estimator $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ with the bandwidth defined by (10). In (10), we set $\mathcal{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ such that $\frac{1}{2^m} \le \frac{1}{n} < \frac{1}{2^{m-1}}$ and c_1 to be a sufficiently large constant. The kernel K is selected from $\mathcal{K}_l(L)$ with a large constant $l \ge \lfloor \beta_0 \vee \beta_1 \rfloor$. Then, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2$$
$$\lesssim \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}\right] \vee \left[\epsilon^2(1\wedge m)^2\right] \vee \left[\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}\right].$$

Lepski's method is known to be adaptive over various nonparametric classes, and it can achieve minimax rates up to a logarithmic factor without knowing the smoothness parameter [18]. Theorem 3.3 shows that this is also the case with contaminated observations. With an adaptive kernel density estimator normalized by $\frac{1}{n(1-\epsilon)}$, the minimax rate (4) is achieved up to a logarithmic factor in Theorem 3.3.

A comparison between the adaptive rate given by Theorem 3.3 and the minimax rate (4) reveals two differences. The first adaptation cost is given by $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}$, compared with $n^{-\frac{2\beta_0}{2\beta_0+1}}$ in (4). Previous work in adaptive nonparametric estimation [1, 18, 2] implies that this cost is unavoidable for adaptation to smoothness. The second adaptation cost is given by $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$, compared with $n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$ in (4). In the next theorem, we show that this adaptations cost is also unavoidable without the knowledge of the smoothness parameters.

Theorem 3.4. *Consider two models* $\mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ *and* $\mathcal{M}(\epsilon, \widetilde{\beta}_0, \widetilde{\beta}_1, \widetilde{L}_0, \widetilde{L}_1, m)$ *with different smoothness parameters. Assume that* $\beta_0 \le \tilde{\beta}_0$, $\beta_1 < \tilde{\beta}_1$, $\beta_0 \ge \beta_1$ *and* $n\epsilon^2 \ge (\log n)^2$. *For any estimator* $\widehat{f}(0)$ *that satisfies*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\widetilde{\beta}_0,\widetilde{\beta}_1,\widetilde{L}_0,\widetilde{L}_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \le C\left(\frac{n}{\log n}\right)^{-\frac{2\widetilde{\beta}_1}{2\widetilde{\beta}_1+1}} \epsilon^{\frac{2}{2\widetilde{\beta}_1+1}},$$

for some constant $C > 0$, *we must have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \gtrsim \left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}.$$

Similar to the statement of Theorem 3.2, Theorem 3.4 shows that it is impossible to achieve a rate that is faster than $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$ across two function classes with different smoothness parameters. We remark that the assumptions $\beta_0 \ge \beta_1$ and $n\epsilon^2 \ge (\log n)^2$ in Theorem 3.4 are necessary conditions for $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}$ to dominate $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}$. Without these two conditions, $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}}$ is the larger term between the two, and the lower bound is already in the literature.

In conclusion, the rate in Theorem 3.3 achieved by Lepski's method cannot be improved unless smoothness parameters are given.

3.4. Unknown contamination proportion and unknown smoothness

When both the contamination proportion and the smoothness are unknown, we consider Lepski's method with a kernel density estimator normalized by $\frac{1}{n}$. Define

$$\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i}{h}\right).$$

Then, a data-driven bandwidth $\widehat{h}$ is selected according to (10). Again, if the set that is maximized over is empty in (10), we will use the convention $\widehat{h} = \frac{1}{n}$. Note that this is a fully data-driven estimator that is adaptive to both the contamination proportion and the smoothness. It enjoys the following guarantee.

Theorem 3.5. *Consider the adaptive kernel density estimator* $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ *with the bandwidth defined by (10). In (10), we set* $\mathcal{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ *such that* $\frac{1}{2^m} \le \frac{1}{n} < \frac{1}{2^{m-1}}$ *and* c_1 *to be a sufficiently large constant. The kernel* K *is selected from* $\mathcal{K}_l(L)$ *with a large constant* $l \ge \lfloor \beta_0 \vee \beta_1 \rfloor$. *Then, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \lesssim \left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^2.$$

Compared with the minimax rate in Theorem 2.1, the rate in Theorem 3.5 can be understood as replacing n and $\epsilon^2(1 \wedge m)^2$ respectively by $n/\log n$ and ϵ^2 in (4). In view of the results in both Section 3.2 and Section 3.3, this rate $\left(\frac{n}{\log n}\right)^{-\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^2$ in Theorem 3.5 cannot be improved by any procedure that is adaptive to both contamination proportion and smoothness.

4. Results for arbitrary contamination

4.1. Minimax rates

In this section, we study the contamination model without any structural assumption on the contamination distribution:

$$X_1, \ldots, X_n \sim (1-\epsilon)P_f + \epsilon G$$

where P_f is a distribution on $\mathbb{R}$ that has a density function f, and G is an arbitrary contamination distribution. This leads to the following model space

$$\begin{aligned}&\mathcal{M}(\epsilon, \beta_0, L_0)\\ &\quad = \left\{(1-\epsilon)P_f + \epsilon G \middle| f \in \mathcal{P}(\beta_0, L_0) \text{ and } G \text{ is an arbitrary distribution}\right\}.\end{aligned}$$

This is often referred to as Huber's ϵ-contamination model [14, 15]. Nonparametric function estimation under Huber's ϵ-contamination model has recently been studied by [6, 12] for global loss functions. In this paper, our focus is on the local estimation of $f(0)$. The corresponding minimax risk is defined by

$$\mathcal{R}(\epsilon, \beta_0, L_0) = \inf_{\widehat{f}(0)} \sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2.$$

In contrast to the minimax rate studied in Section 2.1, we only have one parameter ϵ that indexes the influence of the contamination for $\mathcal{R}(\epsilon, \beta_0, L_0)$.

Theorem 4.1. *Under the setting above, we have*

$$\mathcal{R}(\epsilon, \beta_0, L_0) \asymp [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^{\frac{2\beta_0}{\beta_0+1}}]. \tag{11}$$

The minimax rate given by Theorem 4.1 only involves two terms. The first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$ is the classical minimax rate for nonparametric estimation. The second term $\epsilon^{\frac{2\beta_0}{\beta_0+1}}$ characterizes the influence of contamination. It is worth noticing that the smoothness index of f appears both in $n^{-\frac{2\beta_0}{2\beta_0+1}}$ and $\epsilon^{\frac{2\beta_0}{\beta_0+1}}$. A larger value of β_0 implies a less influence of the contamination. This is in contrast to the rate of $\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ in Theorem 2.1.

The phase transition boundary of $\mathcal{R}(\epsilon, \beta_0, L_0)$ occurs at $\epsilon = n^{-\frac{\beta_0+1}{2\beta_0+1}}$. Below this level, we have $\mathcal{R}(\epsilon, \beta_0, L_0) \asymp n^{-\frac{2\beta_0}{2\beta_0+1}}$, and the contamination has no influence on the classical minimax rate. When ϵ is above $n^{-\frac{\beta_0+1}{2\beta_0+1}}$, the rate becomes

$\epsilon^{\frac{2\beta_0}{\beta_0+1}}$, dominated by the contamination of data. Since we have about $n\epsilon$ contaminated observations in expectation, an optimal procedure can achieve the classical minimax rate $n^{-\frac{2\beta_0}{2\beta_0+1}}$ with at most $n\epsilon \leq n^{\frac{\beta_0}{2\beta_0+1}}$ contaminated data points. Note that the number $n^{\frac{\beta_0}{2\beta_0+1}}$ is an increasing function of β_0.

For the upper bound of the minimax rate, we again consider the kernel density estimator $\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{X_i}{h}\right)$. The error $\widehat{f}_h(0) - f(0)$ can be decomposed as $(\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0)) + (\mathbb{E}_h\widehat{f}(0) - f(0))$. Then, a direct analysis shows that the risk can be bounded by three terms,

$$\mathbb{E}\left(\widehat{f}_h(0) - f(0)\right)^2 \lesssim \frac{1}{nh} \vee h^{2\beta_0} \vee \frac{\epsilon^2}{h^2}, \tag{12}$$

which leads to the optimal choice of bandwidth $h = n^{-\frac{1}{2\beta_0+1}} \vee \epsilon^{\frac{1}{\beta_0+1}}$. It is interesting to note that this choice of bandwidth is always larger than or equal to $n^{-\frac{1}{2\beta_0+1}}$. Recall that when the contamination is smooth, the optimal bandwidth in Theorem 2.2 is smaller than $n^{-\frac{1}{2\beta_0+1}}$. Thus, when there is contamination in the data, one may need to use a larger or smaller bandwidth compared with $n^{-\frac{1}{2\beta_0+1}}$ depending on the assumption of contamination.

The lower bound part of Theorem 4.1 can be viewed as an application of Theorem 5.1 in [7]. A general lower bound for Huber's ϵ-contamination model in [7] reveals a critical quantity called modulus of continuity, defined as

$$\omega(\epsilon) = \sup\left\{|f(0) - \widetilde{f}(0)|^2 : \mathsf{TV}(P_f, P_{\widetilde{f}}) \leq \epsilon/(1-\epsilon), f, \widetilde{f} \in \mathcal{P}(\beta_0, L_0)\right\}.$$

The definition of modulus of continuity goes back to [9, 10], and its relation to Huber's ϵ-contamination model is characterized in [7]. In the current setting, it can be shown that $\omega(\epsilon) \asymp \epsilon^{\frac{2\beta_0}{\beta_0+1}}$, which leads to the lower bound part of Theorem 4.1. In Section 6.5, we will give an alternative self-contained proof of the lower bound.

4.2. Adaptation to either contamination proportion or smoothness

The key to adaptation to either contamination proportion or smoothness is the risk decomposition (12) of the kernel density estimator $\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h}K\left(\frac{X_i}{h}\right)$. We write (12) as the sum of two terms. That is,

$$\frac{1}{nh} \vee h^{2\beta_0} \vee \frac{\epsilon^2}{h^2} \asymp \left(\frac{\epsilon^2}{h^2} + \frac{1}{nh}\right) + h^{2\beta_0}. \tag{13}$$

The first term $\frac{\epsilon^2}{h^2} + \frac{1}{nh}$ is a decreasing function of h with a possibly unknown ϵ, while the second term $h^{2\beta_0}$ is an increasing function of h with a possibly unknown β_0. If we know ϵ but do not know β_0, then we can use Lespki's method with $\frac{\epsilon^2}{h^2} + \frac{1}{nh}$ as a reference curve. On the other hand, if we know β_0 but do

not know ϵ, we can then use a reverse version of Lepski's method with $h^{2\beta_0}$ as a reference curve. Specifically, when ϵ is known but β_0 is unknown, we use

$$\widehat{h} = \max\left\{h \in \mathcal{H} : |\widehat{f}_h(0) - \widehat{f}_l(0)| \le c_1\left(\sqrt{\frac{\log n}{nl}} + \frac{\epsilon}{l}\right), \forall l \le h, l \in \mathcal{H}\right\}. \quad (14)$$

If the set that is maximized over is empty, we take $\widehat{h} = \frac{1}{n}$. When β_0 is known but ϵ is unknown, we use

$$\widehat{h} = \min\left\{h \in \mathcal{H} : |\widehat{f}_h(0) - \widehat{f}_l(0)| \le c_1 l^{\beta_0}, \forall l \ge h, l \in \mathcal{H}\right\}. \quad (15)$$

If the set that is minimized over is empty, we take $\widehat{h} = 1$.

