Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension

Jérôme Dedecker, Bertrand Michel

Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension

Jérôme Dedecker[(1)] and Bertrand Michel[(2)]

February 25, 2013

(1) Laboratoire MAP5 UMR CNRS 8145, Université Paris Descartes, Sorbonne Paris Cité,
(2) Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie - Paris 6

Abstract

The subject of this paper is the estimation of a probability measure on $\mathbb{R}^d$ from data observed with an additive noise, under the Wasserstein metric of order p (with $p \geq 1$). We assume that the distribution of the errors is known and belongs to a class of supersmooth distributions, and we give optimal rates of convergence for the Wasserstein metric of order p. In particular, we show how to use the existing lower bounds for the estimation of the cumulative distribution function in dimension one to find lower bounds for the Wasserstein deconvolution in any dimension.

Keywords: Deconvolution, Wasserstein metrics, supersmooth distributions, minimax rates.
AMS MSC 2010: 62G05, 62C20.

1 Introduction

We observe n random vectors Y_i in $\mathbb{R}^d$ sampled according to the convolution model:

$$Y_i = X_i + \varepsilon_i \tag{1}$$

where the random vectors $X_i = (X_{i,1}, \ldots, X_{i,j}, \ldots, X_{i,d})'$ are i.i.d. and distributed according to an unknown probability measure μ. The random vectors $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,j}, \ldots, \varepsilon_{i,d})'$ are i.i.d. and distributed according to a known probability measure μ_ε. The distribution of the observations Y_i on $\mathbb{R}^d$ is then the convolution $\mu \star \mu_\varepsilon$. Here, we shall assume that there exists an invertible matrix A such that the coordinates of the vector $A\varepsilon_1$ are independent (that is: the image measure of μ_ε by A is the product of its marginals).

This paper is about minimax optimal rates of convergence for estimating the measure μ under Wasserstein metrics. For $p \geq 1$, the Wasserstein distance W_p between μ and μ' is defined by:

$$W_p(\mu, \mu') = \inf_{\pi \in \Pi(\mu, \mu')} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \pi(dx, dy) \right)^{\frac{1}{p}},$$

where $\Pi(\mu, \mu')$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and μ' and p is a real number in $[1, \infty($ (see [RR98] or [Vil08]). The norm $\|.\|$ is the euclidean norm in $\mathbb{R}^d$ corresponding to the inner product $< \cdot, \cdot >$.

The Wasserstein deconvolution problem is interesting in itself since W_p are natural distances for comparing measures (without assuming densities for instance). Moreover, it is also related to recent results in geometric inference. Indeed, in 2011, [CCSM11] have defined a distance function to measures to answer geometric inference problems in a probabilistic setting. According to their result, the topological properties of a shape can be recovered by using the distance to a known measure $\tilde{\mu}$, if $\tilde{\mu}$ is close enough to a measure μ concentrated on this shape with respect to the Wasserstein distance W_2. This fact motivates the study of the Wasserstein deconvolution problem, since in practice the data can be observed with noise.

In the paper [CCDM11], the authors consider a slight modification of the classical kernel deconvolution estimator, and they provide some upper bounds for the rate of convergence of this estimator for the W_2 distance, for several noise distributions. Nevertheless the question of optimal rates of convergence in the minimax sense was left open in this previous work. The main contribution of the present paper is to find optimal rates of convergence for a class of supersmooth distributions, for any dimension under any Wasserstein metric W_p. In particular we prove that the deconvolution estimator of μ under the W_2 metric introduced in [CCDM11] is minimax optimal for a class of supersmooth distributions.

The rates of convergence for deconvolving a density have been deeply studied for other metrics. Minimax rates in the univariate context can be found for instance in [Fan91b, BT08a, BT08b] and in the recent monograph [Mei09]. The multivariate problem has also been investigated in [Tan94, CL11]. All these contributions concern pointwise convergences or $\mathbb{L}^2$ convergences; rates of convergence for the Wasserstein metrics have been studied only by [CCDM11]. In Section 2 of the present paper, we shall see that, in the supersmooth case, lower bounds for the Wasserstein deconvolution problem in any dimension can be deduced from lower bounds for the deconvolution of the cumulative distribution function (c.d.f.) in dimension one.

Another interesting related work is [GPPVW12]. In this recent paper, the authors find lower and upper bounds for the risk of estimating a manifold in Hausdorff distance under several noise assumptions. They consider in particular the additive noise model (1) with a standard multivariate Gaussian noise.

Before giving the main result of our paper, we need some notations. Let ν be a measure on $\mathbb{R}^d$ with density g and let m be another measure on $\mathbb{R}^d$. In the following we shall denote by $m \star g$ the density of $m \star \nu$, that is

$$m \star g(x) = \int_{\mathbb{R}^d} g(x-z)m(dz)\,.$$

We also denote by μ^* (respectively f^*) the Fourier transform of the probability measure μ (respectively of the integrable function f), that is:

$$\mu^*(x) = \int_{\mathbb{R}^d} e^{i<t,x>}\mu(dt) \quad \text{and} \quad f^*(x) = \int_{\mathbb{R}^d} e^{i<t,x>}f(t)dt\,.$$

For $M > 0$ and $p \geq 1$, let $\mathcal{D}_A(M,p)$ be the set of measures μ on $\mathbb{R}^d$ for which

$$\sup_{1\leq j\leq d} \mathbb{E}_\mu\Big((1+|(AX_1)_j|^{2p+2}) \prod_{1\leq \ell\leq d,\, \ell\neq j} (1+(AX_1)_\ell^2)\Big) \leq M < \infty.$$

Let us give the main result of our paper when ε_1 is a non degenerate Gaussian random vector (by non degenerate, we mean that its covariance matrix is not equal to zero).

