Adaptively local 1-dimensional subproblems

Jianqing Fan

November 10, 1989

Abstract

We provide a new insight of the difficulty of nonparametric estimation of a whole function. A new method is invented for finding a minimax lower bound of globally estimating a function. The idea is to adjust automatically the direction to the nearly hardest 1-dimensional subproblem at each location, and to use locally the difficulty of 1-dimensional subproblem. In a variety of contexts, our method can give not only attainable global rates, but also constant factors. Comparing with the existing techniques, our method has the advantages of being easily implemented and understood, and can give constant factors as well.

We illustrate the lower bound by using examples of nonparametric density estimation as well as nonparametric regression. Concise proofs of the lower rates are given. Applying our lower bound to deconvolution setting, we obtain the best attainable global rates of convergence. With the existing techniques, it would be extremely difficult to solve such a problem.

---

[0] *Abbreviated title.* Local 1-d subproblems.

AMS 1980 subject classification. Primary 62G20. Secondary 62G05.

Key words and phrases. Cubical lower bound, 1-dimensional subproblems, global rates of convergence, minimax risks, density estimate, nonparametric regression, deconvolution.

1 Introduction

Nonparametric techniques provide a very useful tool for investigating the structure of some interesting functions. A useful mathematical formulation is to think of estimating some function $T \circ f(x)$ (e.g., density function, regression function) based on a random sample $X_1, ..., X_n$, from a density f with a *priori* $f \in \mathcal{F}$, under some global loss functions. The global loss functions are typically those induced by L_p-norm:

$$L(d, T \circ f) = \left[\int_a^b |T \circ f(x) - d(x)|^p w(x) dx \right]^{1/p}, \tag{1.1}$$

where $w(x)$ is a weight function, and $d(x)$ is a decision function to estimate $T \circ f(x)$.

How can one measure the difficulty of estimating the function $T \circ f$ under the weighted L_p loss? Our approach to the question is that

1. specify a subproblem — given a set of densities $\mathcal{F}_0 \subset \mathcal{F}$, estimating $T \circ f(x)$ with a *priori* $f \in \mathcal{F}_0$; the geometry of $\mathcal{F}_0$ is typically hypercubical (Fan (1989b)).

2. use the difficulty of the subproblem as a lower bound of the difficulty of the full nonparametric problem.

In the second step, we first formulate problems of estimating a *functional* $T \circ f(x_0)$ at each location x_0, then adjust automatically the direction at the location x_0 to the nearly most difficult direction of estimating the *functional* $T \circ f(x_0)$, and finally add the difficulties of 1-dimensional subproblems at all locations together, according to their weights, to find a lower bound. A feature of our approach is to use geometric ideas, which can be easily understood and implemented.

Our approach is related to the illuminating ideas of Donoho and Liu (1987a, 1988) and other approaches (Farrell (1972), Ibragimov *et al.* (1987), Khas'minskii (1979), Stone (1980), etc.) for estimating a statistical *functional* (instead of estimating a whole function). In the context of estimating a linear *functional*, Donoho and Liu (1987a) shows that the difficulty of the hardest 1-dimensional subproblem is hard enough to capture the difficulty

of a full nonparametric problem. However, the hardest one-dimensional subproblem is *not* difficult enough to capture the difficulty of estimating a whole function (e.g., the whole density function). To bridge the gaps, we use a growing number of dimensional subproblem, adjusting directions accordingly, to capture the difficulty of estimating a whole function.

Comparing with the existing methods of Stone (1982) (discriminant analysis based method), Kha'sminskii (1978) (Shannon information based method), Birgé (1987) (Assaud's Lemma based method), our approach is simpler in the second step above. In a sense, the existing approaches attempt to count (see (2.9) of Stone (1982)) how many densities that we can not distinguish at the same time, while our method adopts locally the idea of the 1-dimensional subproblem. Thus, our argument is simpler than the existing ones and is expected to extend to find minimax risks within a few percents (Donoho and Liu (1988), Donoho *et al.* (1987)). Compare also Efroimovich and Pinsker (1982), Nussbaum (1985), where the exact minimax risk is found for the *ellipsoid constraints under L_2 loss.*

The paper specially focuses on finding global rates of convergence, and on introducing the new methodology. As a byproduct, we will give constant factors in lower bounds as well. However, the our primary goal of the paper is not to make much effort to get as sharp constants in lower bounds as we can via our cubical approach. The reason is that doing so might obscure the main idea of the study. Thus, it would be no surprise that one can sharp our constant factors via the cubical approach. Indeed, we believe that the hypercubical approach can be used to find minimax risks within a few percents of error, by a modified version of the cubical approach together with the ideas of Donoho and Liu (1988), Donoho *et al.* (1987). In contrast, with the existing approaches, it would be very difficult to sharp the constant factor.

We demonstrate *within a few lines* that our method can give the best attainable lower rates in a variety of contexts. Examples are estimating density functions, estimating decreasing failure rates, estimating regression functions, estimating of conditional quantile functions, etc. See Keifer (1981) for other combinations of problems. Thus, we provide a

short and simple proof of these deep results, which can be understood at the level of second year graduates.

The cubical method is especially useful for establishing attainable global lower rates of deconvolution problem (indirect observations). We use this to solve the open problem of the optimal global rates for deconvlution. In this context, the cubical method provides the precise description of the difficulty, depending on the tail of characteristic functions of error distributions (Fan (1988a)), of deconvolution.

The paper is organized as follows. Section 2 introduces the cubical lower bound. Section 3 illustrates the method by using examples of nonparametric density estimation and nonparametric regression. Section 4 applies the lower bound to find a global rates of convergence for deconvolution models. Some technical arguments are stated in section 5.

2 Cubical lower bound of global rates

In this section, we give a lower bound for estimating a function $T \circ f(x)$. We discuss the problem for the 1-dimensional case. The higher dimensional results follow naturally.