Before stating the guarantee for $\widehat{f}_{\widehat{h}}(0)$, we want to emphasize that whether the contamination proportion ϵ is known or not is more than a matter of normalization. As a comparison, recall the risk decomposition for a kernel density estimator with structured contamination in (7). There, both $h^{2\beta_0}$ and $\epsilon^2 h^{2\beta_1}$ are increasing functions of h. This implies that simultaneous adaptation to both ϵ and h is possible through Lepski's method, and whether ϵ is given or not only affects the normalization of the kernel density estimator, which is not the case for arbitrary contamination because of (13).

Theorem 4.2. *Consider the adaptive kernel density estimator $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ with the bandwidth $\widehat{h}$ given by (14) or (15). In either case, we set $\mathcal{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ such that $\frac{1}{2^m} \le \frac{1}{n} < \frac{1}{2^{m-1}}$ and c_1 to be a sufficiently large constant. The kernel K is selected from $\mathcal{K}_l(L)$ with a large constant $l \ge \lfloor\beta_0\rfloor$. Then, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim \left(\frac{\log n}{n}\right)^{\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^{\frac{2\beta_0}{\beta_0+1}}.$$

With one of ϵ and β_0 given, Theorem 4.2 guarantees adaptive estimation with the rate $\left(\frac{\log n}{n}\right)^{\frac{2\beta_0}{2\beta_0+1}} \vee \epsilon^{\frac{2\beta_0}{\beta_0+1}}$. Compared with the minimax rate in Theorem 4.1, we have an extra logarithmic factor due to the ignorance of either ϵ or β_0. This logarithmic factor cannot be removed by any adaptive procedure in view of the results of [1, 18, 2].

4.3. Adaptation to both contamination proportion and smoothness?

When both contamination proportion and smoothness are unknown, the adaptation theory with arbitrary contamination is completely different from the case with structured contamination. Since there is no constraint on the contamination distribution, a model with (ϵ, β_0) can also be written as a different model with $(\widetilde{\epsilon}, \widetilde{\beta}_0)$. As a consequence, we can prove the following lower bound.

Lemma 4.1. *For any constants $c_1, c_2 > 0$, there exists a constant c_0, such that for any $\beta_0, \widetilde{\beta}_0 \le c_1$, and any $L_0, \widetilde{L}_0 \ge c_2$, and any estimator $\widehat{f}(0)$, one of the following lower bounds must be true,*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \ge c_0\epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}},$$

$$\sup_{p(0,f,g)\in\mathcal{M}(0,\widetilde{\beta}_0,\widetilde{L}_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \ge c_0\epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}.$$

Lemma 4.1 says that in order for any estimator to adapt to two classes with different contamination proportions and smoothness indices, say $\mathcal{M}(\epsilon, \beta_0, L_0)$ and $\mathcal{M}(0, \widetilde{\beta}_0, \widetilde{L}_0)$, it is impossible to achieve a rate that is better than $\epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}$ across both classes. The lower bound $\epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}$ is a function of both ϵ, the contamination proportion of the first class $\mathcal{M}(\epsilon, \beta_0, L_0)$, and $\widetilde{\beta}_0$, the smoothness index of the second class $\mathcal{M}(0, \widetilde{\beta}_0, \widetilde{L}_0)$. As we will show in the following, this specific form has a profound implication, in that an adaptive estimation rate that is a function of an individual class is impossible!

As a first step, the following definition formulates what adaptivity means in our specific setting.

Definition 4.1. *An estimator $\widehat{f}(0)$ is called $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ rate adaptive if the following holds: for any $n \ge 1$, any $\epsilon \le 1/2$, any $\beta_0 \le c_1$ and any $L_0 \le c_2$, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \le c_3 n^{-r_1(\beta_0)} \vee \epsilon^{r_2(\beta_0)}. \tag{16}$$

As concrete examples, when the contamination distribution is restricted to those with density functions that are Hölder smooth, it is shown in Theorem 3.5 that adaptive estimation is possible with some $r_1(\beta_0) < \frac{2\beta_0}{2\beta_0+1}$ and $r_2(\beta_0) = 2$. When the contamination distribution is arbitrary, Theorem 4.2 shows that adaptive estimation is possible over (ϵ, β_0) if either ϵ or β_0 is fixed (known) with some $r_1(\beta_0) < \frac{2\beta_0}{2\beta_0+1}$ and $r_2(\beta_0) = \frac{2\beta_0}{\beta_0+1}$. In contrast, the following theorem shows that such a goal is impossible for any $r_1(\cdot)$ and $r_2(\cdot)$ when both ϵ and β_0 are unknown.

Theorem 4.3. *For any constants $c_1, c_2, c_3 > 0$ and any positive functions $r_1(\cdot)$ and $r_2(\cdot)$, there is no estimator $\widehat{f}(0)$ that is $(c_1, c_2, c_3, r_1(\cdot), r_2(\cdot))$ rate adaptive.*

The impossibility result of Theorem 4.3 is a consequence of Lemma 4.1. The lower bound $\epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}$ in Lemma 4.1 involves an ϵ and a $\widetilde{\beta}$ from two different classes. This leads to a contradiction given the definition of adaptivity in (16). A rigorous proof of this argument will given in Section 6.7.

In conclusion, when the contamination is arbitrary, the theory of adaptation to both contamination proportion and smoothness is qualitatively different from adaptation to only one of them. In comparison, when the contamination

is structured, that difference is just quantitative according to the results in Section 3. Therefore, in order to achieve sensible error rates adaptively in a robust density estimation context, we need to either assume a given contamination proportion, a given smoothness index, or a structured contamination distribution.

5. Discussion

5.1. Extensions to multivariate settings

The results in the paper can all be extended to robust multivariate density estimation. We define a d-dimensional isotropic Hölder class as follows,

$$\Sigma_d(\beta, L) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \middle| \max_{l \in I(\beta)} |\nabla_l f(x_1) - \nabla_l f(x_2)| \le L\|x_1 - x_2\|^{\beta - \lfloor \beta \rfloor} \right. \\ \left. \text{for any } x_1, x_2 \in \mathbb{R}^d \right\},$$

where we use $I(\beta)$ to denote the set of multi-indices $\{l = (l_1, ..., l_d) \big| l_1 + \cdots + l_d = \lfloor \beta \rfloor\}$. The class of density functions is defined as

$$\mathcal{P}_d(\beta, L) = \left\{ f : \mathbb{R}^d \to [0, \infty) \middle| f \in \Sigma_d(\beta, L), \int f = 1 \right\}.$$

Note that the dimension d is assumed to be a constant. Then, the two contamination models considered in the paper are extended as

$$\mathcal{M}_d(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \\ = \left\{ (1-\epsilon) f + \epsilon g \middle| f \in \mathcal{P}_d(\beta_0, L_0), g \in \mathcal{P}_d(\beta_1, L_1), g(0) \le m \right\},$$

and

$$\mathcal{M}_d(\epsilon, \beta_0, L_0) \\ = \left\{ (1-\epsilon) P_f + \epsilon G \middle| f \in \mathcal{P}_d(\beta_0, L_0) \text{ and } G \text{ is an arbitrary distribution} \right\}.$$

Similarly, we can define the corresponding minimax rates $\mathcal{R}_d(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ and $\mathcal{R}_d(\epsilon, \beta_0, L_0)$.

Theorem 5.1. *For the two contamination models on $\mathbb{R}^d$, we have*

$$\mathcal{R}_d(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \asymp [n^{-\frac{2\beta_0}{2\beta_0 + d}}] \vee [\epsilon^2 (1 \wedge m)^2] \vee [n^{-\frac{2\beta_1}{2\beta_1 + d}} \epsilon^{\frac{2d}{2\beta_1 + d}}],$$

and

$$\mathcal{R}_d(\epsilon, \beta_0, L_0) \asymp [n^{-\frac{2\beta_0}{2\beta_0 + d}}] \vee [\epsilon^{\frac{2\beta_0}{\beta_0 + d}}].$$

The extra factor of dimension d makes the interpretation of results even more interesting. For example, the phase transition boundary of $\mathcal{R}_d(\epsilon, \beta_0, L_0)$ now occurs at $\epsilon = n^{-\frac{\beta_0+d}{2\beta_0+d}}$. This implies that the influence of contamination becomes more severe as the dimension grows. In contrast, the minimax rate of $\mathcal{R}_d(\epsilon, \beta_0, \beta_1, L_0, L_1, m)$ leads to a completely different interpretation. For example, when $m \geq 1$, we have

$$\mathcal{R}_d(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \asymp n^{-\frac{2\beta_0}{2\beta_0+d}} \vee \epsilon^2.$$

The second term ϵ^2 does not change with the dimension d, and the phase transition boundary between $n^{-\frac{2\beta_0}{2\beta_0+d}}$ and ϵ^2 is at $\epsilon = n^{-\frac{\beta_0}{2\beta_0+d}}$, which increases with respect to d. This suggests that the influence of contamination becomes less severe as d grows. In short, the contamination influence on density estimation can be drastically different in a multivariate setting, depending on whether the contamination distribution is structured or arbitrary.

5.2. Consistency in the hardest scenario

When there is no constraint on the contamination distribution, adaptation is impossible over both contamination proportion and smoothness in the sense of (16). One may wonder whether there is still anything to do in such a scenario with almost nothing is assumed. In this section, we show that consistency is still possible under this hardest scenario.

Before introducing the procedure, we remark that achieving consistency without knowing ϵ and β_0 is a non-trivial problem due to the risk decomposition (12) for a kernel density estimator. According to (12), a choice of bandwidth that leads to consistency must satisfy $nh \to \infty$, $h \to 0$ and $h/\epsilon \to \infty$. Note that the first and the second requirements can be satisfied easily with a choice of h that does not depend on any model parameter. For example, one can choose $h = n^{-1/2}$. However, the third requirement $h/\epsilon \to \infty$ is problematic without the knowledge of ϵ. For any choice of $h \to 0$, there is an adversarial ϵ to make $h/\epsilon \to \infty$ fail.