Theorem 1. *Assume that we observe $Y_1,\dots,Y_n$ in the multivariate convolution model (1), where ε_1 is a non degenerate Gaussian random vector. Let A be an invertible matrix such that the coordinates of $A\varepsilon_1$ are independent. Let $M > 0$ and $p \geq 1$. Then*

1. *There exists a constant $C > 0$ such that for any estimator $\tilde{\mu}_n$ of the measure μ:*

$$\liminf_{n\to\infty} (\log n)^{p/2} \sup_{\mu\in\mathcal{D}_A(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_p^p(\tilde{\mu}_n,\mu)) \geq C.$$

2. *One can build an estimator $\hat{\mu}_n$ of μ such that*

$$\sup_{n\geq 1} \sup_{\mu\in\mathcal{D}_A(M,p)} (\log n)^{p/2} \, \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n,\mu)) \leq K\,,$$

for some positive constant K.

Note that in Theorem 1 the random vector ε_1 may have all its coordinates, excepts one, equal to zero almost surely. In other words, a Gaussian noise in one direction leads to the same rate of convergence as an isotropic Gaussian noise.

The paper is organized as follows. The proof of the lower bound is given in Section 2. In Section 3 we then give the corresponding upper bound in the same context by generalizing the results of [CCDM11] for all $p \geq 1$. We finally discuss the W_p deconvolution problem for ordinary smooth case in Section 4. Some additional technical results are given in Appendix.

2 Lower bounds

2.1 Main result

The following theorem is the main result of this section. It gives a lower bound on the rates of convergence of measure estimators in the supersmooth case for any dimension and under any metric W_p.

Theorem 2. *Let $M > 0$, $p \geq 1$. Assume that we observe $Y_1, \dots, Y_n$ in the multivariate convolution model (1). Assume that there exists $j_0 \in \{1, \dots, d\}$ such that the coordinate $(A\varepsilon_1)_{j_0}$ has a density g with respect to the Lebesgue measure satisfying for all $w \in \mathbb{R}$:*

$$|g^*(w)|(1+|w|)^{-\tilde{\beta}} \exp(|w|^\beta/\gamma_1) \leq c_1 \tag{2}$$

for some $\beta > 0$ and some $\tilde{\beta} \in \mathbb{R}$. Also assume that there exist some constants $\kappa_1 \in (0,1)$ and $\kappa_2 > 1$ such that

$$P(|(A\varepsilon_1)_{j_0} - t| \leq |t|^{\kappa_1}) = O(|t|^{-\kappa_2}) \quad \text{as } |t| \to \infty \tag{3}$$

and

$$\max\left(p+2\,,\, \frac{\kappa_2}{2\kappa_1} + \frac{1}{2}\right) < \kappa_2. \tag{4}$$

Then there exists a constant $C > 0$ such that for all estimator $\tilde{\mu}_n$ of the measure μ:

$$\liminf_{n\to\infty} (\log n)^{p/\beta} \sup_{\mu\in\mathcal{D}_A(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_p^p(\tilde{\mu}_n,\mu) \geq C.$$

The assumption about the random variable $(A\varepsilon_1)_{j_0}$ means that the noise is supersmooth in at least one direction. Indeed, as shown in Section 2.1.1, the lower bound for the multivariate problem can be deduced from the lower bound for the $\mathbb{L}^1$ estimation of the c.d.f. of $(A\varepsilon_1)_{j_0}$. If the distribution of the noise is supersmooth in several directions then one may choose the direction with the greatest coefficient β.

The assumption (3) is classical in the deconvolution setting, see for instance [Fan91b, Fan92]. The technical assumption (4) summarizes the conditions on p and κ_2. The condition $\frac{\kappa_2}{2\kappa_1}+\frac{1}{2} < \kappa_2$ is also required in [Fan91b] and [Fan92]. The additional condition $p+2 < \kappa_2$ is a consequence of the moment assumption on μ.

If g has a polynomial decay rate at infinity, we can state the following lemma:

Lemma 1. *Assume that the density g of $(A\varepsilon_1)_{j_0}$ satisfies $|g(t)| = O(|t|^{-a})$ as t tends to infinity, for some $a \geq p+3$. Then one can find $\kappa_1 \in (0,1)$ and $\kappa_2 > 1$ such that Conditions (3) and (4) are satisfied.*

Proof. For $t > 0$ large enough

$$
\begin{aligned}
P(|(A\varepsilon_1)_{j_0} - t| \leq t^{\kappa_1}) &= \int_{t-t^{\kappa_1}}^{t+t^{\kappa_1}} g(u)du \\
&= O\left((t-t^{\kappa_1})^{-p-2} - (t+t^{\kappa_1})^{-p-2}\right) = O\left(t^{-(p+3-\kappa_1)}\right).
\end{aligned}
$$

We choose $\kappa_1 = 3/5$ and we take $\kappa_2 = p+3-\kappa_1 = p+12/5$. Note that $\kappa_1 \in (0,1)$, $\kappa_2 > 1$ and that (3) is satisfied. Moreover, $\kappa_2 > p+2$ and $\kappa_2 - \kappa_2/(2\kappa_1) - \frac{1}{2} \geq 1/3 > 0$ and thus Condition (4) is also satisfied.

□

2.1.1 Wasserstein deconvolution and c.d.f. deconvolution

It is well known that the Wasserstein distance W_1 between two measures μ and μ' on $\mathbb{R}$ can be computed using the cumulative distribution functions: Let μ and μ' be two probability measures on $\mathbb{R}$, then

$$
W_1(\mu,\mu') = \int_{\mathbb{R}} |F_\mu(x) - F_{\mu'}(x)|\, dx.
$$

According to this property, lower bounds on the rates of convergence for estimating μ in the one dimensional convolution model (1) for the metric W_1 can be directly deduced from lower bounds on the rates of convergence for the estimation of the c.d.f. of μ using the integrated risk $\mathcal{R}(\hat{F}) := \int_{\mathbb{R}} |F_\mu(t) - \hat{F}(t)|dt$. This last problem has been less studied than pointwise rates in the deconvolution context but some results can be found in the literature. For instance [Fan92] gives the optimal rate of convergence in the supersmooth case for an integrated (weighted) L_p risk under similar smoothness conditions as for the pointwise case (studied in [Fan91b]). The *cubical method* followed in [Fan92] to compute the integrated lower bound is also detailed in [Fan93]. It is based on a multiple hypothesis strategy, see [Tsy09] for other examples of using multiple hypothesis schema for computing lower bounds for integrated risks.