Let $[a, b]$ be an interval on the line, and $x_{n,j} = a + j(b-a)/m_n$, $j = 1, ..., m_n$. Denote $\boldsymbol{\theta}_{m_n} = (\theta_1, \ ..., \ \theta_{m_n})$, and

$$f_{\boldsymbol{\theta}_{m_n}}(t) = f_0(t) + a_n^{-1} \sum_{j=1}^{m_n} \theta_j H(m_n(t - x_{n,j})), \tag{2.1}$$

where $f_0(t)$ is a density function, $H(\cdot)$ is a bounded function whose integral on the line is 0, and m_n and a_n are sequences tending to infinity. By suitable choices of m_n, a_n, f_0, and H, the function $f_{\boldsymbol{\theta}_{m_n}}$ will be a density function. Denote a class of densities by

$$\mathcal{F}_n = \{f_{\boldsymbol{\theta}_{m_n}} : \ \boldsymbol{\theta}_{m_n} \text{is a sequence of 0's and 1's}\} \tag{2.2}$$

and denote

$$\boldsymbol{\theta}_{j0} = (\theta_1, \ ..., \ \theta_{j-1}, 0, \theta_{j+1}, \ ..., \theta_{m_n}), \ \ \boldsymbol{\theta}_{j1} = (\theta_1, \ ..., \ \theta_{j-1}, 1, \theta_{j+1}, \ ..., \theta_{m_n}). \tag{2.3}$$

Suppose that a_n and m_n are chosen so that

$$\max_{1\le j\le m_n}\max_{\boldsymbol{\theta}_{m_n}\in\{0,1\}^{m_n}}\chi^2(f_{\boldsymbol{\theta}_{j0}},f_{\boldsymbol{\theta}_{j1}})\le c/n, \tag{2.4}$$

for some constant c, where $\chi^2(f,g)=\int(f-g)^2/f dx$. Then, we have the following lower bound of estimating $T\circ f(x)$.

Theorem 1. *Suppose that $T\circ f(x)$ satisfies*

$$|T\circ f_{\boldsymbol{\theta}_{j0}}(x)-T\circ f_{\boldsymbol{\theta}_{j1}}(x)|\ge|T_{nH}(m_n(x-x_{n,j}))|, \tag{2.5}$$

for some function T_{nH}, depending on n and H. Let $w(x)$ be a non-negative, continuous function on $[a,\ b]$. If the condition (2.4) is satisfied, then for any estimate $\hat{T}_n(x)$ based on n i.i.d. observations from the unknown density f, we have

$$\begin{aligned}&\inf_{\hat{T}_n(x)}\sup_{f\in\mathcal{F}_n}E_f\int_a^b|\hat{T}_n(x)-T\circ f(x)|^p w(x)dx\\ \ge\ &\frac{1-\sqrt{1-e^{-c}}}{2^{p+1}(b-a)}\int_a^b w(x)dx\int_a^b|T_{nH}(x-a)|^p dx(1+o(1)).\end{aligned} \tag{2.6}$$

Let's give a proof here to illustrate the simple and basic idea inside.

Proof. Assign the prior $\theta_1,\ ...,\ \theta_{n_m}$ to be i.i.d. with

$$P(\theta_j=0)=P(\theta_j=1)=1/2,\ \text{for } j=1,...,m_n$$

Denote $E_\theta g(\boldsymbol{\theta}_{m_n})$ by the expectation of $g(\boldsymbol{\theta}_{m_n})$ with respect to the prior distribution of $\theta_1,\ ...,\ \theta_{m_n}$. By Fubini's Theorem,

$$\begin{aligned}&\inf_{\hat{T}_n(x)}\sup_{f\in\mathcal{F}_n}E_f\int_a^b|\hat{T}_n(x)-T\circ f(x)|^p w(x)dx\\ \ge\ &\inf_{\hat{T}_n(x)}E_\theta E_{f_\theta}\int_a^b|\hat{T}_n(x)-T\circ f(x)|^p w(x)dx\\ \ge\ &\int_a^b\inf_{\hat{T}_n(x)}E_\theta E_{f_\theta}|\hat{T}_n(x)-T\circ f(x)|^p w(x)dx.\end{aligned} \tag{2.7}$$

Let $a_{nj}(x) = |T \circ f_{\boldsymbol{\theta}_{j_0}}(x) - T \circ f_{\boldsymbol{\theta}_{j_1}}(x)|/2$, where $\boldsymbol{\theta}_{j_0}, \boldsymbol{\theta}_{j_1}$ are given by (2.3). Then, it follows that

$$\begin{aligned} & \inf_{\hat{T}_n(x)} E_\theta E_{f_\theta} |\hat{T}_n(x) - T \circ f(x)|^p \\ \geq & \max_{1 \leq j \leq m_n} \inf_{\hat{T}_n(x)} E_\theta \left\{ E_{\theta_j} E_{f_\theta} |\hat{T}_n(x) - T \circ f(x)|^p \right\}. \end{aligned} \tag{2.8}$$

Denote $P_{\theta_j 0}$ and $P_{\theta_j 1}$ by the probability measures generated by the density functions $f_{\boldsymbol{\theta}_{j_0}}$ and $f_{\boldsymbol{\theta}_{j_1}}$, respectively. Then, the last quantity (2.8) is no smaller than

$$\begin{aligned} & \max_{1 \leq j \leq m_n} E_\theta a_{nj}^p(x) \frac{1}{2} \inf_{\hat{T}_n(x)} [P_{\theta_j 0}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_0}}(x)| \geq a_{nj}(x)\} \\ & \qquad + P_{\theta_j 1}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_1}}(x)| \geq a_{nj}(x)\}] \\ \geq & \max_{1 \leq j \leq m_n} a_{nj}^p(x)/2 E_\theta \inf_{\hat{T}_n(x)} [P_{\theta_j 0}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_0}}(x)| \geq a_{nj}(x)\} \\ & \qquad + P_{\theta_j 1}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_0}}(x)| \leq a_{nj}(x)\}]. \end{aligned} \tag{2.9}$$

The terms in the square blanket can be viewed as the sum of type I and type II errors of a testing procedure for the testing problem:

$$H_0 : f = f_{\boldsymbol{\theta}_{j_0}} \longleftrightarrow H_1 : f = f_{\boldsymbol{\theta}_{j_1}}. \tag{2.10}$$