Despite the above difficulty, we show that a data-driven bandwidth leads to consistency if we know that the smoothness β_0 has a lower bound $\widetilde{\beta}_0$. We consider a kernel density estimator $\widehat{f}_h(0) = \frac{1}{n}\sum_{i=1}^n \frac{1}{h} K\left(\frac{X_i}{h}\right)$. Then, we choose h by the reverse version of Lepskis' method that is similar to (15). We define $\widehat{h}$ by

$$\widehat{h} = \min\left\{h \in \mathcal{H} : |\widehat{f}_h(0) - \widehat{f}_l(0)| \leq c_1 l^{\widetilde{\beta}_0}, \forall l \geq h, l \in \mathcal{H}\right\}. \tag{17}$$

Again, we use the convention that if the set that is minimized over is empty, we take $\widehat{h} = 1$.

Theorem 5.2. *Consider the kernel density estimator $\widehat{f}(0) = \widehat{f}_{\widehat{h}}(0)$ with the bandwidth $\widehat{h}$ given by (17). We set $\mathcal{H} = \left\{1, \frac{1}{2}, \cdots, \frac{1}{2^m}\right\}$ such that $\frac{1}{2^m} \leq \frac{1}{n} <$*

$\frac{1}{2^{m-1}}$ *and* c_1 *to be a sufficiently large constant. The kernel* K *is selected from* $\mathcal{K}_l(L)$ *with a large constant* $l \geq \lfloor \beta_0 \rfloor$. *Then, as* $n \to \infty$ *and* $\epsilon \to 0$. *we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \to 0,$$

if $\beta_0 \geq \widetilde{\beta}_0$.

Note that the requirements $n \to \infty$ and $\epsilon \to 0$ are necessary conditions of consistency given the minimax rate (11). The procedure does not require knowledge of ϵ or β_0, and thus consistency can be achieved without knowing ϵ and β_0 even if adaptation is impossible. The procedure (17) uses a conservative $\widetilde{\beta}_0$ in the reverse version of Lepski's method, and can be viewed as an extension of (15) that uses the true smoothness index β_0.

6. Proofs

6.1. Proofs of Theorem 2.2 and Theorem 3.1

Proof of Theorem 2.2. Decompose the error as

$$\widehat{f}(0) - f(0) = (\widehat{f}(0) - \mathbb{E}\widehat{f}(0)) + \left(\mathbb{E}\widehat{f}(0) - f(0) - \frac{\epsilon}{1-\epsilon}g(0)\right) + \frac{\epsilon}{1-\epsilon}g(0),$$

where the first term is the stochastic error, the second term stands for bias, and the third term is the misspecification error caused by contamination.

For the variance term, we have

$$\mathbb{E}(\widehat{f}(0) - \mathbb{E}\widehat{f}(0))^2 = \mathsf{Var}\left(\frac{\sum_{i=1}^n \frac{1}{h}K\left(\frac{X_i}{h}\right)}{n(1-\epsilon)}\right) = \frac{\mathsf{Var}(\frac{1}{h}K(\frac{X}{h}))}{n(1-\epsilon)^2},$$

where

$$\begin{aligned}\mathsf{Var}\left(\frac{1}{h}K\left(\frac{X}{h}\right)\right) &\leq \int \frac{1}{h^2}K^2\left(\frac{x}{h}\right)((1-\epsilon)f(x) + \epsilon g(x))dx \\ &\lesssim \frac{1}{h}\int \frac{1}{h}K^2\left(\frac{x}{h}\right)dx \lesssim \frac{1}{h}.\end{aligned}$$

This gives the variance bound

$$\mathbb{E}(\widehat{f}(0) - \mathbb{E}\widehat{f}(0))^2 \lesssim \frac{1}{nh}. \tag{18}$$

For the bias term we have

$$\mathbb{E}\widehat{f}(0) = \int \frac{1}{h}K\left(\frac{x}{h}\right)f(x)dx + \frac{\epsilon}{1-\epsilon}\int \frac{1}{h}K\left(\frac{x}{h}\right)g(x)dx.$$

Since $f \in \mathcal{P}(\beta_0, L_0)$ and $g \in \mathcal{P}(\beta_1, L_1)$, we have $|\int \frac{1}{h} K\left(\frac{x}{h}\right)(f(x) - f(0))dx| \lesssim h^{\beta_0}$ and $|\int \frac{1}{h} K\left(\frac{x}{h}\right)(g(x) - g(0))dx| \lesssim h^{\beta_1}$. See [26, Chapter 1.2] for an explicit bias calculation. Adding up the two bias bounds, we get

$$\left|\mathbb{E}\widehat{f}(0) - f(0) - \frac{\epsilon}{1-\epsilon} g(0)\right| \lesssim h^{\beta_0} + \epsilon h^{\beta_1}. \tag{19}$$

For the last term, it is easy to see that

$$\left(\frac{\epsilon}{1-\epsilon} g(0)\right)^2 \lesssim \epsilon^2 (m \wedge 1)^2, \tag{20}$$

since $g(0) \leq m$ by the assumption and $g(0) \lesssim 1$ by the fact that $g \in \mathcal{P}(\beta_1, L_1)$.

With the relation $\mathbb{E}(A_1 + A_2 + A_3)^2 \lesssim \mathbb{E}A_1^2 + \mathbb{E}A_2^2 + \mathbb{E}A_3^2$ and the three bounds in (18), (19) and (20), we conclude the proof by the specific choice of $h = n^{-\frac{1}{2\beta_0+1}} \wedge n^{-\frac{1}{2\beta_1+1}} \epsilon^{-\frac{2}{2\beta_1+1}}$. □

Proof of Theorem 3.1. The error decomposes as

$$\widehat{f}(0) - f(0) = (\widehat{f}(0) - \mathbb{E}\widehat{f}(0)) + (\mathbb{E}\widehat{f}(0) - (1-\epsilon)f(0) - \epsilon g(0)) + \epsilon(g(0) - f(0)).$$

Using the same argument that leads to (18), we have $\mathbb{E}(\widehat{f}(0) - \mathbb{E}\widehat{f}(0))^2 \lesssim \frac{1}{nh}$ for the variance term. The bias term $(\mathbb{E}\widehat{f}(0) - (1-\epsilon)f(0) - \epsilon g(0))$ can be further decomposed as

$$(1-\epsilon)\int \frac{1}{h} K\left(\frac{x}{h}\right)(f(x) - f(0))dx + \epsilon \int \frac{1}{h} K\left(\frac{x}{h}\right)(g(x) - g(0))dx.$$

Therefore, the same argument that leads to (19) also gives the bound

$$|\mathbb{E}\widehat{f}(0) - (1-\epsilon)f(0) - \epsilon g(0)| \lesssim h^{\beta_0} + \epsilon h^{\beta_1}.$$

For the last term, we have $\epsilon|g(0) - f(0)| \lesssim \epsilon$. Combining the three bounds above, we have

$$\mathbb{E}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim \frac{1}{nh} + h^{2\beta_0} + \epsilon^2.$$

Choose $h = n^{-\frac{1}{2\beta_0+1}}$, and the proof is complete. □

6.2. Proof of Theorem 2.3

The proof of Theorem 2.3 mainly relies on Le Cam's two-point argument. The method is summarized by the following lemma.

Lemma 6.1. *Consider two distributions P_{θ_0} and P_{θ_1} whose parameters of interest are separated by $\Delta = |T_{\theta_0} - T_{\theta_1}|$. Assume $\chi^2(P_{\theta_0}, P_{\theta_1}) \leq \alpha$. Then, we have*

$$\inf_{\widehat{T}} \sup_{\theta \in \{\theta_0, \theta_1\}} \mathbb{E}_\theta \left(\widehat{T} - T_\theta\right)^2 \geq \frac{1}{8} e^{-\alpha} \Delta^2.$$

We refer the readers to [27] and [26, Chapter 2.3] for rigorous proofs. In the setting of Theorem 2.3, we need to find two pairs of density functions (f, g) and $(\widetilde{f}, \widetilde{g})$ that satisfy $f, \widetilde{f} \in \mathcal{P}(\beta_0, L_0)$, $g, \widetilde{g} \in \mathcal{P}(\beta_1, L_1)$ and $g(0) \vee \widetilde{g}(0) \leq m$. Since we are working with i.i.d. observations, it is sufficient to show that

$$\chi^2\left(p(\epsilon, \widetilde{f}, \widetilde{g}), p(\epsilon, f, g)\right) \lesssim n^{-1}.$$

Then, Lemma 6.1 implies $\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim |f(0) - \widetilde{f}(0)|^2$.

The lower bound of Theorem 2.3 contains three terms. We thus split the proof into three parts, and then combine the three arguments in the end.

Lemma 6.2. *We have*

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim n^{-\frac{2\beta_0}{2\beta_0+1}}.$$

Proof. The proof uses a similar argument in [26, Chapter 2.5]. Since we are dealing with a setting with contamination, we still give a proof to be self contained. We define the following four functions,

$$\begin{aligned} g(x) &= \widetilde{g}(x) = c_1 a(c_1 x), \\ f(x) &= f_0(x), \\ \widetilde{f}(x) &= f_0(x) + c_2 h^{\beta_0} b\left(\frac{x}{h}\right). \end{aligned}$$

Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The functions $a(x)$ and $b(x)$ are given by Lemma 2.1 and Lemma 2.2. We first verify that for appropriate choices of c_1, c_2 and $h \leq 1$, the constructed functions are well-defined densities in the desired parameter spaces.

- We have $f \in \mathcal{P}(\beta_0, L_0)$ by construction. Since $h \leq 1$, $b(x/h)$ is compactly supported on an area where f_0 is lower bounded by some positive constant. Thus, with a $c_2 > 0$ that is sufficiently small, $\widetilde{f}$ is nonnegative. The fact $\int \widetilde{f} = 1$ can be derived from the property of b in Lemma 2.2. Hence, $\widetilde{f} \in \mathcal{P}(\beta_0, L_0)$ when c_2 is small enough.
- With a sufficiently small $c_1 > 0$, we have $g, \widetilde{g} \in \mathcal{P}(\beta_1, L_1)$.
- By $a(0) = 0$ according to Lemma 2.1, we get $|g(0)| \vee |\widetilde{g}(0)| \leq m$.