For $M > 0$ and $p \geq 1$, we consider the set $\mathcal{C}_A(M,p)$ of the measures μ in $\mathcal{D}_A(M,p)$ for which the coordinates of AX_1 are independent. Thus, for $\mu \in \mathcal{C}_A(M,p)$:

$$
\sup_{1\leq j\leq d} \left(\mathbb{E}_\mu(1+|(AX_1)_j|^{2p+2}) \prod_{1\leq \ell\leq d,\, \ell\neq j} \mathbb{E}_\mu(1+(AX_1)_\ell^2)\right) \leq M < \infty.
$$

Moreover we simply use the notation $\mathcal{C}(M,p)$ if $A = \mathrm{I}_d$.

The following theorem gives lower bounds for $W_1(\tilde{\mu}_n,\mu)$ in the d-dimensional case, which are derived from lower bounds on the rates of convergence of c.d.f. estimators in $\mathbb{R}$.

Theorem 3. *Under the same assumptions as in Theorem 2, there exists $C > 0$ such that for all estimator $\tilde{\mu}_n$ of the measure μ:*

$$
\liminf_{n\to\infty} (\log n)^{1/\beta} \sup_{\mu\in\mathcal{C}_A(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1(\tilde{\mu}_n,\mu) \geq C.
$$

Theorem 2 is a corollary of Theorem 3 because

1. $\mathcal{C}_A(M,p)$ is a subset of $\mathcal{D}_A(M,p)$.
2. For any $p \geq 1$, $\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1^p(\tilde{\mu}_n,\mu) \geq \left(\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1(\tilde{\mu}_n,\mu)\right)^p$.
3. W_1 is the smallest among all the Wasserstein distances: for any $p \geq 1$ and any measures μ and μ' on $\mathbb{R}^d$: $W_p(\mu,\mu') \geq W_1(\mu,\mu')$.

2.2 Proof of Theorem 3

Since the works of Le Cam, it is well known that rates of convergence of estimators on some probably measure space $\mathcal{P}$ can be lower bounded by introducing some convenient finite subset of $\mathcal{P}$ whose elements are close enough for the total variation distance or for the Hellinger distance. In the deconvolution setting, χ^2 distance are preferable to these last metrics. Here, the following definition of the χ^2 distance will be sufficient: for two positive densities h_1 and h_2 with respect to the Lebesgue measure on $\mathbb{R}^d$, the χ^2 distance between h_1 and h_2 is defined by

$$\chi^2(h_1, h_2) = \int_{\mathbb{R}^d} \frac{\{(h_1(x) - h_2(x)\}^2}{h_1(x)} dx.$$

The main arguments for proving Theorem 3 comes from [Fan91b, Fan92, Fan93]. However some modifications are necessary to compute the lower bounds under the moment assumption $\mathcal{C}_A(M,p)$. Furthermore, we note that Theorem 1 in [Fan93] cannot be directly applied in this multivariate context.

Without loss of generality, we take $j_0 = 1$. We shall first prove Theorem 3 in the case where ε_1 has independent coordinates.

2.2.1 Errors with independent coordinates

In this section, we observe $Y_1, \ldots, Y_n$ in the multivariate convolution model (1) and we assume that the random variables $(\varepsilon_{1,j})_{1 \le j \le d}$ are independent. This means that $A = \mathrm{I}_d$ and that ε_1 has the distribution $\mu_\varepsilon = \mu_{\varepsilon,1} \otimes \mu_{\varepsilon,2} \otimes \cdots \otimes \mu_{\varepsilon,d}$.

Definition of a finite family in $\mathcal{C}(M,p)$. Let us introduce a finite class of probability measures in $\mathcal{C}(M,p)$ which are absolutely continuous with respect to the Lebesgue measure λ_d. First, we define some densities

$$f_{0,r}(t) := C_r(1 + t^2)^{-r} \tag{5}$$

with some $r > 0$ such that

$$\max\left(p + \frac{3}{2}, \frac{1}{2}\frac{\kappa_2}{\kappa_1}\right) < r < \kappa_2 - \frac{1}{2}. \tag{6}$$

Note that this is possible according to (4). Moreover, $f_{0,r}$ has a finite $(2p+2)$-th moment.

Next, let b_n be the sequence

$$b_n := \left[\left(\frac{1}{\eta}\log n\right)^{1/\beta}\right] \vee 1, \tag{7}$$

where $[\cdot]$ is the integer part, and $\eta = \left(1 - \frac{2r}{2\kappa_2 - 1}\right)/\gamma$. Note that b_n is correctly defined in this way since $\kappa_2 - \frac{1}{2} > r$. For any $\theta \in \{0,1\}^{b_n}$, let

$$f_\theta(t) = f_{0,r}(t) + A\sum_{s=1}^{b_n} \theta_s H\left(b_n(t - t_{s,n})\right), \quad t \in \mathbb{R}, \tag{8}$$

where A is a positive constant and $t_{s,n} = (s-1)/b_n$. The function H is a bounded function whose integral on the line is 0. Moreover, we may choose a function H such that (see for instance [Fan91b] or [Fan93]):

(A1) $\int_{-\infty}^{+\infty} H\,dt = 0$ and $\int_0^1 |H^{(-1)}|\,dt > 0$,

(A2) $|H(t)| \le c(1+t^2)^{-r}$,

(A3) $H^*(z) = 0$ outside $[1,2]$

where $H^{(-1)}(t) := \int_{-\infty}^t H(u)\,du$ is a primitive of H.