Since the χ^2-distance for the pair of densities is no large than c/n, it follows that (see e.g. Lemma 1.3, Fan (1989), page 14)

$$\begin{aligned} & P_{\theta_j 0}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_0}}(x)| \geq a_{nj}(x)\} + P_{\theta_j 1}\{|\hat{T}_n(x) - T \circ f_{\boldsymbol{\theta}_{j_0}}(x)| \leq a_{nj}(x)\} \\ \geq & \ 1 - \sqrt{1 - e^{-c}}. \end{aligned} \tag{2.11}$$

Consequently, by (2.5), (2.7), (2.9), (2.11), we have

$$\begin{aligned} & \inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_n} E_f \int_a^b |\hat{T}_n(x) - T \circ f_{j0}(x)|^p w(x) dx \\ \geq & \frac{1 - \sqrt{1 - e^{-c}}}{2} \int_a^b \max_{1 \leq j \leq m_n} a_{nj}^p(x) w(x) dx \\ \geq & \frac{1 - \sqrt{1 - e^{-c}}}{2^{p+1}} \sum_{j=1}^{m_n} \int_{x_{nj}}^{x_{nj+1}} |T_{nH}(m_n(x - x_{n,j}))|^p w(x) dx. \end{aligned} \tag{2.12}$$

Now, we need to calculate the summation in the last expression is no smaller than the quantity indicated in the Theorem 1. By the uniform continuity of $w(x)$, it follows that for any given ε, there exists an n_0 such that when $n \geq n_0$,

$$\inf_{x \in [x_{n,j},\, x_{n,j+1}]} w(x) \geq (1-\varepsilon) w(x_{n,j}).$$

Consequently, when $n \geq n_0$, the summation in (2.12) is no smaller than

$$\begin{aligned} & (1-\varepsilon)/m_n \sum_{j=1}^{m_n} w(x_{n,j}) \int_a^b |T_{nH}(x-a)|^p dx \\ = \; & (1-\varepsilon)/(b-a) \sum_{j=1}^{m_n} w(x_{n,j})(x_{n,j+1}-x_{n,j}) \int_a^b |T_{nH}(x-a)|^p dx. \end{aligned} \tag{2.13}$$

The conclusion follows by letting $n \to \infty$ and then letting $\varepsilon \to 0$.

Specially, when $T \circ f = f^{(k)}(x)$, the k^{th} derivative of the unknown density function, the condition (2.5) (f_0 and H are chosen to have the k^{th} derivative) satisfies with

$$|T_{nH}(x)| = |m_n^k / a_n H^{(k)}(x)|. \tag{2.14}$$

We have the following lower bounded for estimating $f^{(k)}(x)$.

Corollary 1. *Under the assumption of (2.4),*

$$\begin{aligned} & \inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_n} E_f \int_a^b |\hat{T}_n(x) - f^{(k)}(x)|^p w(x) dx \\ \geq \; & \frac{1-\sqrt{1-e^{-c}}}{2^{p+1}(b-a)} \int_a^b w(x) dx \int_a^b |H^{(k)}(x-a)|^p dx \\ & (m_n^k / a_n)^p (1+o(1)). \end{aligned} \tag{2.15}$$

One may wonder whether condition (2.4) is easy to check or not. The following Lemma gives an easy sufficient condition for (2.4).

Lemma 1. *If H has bounded support on $[0,\, b-a]$, then*

$$\max_{1 \leq j \leq m_n} \max_{\boldsymbol{\theta}_{m_n} \in \{0,1\}^{m_n}} \chi^2(f_{\boldsymbol{\theta}_{j_0}}, f_{\boldsymbol{\theta}_{j_1}}) \leq \frac{1}{D m_n a_n^2} \int_a^b |H(x-a)|^2 dx, \tag{2.16}$$

where

$$D = \min_{a \le x \le b} f_0(x) - \max_{0 \le x \le b-a} |H(x)|/a_n.$$

Remark 2.1. Geometrically speaking, if we take $\{H(m_n(t - x_{n,j})) : j = 1, ..., m_n\}$ (H has the support $[0, b-a]$) as a part of the orthogonal basis, then our class of densities $\mathcal{F}_n$ corresponds the vertices of an m_n-dimensional hypercube centered at the point θ_0, the Fourier-Bessel coefficients of the density f_0, in R^∞. The condition (2.4) essentially says that any two vertices of the hypercube can not be tested consistently. Thus, the χ^2-distance in (2.4) can be replaced by Hellinger distance, and Theorem 1 holds by replacing $(1-\sqrt{1-e^{-c}})$ by $(1-\sqrt{1-e^{-2c}})$ (see (1.3.9) of Fan (1989a), page 46-47, 475-477 of Le Cam (1985)). We state Theorem in terms of χ^2-distance because the condition 2.4 is easier to verify (see Lemma 1).

Remark 2.2. Comparing with the approach of Donoho and Liu (1987a), we use the difficulty of 1-dimensional subproblems *locally* (see (2.7)) as the difficulty of the full nonparametric problem. The idea of our approach adjusts automatically (see (2.8) and (2.9)) the direction of a 1-dimensional subproblem at each location to the direction of the *nearly hardest 1-dimensional* subproblem at that location, and then add the difficulty up according to the weight of each location (see (2.12)). The notion here is also different from Donoho and Liu (1987b), Bickel and Ritov (1988) and cubical method of Fan (1989b), where only two highly composite sets of densities are tested. Our idea here is to test a variety of simple hypotheses.

Remark 2.3. Discrimant analysis based methods (Stone (1982)), Shannon information based methods (Khas'minskii (1978)), and Assouad's Lemma (Lemma 9, P524, LeCam (1985)) based methods (Birgé (1987)) essentially count how many densities totally we can not distinguish at the same time. Thus, the analysis would be more complicated than ours.

Our method does not try to count the number of densities that we can not distinguish, but adjust directions of pairs of densities accordingly.