We use the notation $p = (1-\epsilon)f + \epsilon g$ and $q = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$. Note that p can be lower bounded by a positive constant on the interval $[-1, 1]$ according to its definition. Moreover, we have

$$p(x) - q(x) = -(1-\epsilon)c_2 h^{\beta_0} b\left(\frac{x}{h}\right),$$

and the support of $b\left(\frac{x}{h}\right)$ is $[-h, h] \subset [-1, 1]$. This leads to the bound

$$\chi^2(q, p) = \int_{-1}^{1} \frac{(p-q)^2}{p} \lesssim \int (p-q)^2 \asymp h^{2\beta_0} \int b^2\left(\frac{x}{h}\right) \asymp h^{2\beta_0+1}.$$

In order that $n\chi^2(q,p) \lesssim 1$, we can choose $h = n^{-\frac{1}{2\beta_0+1}}$. This leads to

$$|f(0) - \widetilde{f}(0)| \asymp n^{-\frac{\beta_0}{2\beta_0+1}}.$$

Use Lemma 6.1, and the proof is complete. □

Lemma 6.3. *We have*

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim \epsilon^2(1 \wedge m)^2.$$

Proof. By [26], for any $p \in \mathcal{P}(\beta, L)$, there exists a constant p_{max} such that $\sup_x |p(x)| \leq p_{max}$. Therefore, it is sufficient to consider m that is bounded by some constant, say $m \leq 1$. Consider the following four functions,

$$\begin{aligned} f(x) &= f_0(x), \\ \widetilde{f}(x) &= f_0(x) + c_1 \frac{\epsilon}{1-\epsilon} m b(x), \\ g(x) &= c_2 a(c_2 x) + c_1 m b(x), \\ \widetilde{g}(x) &= c_2 a(c_2 x). \end{aligned}$$

Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The functions $a(x)$ and $b(x)$ are given by Lemma 2.1 and Lemma 2.2. With appropriate choices of the constants $c_1, c_2 > 0$, $f, \widetilde{f}, g, \widetilde{g}$ are well-defined density functions that belong to the desired function classes.

- By Lemma 2.1, We have $f_0 \in \mathcal{P}(\beta_0, L_0/2) \subset \mathcal{P}(\beta_0, L_0)$ by construction. Since f_0 is strictly positive on $[-1,1]$ and b is compactly supported on $[-1,1]$, we have $\widetilde{f} \in \mathcal{P}(\beta_0, L_0)$ for some sufficiently small constant $c_1 > 0$ according to the properties of b listed in Lemma 2.2.
- By definition of a, we have $\widetilde{g} \in \mathcal{P}(\beta_1, L_1/2)$ for some sufficiently small $c_2 > 0$ according to Lemma 2.1. Since $b(x)$ only takes negative values when $c_2 a(c_2 x)$ is lower bounded by a positive constant, g is nonnegative and $g \in \mathcal{P}(\beta_1, L_1)$ when c_1 is small enough.
- We also have $|g(0)| \vee |\widetilde{g}(0)| \leq m$ for a sufficiently small c_1 because $a(0) = 0$ and $|b(0)|$ is bounded by a constant according to Lemma 2.1 and Lemma 2.2.

In summary, we have

$$(1-\epsilon)f + \epsilon g, (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g} \in \mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m).$$

Moreover, according to our construction, we have

$$(1-\epsilon)f + \epsilon g = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g},$$

and

$$|f(0) - \widetilde{f}(0)| = c_1 \frac{\epsilon}{1-\epsilon} m|b(0)| \gtrsim m\epsilon,$$

where we have used $|b(0)| \gtrsim 1$ by Lemma 2.2. Finally, using Lemma 6.1, we obtain the desired lower bound result. □

Lemma 6.4. *Assume $\beta_1 \le \beta_0$ and $n\epsilon^2 \ge 1$. Then, we have*

$$\mathcal{R}(\epsilon, \beta_0, \beta_1, L_0, L_1, m) \gtrsim n^{-\frac{2\beta_1}{2\beta_1+1}} \epsilon^{\frac{2}{2\beta_1+1}}.$$

Proof. Consider the following four functions,

$$\begin{aligned}
f(x) &= f_0(x), \\
\widetilde{f}(x) &= f_0(x) + c_2 \frac{\epsilon}{1-\epsilon} \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right) \right], \\
g(x) &= c_1 a(c_1 x) + c_2 \left[h^{\beta_0} l\left(\frac{x}{h}\right) - h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right) \right] \\
&\quad - c_3 \widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right), \\
\widetilde{g}(x) &= c_1 a(c_1 x).
\end{aligned}$$

Since the proof relies on perturbing a density at a point where it is 0, the verification of nonnegativity is more delicate, which motivates another tuning constant controlling the center of the negative part of the perturbation. Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The functions $a(x)$ and $b(x)$ are given by Lemma 2.1 and Lemma 2.2. The numbers h and $\widetilde{h}$ are chosen so that the following equation is satisfied:

$$c_2 h^{\beta_0} l(0) = c_3 \widetilde{h}^{\beta_1} b(0). \tag{21}$$

Now, we verify that with appropriate choices of constants c_1, c_2, c_3, c_4, the constructed functions belong to the parameter spaces.

- The functions f and $\widetilde{g}$ are automatically density functions by definition. Note that we can choose a small constant c_4 so that the negative perturbation $-h^{\beta_0} l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0} l\left(\frac{2(x+c_4)}{h}\right)$ has a support in a region where both f_0 and $c_1 a(c_1 x)$ are bounded below by a positive constant. This immediately implies that $\widetilde{f}(x) \ge 0$ for all x with a sufficiently small constant c_2. Similarly, the support of $-c_3 \widetilde{h}^{\beta_1} b\left(\frac{x}{\widetilde{h}}\right)$ is $[-\widetilde{h}, \widetilde{h}]$, which is contained in a region where $c_1 a(c_1 x)$ is bounded below by a positive constant for a sufficiently small $\widetilde{h}$. Therefore, $g(x) \ge 0$ for all x with a sufficiently small constant c_3. We also note that $\int \widetilde{f} = \int g = 1$ according to the definitions.
- When c_1, c_2, c_3 are chosen small enough, we have $f, \widetilde{f} \in \Sigma(\beta_0, L_0)$ and $g, \widetilde{g} \in \Sigma(\beta_1, L_1)$. Here $g \in \Sigma(\beta_1, L_1)$ is a consequence of the assumption that $\beta_1 \le \beta_0$.
- Finally, we have $l(2c_4/h) = l(-2c_4/h) = 0$ for a sufficiently small h. This implies $g(0) = \widetilde{g}(0) = 0$ because of (21). Therefore, $|g(0) \vee \widetilde{g}(0)| \le m$.

In summary, we have

$$(1-\epsilon) f + \epsilon g, (1-\epsilon)\widetilde{f} + \epsilon \widetilde{g} \in \mathcal{M}(\epsilon, \beta_0, \beta_1, L_0, L_1, m).$$

Besides the properties listed above, we also note that both f and g can be bounded from below by some positive constant on the interval $[-1,1]$, if the constants c_2, c_3 are sufficiently small. This implies that the density $(1-\epsilon)f+\epsilon g$ is lower bounded by some positive constant on the interval $[-1,1]$.

Now, according to the above construction, for $p=(1-\epsilon)f+\epsilon g$ and $q=(1-\epsilon)\widetilde{f}+\epsilon\widetilde{g}$, we have

$$p(x)-q(x)=-\epsilon c_3\widetilde{h}^{\beta_1}b\left(\frac{x}{\widetilde{h}}\right).$$

Given that the support of $b\left(\frac{x}{\widetilde{h}}\right)$ is within $[-\widetilde{h},\widetilde{h}]\subset[-1,1]$ with a sufficiently small $\widetilde{h}$, we have

$$\chi^2(q,p)=\int_{-1}^{1}\frac{(p-q)^2}{p}\lesssim\int(p-q)^2\asymp\epsilon^2\widetilde{h}^{2\beta_1}\int b^2\left(\frac{x}{\widetilde{h}}\right)\asymp\epsilon^2\widetilde{h}^{2\beta_1+1}.$$

In order that $n\chi^2(q,p)\lesssim 1$, it is sufficient to choose $\widetilde{h}\asymp\left(n\epsilon^2\right)^{-\frac{1}{2\beta_1+1}}$. The condition $n\epsilon^2\geq 1$ implies that $\widetilde{h}$ can be picked sufficiently small. Moreover, with the relation (21), we have

$$|f(0)-\widetilde{f}(0)|=c_2\frac{\epsilon}{1-\epsilon}h^{\beta_0}l(0)\asymp\epsilon h^{\beta_0}\asymp\epsilon\widetilde{h}^{\beta_1}\asymp\epsilon^{\frac{1}{2\beta_1+1}}n^{-\frac{\beta_1}{2\beta_1+1}}.$$

Finally, using Lemma 6.1, we obtain the desired lower bound result. □

We combine the results of Lemma 6.2, Lemma 6.3 and Lemma 6.4.

Proof of Theorem 2.3. In order that the third term $n^{-\frac{2\beta_1}{2\beta_1+1}}\epsilon^{\frac{2}{2\beta_1+1}}$ dominates the other two, it is necessary that $\epsilon^2\geq n^{\frac{2\beta_1-2\beta_0}{2\beta_0+1}}$. This implies both $\beta_1\leq\beta_0$ and $n\epsilon^2\geq 1$. By Lemma 6.4, we have

$$\mathcal{R}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)\gtrsim n^{-\frac{2\beta_1}{2\beta_1+1}}\epsilon^{\frac{2}{2\beta_1+1}}.$$

When the first or the second term dominate, we use Lemma 6.2 and Lemma 6.3, and obtain

$$\mathcal{R}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)\gtrsim[n^{-\frac{2\beta_0}{2\beta_0+1}}]\vee[\epsilon^2(1\wedge m)^2].$$

Hence, the proof is complete. □

6.3. *Proofs of Theorem 3.2 and Theorem 3.4*

The proofs of both theorems rely on the following constrained risk inequality by [1].

Lemma 6.5. *Consider two distributions P_{θ_0} and P_{θ_1} whose parameters of interest are separated by $\Delta = |T_{\theta_0} - T_{\theta_1}|$. For any estimator $\widehat{T}$, assume*

$$\mathbb{E}_{\theta_0}(\widehat{T} - T_{\theta_0})^2 \le \delta^2.$$

Then, whenver $\delta I \le \Delta$, we have

$$\mathbb{E}_{\theta_1}(\widehat{T} - T_{\theta_1})^2 \ge (\Delta - \delta I)^2,$$

where $I = \sqrt{\int \frac{dP_{\theta_1}^2}{dP_{\theta_0}}}$.

Proof of Theorem 3.2. We consider the following four functions,

$$\begin{aligned} f &= f_0, \\ g &= c_1 a(c_1 x), \\ \widetilde{f} &= \frac{1-\epsilon}{1-\widetilde{\epsilon}} f_0 + \frac{\epsilon - \widetilde{\epsilon}}{1-\widetilde{\epsilon}} c_1 a(c_1 x), \\ \widetilde{g} &= c_1 a(c_1 x). \end{aligned}$$

Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The function $a(\cdot)$ is given by Lemma 2.1. The constant c_1 is sufficiently small so that $c_1 a(c_1 x)$ belongs to both $\mathcal{P}(\beta_0, L_0/2)$ and $\mathcal{P}(\beta_1, L_1/2)$. Now it is easy to check that $f, \widetilde{f} \in \mathcal{P}(\beta_0, L_0)$, $g, \widetilde{g} \in \mathcal{P}(\beta_1, L_1)$ and $g(0) \vee \widetilde{g}(0) = 0 \le m$, so that the constructed functions are well-defined densities in the parameter spaces.