Using (A2) and Lemma 3 of Appendix A, we choose $A > 0$ small enough in such a way that f_θ is a density on $\mathbb{R}$. Note that by replacing H by H/A in the following, we finally can take $A = 1$ in (8). Using (A2) and Lemma 3 again, we get that for M large enough, for all $\theta \in \{0,1\}^{b_n}$:

$$\int_{\mathbb{R}} \left(1 + t^2 \vee t^{2p+2}\right) f_\theta(t)\,dt \le M^{1/d}. \tag{9}$$

We finally use these univariate densities f_θ to define a finite family of probability measures on $\mathbb{R}^d$ which is included in $\mathcal{C}(M,p)$. For $\theta \in \{0,1\}^{b_n}$, let us define the probability measure on $\mathbb{R}^d$:

$$\mu_\theta := (f_\theta \cdot d\lambda) \otimes (f_{0,r} \cdot d\lambda) \otimes \cdots \otimes (f_{0,r} \cdot d\lambda)\,. \tag{10}$$

For any $j \in \{1,\dots,d\}$, according to (9) :

$$\left(\mathbb{E}_{\mu_\theta}(1 + X_{1,j}^{2p+2}) \prod_{2\le \ell \le d} \mathbb{E}_{\mu_\theta}(1 + X_{1,\ell}^2)\right) \le M$$

and thus $\mu_\theta \in \mathcal{C}(M,p)$.

Lower bound. Let $\tilde\mu_n$ be an estimator of μ and let $(\tilde\mu_n)_1$ be the marginal distribution of $\tilde\mu_n$ on the first coordinate (conditionally to the sample $Y_1,\dots,Y_n$). According to Lemma 6 of Appendix B:

$$\begin{aligned}
\sup_{\mu\in\mathcal{C}(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1(\mu, \tilde\mu_n) &\ge \sup_{\theta\in\{0,1\}^n} \mathbb{E}_{(\mu_\theta\star\mu_\varepsilon)^{\otimes n}} W_1(\mu_\theta,\, \tilde\mu_n) \\
&\ge \sup_{\theta\in\{0,1\}^n} \mathbb{E}_{(\mu_\theta\star\mu_\varepsilon)^{\otimes n}} W_1(f_\theta\cdot d\lambda\,,\, (\tilde\mu_n)_1) \\
&\ge \inf_{\hat f_n} \sup_{\theta\in\{0,1\}^n} \mathbb{E}_{(\mu_\theta\star\mu_\varepsilon)^{\otimes n}} W_1\left(f_\theta\cdot d\lambda\,,\, \hat f_n\right)
\end{aligned}$$

where the infimum of the last line is taken over all the probability measure estimators of $f_\theta \cdot d\lambda$.

Following [Fan93] (see also the proof of Theorem 2.14 in [Mei09]), we now introduce a random vector $\tilde\theta$ whose components $\tilde\theta_s$ are i.i.d. Bernoulli random variables $\tilde\theta_1,\dots,\tilde\theta_{b_n}$ such that $P(\tilde\theta_s = 1) = \frac12$. The density $f_{\tilde\theta}$ is thus a random density taking its values in the set of densities defined by (8). Let $\mathbb{E}$ be the expectation according to the law of $\tilde\theta$. For any probability estimator $\hat f_n$:

$$\begin{aligned}
\sup_{\mu\in\mathcal{C}(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1(\mu, \tilde\mu_n) &\ge \mathbb{E}\,\mathbb{E}_{(\mu_{\tilde\theta}\star\mu_\varepsilon)^{\otimes n}} \left[W_1\left(f_{\tilde\theta}\cdot d\lambda\,,\, \hat f_n\right)\right] \\
&\ge \int_{\mathbb{R}} \mathbb{E}\,\mathbb{E}_{(\mu_{\tilde\theta}\star\mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde\theta}(t) - \hat F_n(t)|\right)\,dt
\end{aligned} \tag{11}$$

where $\hat F$ and F_θ are the c.d.f. of the distributions $\hat f_n$ and $f_\theta \cdot d\lambda$. For $\theta \in \{0,1\}^{b_n}$ and $s \in \{1,\dots,b_n\}$, let us define

$$f_{\theta,s,0} := f_{(\theta_1,\dots,\theta_{s-1},0,\theta_{s+1},\dots,\theta_{b_n})} \quad \text{and } f_{\theta,s,1} := f_{(\theta_1,\dots,\theta_{s-1},1,\theta_{s+1},\dots,\theta_{b_n})}$$

and the corresponding probability measures $\mu_{\theta,s,0}$ and $\mu_{\theta,s,1}$ on $\mathbb{R}^d$ defined by (10) for $f_\theta = f_{\theta,s,0}$ or $f_{\theta,s,1}$. Let $\bar h_{\theta,s,0}$ and $\bar h_{\theta,s,1}$ be the densities of $\mu_{\theta,s,0}\star\mu_\varepsilon$ and $\mu_{\theta,s,1}\star\mu_\varepsilon$ for the Lebesgue measure on $\mathbb{R}^d$. Since the margins of μ_θ and μ_ε are independent, for any $y_i = (y_{i,1}, \dots, y_{i,j}, \dots, y_{i,d}) \in \mathbb{R}^d$ ($u = 0$ or 1), we have:

$$\bar{h}_{\theta,s,u}(y_i) = h_{\theta,s,u}(y_{i,1}) \prod_{j=2,\dots,d} f_{0,r} \star \mu_{\varepsilon,j}(y_{i,j}) \tag{12}$$

where $h_{\theta,s,u} = f_{\theta,s,u} \star g$. Let $F_{\theta,s,0}$ and $F_{\theta,s,1}$ be the c.d.f of $f_{\theta,s,0}$ and $f_{\theta,s,1}$. For $t \in [t_{s,n}, t_{s+1,n}]$ where s in $\{1, \dots, b_n\}$, by conditioning by $\tilde{\theta}_s$, we find that

$$\mathbb{E}\, \mathbb{E}_{(\mu_{\tilde{\theta}} \star \mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde{\theta}}(t) - \hat{F}_n(t)| \right) = \frac{1}{2} \mathbb{E} \left[\mathbb{E}_{\bar{h}^{\otimes n}_{\tilde{\theta},s,0}} \left(|F_{\tilde{\theta},s,0}(t) - \hat{F}_n(t)| \right) + \mathbb{E}_{\bar{h}^{\otimes n}_{\tilde{\theta},s,1}} \left(|F_{\tilde{\theta},s,1}(t) - \hat{F}_n(t)| \right) \right].$$