Remark 2.4. Choose $a_n = m_n^{m+\alpha}$, and the functions $H, f_0(t)$ so that they have bounded $m+1$ derivatives. The class of $\mathcal{F}_n$ defined by (2.1) will be a subset of the smoothness constraints:

$$\mathcal{F}_{m,\alpha,B} = \{f : |f^{(m)}(x) - f^{(m)}(y)| \leq B|x-y|^{\alpha}\}, \tag{2.17}$$

for some $0 \leq \alpha < 1$. Thus, (2.6) is also a minimax lower bound for $\mathcal{F}_{m,\alpha,B}$.

Remark 2.5. It is not hard to get a lower bound of the minimax risk

$$\inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_n} E\left(\int_a^b |\hat{T}_n(x) - T \circ f(x)|^p w(x)dx\right)^{1/p} \tag{2.18}$$

(Direct use our method would fail). By normalizing of $w(x)$, without loss of generality, assume that the total weight of $w(x)$ on $[a, b]$ is 1. Thus using the fact that

$$\left(\int_a^b |\hat{T}_n(x) - T \circ f(x)|^p w(x)dx\right)^{1/p} \geq \int_a^b |\hat{T}_n(x) - T \circ f(x)| w(x)dx,$$

for $p \geq 1$, we can easily get a lower bound of (2.18).

Remark 2.6. We attempt only to sharp the lower bound in terms of rates of convergence. However, taking any function H and f_0 would give us a constant factor in the lower bound as well. Nevertheless, our method can provide a nonasymptotic lower bound, if we bound (2.12) from below by using the minimum of $w(x)$. To find a sharper constant factor, we adopt the same idea by using the difficulty of the whole 1-dimensional ($0 \leq \theta_j \leq 1$) subproblem instead of only a 2-point ($\theta = 0, 1$) subproblem *locally*.

3 Applications to Nonparametric regression and density estimation

In this section, we apply our lower bounds to the setting of nonparametric regression and density estimation. For the simplicity of notation, we use 1-dimensional version of our Theorem 1 again.

3.1 Density Estimation

Suppose that we have n i.i.d. observations $X_1, ..., X_n$ based on an unknown density function f, with a smoothness *priori* $\mathcal{F}_{m,\alpha,B}$ defined by (2.17).

Let $a_n = m_n^{m+\alpha}$, and choose the density function $f_0(x)$ and the function $H(x)$ supported on $[0, b-a]$ such that $\mathcal{F}_n \subset \mathcal{F}_{m,\alpha,B}$. With $m_n = (n/d)^{\frac{1}{2(m+\alpha)+1}}$ and Lemma 1, the condition (2.4) is satisfied with

$$c = \frac{d}{min_{a \le x \le b} f_0(x)} \int_a^b |H(x-a)|^2 dx,$$

by Lemma 1. Thus, by Corollary 1, the minimax lower bound is as follows.

Theorem 2. *For estimating the kth derivative of the unknown density, we have*

$$\begin{aligned} &\inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_n} E_f \int_a^b |\hat{T}_n(x) - f^{(k)}(x)|^p w(x) dx \\ \ge\ & \frac{1-\sqrt{1-e^{-c}}}{2^{p+1}(b-a)} \int_a^b w(x) dx \int_a^b |H^{(k)}(x-a)|^p dx d^{\frac{p(m+\alpha-k)}{2(m+\alpha)+1}} n^{-\frac{p(m+\alpha-k)}{2(m+\alpha)+1}}. \end{aligned}$$

Thus, we give an short and easily understood proof of global lower rates. A byproduct of our proof is that we could give a constant factor of the lower bound. To find a sharper lower bound in constant factor, we may want to maximize the last term over $d > 0$, and the function H subject to $f_{\boldsymbol{\theta}_{m_n}} \in \mathcal{F}_{m,\alpha,B}$.

Specially, for estimating density function, we have

$$\inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_{m,\alpha,B}} E_f \int_0^1 |\hat{T}_n(x) - f(x)|^2 dx$$

$$\geq \int_0^1 |H(x)|^2 dx \left(\sup_{d>0} \frac{(1-\sqrt{1-e^{-c}})}{8} d^{\frac{2(m+\alpha)}{2(m+\alpha)+1}} \right) n^{-\frac{2(m+\alpha)}{2(m+\alpha)+1}}, \tag{3.1}$$

and

$$\begin{aligned} &\inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_{m,\alpha,B}} E_f \int_0^1 |\hat{T}_n(x) - f(x)| dx \\ &\geq \int_0^1 |H(x)| dx \left(\sup_{d>0} \frac{(1-\sqrt{1-e^{-c}})}{4} d^{\frac{m+\alpha}{2(m+\alpha)+1}} \right) n^{-\frac{m+\alpha}{2(m+\alpha)+1}}, \end{aligned} \tag{3.2}$$

where $c = d \int_0^1 H^2(x)dx / \min_{0 \leq x \leq 1} f_0(x)$.

Remark 3.1. Taking $f_0(x)$ to be a strictly decreasing function as Kiefer (1981) (e.g. $f_0(x) = L(1+x)^{-L-1}, x \geq 0$), and $H(x)$ to have a small first derivative, then the class of $\mathcal{F}_n$ is a subclass of decreasing densities. Thus, the rates above are also the lower rates for estimating a decreasing densities as well as decreasing failure rates (Kiefer (1981)).

3.2 Nonparametric Regression

A useful mathematical model of nonparametric regression is to think of estimating a conditional mean

$$T \circ f(x) = \int_{-\infty}^{+\infty} y f(x,y) dy / \int_{-\infty}^{+\infty} f(x,y) dy, \tag{3.3}$$

using the random sample $(X_1, Y_1), ..., (X_n, Y_n)$ i.i.d. from the density $f(x,y)$. To find a lower bound, we use a simple normal submodel:

$$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-(y - m(x))^2 / 2\sigma^2) f_X(x), \tag{3.4}$$

where f_X is the marginal density of covariate X.

Define

$$f_{\boldsymbol{\theta}_{m_n}}(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \left[y - m_0(x) - a_n^{-1} \sum_1^{m_n} \theta_j H(m_n(x - x_{n,j})) \right]^2 \right) f_X(x). \tag{3.5}$$

Then, the condition (2.5) is satisfied with

$$|T \circ f_{\boldsymbol{\theta}_{j_0}}(x) - T \circ f_{\boldsymbol{\theta}_{j_1}}(x)| = |H(m_n(x - x_{n,j}))| / a_n.$$

For this model, it is more informative to compute directly the sum of type I and type II errors instead of computing the χ^2-distance defined by (2.4).