It is easy to check that

$$(1-\epsilon)f + \epsilon g = (1-\widetilde{\epsilon})\widetilde{f} + \widetilde{\epsilon}\widetilde{g}.$$

This implies $\int q^2/p = 1$ for $p = (1-\epsilon)f + \epsilon g$ and $q = (1-\widetilde{\epsilon})\widetilde{f} + \widetilde{\epsilon}\widetilde{g}$. We also have

$$\left| f(0) - \widetilde{f}(0) \right| = \frac{\epsilon - \widetilde{\epsilon}}{1-\widetilde{\epsilon}} f(0).$$

According to Lemma 6.5, suppose there is an estimator $\widehat{f}(0)$ that satisfies $\mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \le C\widetilde{\epsilon}^2$, we must have

$$\mathbb{E}_{q^n}(\widehat{f}(0) - \widetilde{f}(0))^2 \ge \left(\frac{\epsilon - \widetilde{\epsilon}}{1-\widetilde{\epsilon}} f(0) - C^{1/2}\widetilde{\epsilon} \right)^2.$$

Therefore, there exists a constant $C' > 0$, such that for $\epsilon \ge C'\widetilde{\epsilon}$, $\mathbb{E}_{q^n}(\widehat{f}(0) - \widetilde{f}(0))^2 \gtrsim \epsilon^2$.

Fix a constant $C_1 > 0$ that is small enough. Then according to the above reasoning, there exist two constants $C_2 > 0$, $C_3 > 0$ depending on C_1 such

that if two contamination proportions ϵ and $\widetilde{\epsilon}$ satisfy $C_2\widetilde{\epsilon} \le \epsilon < 1/2$, then the following two statements cannot be true at the same time:

$$\sup_{p(\widetilde{\epsilon},f,g)\in\mathcal{M}(\widetilde{\epsilon},\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \le C_1\widetilde{\epsilon}^2,$$
$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \le C_3\epsilon^2.$$

This implies that

$$\sup_{0<\epsilon<1/2} \epsilon^{-2} \sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,\beta_1,L_0,L_1,m)} \mathbb{E}_{p^n}\left(\widehat{f}(0)-f(0)\right)^2 \gtrsim 1,$$

for any estimator $\widehat{f}$, which concludes the proof. □

Proof of Theorem 3.4. We construct the following four functions

$$\begin{aligned}
\widetilde{f}(x) &= f_0(x),\\
f(x) &= f_0(x) - c_2\frac{\epsilon}{1-\epsilon}\left[h^{\beta_0}l\left(\frac{x}{h}\right) - h^{\beta_0}l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0}l\left(\frac{2(x+c_4)}{h}\right)\right],\\
\widetilde{g}(x) &= c_1a(c_1x),\\
g(x) &= c_1a(c_1x) + c_2\left[h^{\beta_0}l\left(\frac{x}{h}\right) - h^{\beta_0}l\left(\frac{2(x-c_4)}{h}\right) - h^{\beta_0}l\left(\frac{2(x+c_4)}{h}\right)\right]\\
&\quad - c_3\widetilde{h}^{\beta_1}b\left(\frac{x}{\widetilde{h}}\right).
\end{aligned}$$

The construction is similar to that in the proof of Lemma 6.4. The difference is that the perturbation is now put on both f and g. Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The functions $a(x)$ and $b(x)$ are given by Lemma 2.1 and Lemma 2.2. The numbers h and $\widetilde{h}$ are chosen so that the following equation is satisfied:

$$c_2h^{\beta_0}l(0) = c_3\widetilde{h}^{\beta_1}b(0). \tag{22}$$

Similar to the argument used in Lemma 6.4, it is not hard to check that with appropriate choices of the constants c_1, c_2, c_3, we have $\widetilde{f} \in \mathcal{P}(\widetilde{\beta}_0, \widetilde{L}_0)$, $\widetilde{g} \in \mathcal{P}(\widetilde{\beta}_1, \widetilde{L}_1)$, $f \in \mathcal{P}(\beta_0, L_0)$ and $g \in \mathcal{P}(\beta_1, L_1)$, given that $\widetilde{\beta}_0 \ge \beta_0 \ge \beta_1$ and $\widetilde{\beta}_1 > \beta_1$. The numbers h and $\widetilde{h}$ are both required to be sufficiently small. We also have $g(0) = \widetilde{g}(0) = 0$ according to the definition with an appropriate choice of c_4. Then, the constructed functions are well-defined densities in the parameter spaces.

With the notation $p = (1-\epsilon)f + \epsilon g$ and $q = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$, we check the quantities in Lemma 6.5. Note that

$$|p(x) - q(x)| = c_3\epsilon\widetilde{h}^{\beta_1}b\left(\frac{x}{\widetilde{h}}\right).$$

With a similar argument in the proof of Lemma 6.4, the function $b\left(\frac{x}{\widetilde{h}}\right)$ is supported within $[-\widetilde{h},\widetilde{h}] \subset [-1,1]$, and $p(x)$ is lower bounded by some constant uniformly over $x \in [-1,1]$. This implies,

$$I = \left(\int \frac{q^2}{p}\right)^{\frac{n}{2}} = \left(1 + \int \frac{(q-p)^2}{p}\right)^{\frac{n}{2}} \leq \exp\left(\frac{C_1 n}{2}\int (p-q)^2\right)$$
$$\leq \exp\left(C_1' n \epsilon^2 \widetilde{h}^{2\beta_1+1}\right).$$

Moreover, we also have

$$\Delta = |f(0) - \widetilde{f}(0)| = c_2 \frac{\epsilon}{1-\epsilon} h^{\beta_0} l(0),$$

and

$$\delta = C^{1/2}\left(\frac{n}{\log n}\right)^{-\frac{\widetilde{\beta}_1}{2\widetilde{\beta}_1+1}} \epsilon^{\frac{1}{2\widetilde{\beta}_1+1}}.$$

In order that $I \leq \left(\frac{n\epsilon^2}{\log n}\right)^c$ for some sufficiently small constant $c > 0$, we can choose $\widetilde{h} \asymp \left(\frac{n\epsilon^2}{\log n}\right)^{-\frac{1}{2\beta_1+1}}$, which is always possible with the condition $n\epsilon^2 \geq (\log n)^2$. According to the relation (22), we have $\Delta \asymp \epsilon^{\frac{1}{2\beta_1+1}}\left(\frac{n}{\log n}\right)^{-\frac{\beta_1}{2\beta_1+1}}$. Plugging these quantities into the constrained risk inequality in Lemma 6.5 and using $\beta_1 < \widetilde{\beta}_1$, we get the desired lower bound. □

6.4. Proofs of Theorem 3.3 and Theorem 3.5

The proofs of the two theorems are similar. Thus, we give a detailed proof of Theorem 3.5 first, and then sketch the proof of Theorem 3.3.

Proof of Theorem 3.5. For every bandwidth h, the error decomposes as

$$\widehat{f}_h(0) - f(0) = (\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0)) + (\mathbb{E}\widehat{f}_h(0) - (1-\epsilon)f(0) - \epsilon g(0)) + \epsilon(g(0) - f(0)), \tag{23}$$

where the three terms correspond to a stochastic part that depends on h, a deterministic part that depends on h, and a deterministic part that does not depend on h. With the same argument in the proof of Theorem 3.1, we have

$$\mathbb{E}(\widehat{f}(0) - \mathbb{E}\widehat{f}(0))^2 \lesssim \frac{1}{nh},$$
$$|\mathbb{E}\widehat{f}(0) - (1-\epsilon)f(0) - \epsilon g(0)| \lesssim h^{\beta_0} + \epsilon h^{\beta_1},$$

and

$$\epsilon|g(0) - f(0)| \lesssim \epsilon.$$

Define the oracle bandwidth h_* to be the largest $h \in \mathcal{H}$ such that

$$h^{\beta_0} + \epsilon h^{\beta_1} \leq c\sqrt{\frac{\log n}{nh}},$$

where the constant $c > 0$ will be determined later. Then it is easy to see that h_* satisfies

$$c'\sqrt{\frac{\log n}{nh_*}} \le h_*^{\beta_0} + \epsilon h_*^{\beta_1} \le c\sqrt{\frac{\log n}{nh_*}}, \tag{24}$$

for some constant c' that only depends on c.

We proceed to prove that $\widehat{h} \ge h_*$ with high probability. By the definition of $\widehat{h}$, we have

$$\begin{aligned}\mathbb{P}(\widehat{h} < h_*) &\le \mathbb{P}\left(\exists l \le h_* \text{ and } l \in \mathcal{H} \text{ s.t. } |\widehat{f}_{h_*}(0) - \widehat{f}_l(0)| > c_1\sqrt{\frac{\log n}{nl}}\right) \\ &\le \sum_{l \le h_*, l \in \mathcal{H}} \mathbb{P}\left(|\widehat{f}_{h_*}(0) - \widehat{f}_l(0)| > c_1\sqrt{\frac{\log n}{nl}}\right).\end{aligned}$$

We derive a bound for $\mathbb{P}\left(|\widehat{f}_{h_*}(0) - \widehat{f}_l(0)| > c_1\sqrt{\frac{\log n}{nl}}\right)$ for each $l \le h_*$ and $l \in \mathcal{H}$. Due to the error decomposition (23), we have:

$$|\widehat{f}_{h_*}(0) - \widehat{f}_l(0)| \le C(h_*^{\beta_0} + \epsilon h_*^{\beta_1}) + |\widehat{f}_{h_*}(0) - \mathbb{E}\widehat{f}_{h_*}(0)| + |\widehat{f}_l(0) - \mathbb{E}\widehat{f}_l(0)|,$$

for some constant $C > 0$. By (24), the bias term can be controlled as

$$C(h_*^{\beta_0} + \epsilon h_*^{\beta_1}) \le C \times c\sqrt{\frac{\log n}{nh_*}} \le \frac{c_1}{2}\sqrt{\frac{\log n}{nl}},$$

for a sufficiently small $c > 0$. Thus, we have

$$\begin{aligned}\mathbb{P}(\widehat{h} < h_*) &\le \sum_{l \le h_*, l \in \mathcal{H}} \mathbb{P}\left(|\widehat{f}_{h_*}(0) - \mathbb{E}\widehat{f}_{h_*}(0)| + |\widehat{f}_l(0) - \mathbb{E}\widehat{f}_l(0)| \ge \frac{c_1}{2}\sqrt{\frac{\log n}{nl}}\right) \\ &\le \sum_{l \le h_*, l \in \mathcal{H}} \mathbb{P}\left(|\widehat{f}_{h_*}(0) - \mathbb{E}\widehat{f}_{h_*}(0)| \ge \frac{c_1}{4}\sqrt{\frac{\log n}{nh^*}}\right) \\ &\quad + \sum_{l \le h_*, l \in \mathcal{H}} \mathbb{P}\left(|\widehat{f}_l(0) - \mathbb{E}\widehat{f}_l(0)| \ge \frac{c_1}{4}\sqrt{\frac{\log n}{nl}}\right).\end{aligned}$$