Hence

$$\mathbb{E}\, \mathbb{E}_{(\mu_{\tilde{\theta}} \star \mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde{\theta}}(t) - \hat{F}_n(t)| \right) \geq \frac{1}{2} \mathbb{E} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} \left\{ |F_{\tilde{\theta},s,0}(t) - \hat{F}_n(t)| + |F_{\tilde{\theta},s,1}(t) - \hat{F}_n(t)| \right\} \min \left(\prod_{i=1}^n \bar{h}_{\tilde{\theta},s,0}(y_i), \prod_{i=1}^n \bar{h}_{\tilde{\theta},s,1}(y_i) \right) dy_1 \dots dy_n,$$

and consequently, according to (12),

$$\mathbb{E}\, \mathbb{E}_{(\mu_{\tilde{\theta}} \star \mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde{\theta}}(t) - \hat{F}_n(t)| \right) \geq \frac{1}{2} \mathbb{E} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} |F_{\tilde{\theta},s,0}(t) - F_{\tilde{\theta},s,1}(t)| \min \left(\prod_{i=1}^n h_{\tilde{\theta},s,0}(y_{i,1}), \prod_{i=1}^n h_{\tilde{\theta},s,1}(y_{i,1}) \right) \left\{ \prod_{j=2}^d f_{0,r} \star \mu_{\varepsilon,j}(y_{i,j}) \right\} dy_1 \dots dy_n.$$

By using Fubini, it follows that

$$\mathbb{E}\, \mathbb{E}_{(\mu_{\tilde{\theta}} \star \mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde{\theta}}(t) - \hat{F}_n(t)| \right) \geq \frac{1}{2} \mathbb{E} \int_{\mathbb{R}} \dots \int_{\mathbb{R}} |F_{\tilde{\theta},s,0}(t) - F_{\tilde{\theta},s,1}(t)| \min \left(\prod_{i=1}^n h_{\tilde{\theta},s,0}(y_{i,1}), \prod_{i=1}^n h_{\tilde{\theta},s,1}(y_{i,1}) \right) dy_{1,1} \dots dy_{n,1}.$$

Note that for any $\theta \in \{0,1\}^{b_n}$, $|F_{\tilde{\theta},s,0}(t) - F_{\tilde{\theta},s,1}(t)| = b_n^{-1} \left| H^{(-1)} \left(b_n (t - t_{s,n}) \right) \right|$, thus

$$\mathbb{E}\, \mathbb{E}_{(\mu_{\tilde{\theta}} \star \mu_\varepsilon)^{\otimes n}} \left(|F_{\tilde{\theta}}(t) - \hat{F}_n(t)| \right) \geq \frac{\left| H^{(-1)} \left(b_n (t - t_{s,n}) \right) \right|}{2 b_n} \mathbb{E} \int_{\mathbb{R}^n} \min \left(\prod_{i=1}^n h_{\tilde{\theta},s,0}(y_{i,1}), \prod_{i=1}^n h_{\tilde{\theta},s,1}(y_{i,1}) \right) dy_{1,1} \dots dy_{n,1}. \tag{13}$$

According to Le Cam's Lemma (see Lemma 7 of Appendix B), for any $\theta \in \{0,1\}^{b_n}$:

$$\begin{aligned}
\int_{\mathbb{R}^n} \min \left(\prod_{i=1}^n h_{\theta,s,0}(y_{i,1}), \prod_{i=1}^n h_{\theta,s,1}(y_{i,1}) \right) dy_{1,1} \dots dy_{n,1} \\
\geq \frac{1}{2} \left[\int_{\mathbb{R}^n} \left\{ \prod_{i=1}^n h_{\theta,s,0}(y_{i,1}) \right\}^{\frac{1}{2}} \left\{ \prod_{i=1}^n h_{\theta,s,1}(y_{i,1}) \right\}^{\frac{1}{2}} dy_{1,1} \dots dy_{n,1} \right]^2 \\
\geq \frac{1}{2} \left[\int_{\mathbb{R}} \sqrt{h_{\theta,s,0}(y_{1,1})\, h_{\theta,s,1}(y_{1,1})} dy_{1,1} \right]^{2n} \\
\geq \frac{1}{2} \left[1 - \frac{1}{2} \chi^2 \left(h_{\theta,s,0}\,,\, h_{\theta,s,1} \right) \right]^{2n}
\end{aligned} \tag{14}$$

where we have used Lemma 8 of Appendix B for the last inequality. Assume for the moment that there exists a constant $c > 0$ such that for any $\theta \in \{0,1\}^{b_n}$:

$$\chi^2\left(h_{\theta,s,0}\,,\, h_{\theta,s,1}\right) \le \frac{c}{n}. \tag{15}$$

Then, using (11), (13), (14) and (15), we find that there exists a constant $C > 0$ such that

$$\sup_{\mu\in\mathcal{C}(M,p)} \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}} W_1\left(\mu, \tilde{\mu}_n\right) \ge \frac{C}{b_n}\sum_{s=1}^{b_n}\int_{t_{s,n}}^{t_{s+1,n}} \left|H^{(-1)}\left(b_n(t-t_{s,n})\right)\right| dt \ge \frac{C}{b_n}\int_0^1 \left|H^{(-1)}(u)\right| du\,.$$

Take b_n as in (7) and the theorem is thus proved (for $A = \mathrm{I}_d$) since the last term is positive according to (A1).

Proof of (15). Let C be a positive constant which may vary from line to line. We follow [Fan92] to show that (15) is valid for b_n chosen as in (7). Recall that we have chosen the function H such that, by Lemma 3 of Appendix A, $f_\theta \ge C f_{0,r}$. Thus,

$$\begin{aligned}
\chi^2\left(h_{\theta,s,0}\,,\, h_{\theta,s,1}\right) &\le \int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H\left[b_n(t-u-t_{s,n})\right] g(u)\,du\right\}^2}{f_{\theta,s,0}\star g(t)} dt\\
&\le C\int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H\left[b_n(t-u-t_{s,n})\right] g(u)\,du\right\}^2}{f_{0,r}\star g(t)} dt\\
&\le C\int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H\left[b_n(t'-u)\right] g(u)\,du\right\}^2}{\int_{-\infty}^{+\infty} f_{0,r}(t'+t_{s,n}-u)g(u)\,du} dt'.
\end{aligned}$$