Lemma 2. *Suppose that the function $H(\cdot)$ is bounded and squared integrable. Then for testing procedure T_n of the testing problem (2.10), we have*

$$\min_{1\le j\le m_n}\left[P_{\boldsymbol{\theta}_{j0}}(T_n>0)+P_{\boldsymbol{\theta}_{j1}}(T_n\le 0)\right]\ge 2\Phi(-c_H)(1+o(1)), \tag{3.6}$$

provided $n/(m_n a_n^2)\le d$, where $\boldsymbol{\theta}_{j0}$ and $\boldsymbol{\theta}_{j1}$ are defined by (2.3)

$$c_H=(2\sigma)^{-1}\sqrt{d\max f_X(x)\int_{-\infty}^{+\infty}H^2(y)dy}. \tag{3.7}$$

Thus, by Corollary 1 and Lemma 2, we have for regression estimator $\hat{T}_n(x)$ of estimating the kth derivative of the regression function $m^{(k)}(x)$,

$$\begin{aligned}&\inf_{\hat{T}_n(x)}\sup_{f\in\mathcal{F}_n}E_f\int_a^b|\hat{T}_n(x)-m^{(k)}(x)|^p w(x)dx\\ \ge\ &\frac{\Phi(-c_H)}{2^p(b-a)}\int_a^b w(x)dx\int_a^b|H^{(k)}(x-a)|^p dx(m_n^k/a_n)^p(1+o(1)),\end{aligned} \tag{3.8}$$

where $\mathcal{F}_n$ is defined by (2.2) with $f_{\boldsymbol{\theta}_{m_n}}$ (3.5). Specially, with $a_n=m_n^{m+\alpha}$, $\mathcal{F}_n$ is a subset of the constraint

$$\mathcal{F}_{m,\alpha,B}=\{f(x,y):|m^{(m)}(x)-m^{(m)}(y)|\le B|x-y|^\alpha\}, \tag{3.9}$$

provided $H^{(m)}(x)\le B$ and $H^{(m+1)}(x)\le B$. Consequently, with $m_n=(n/d)^{1/(2m+2\alpha+1)}$, the condition of Lemma 1 is satisfied, and the optimal rate of estimating $m^{(k)}(x)$ is $n^{-\frac{m+\alpha-k}{2(m+\alpha)+1}}$ with the constant factor (see (3.8)) specified by

$$\frac{\Phi(-c_H)}{2^p(b-a)}\int_a^b w(x)dx\int_a^b|H^{(k)}(x-a)|^p dx d^{\frac{(m+\alpha-k)p}{2m+2\alpha+1}}. \tag{3.10}$$

Thus, our cubical method can not only give an attainable lower rate, but also can give a constant factor in the lower bound. Perhaps, the constant factors are the first one we know of.

Theorem 3. *The minimax lower bound of estimating a regression function $m(x)$ is*

$$\begin{aligned} &\inf_{\hat{T}_n(x)} \sup_{f \in \mathcal{F}_n} E_f \int_a^b |\hat{T}_n(x) - m(x)|^p w(x) dx \\ &\geq \frac{\Phi(-c_H) w}{2^p (b-a)} \int_a^b |H(x-a)|^p dx (d/n)^{\frac{p(m+\alpha)}{2(m+\alpha)+1}}, \end{aligned} \tag{3.11}$$

where $w = \int_a^b w(x)dx$, and c_H is defined by (3.7).

Remark 3.2. For the normal submodel, the conditional quantile functions of Y given $X = x$ is a constant multiplying $m(x)$. Thus, the lower bounds above is also applicable to the conditional quantile functions.

Remark 3.3. Constructing $m_0(x)$, and $H(x)$ similar to Remark 3.1, the lower rates above is applicable to estimate the conditional mean and conditional quantile functions under additional monotonic constraints.

4 Optimal global rates for deconvolution

In this section, we use our cubical lower bound to find attainable global rates for deconvolution problems. An advantage of our method is that the underlying structure of the problem can be easily seen.

4.1 Introduction

The deconvolution problem arises when direct observation is not possible. The basic model is as follows. We want to estimate the unknown density of a random variable X, but the only data available are observations $Y_1, ..., Y_n$, which are contaminated with independent additive error ε, from the model

$$Y = X + \varepsilon. \tag{4.1}$$

In density function terms, we have realizations $Y_1, \ldots, Y_n$ from the density

$$f_Y(y) = \int f_X(y-x) dF_\varepsilon(x), \tag{4.2}$$

and we want to estimate the density f_X of the random variable X, where F_ε is the cumulative distribution function of the random error ε.

In practice, it may be more interesting to understand how to estimate a whole density function and how well an estimator behaves. A kernel type of estimate is typically used to estimate the unknown density. The global rates of the kernel density estimator have been derived for some special cases (e.g., the error distribution is normal, Cauchy) (Fan (1989a), Zhang (1989)). The question arises whether these rates are the best attainable ones. Based on a Farrell-Stone type of argument, Fan (1989a) and Zhang (1989) prove that the best attainable rate is achieved by a kernel density estimate for supersmooth errors (see Fan (1988a) for a definition), the rates of which are extremely slow ($O((\log n)^{-k})$, for some k depending on the smoothness of the unknown density and the error distribution). Thus, it is impractical to deconvolve an unknown density with supersmooth errors (Fan (1988a)). A more practical question is how well we can deconvolve a whole density when the error distribution is not too smooth: the tail of characteristic function of ε satisfies

$$d_0 \le |\phi_\varepsilon(t)t^\beta| \le d_1 \ \ (\text{as } t \to \infty), \tag{4.3}$$

which is called an ordinary smooth distribution of order β (Fan (1988a)). The examples of distributions satisfying (4.3) are double exponential distributions, gamma distributions, symmetric gamma distributions. However, the approaches of Fan (1989a) and Zhang (1989) fail to answer the above question.