For any $l \le h_*$ and $l \in \mathcal{H}$, we use Bernstein's inequality, and get

$$\begin{aligned}&\mathbb{P}\left(|\widehat{f}_l(0) - \mathbb{E}\widehat{f}_l(0)| \ge t\right) \\ \le\ &\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n l^{-1}K(X_i/l) - \mathbb{E}l^{-1}K(X/l)\right| \ge t\right) \\ \le\ &2\exp\left(-\frac{nt^2/2}{\sigma^2 + Mt/3}\right),\end{aligned}$$

where we choose $t = \frac{c_1}{4}\sqrt{\frac{\log n}{nl}}$, and σ^2 and M have bounds

$$\sigma^2 \le \mathbb{E} l^{-2} K^2(X/l) \lesssim l^{-1} \text{ and } M \lesssim l^{-1}.$$

This implies the bound

$$\mathbb{P}\left(|\widehat{f}_l(0) - \mathbb{E}\widehat{f}_l(0)| \ge \frac{c_1}{4}\sqrt{\frac{\log n}{nl}}\right) \le 2\exp\left(-C' \log n\right), \tag{25}$$

where the constant $C' > 0$ can be arbitrarily large given a sufficiently large $c_1 > 0$. For example, we set a large enough $c_1 > 0$ so that $C' = 3$. This gives

$$\mathbb{P}(\widehat{h} < h_*) \le 4|\mathcal{H}| n^{-3} \lesssim n^{-3} \log n.$$

Now, on the event $\{\widehat{h} \ge h_*\}$, the risk decomposes as

$$|\widehat{f}_{\widehat{h}}(0) - f(0)| \le |\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0)| + |\widehat{f}_{h_*}(0) - f(0)|.$$

Due to the definition of $\widehat{h}$, the first term satisfies

$$|\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0)| \le c_1\sqrt{\frac{\log n}{nh_*}}. \tag{26}$$

For the second term, the error decomposition and the relation (24) implies

$$\mathbb{E}|\widehat{f}_{h_*}(0) - f(0)|^2 \lesssim \frac{\log n}{nh_*} + \epsilon^2.$$

Therefore, we have

$$\begin{aligned}
& \mathbb{E}(\widehat{f}_{\widehat{h}}(0) - f(0))^2 \\
\le\ & \mathbb{E}((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \widehat{h} \ge h_*) + \mathbb{E}((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \widehat{h} < h_*) \\
\le\ & 2\mathbb{E}((\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0))^2 : \widehat{h} \ge h_*) + 2\mathbb{E}((\widehat{f}_{h_*}(0) - f(0))^2 : \widehat{h} \ge h_*) \\
& + O\left(n^2 \mathbb{P}(\widehat{h} < h_*)\right) \\
\lesssim\ & \frac{\log n}{nh_*} + \epsilon^2 + \frac{\log n}{n} \\
\lesssim\ & \left(\frac{\log n}{n}\right)^{\frac{2\beta_0}{2\beta_0+1}} + \epsilon^2.
\end{aligned}$$

The last inequality above is by realizing that $h_* \asymp \left(\frac{n}{\log n}\right)^{-\frac{1}{2\beta_0+1}}$ from the relation (24). The proof is complete. □

Proof of Theorem 3.3. The proof for Theorem 3.3 follows the same argument as that of Theorem 3.5. The only difference lies in the normalization, which leads to the error decomposition

$$\widehat{f}_h(0) - f(0) = (\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0)) + \left(\mathbb{E}\widehat{f}_h(0) - f(0) - \frac{\epsilon}{1-\epsilon}g(0)\right) + \frac{\epsilon}{1-\epsilon}g(0).$$

The rest of the details are the same and is omitted. □

6.5. Proof of Theorem 4.1

We split the proof into upper and lower bounds. We first prove the following upper bound.

Theorem 6.1. *For the estimator $\widehat{f}(0) = \widehat{f}_h(0)$ with some $K \in \mathcal{K}_{\lfloor\beta_0\rfloor}(L)$ and $h = n^{-\frac{1}{2\beta_0+1}} \vee \epsilon^{\frac{1}{\beta_0+1}}$, we have*

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \lesssim [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^{\frac{2\beta_0}{\beta_0+1}}].$$

Proof. Decompose the error as

$$\widehat{f}_h(0) - f(0) = (\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0)) + (\mathbb{E}_h\widehat{f}(0) - f(0)),$$

where the first term is the stochastic error and the second term is the bias. For the first term, we have

$$\mathbb{E}(\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0))^2 = \frac{1}{n}\mathsf{Var}\left(\frac{1}{h}K\left(\frac{X}{h}\right)\right),$$

and

$$\begin{aligned}\mathsf{Var}\left(\frac{1}{h}K\left(\frac{X}{h}\right)\right) &\leq (1-\epsilon)\int \frac{1}{h^2}K^2\left(\frac{x}{h}\right)f(x)dx + \epsilon\int \frac{1}{h^2}K^2\left(\frac{x}{h}\right)dG(x)\\ &\lesssim \frac{1}{h}\int \frac{1}{h}K^2\left(\frac{x}{h}\right)dx + \frac{\epsilon}{h^2}\int dG(x)\\ &\lesssim \frac{1}{h} + \frac{\epsilon}{h^2}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}(\widehat{f}_h(0) - \mathbb{E}\widehat{f}_h(0))^2 \lesssim \frac{1}{nh} + \frac{\epsilon}{nh^2}. \tag{27}$$

For the bias term, we have

$$\begin{aligned}&\mathbb{E}\widehat{f}_h(0) - f(0)\\ &\quad= (1-\epsilon)\int \frac{1}{h}K\left(\frac{x}{h}\right)(f(x) - f(0))dx + \epsilon\int \frac{1}{h}K\left(\frac{x}{h}\right)dG(x) - \epsilon f(0),\end{aligned}$$

where the first term has bound

$$\left|\int \frac{1}{h}K\left(\frac{x}{h}\right)(f(x) - f(0))dx\right| \lesssim h^{\beta_0},$$

by [26, Chapter 1.2], and the next two terms can be bounded as

$$\left|\epsilon\int \frac{1}{h}K\left(\frac{x}{h}\right)dG(x) - \epsilon f(0)\right| \lesssim \frac{\epsilon}{h}\int dG(x) + \epsilon f(0) \lesssim \frac{\epsilon}{h}.$$

Therefore, we have

$$|\mathbb{E}\widehat{f}_h(0) - f(0)| \lesssim h^{\beta_0} + \frac{\epsilon}{h}. \tag{28}$$

Combine the two bounds (27) and (28), choose $h = n^{-\frac{1}{2\beta_0+1}} \vee \epsilon^{\frac{1}{\beta_0+1}}$, and then we complete the proof. □

Now we state the lower bound.

Theorem 6.2. *We have*

$$\mathcal{R}(\epsilon, \beta_0, L_0) \gtrsim [n^{-\frac{2\beta_0}{2\beta_0+1}}] \vee [\epsilon^{\frac{2\beta_0}{\beta_0+1}}].$$

Before proving this theorem, we need the following lemma.

Lemma 6.6. *A function $d(x)$ can be written as the difference of two density functions if and only if*

$$\int d = 0 \quad \text{and} \quad \int |d| \le 2.$$

Proof. The "only if" part is obvious. Now assume the two conditions hold, and then for any density function f, we have the following decomposition for d,

$$d = \left[d_+ + \left(1 - \frac{1}{2}\int |d|\right) f\right] - \left[d_- + \left(1 - \frac{1}{2}\int |d|\right) f\right],$$

where d_+ and d_- are the positive and negative parts of the function. The first condition implies $\int d_+ = \int d_- = \frac{1}{2}\int |d|$. Thus,

$$\int \left[d_+ + \left(1 - \frac{1}{2}\int |d|\right) f\right] = \int \left[d_- + \left(1 - \frac{1}{2}\int |d|\right) f\right] = 1.$$

The second condition guarantees that both $d_+ + \left(1 - \frac{1}{2}\int |d|\right) f$ and $d_- + \left(1 - \frac{1}{2}\int |d|\right) f$ are nonnegative. Thus, the proof is complete. □

Proof of Theorem 6.2. For the lower bound of the first term $n^{-\frac{2\beta_0}{2\beta_0+1}}$, see the proof of Lemma 6.2. We give a proof for the second term. Consider the following two functions

$$f = f_0,$$
$$\widetilde{f} = f_0 + ch^{\beta_0} b\left(\frac{x}{h}\right).$$

Here, we take f_0 as the density function of some normal distribution with mean zero so that $f_0 \in \mathcal{P}(\beta_0, L_0/2)$. The function b is defined in Lemma 2.2. The constant c is chosen small enough so that $f \in \mathcal{P}(\beta_0, L_0)$. In order that there exist g and $\widetilde{g}$ so that

$$(1-\epsilon)f + \epsilon g = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g},$$

it suffices to verify the existence of densities g and $\widetilde{g}$ such that

$$g(x) - \widetilde{g}(x) = \frac{1-\epsilon}{\epsilon}\left(\widetilde{f}(x) - f(x)\right) = c\frac{1-\epsilon}{\epsilon}h^{\beta_0}b\left(\frac{x}{h}\right).$$

By Lemma 6.6, it further suffices to verify the condition

$$c\frac{1-\epsilon}{\epsilon}\int h^{\beta_0}\left|b\left(\frac{x}{h}\right)\right|dx \leq 2,$$

and this is guaranteed by taking some $h \asymp \epsilon^{\frac{1}{\beta_0+1}}$. Now we have g and $\widetilde{g}$ such that $(1-\epsilon)f + \epsilon g = (1-\epsilon)\widetilde{f} + \epsilon\widetilde{g}$ holds. Moreover,

$$|f(0) - \widetilde{f}(0)| = ch^{\beta_0}b(0) \asymp \epsilon^{\frac{\beta_0}{\beta_0+1}}.$$

Apply Lemma 6.1, and the proof is complete. □

6.6. *Proofs of Theorem 4.2 and Theorem 5.2*

We first prove Theorem 5.2. Then, the proof of Theorem 4.2 will be sketched using arguments in the proofs of Theorem 3.5 and Theorem 5.2.