Moreover, there exists a positive constant C such that for any $t \in \mathbb{R}$ and any $s \in \{1,\dots,b_n\}$, $f_{0,r}(t+t_{s,n}) \ge C f_{0,r}(t)$. Then,

$$\begin{aligned}
\chi^2\left(h_{\theta,s,0}\,,\, h_{\theta,s,1}\right) &\le C\int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H\left[b_n(t'-u)\right] g(u)\,du\right\}^2}{\int_{-\infty}^{+\infty} f_{0,r}(t'-u)g(u)\,du} dt'\\
&\le C b_n^{-1}\int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2}{f_{0,r}\star g(v/b_n)} dv.
\end{aligned} \tag{16}$$

The right side of (16) is typically the kind of χ^2 divergence that is upper bounded in the proof of Theorem 4 in [Fan91b] for computing pointwise rates of convergence. However, a slight modification of the proof of Fan is necessary since we can not assume here that $r < \min(1, \kappa_2 - 0.5)$ (because $r > p + 3/2$). It is shown in the proof of Theorem 4 in [Fan91b] that

$$\int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2 dv = O\left(b_n^{2\tilde{\beta}}\exp(-2b_n^{\beta}/\gamma)\right). \tag{17}$$

According to Lemma 4 of Appendix A, there exist $t_0 > 0$, $C_1 > 0$ and $C_2 > 0$ such that for any $t \in \mathbb{R}$:

$$f_{0,r}\star g(t) \ge C_1 \mathbb{1}_{|t|\le t_0} + \frac{C_2}{t^{2r}}\mathbb{1}_{|t|>t_0} \tag{18}$$

Note that we can apply Lemma 5 of appendix A since r satisfies (6). Then, using (17), (18) and Lemma 5 of Appendix A, for $T > t_0$ we have:

$$
\begin{aligned}
&\int_{-\infty}^{+\infty} \frac{\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2}{f_{0,r}\star g(v/b_n)}dv \\
&= \int_{|v|/b_n\le T} \frac{\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2}{f_{0,r}\star g(v/b_n)}dv + \int_{|v|/b_n> T} \frac{\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2}{f_{0,r}\star g(v/b_n)}dv \\
&\le (C_1\wedge C_2T^{-2r})^{-1}\int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\,dy/b_n\right\}^2 dv + C_r\int_{|v|/b_n>T}\frac{(|v|/b_n)^{-2\kappa_2}}{(|v|/b_n)^{-2r}}dv \\
&\le O\left(T^{2r}b_n^{2\tilde{\beta}}\exp(-2b_n^{\beta}/\gamma)\right) + O\left(b_n^{2(r-\kappa_2)}T^{2(r-\kappa_2)+1}\right)
\end{aligned}
$$

for T large enough. By taking $T = T_n = b_n^{\frac{2r-2\kappa_2-2\tilde{\beta}}{2\kappa_2-1}}\exp\left(\frac{2b_n^{\beta}}{\gamma(2\kappa_2-1)}\right)$ in this bound and according to (16), we find that for n large enough:

$$
\begin{aligned}
\chi^2(h_{\theta,s,0}\,,\,h_{\theta,s,1}) &= O\left(b_n^{2\tilde{\beta}+4r\frac{r-\kappa_2-\tilde{\beta}}{2\kappa_2-1}}\exp\left\{-\frac{2b_n^{\beta}}{\gamma}\left[1-\frac{2r}{2\kappa_2-1}\right]\right\}\right) \\
&= O\left(\exp(-\eta b_n^{\beta})\right) = O\left(\frac{1}{n}\right)
\end{aligned}
$$

for b_n defined by (7).

2.2.2 The general case

We now assume, as in the introduction, that there exists a invertible matrix A such that the coordinates of the vector $A\varepsilon_1$ are independent. Let $\mu \in \mathcal{C}_A(M,p)$ and let $\hat{\mu}_n$ be an estimator of the probability measure μ. Let μ^A and $\hat{\mu}_n^A$ be the image measures of μ and $\hat{\mu}$ by A. Then,

$$
\begin{aligned}
W_1(\hat{\mu}_n^A,\mu^A) &= \min_{\pi\in\Pi(\hat{\mu}_n^A,\mu^A)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|\pi(dx,dy) \\
&= \min_{\tau\in\Pi(\hat{\mu}_n,\mu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|Ax-Ay\|\,\tau(dx,dy) \\
&\le \|A\|\,W_1\left(\hat{\mu}_n,\mu\right)\,,
\end{aligned}
$$

where $\|A\| = \sup_{\|x\|=1}\|Ax\|$. Consequently $W_1\left(\hat{\mu}_n,\mu\right) \ge \|A\|^{-1}W_1(\hat{\mu}_n^A,\mu^A)$.

The image measure of $\mu\star\mu_\varepsilon$ by A is equal to $\mu^A\star\mu_\varepsilon^A$, where μ_ε^A is the image measure of μ_ε by A. Moreover, the probability measure estimator $\hat{\mu}_n^A$ can be written $\hat{\mu}_n^A = m(Z_1,\ldots,Z_n)$ where $Z_i = AY_i$ and m is a measurable function from $(\mathbb{R}^d)^n$ into the set of probability measures on $\mathbb{R}^d$. Thus,

$$
\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}\,W_1\left(\hat{\mu}^A,\mu^A\right) = \mathbb{E}_{(\mu^A\star\mu_\varepsilon^A)^{\otimes n}}\,W_1\left(m(Z_1,\ldots,Z_n),\mu^A\right).
$$

Since $\mu\in\mathcal{C}_A(M,p)\Leftrightarrow\mu^A\in\mathcal{C}(M,p)$, we obtain that

$$
\sup_{\mu\in\mathcal{C}_A(M,p)}\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}\,W_1(\hat{\mu}_n,\mu) \ge \|A\|^{-1}\sup_{\mu^A\in\mathcal{C}(M,p)}\mathbb{E}_{(\mu^A\star\mu_\varepsilon^A)^{\otimes n}}\,W_1\left(m(Z_1,\ldots,Z_n)\,,\,\mu^A\right). \quad (19)
$$

Note that, in the model $Z_i = AX_i + A\varepsilon_i$, the error $\eta = A\varepsilon$ has independent coordinates and satisfies the assumptions of Theorem 2 for $A = \mathrm{I}_d$.