4.2 Lower bounds for Deconvolution

Suppose that we have n i.i.d. observations from the model (4.2) with f_X satisfying the nonparametric constraint $\mathcal{F}_{m,\alpha,B}$ defined by (2.17). Take $f_0(x) = C_r(1+x^2)^{-r}$ $(r > 0.5)$, and a $(k+1)$-time bounded differentiable function $H(x)$ (to be specified in the proof of Theorem 4), $a_n = m_n^{m+\alpha}$. By Remark 2.4, $\mathcal{F}_n \subset \mathcal{F}_{m,\alpha,B}$. By corollary 1, if

$$\max_{1\le j\le m_n} \max_{\boldsymbol{\theta}_{m_n}\in\{0,1\}^{m_n}} \chi^2(f_{\boldsymbol{\theta}_{j_0}} * F_\varepsilon, f_{\boldsymbol{\theta}_{j_1}} * F_\varepsilon) \le c/n, \tag{4.4}$$

then (2.15) holds with $f \equiv f_X$. It will prove by Lemma 3 (Section 5) that under the assumptions of Theorem 4, when $m_n = n^{1/[2(m+\alpha+\beta)+1]}$, the condition (4.4) is satisfied for some $H(x)$ (see the proof of Theorem 4 for detail). By Corollary 1, we have the following lower bound.

Theorem 4. *Suppose that the tail of the characteristic function ϕ_ε of the random variable ε satisfies*

$$|t^{\beta+j}\phi_\varepsilon^{(j)}(t)| \le d_j, j = 0, 1, 2, \ (\text{as } t \to \infty), \tag{4.5}$$

where d_j is a nonnegative constant, and $\phi_\varepsilon^{(j)}$ is the j^{th} derivative of ϕ_ε. Then no estimator can estimate $T \circ f(x) = f_X^{(k)}(x)$, under the constraint that $f_X \in \mathcal{F}_{m,\alpha,B}$ defined by (2.17), faster than the rate $O(n^{-(m+\alpha-k)/(2m+2\alpha+2\beta+1)})$ in the sense that for any $0 \le p < \infty$,

$$\begin{aligned} &\inf_{\hat{T}_n(x)} \sup_{f_X \in \mathcal{F}_{m,\alpha,B}} E_{f_X} \int_a^b |\hat{T}_n(x) - f_X^{(k)}(x)|^p w(x)dx \\ \ge\ & D_p \int_a^b w(x)dx n^{-\frac{p(m+\alpha-k)}{2(m+\alpha+\beta)+1}}, \end{aligned} \tag{4.6}$$

where D_p is a positive constant.

Remark 4.1. Combining with upper bound results (Fan (1989c)), we have demonstrate that the global rates in Theorem 4 are the best attainable ones for the ordinary smooth error distributions under L_p-norm ($1 \le p < \infty$). Specifically, for estimating $f_X^{(k)}(x)$ under the constraint $\mathcal{F}_{m,\alpha,B}$, we have the following rates of convergence ($l = m + \alpha$):

error distributions	$\varepsilon \sim$ Gamma(β)	$\varepsilon \sim$ symmetric Gamma(β)	
		$\beta \ne 2j+1$, (j integer)	$\beta = 2j+1$, (j integer)
optimal global rates	$O(n^{-\frac{l-k}{2(l+\beta)+1}})$	$O(n^{-\frac{l-k}{2(l+\beta)+1}})$	$O(n^{-\frac{l-k}{2(l+\beta)+3}})$

Thus, the optimal global rates for estimating $f_X^{(k)}(x)$ is $O(n^{-\frac{l-k}{2l+5}})$, when error is double exponential.

Remark 4.2. For deconvolving with a supersmooth distribution (e.g. normal, Cauchy), the difficulty of estimating a whole density function is captured by a 1-dimensional subproblem (Fan (1989a)). In contrast, for deconvolving with an ordinary smooth error distribution, a 1-dimensional subproblem is *not* difficult enough to capture the difficulty of the full nonparametric problem of estimating *the whole density.* Our arguments indicate that the difficulty is captured at a growing number m_n-dimensional subproblem. Moreover, our results indicate that the difficulty of deconvolution depends on both smoothness constraints and the smoothness of error distributions.

5 Proofs

Proof of Lemma 1. As H vanishes outside $[0,\ b-a]$, it follows that

$$\begin{aligned}\chi^2(f_{\boldsymbol{\theta}_{j_0}}, f_{\boldsymbol{\theta}_{j_1}}) &= \int_{x_{n,j}}^{x_{n,j+1}} \frac{a_n^{-2} H^2(m_n(t-x_{n,j}))}{f_0(t) + H(m_n(t-x_{n,j}))/a_n} dt \\ &\leq (\min_{x_{n,j} \leq t \leq x_{n,j+1}} f_0(t) - C/a_n)^{-1} \int_{x_{n,j}}^{x_{n,j+1}} a_n^{-2} H^2(m_n(t-x_{n,j}))dt,\end{aligned}$$

where $C = \max_{0 \leq t \leq b-a} |H(t)|$. The conclusion follows.

Proof of Lemma 2. Denote

$$H_j(x) = m_0(x) + \sum_{i \neq j} \theta_j H(m_n(x - x_{n,i})) + jH(m_n(x - x_{n,j}))\ \ (j = 0, 1).$$

Then, the log-likelihood ratio test statistics is

$$\hat{T}_n = -\sum_1^n (Y_i - H_0(X_i))(H_0(X_i) - H_1(X_i))/\sigma^2 - \sum_1^n (H_0(X_i) - H_1(X_i))^2/(2\sigma^2). \quad (5.1)$$

The sum of type I and type II errors of the best testing procedure for the test problem (2.10) is

$$P_{H_0}\{\hat{T}_n > 0\} + P_{H_1}\{\hat{T}_n \leq 0\}, \quad (5.2)$$

where P_{H_0} and P_{H_1} are the probability measures generated under the hypotheses H_0 and H_1, respectively. Note that under the hypothesis H_0, given $X_1, ..., X_n$, the conditional

distribution of $\sigma^2\hat{T}_n$ is

$$N\left(-\sum_1^n (H_0(X_i)-H_1(X_i))^2/2, \sigma^2\sum_1^n (H_0(X_i)-H_1(X_i))^2\right).$$