Proof of Theorem 5.2. We consider observations $X_1, ..., X_n$. We assume that $X_1, ..., X_a$ are generated from the density f with some integer a, and the remaining observations $X_{a+1}, ..., X_n$ are generated from contamination. The number a follows Binomial$(n, 1-\epsilon)$. This is without loss of generality, because the definition of $\widehat{f}$ does not depend on the order of the data $X_1, ..., X_n$. Apply Bernstein's inequality, and we get

$$\mathbb{P}\left(\frac{n-a}{n} \geq 2\epsilon\right) \leq \exp\left(-\frac{3}{8}n\epsilon\right).$$

From now on, we assume that $\epsilon \geq \frac{8\log n}{n}$, so that $\frac{n-a}{n} \leq 2\epsilon$ with probability at least $1 - n^{-3}$. The case $\epsilon < \frac{8\log n}{n}$ will be considered in the end of the proof. Moreover, the following analysis conditions on the event $\left\{\frac{n-a}{n} \leq 2\epsilon\right\}$, and we use $\bar{\mathbb{P}}$ and $\bar{\mathbb{E}}$ to denote probability and expectation conditioning on the random variable a.

We start by the following error decomposition,

$$\begin{aligned}\widehat{f}_h(0) - f(0) \quad = \quad & \frac{1}{n}\sum_{i=1}^{a}\left(h^{-1}K(X_i/h) - \mathbb{E}_{X\sim f}h^{-1}K(X/h)\right) \\ & + \frac{a}{n}\left(\mathbb{E}_{X\sim f}h^{-1}K(X/h) - f(0)\right) \\ & + \frac{1}{n}\sum_{i=a+1}^{n} h^{-1}K(X_i/h) - \frac{n-a}{n}f(0).\end{aligned}$$

With similar arguments used in the proof of Theorem 2.2, we have

$$\bar{\mathbb{E}}\left(\frac{1}{n}\sum_{i=1}^{a}\left(h^{-1}K(X_i/h)-\mathbb{E}_{X\sim f}h^{-1}K(X/h)\right)\right)^2 \lesssim \frac{1}{nh},$$

and

$$\left|\mathbb{E}_{X\sim f}h^{-1}K(X/h)-f(0)\right| \lesssim h^{\beta_0}.$$

Moreover, $\frac{n-a}{n} \le 2\epsilon$ implies that

$$\frac{1}{n}\sum_{i=a+1}^{n} h^{-1}K(X_i/h) \lesssim \frac{\epsilon}{h},$$

and $\frac{n-a}{n}f(0) \lesssim \epsilon$. These bounds motivate us to define an oracle bandwidth h_* that is the smallest $h \in \mathcal{H}$ such that

$$\frac{\epsilon}{h}+\sqrt{\frac{\log n}{nh}} \le h^{\widetilde{\beta}_0}.$$

Then it is obvious that h_* satisfies

$$ch_*^{\widetilde{\beta}_0} \le \frac{\epsilon}{h_*}+\sqrt{\frac{\log n}{nh_*}} \le h_*^{\widetilde{\beta}_0}, \tag{29}$$

with some constant $c>0$. Now we prove that $\widehat{h} \le h_*$ holds with high probability. According to the definition of $\widehat{h}$, we have

$$\begin{aligned}\bar{\mathbb{P}}(\widehat{h}>h_*) &\le \bar{\mathbb{P}}\left(\exists l \ge h_* \text{ and } l\in\mathcal{H} \text{ s.t. } |\widehat{f}_{h_*}(0)-\widehat{f}_l(0)| \ge c_1 l^{\widetilde{\beta}_0}\right)\\ &\le \sum_{l\ge h_*, l\in\mathcal{H}} \bar{\mathbb{P}}\left(|\widehat{f}_{h_*}(0)-\widehat{f}_l(0)| \ge c_1 l^{\widetilde{\beta}_0}\right).\end{aligned}$$

By the risk decomposition, for $l \ge h_*$ and $l \in \mathcal{H}$, the difference $|\widehat{f}_{h^*}(0)-\widehat{f}_l(0)|$ is bounded as

$$\begin{aligned}|\widehat{f}_{h_*}(0)-\widehat{f}_l(0)| \le{}& \left|\frac{1}{n}\sum_{i=1}^{a}\left(h_*^{-1}K(X_i/h_*)-\mathbb{E}_{X\sim f}h_*^{-1}K(X/h_*)\right)\right|\\ &+\left|\frac{1}{n}\sum_{i=1}^{a}\left(l^{-1}K(X_i/l)-\mathbb{E}_{X\sim f}l^{-1}K(X/l)\right)\right|\\ &+C\left(\frac{\epsilon}{h_*}+l^{\beta_0}\right),\end{aligned}$$

for some constant $C>0$. According to (29) and the condition $\widetilde{\beta}_0 \le \beta_0$, we have

$$C\left(\frac{\epsilon}{h_*}+l^{\beta_0}\right) \le C\left(h_*^{\widetilde{\beta}_0}+l^{\widetilde{\beta}_0}\right) \le \frac{c_1}{4}h_*^{\widetilde{\beta}_0}+\frac{c_1}{4}l^{\widetilde{\beta}_0},$$

where the last inequality holds for a sufficiently large c_1. Thus, we have the bound

$$\begin{aligned}
&\bar{\mathbb{P}}(\widehat{h} > h_*) \\
&\quad \le \sum_{l \ge h_*, l \in \mathcal{H}} \bar{\mathbb{P}}\left(|\widehat{f}_{h_*}(0) - \widehat{f}_l(0)| \ge \frac{c_1}{2} l^{\tilde{\beta}_0} + \frac{c_1}{2} h_*^{\tilde{\beta}_0} \right) \\
&\quad \le \sum_{l \ge h_*, l \in \mathcal{H}} \bar{\mathbb{P}}\left(\left| \frac{1}{n} \sum_{i=1}^{a} \left(h_*^{-1} K(X_i/h_*) - \mathbb{E}_{X\sim f} h_*^{-1} K(X/h_*) \right) \right| \ge \frac{c_1}{4} h_*^{\tilde{\beta}_0} \right) \\
&\quad + \sum_{l \ge h_*, l \in \mathcal{H}} \bar{\mathbb{P}}\left(\left| \frac{1}{n} \sum_{i=1}^{a} \left(l^{-1} K(X_i/l) - \mathbb{E}_{X\sim f} l^{-1} K(X/l) \right) \right| \ge \frac{c_1}{4} l^{\tilde{\beta}_0} \right) \\
&\quad \le \sum_{l \ge h_*, l \in \mathcal{H}} \bar{\mathbb{P}}\left(\left| \frac{1}{n} \sum_{i=1}^{a} \left(h_*^{-1} K(X_i/h_*) - \mathbb{E}_{X\sim f} h_*^{-1} K(X/h_*) \right) \right| \ge \frac{c_1}{4} \sqrt{\frac{\log n}{n h_*}} \right) \\
&\quad + \sum_{l \ge h_*, l \in \mathcal{H}} \bar{\mathbb{P}}\left(\left| \frac{1}{n} \sum_{i=1}^{a} \left(l^{-1} K(X_i/l) - \mathbb{E}_{X\sim f} l^{-1} K(X/l) \right) \right| \ge \frac{c_1}{4} \sqrt{\frac{\log n}{n l}} \right),
\end{aligned}$$

where the last inequality is by (29) and the observation that

$$l^{\tilde{\beta}_0} \ge h_*^{\tilde{\beta}_0} \ge \sqrt{\frac{\log n}{n h_*}} \ge \sqrt{\frac{\log n}{n l}}.$$

Use Bernstein's inequality in a similar way that derives (25), we obtain the bound

$$\bar{\mathbb{P}}\left(\left| \frac{1}{n} \sum_{i=1}^{a} \left(l^{-1} K(X_i/l) - \mathbb{E}_{X\sim f} l^{-1} K(X/l) \right) \right| \ge \frac{c_1}{4} \sqrt{\frac{\log n}{n l}} \right) \le 2n^{-3},$$

when the constant c_1 is chosen to be sufficiently large. Then, we have

$$\bar{\mathbb{P}}(\widehat{h} > h_*) \lesssim n^{-3} \log n.$$

On the event $\widehat{h} \le h_*$, the error decomposes as

$$|\widehat{f}_{\widehat{h}}(0) - f(0)| \le |\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0)| + |\widehat{f}_{h_*}(0) - f(0)|.$$

Due to definition of $\widehat{h}$, the first term is bounded as

$$|\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0)| \le c_1 h_*^{\tilde{\beta}_0}.$$

The second term uses the oracle bandwidth h_*. Then, we have

$$\begin{aligned}
&\bar{\mathbb{E}}(\widehat{f}_{\widehat{h}}(0) - f(0))^2 \\
&\quad \le \bar{\mathbb{E}}((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \widehat{h} \le h_*) + \bar{\mathbb{E}}((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \widehat{h} > h_*) \\
&\quad \lesssim \bar{\mathbb{E}}((\widehat{f}_{\widehat{h}}(0) - \widehat{f}_{h_*}(0))^2 : \widehat{h} \le h_*) + \bar{\mathbb{E}}((\widehat{f}_{h_*}(0) - f(0))^2 : \widehat{h} \le h_*) + n^2 \bar{\mathbb{P}}(\widehat{h} > h_*)
\end{aligned}$$

$$\lesssim h_*^{2\tilde{\beta}_0} + \frac{1}{nh_*} + h_*^{2\beta_0} + \frac{\epsilon^2}{h_*^2} + n^{-1}\log n$$
$$\lesssim \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \vee \epsilon^{\frac{2\tilde{\beta}_0}{\tilde{\beta}_0+1}},$$

where we have used (29) in the last inequality. Integrating over the random variable a, we have

$$\begin{aligned}
&\mathbb{E}(\widehat{f}_{\widehat{h}}(0) - f(0))^2 \\
&\quad \le \mathbb{E}\left((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \frac{n-a}{n} < 2\epsilon\right) + \mathbb{E}\left((\widehat{f}_{\widehat{h}}(0) - f(0))^2 : \frac{n-a}{n} \ge 2\epsilon\right) \\
&\quad \lesssim \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \vee \epsilon^{\frac{2\tilde{\beta}_0}{\tilde{\beta}_0+1}} + n^2\mathbb{P}\left(\frac{n-a}{n} \ge 2\epsilon\right) \\
&\quad \lesssim \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \vee \epsilon^{\frac{2\tilde{\beta}_0}{\tilde{\beta}_0+1}}. \qquad (30)
\end{aligned}$$