We now apply the lower bound obtained in Section 2.2.1 , which gives that there exists a positive constant C such that

$$\liminf_{n\to\infty} (\log n)^{1/\beta} \sup_{\nu\in\mathcal{C}(M,p)} \mathbb{E}_{(\nu\star\mu_\eta)^{\otimes n}} W_1\left(m(Z_1,\ldots,Z_n)\,,\,\nu\right) \geq C. \tag{20}$$

The result follows from (20) and (19).

3 Upper bounds

In this section, we generalize the results of [CCDM11] by proving an upper bound on the rates of convergence for the estimation of the probability μ under any metric W_p.

3.1 Errors with independent coordinates

In this section, we assume that the random variables $(\varepsilon_{1,j})_{1\leq j\leq d}$ are independent, which means that ε_1 has the distribution $\mu_\varepsilon = \mu_{\varepsilon,1}\otimes\cdots\otimes\mu_{\varepsilon,d}$.

Let $p\in[1,\infty[$ and denote by $\lceil p\rceil$ the smallest integer greater than p. We first define a kernel k whose Fourier transform is smooth enough and compactly supported over $[-1,1]$. Such kernels can be defined by considering powers of the sinc function. More precisely, let

$$k(x) = c_p\left\{\frac{(2\lceil p/2\rceil+2)\sin\frac{x}{2\lceil p/2\rceil+2}}{x}\right\}^{2\lceil p/2\rceil+2}.$$

where c_p is such that $\int k(x)dx = 1$. The kernel k is a symmetric density, and k^* is supported over $[-1,1]$. Moreover k^* is $\lceil p\rceil$ times differentiable with Lipschitz $\lceil p\rceil$-th derivative. For any $j\in\{1,\cdots,d\}$ and any $h_j>0$, let

$$\tilde{k}_{j,h_j}(x) = \frac{1}{2\pi}\int e^{iux}\frac{k^*(u)}{\mu_j^*(u/h_j)}du\,.$$

A preliminary estimator $\hat{f}_n$ is given by

$$\hat{f}_n(x_1,\ldots,x_d) = \frac{1}{n}\sum_{i=1}^n\prod_{j=1\ldots d}\frac{1}{h_j}\tilde{k}_{j,h_j}\Big(\frac{x_j-Y_{i,j}}{h_j}\Big)\,. \tag{21}$$

The estimator (21) is the multivariate version of the standard deconvolution kernel density estimator which was first introduced in [CH88]. This estimator has been the subject of many works in the one dimensional case, but only few authors have studied the multidimensional deconvolution problem, see [Tan94], [CL11] and [CCDM11].

The estimator $\hat{f}_n$ is not necessarily a density, since it has no reason to be non negative. Since our estimator has to be a probability measure, we define

$$\hat{g}_n(x) = \alpha_n\hat{f}_n^+(x), \quad \text{where} \quad \alpha_n = \frac{1}{\int_{\mathbb{R}^d}\hat{f}_n^+(x)dx} \quad \text{and} \quad \hat{f}_n^+ = \max\{0,\hat{f}_n\}\,.$$

The estimator $\hat{\mu}_n$ of μ is then the probability measure with density $\hat{g}_n$.

The next theorem gives the rates of convergence of the estimator $\hat{\mu}_n$ under some assumptions on the derivatives of the functions $r_j := 1/\mu_{\varepsilon,j}^*$.

Theorem 4. *Let $M > 0$. Assume that we observe a n-sample $Y_1 \dots, Y_n$ in the multivariate convolution model (1). Also assume that there exists $\beta > 0$, $\tilde{\beta} \geq 0$, $\gamma_2 > 0$ and $c_2 > 0$ such that for every $j \in \{1, \dots, d\}$, every $\ell \in \{0, 1 \dots, \lceil p \rceil + 1\}$ and every $t \in \mathbb{R}$:*

$$\left| r_j^{(\ell)}(t) \right| \leq c_2 (1 + |t|^{\tilde{\beta}}) \exp\left(|t|^{\beta} / \gamma_2 \right). \tag{22}$$

Taking $h_1 = \dots = h_d = (4d/(\gamma_2 \log(n))^{1/\beta}$, there exists a positive constant C such that

$$\sup_{\mu \in \mathcal{D}(M,p)} \mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} \left(W_p^p (\mu, \hat{\mu}_n) \right) \leq C \left(\log n \right)^{-\frac{p}{\beta}} .$$

3.1.1 Proof of Theorem 4

Let $H = (h_1, h_2, \dots, h_d)$. We follow the proof of Proposition 2 in [CCDM11]. First we have the bias-variance decomposition

$$\mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} (W_p^p(\hat{\mu}_n, \mu)) \leq 2^{p-1} B(H) + 2^{2(p-1)} \int_{\mathbb{R}^d} (2^{p-1} C(H) + \|x\|^p) \sqrt{\mathrm{Var}(\hat{f}_n(x))} dx \,,$$

where

$$B(H) = \int \|H^t x\|^p K(x) dx \quad \text{and} \quad C(H) = B(H) + \int \|x\|^p \mu(dx) \,.$$

The proof of this inequality is the same as that of Proposition 1 in [CCDM11], by using Theorem 6.15 in [Vil08].