Similarly, under the hypothesis H_1, given $X_1,...,X_n$, the conditional distribution of $\sigma^2\hat{T}_n$ is

$$N\left(\sum_1^n (H_0(X_i)-H_1(X_i))^2/2, \sigma^2\sum_1^n (H_0(X_i)-H_1(X_i))^2\right).$$

Thus, by (5.1),

$$\begin{aligned} & P_{H_0}\{\hat{T}_n > 0 | X_1,...,X_n\} + P_{H_1}\{\hat{T}_n < 0 | X_1,...,X_n\}, \\ = \; & 2\Phi\left(-(2\sigma)^{-1}\sqrt{\sum_1^n (H_0(X_i)-H_1(X_i))^2}\right), \end{aligned} \tag{5.3}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Note that the expected value of the above quantity is

$$\begin{aligned} E\sum_1^n (H_0(X_i)-H_1(X_i))^2 &= \frac{n}{a_n^2}\int_{-\infty}^{+\infty} H^2(m_n(x-x_{n,j}))f_X(x)dx \\ &= \frac{n}{m_n a_n^2}\int_{-\infty}^{+\infty} H^2(y)f_X(x_{n,j}+y/m_n)dy \\ &\le n/(m_n a_n^2)\max f_X(x)\int_{-\infty}^{+\infty} H^2(y)dy, \end{aligned} \tag{5.4}$$

and the variance of the above random quantity is

$$\begin{aligned} \mathrm{var}(\sum_1^n (H_0(X_i)-H_1(X_i))^2) &= n\mathrm{var}(H^2(m_n(X_1-X_{n,j}))/a_n^2) \\ &\le \frac{n}{m_n a_n^4}\max f_X(x)\int_{-\infty}^{+\infty} H^4(y)dy \\ &\to 0, \end{aligned} \tag{5.5}$$

where f_X is the marginal density of the random variable X_1. Consequently, combining the last three displays (5.3), (5.4), and (5.5), by Lebesque's dominated convergence theorem, we have

$$\max_{1\le j\le n}\left[P_{H_0}\{\hat{T}_n > 0\} + P_{H_1}\{\hat{T}_n < 0\}\right] \ge 2\Phi(-c_H)(1+o(1)).$$

The conclusion follows.

Proof of Theorem 4. We only prove the case that $m + \alpha > 1$; the other case follow the same idea expect more detailed technical arguments are required. We first construct the class of densities $\mathcal{F}_n$ defined by (2.2). Take a real function $H(\cdot)$ satisfying the following conditions:

1. $H(x) \not\equiv 0$, having all order bounded derivatives,

2. $\int_{-\infty}^{+\infty} H(x)dx = 0$,

3. $H(x) = O(x^{-4}), \text{as } x \to \infty$,

4. $\phi_H(t) = 0$, when $|t| \notin [1, 2]$, where ϕ_H is the Fourier transform of H.

It is easy to argue (Fan (1988a)) that such a function does exist.

Let $l = m + \alpha$ and $f_0(t) = C_r(1 + x^2)^{-r}(1.5 > r > 0.5)$ be a density function. Then, it is easy to see that with $a_n = m_n^l$, the class of densities defined by (2.1) is a subset of $\mathcal{F}_{m,\alpha,B}$, when n is large enough. Thus, by Corollary 1,

$$\inf_{\hat{T}_n(x)} \sup_{f_X \in \mathcal{F}_{m,\alpha,B}} E_{f_X} \int_a^b |\hat{T}_n(x) - f_X^{(k)}(x)|^p w(x)dx > c_0 m_n^{-(l-k)p}, \tag{5.6}$$

for some constant $c_0 > 0$. To complete the proof, we need only to check that the condition (4.4) is satisfied when $m_n = c_1 n^{-1/(2l+2\beta+1)}$ for some $c_1 > 0$, which will be proved in Lemma 3. The basic ideas of proving Lemma 3 are that the χ^2–distances for pairs of densities in (4.4) are equivalent to L_2-distances, and then use Parseval's identity to conclude the result.

Lemma 3. *Under the assumptions of Theorem 4,*

$$\max_{1 \le j \le m_n} \max_{\boldsymbol{\theta}_{m_n} \in \{0,1\}^{m_n}} \chi^2(f_{\boldsymbol{\theta}_{j_0}} * F_\varepsilon, f_{\boldsymbol{\theta}_{j_1}} * F_\varepsilon) = O(m_n^{-(2l+2\beta+1)}). \tag{5.7}$$

Proof. By changing variables,

$$\chi^2(f_{\boldsymbol{\theta}_{j_0}} * F_\varepsilon, f_{\boldsymbol{\theta}_{j_1}} * F_\varepsilon)$$

$$= m_n^{-(2l+1)} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} H(x-y) dF_\varepsilon(y/m_n) \right]^2 / g_n(x/m_n + x_{n,j}) dx, \qquad (5.8)$$

where $g_n(x) = [f_0(x) + m_n^{-l} \sum_{i \neq j} \theta_i H(m_n(x - x_{n,j}))] * F_\varepsilon$, and $l = m + \alpha$.

Note that the minimum value of $f_0(x)$ on any bounded intervals is positive, and the function H is bounded. Hence, it follows that when n is large

$$\inf_{1 \leq j \leq m_n} \inf_{|x| \leq 1} g_n(x/m_n + x_{n,j}) > c_2 > 0. \qquad (5.9)$$