Finally, we consider the situation when $\epsilon < \frac{8\log n}{n}$. In this case, for any contamination distribution g, there is another $\widetilde{g}$ such that

$$(1-\epsilon)f + \epsilon g = \left(1 - \frac{8\log n}{n}\right) f + \frac{8\log n}{n}\widetilde{g}.$$

See [7] for a rigorous argument of the above equality. Then, we can equivalently analyze the risk with contamination proportion $\frac{8\log n}{n}$. This leads to the error bound

$$\left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \vee \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{\tilde{\beta}_0+1}} \asymp \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \asymp \left(\frac{\log n}{n}\right)^{\frac{2\tilde{\beta}_0}{2\tilde{\beta}_0+1}} \vee \epsilon^{\frac{2\tilde{\beta}_0}{\tilde{\beta}_0+1}}. \quad (31)$$

Hence, we let $n \to \infty$ and $\epsilon \to 0$, and the proof is complete. □

Proof of Theorem 4.2. For the estimator that uses (15), the result is a special case of Theorem 5.2 by letting $\widetilde{\beta}_0 = \beta_0$ in view of the bounds (30) and (31). For the estimator that uses (14), the result follows the same argument as the proof of Theorem 3.5. □

6.7. Proofs of Lemma 4.1 and Theorem 4.3

Proof of Lemma 4.1. We use $\phi(\cdot)$ to denote the density of $N(0,1)$. Then, define

$$f(x) = c_3\phi(c_3 x),$$
$$g(x) = \frac{c_4}{\epsilon^{\frac{1}{\tilde{\beta}_0+1}}}\phi\left(\frac{c_4 x}{\epsilon^{\frac{1}{\tilde{\beta}_0+1}}}\right),$$

$$\widetilde{f}(x) = (1-\epsilon)f(x) + \epsilon g(x),$$
$$\widetilde{g}(x) = \phi(x).$$

First, there exists a constant c_3 depending on c_1, c_2 such that for any $\beta_0, \widetilde{\beta}_0 \le c_1$ and $L_0, \widetilde{L}_0 \ge c_2$, we have $f \in \mathcal{P}(\beta_0, L_0) \cap \mathcal{P}(\widetilde{\beta}_0, \widetilde{L}_0/2)$. This is due to the fact that $\phi^{(\alpha)}(x)$ is uniformly bounded for all $\alpha \le c$ when c is some constant. By definition,

$$\epsilon g(x) = c_4 \epsilon^{\frac{\widetilde{\beta}_0}{\widetilde{\beta}_0+1}} \phi\left(\frac{c_4 x}{\epsilon^{\frac{1}{\widetilde{\beta}_0+1}}}\right),$$

For the same reason as above, there exists a constant c_4 depending on c_1, c_2 such that for any $\widetilde{\beta}_0 \le c_1$ and $\widetilde{L}_0 \ge c_2$, we have $\epsilon g \in \Sigma(\widetilde{\beta}_0, \widetilde{L}_0/2)$, which then implies $\widetilde{f} \in \mathcal{P}(\widetilde{\beta}_0, \widetilde{L}_0)$. Now we note that

$$(1-\epsilon)f + \epsilon g = (1-0)\widetilde{f} + 0\widetilde{g},$$

and

$$\left|\widetilde{f}(0) - f(0)\right| = \epsilon |f(0) - g(0)| \ge c_0 \epsilon^{\frac{\widetilde{\beta}_0}{\widetilde{\beta}_0+1}},$$

when ϵ smaller than a constant and where c_0 is a constant depending on c_3, c_4. Thus for any estimator $\widehat{f}(0)$,

$$\left[\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2\right] \vee \left[\sup_{p(0,f,g)\in\mathcal{M}(0,\widetilde{\beta}_0,\widetilde{L}_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2\right] \ge c_0 \epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}},$$

by applying Lemma 6.1. □

Proof of Theorem 4.3. For any constants c_1, c_2, let c_0 be the constant as guaranteed to exist in Theorem 4.1, and assume there exists an estimator $\widehat{f}(0)$ which is $(c_1, c_2, c_3, r_1(\beta_0), r_2(\beta_0))$ rate adaptive. With $L_0 = c_2$, we consider two models respectively with parameters $(n, \epsilon, \beta_0, L_0)$ and $(n, 0, \widetilde{\beta}_0, L_0)$, with the specific values of $n, \epsilon, \beta_0, \widetilde{\beta}_0$ to be chosen later. By the definition of rate adaptivity (16), we have:

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \le c_3[\epsilon^{r_2(\beta_0)} \vee n^{-r_1(\beta_0)}],$$
$$\sup_{p(0,f,g)\in\mathcal{M}(0,\widetilde{\beta}_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \le c_3 n^{-r_1(\widetilde{\beta}_0)}.$$

On the other hand, Theorem 4.1 claims that for any small enough ϵ, any large enough n, any $\beta_0, \widetilde{\beta}_0 \le c_1$, we have

$$\sup_{p(\epsilon,f,g)\in\mathcal{M}(\epsilon,\beta_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2 \vee \sup_{p(0,f,g)\in\mathcal{M}(0,\widetilde{\beta}_0,L_0)} \mathbb{E}_{p^n}\left(\widehat{f}(0) - f(0)\right)^2$$

$\geq c_0 \epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}$.

Together this yields

$$c_3[\epsilon^{r_2(\beta_0)} \vee n^{-r_1(\beta_0)} \vee n^{-r_1(\widetilde{\beta}_0)}] \geq c_0 \epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}. \tag{32}$$

Now we choose $n, \beta_0, \widetilde{\beta}_0, \epsilon$ in a legitimate range so that this inequality becomes a contradiction. First we fix some $\beta_0 \leq c_1$. Then we choose ϵ to be small enough such that $c_3 \epsilon^{r_2(\beta_0)} \leq c_0 \epsilon^a$ for some $a > 0$. Indeed this holds as long as $\epsilon^{r_2(\beta_0)} < \frac{c_0}{c_3}$. Now since $a > 0$, we can choose $\widetilde{\beta}_0$ to be small enough such that $\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1} < a$. Finally since $r_1(\beta_0), r_1(\widetilde{\beta}_0) > 0$, we can choose n large enough such that $c_3[n^{-r_1(\beta_0)} \vee n^{-r_1(\widetilde{\beta}_0)}] < c_0 \epsilon^{\frac{2\widetilde{\beta}_0}{\widetilde{\beta}_0+1}}$. With these choices, it is obvious that equation (32) becomes a contradiction, as desired. □

6.8. Proof of Theorem 5.1

The proofs are exactly the same as in the one-dimensional case. For the lower bounds, we only need to replace the mollifier function $l(x)$ by its multivariate extension $l_d(x) = l(\|x\|)$. The upper bounds are achieved by $\widehat{f}_h(0) = \frac{1}{n(1-\epsilon)} \sum_{i=1}^n \frac{1}{h^d} K_d\left(\frac{X_i}{h}\right)$, where the bandwidth is $h = n^{-\frac{1}{2\beta_0+d}} \wedge n^{-\frac{1}{2\beta_0+d}} \epsilon^{-\frac{2}{2\beta_0+d}}$ for structured contamination and is $h = n^{-\frac{1}{2\beta_0+d}} \vee \epsilon^{\frac{1}{\beta_0+d}}$ for arbitrary contamination. We can use a product kernel for K_d. See [8, Chapter 12] for details.

Acknowledgement

The authors thank Zhao Ren for reading the manuscript and for his helpful comments. They also thank an associate editor and a referee for their suggestions that lead to the improvement of the paper.

References

[1] Lawrence D. Brown and Mark G. Low. A constrained risk inequality with applications to nonparametric functional estimation. *The Annals of Statistics*, 24(6):2524–2535, 1996. MR1425965

[2] T. Tony Cai. Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statistica Sinica*, 13:881–902, 2003. MR1997178

[3] T. Tony Cai and Jiashun Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics*, 38(1):100–145, 2010. MR2589318

[4] T. Tony Cai and Mark G. Low. On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343, 2005. MR2211088

[5] T. Tony Cai and Mark G. Low. Optimal adaptive estimation of a quadratic functional. *The Annals of Statistics*, 34(5):2298–2325, 2006. MR2291501

[6] Mengjie Chen, Chao Gao, and Zhao Ren. A general decision theory for Huber's ϵ-contamination model. *Electronic Journal of Statistics*, 10(2):3752–3774, 2016. MR3579675

[7] Mengjie Chen, Chao Gao, Zhao Ren, et al. Robust covariance and scatter matrix estimation under Huber's contamination model. *The Annals of Statistics*, 46(5):1932–1960, 2018. MR3845006

[8] L. Devroye and G. Lugosi. Combinatorial methods in density estimation, 2001. MR1843146

[9] David L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994. MR1272082

[10] David L. Donoho and Richard C. Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, 19(2):668–701, 1991. MR1105839

[11] Bradley Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104, 2004. MR2054289

[12] Chao Gao. Robust regression via mutivariate regression depth. *Bernoulli (to appear)*, 2017.

[13] Christian H. Hesse. Deconvolving a density from partially contaminated observations. *Journal of Multivariate Analysis*, 55(2):246–260, 1995. MR1370403

[14] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. MR0161415

[15] Peter J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36(6):1753–1758, 1965. MR0185747

[16] Jiashun Jin and T. Tony Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007. MR2325113

[17] Iain M. Johnstone. Chi-square oracle inequalities. *Lecture Notes-Monograph Series*, pages 399–418, 2001. MR1836572

[18] O.V. Lepski and V.G. Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. *The Annals of Statistics*, 25(6):2512–2546, 1997. MR1604408

[19] O.V. Lepski and T. Willer. Oracle inequalities and adaptive estimation in the convolution structure density model. *The Annals of Statistics*, 47(1):233–287, 2019. MR3909933

[20] O.V. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability & Its Applications*, 35(3):454–466, 1991. MR1091202

[21] O.V. Lepskii. Asymptotically minimax adaptive estimation. I: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992. MR1147167

[22] O.V. Lepskii. Asymptotically minimax adaptive estimation. II. schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications*, 37(3):433–448, 1993. MR1214353

[23] Rostyslav Maiboroda and Olena Sugakova. Nonparametric density estimation for symmetric distributions by contaminated data. *Metrika*, 75(1):109–126, 2012. MR2878111

[24] Bernard W. Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC Press, 1986. MR0848134

[25] Karine Tribouley. Adaptive estimation of integrated functionals. *Mathematical Methods of Statistics*, 9(1):19–38, 2000. MR1772223

[26] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer, 2009. MR2724359

[27] Bin Yu. Assouad, fano, and le cam. *Festschrift for Lucien Le Cam*, 423:435, 1997. MR1462963

[28] Ming Yuan and Jiaqin Chen. Deconvolving multidimensional density from partially contaminated observations. *Journal of Statistical Planning and Inference*, 104(1):147–160, 2002. MR1900523