Let I_1 be the integration of (5.8) over $x \in [-1, 1]$, and I_2 be the integration of (5.8) over $|x| > 1$. Then, by Parseval's identity and (5.9), we have

$$\begin{aligned} I_1 &\leq c_2^{-1} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} H(x-y) dF_\varepsilon(y/m_n) \right]^2 dx \\ &= c_2^{-1} \int_{-\infty}^{+\infty} |\phi_H(t)|^2 |\phi_\varepsilon(m_n t)|^2 dt \\ &= 2c_2^{-1} \int_1^2 |\phi_H(t)|^2 |\phi_\varepsilon(m_n t)|^2 dt, \end{aligned} \qquad (5.10)$$

as $|\phi_H(t)|$ is symmetric, having a bounded support. By the assumption (4.5), we conclude that

$$I_1 = O(m_n^{-2\beta}).$$

Now, let's evaluate I_2; the Lemma follows if we show that $I_2 = O(m_n^{-2\beta})$. Denote

$$\phi_n(t) = \frac{d^2[\phi_H(t)\phi_\varepsilon(m_n t)]}{dt^2}.$$

It follows by the Fourier inversion and the integration by parts that

$$\begin{aligned} \int_{-\infty}^{+\infty} H(x-y) dF_\varepsilon(y/m_n) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(-itx) \phi_H(t) \phi_\varepsilon(m_n t) dt \\ &= -\frac{1}{2\pi x^2} \int_{1 \leq |t| \leq 2} \exp(-itx) \phi_n(t) dt, \end{aligned} \qquad (5.11)$$

because $\phi_H(t)$ vanishes when $|t| \notin [1, 2]$. By the assumption (4.5) and (5.11), we have that when n is large enough,

$$m_n^\beta \left| \int_{-\infty}^{+\infty} H(x-y) dF_\varepsilon(y/m_n) \right| \leq c_3/x^2,$$

for some constant $c_3 > 0$. Thus (see (5.8)),

$$I_2 \leq (c_3 m_n^{-\beta})^2 \int_{|x|\geq 1} \frac{1}{x^4 g_n(x/m_n + x_{n,j})} dx. \tag{5.12}$$

As the sequence of $x_{n,j}$ is bounded, by the property 3 of the function $H(\cdot)$, there exists an M such that when $|x| > M$,

$$f_0(x) + m_n^{-l} \sum_{i \neq j} \theta_i H(m_n(x - x_{n,i})) \geq C_r(1 + x^2)^{-r} - c_4 m_n^{-l+1} (m_n x)^{-4},$$

for some $c_4 > 0$. When $|x| < M$

$$f_0(x) + m_n^{-l} \sum_{i \neq j} \theta_i H(m_n(x - x_{n,i})) \geq C_r(1 + x^2)^{-r} - m_n^{-l+1} C,$$

where C is an upper bound of the function H. Thus, when n is large enough,

$$f_0(x) + m_n^{-l} \sum_{i \neq j} \theta_i H(m_n(x - x_{n,i})) \geq f_0(x)/2.$$

Hence,

$$g_n(x) \geq 1/2 \int_{-\infty}^{+\infty} C_r(1 + (x - y)^2)^{-r} dF_\epsilon(y) \stackrel{\text{def}}{=} g(x).$$

Note that $g(x) \geq \min\{c_5, c_6 x^{-2r}\}$, for some $c_5, c_6 > 0$, as when x is large, the convolution above is of order x^{-2r}, and when x is bounded, $g(x)$ is bounded away from 0. By (5.12), it follows that

$$I_2 \leq 2(c_3 m_n^{-\beta})^2 \int_{|x|\geq 1} \frac{dx}{x^4 g(x/m_n + x_{n,j})} = O(m_n^{-\beta}),$$

as $4 - 2r > 1$. The conclusion follows.

References

[1] Bickel, P. J. and Ritov, Y. (1988). Estimating integrated squared density derivatives. *Tech. Report 146*, Department of Statistics, University of California, Berkeley.

[2] Birgé, L. (1987). Estimating a density under the order restrictions: Nonasymptotic minimax risk. *Ann. Statist.*, **15**, 995-1012.

[3] Donoho, D. L. and Liu, R. C. (1987a). Hardest 1-dimensional subproblems. *Tech. Report 105*, Dept. of Statist., University of California, Berkeley.

[4] Donoho, D. L. and Liu, R. C. (1987b). Geometrizing rates of convergence II. *Tech. Report 120*, Dept. of Statist., University of California, Berkeley.

[5] Donoho, D. L. and Liu, R. C. (1988). Geometrizing rate of convergence III. *Tech. Report 138*, Dept. of Statist., University of California, Berkeley.

[6] Donoho, D. L., MacGibbon, B., and Liu, R.C. (1987). Minimax risk for hyperrectangles. *Tech. Report 123*, Dept. of Statist., University of California, Berkeley.

[7] Efroimovich, S.Y. and Pinsker, M.S. (1982). Estimation of square-integrable probability density of a random variable. *Problems of Information Transmission*, 1775-189.

[8] Fan, J. (1988a). On the optimal rates of convergence for nonparametric deconvolution problem. *Tech. Report 157*, Dept. of Statistics, University of California, Berkeley.

[9] Fan, J. (1989a). *Contributions to the Estimation of Nonregular Functionals.* Dissertation, Dept. of Statist., Univ. of California, Berkeley.

[10] Fan, J. (1989b). On the estimation of quadratic functionals. *Institute of Statistics Mimeo Series #2005*, Univ. of North Carolina, Chapel Hill.

[11] Fan, J. (1989c). Global behavior of deconvolution kernel estimates. *Preprint*

[12] Farrell, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.*, **43**, 170-180.

[13] Ibragimov, I.A., Nemirovskii, A.S. and Kha'sminskii, R.Z. (1987). Some problems on nonparametric estimation in Gaussian white noise. *Theory Prob. Appl.*, **31**, 391-406.

[14] Khas'minskii, R.Z. (1978). A lower bound on the risks of nonparametric estimates densities in the uniform metric. *Theory. Probab. Appl.*, **23**, 794-798.

[15] Khas'minskii, R.Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. In *Contribution to Statistics* (J. Hajek memorial volume), Academia, Prague, 91-97.

[16] Kiefer, J. (1981). Optimum rates for non-parametric density and regression estimates, under order restrictions. *Statistics and Probability: Essays in Honor of C. R. Rao* (Kallianpur, *et al.* eds.), 419–427.

[17] Le Cam, L. (1985). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York-Berlin-Hei

[18] Neusbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.*, **13**, 984-997,

[19] Stone, C. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**, 1348-1360.

[20] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040-1053.

[21] Zhang, C. H. (1989). Fourier methods for estimating mixing densities and distributions. *Manuscript* .