Note that $B(H)$ is such that $B(H) \leq d^{p-1} \beta (h_1^p + \dots + h_d^p)$, with $\beta = \int |u|^p k(u) du$. To ensure the consistency of the estimator, the bias term $B(H)$ has to tend to zero as n tends to infinity. Without loss of generality, we assume in the following that H is such that $B(H) \leq 1$. Hence, the variance term

$$V_n = 2^{2(p-1)} \int_{\mathbb{R}^d} (2^{p-1} C(H) + \|x\|^p) \sqrt{\mathrm{Var}(\hat{f}_n(x))} dx$$

is such that

$$V_n \leq C \int_{\mathbb{R}^d} \left(1 + \sum_{j=1}^{d} |x_j|^p \right) \sqrt{\mathrm{Var}(\hat{f}_n(x_1, \dots, x_n))} \; dx_1 \dots dx_d$$

for some positive constant C. Now

$$\sqrt{\mathrm{Var}(\hat{f}_n(x_1, \dots, x_n))} \leq \frac{1}{\sqrt{n}} \sqrt{\mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} \left[\left\{ \prod_{j=1}^{d} \frac{1}{h_j} \tilde{k}_{j,h_j} \left(\frac{x_j - Y_{1,j}}{h_j} \right) \right\}^2 \right]} .$$

Applying Cauchy-Schwarz's inequality d-times, we obtain that

$$\begin{aligned} &\int_{\mathbb{R}^d} \sqrt{\mathrm{Var}(\hat{f}_n(x_1, \dots, x_n))} \; dx_1 \dots dx_d \\ &\qquad \leq \frac{D_1}{\sqrt{n}} \sqrt{\mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} \Big(\prod_{j=1}^{d} \int (1 \vee x_j^2) \Big(\frac{1}{h_j} \tilde{k}_{j,h_j} \Big(\frac{x_j - Y_{1,j}}{h_j} \Big) \Big)^2 dx_j \Big)} \\ &\qquad\qquad \leq \frac{D_2}{\sqrt{n}} \sqrt{\mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} \Big(\prod_{j=1}^{d} (1 \vee Y_{1,j}^2) \Big) \prod_{j=1}^{d} \int (1 + u_j^2 h_j^2) \frac{1}{h_j} (\tilde{k}_{j,h_j}(u_j))^2 du_j} \end{aligned}$$

where D_1 and D_2 are positive constants depending on d. Now, by independence of X_1 and ε_1, and by independence of the coordinates of ε_1,

$$\mathbb{E}_{\mu\star\mu_\varepsilon}\Big(\prod_{j=1}^d (1\vee Y_{1,j}^2)\Big) \le \mathbb{E}_\mu\Big(\prod_{j=1}^d (1+X_{1,j}^2)\Big)\prod_{j=1}^d (1+\mathbb{E}_{\mu_\varepsilon}(\varepsilon_{1,j}^2)).$$

Since $\mu \in D(M,p)$, it follows that

$$\int_{\mathbb{R}^d} \sqrt{\mathrm{Var}(\hat{f}_n(x))}\, dx \le \frac{A_0}{\sqrt{n}}\sqrt{\prod_{j=1}^d \int (1+u_j^2h_j^2)\frac{1}{h_j}(\tilde{k}_{j,h_j}(u_j))^2 du_j}\,. \tag{23}$$

In the same way, using again that $\mu \in D(M,p)$, we obtain that

$$\begin{aligned}&\int_{\mathbb{R}^d} |x_\ell|^p \sqrt{\mathrm{Var}(\hat{f}_n(x_1,\ldots,x_n))}\, dx_1\ldots dx_d \\ &\quad\le \frac{A_\ell}{\sqrt{n}}\sqrt{\int (1+|u_\ell|^{2p+2}h_\ell^{2p+2})\frac{1}{h_\ell}(\tilde{k}_{\ell,h_\ell}(u_\ell))^2 du_\ell \prod_{j\ne\ell}\int (1+u_j^2h_j^2)\frac{1}{h_j}(\tilde{k}_{j,h_j}(u_j))^2 du_j}.\end{aligned}$$

Starting from these computations, one can prove the following Proposition.

Proposition 1. *Let $(h_1,\ldots,h_d)\in[0,1]^d$. The following upper bound holds*

$$\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n,\mu)) \le (2d)^{p-1}\beta(h_1^p+\cdots+h_d^p) + \frac{L}{\sqrt{n}}\left(\prod_{j=1}^d I_j(h_j) + \sum_{\ell=1}^d J_\ell(h_\ell)\Big(\prod_{j=1,j\ne\ell}^d I_j(h_j)\Big)\right)$$

where L is some positive constant L and

$$\begin{aligned} I_j(h) &\le \sqrt{\int_{-1/h}^{1/h} (r_j(u))^2 + (r_j'(u))^2 du}\,, \\ J_j(h) &\le \sqrt{\int_{-1/h}^{1/h} (r_j(u))^2 + (r_j^{(\lceil p\rceil+1)}(u))^2 du} \\ &\quad + \sum_{k=1}^{\lceil p\rceil} h^{\lceil p\rceil+1-k}\sqrt{\int_{-1/h}^{1/h} (r_j^{(k)}(u))^2 du}\,. \end{aligned}$$

Let us finish the proof of Theorem 4 before proving Proposition 1. Take $h_1=\ldots=h_d=h$. The condition (22) on the derivatives of r_j leads to the upper bounds

$$\mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n,\mu)) \le C\Big(h^p + \frac{1}{\sqrt{n}h^{d(2\tilde{\beta}+1)/2}}\exp(d/(\gamma_2 h^\beta))\Big)\,.$$

The choice $h=(4d/(\gamma_2\log(n))^{1/\beta}$ gives the desired result.

Proof of Proposition 1. It follows the proof of Proposition 2 in [CCDM11]. By Plancherel's identity,

$$\begin{aligned}\int \frac{1}{h}(\tilde{k}_{j,h}(u))^2 du = \frac{1}{2\pi}\int \frac{1}{h}\frac{(k^*(u))^2}{(\mu_{\varepsilon,j}^*(u/h))^2}du &= \frac{1}{2\pi}\int \frac{(k^*(hu))^2}{(\mu_{\varepsilon,j}^*(u))^2}du \\ &\le \frac{1}{2\pi}\int_{-1/h}^{1/h} r_j^2(u) du\,.\end{aligned}$$

the last upper bound being true because k^* is supported over $[-1,1]$ and bounded by 1.

Let C be a positive constant, which may vary from line to line. Let $q_{j,h}(u) = r_j(u/h)k^*(u)$. Since $q_{j,h}$ is differentiable with compactly supported derivative, we have that

$$-iu2\pi\tilde{k}_{j,h}(u) = (q'_{j,h})^*(u)\,.$$

Applying Plancherel's identity again,

$$\begin{aligned}\int hu^2(\tilde{k}_{j,h}(u))^2du &= \frac{1}{2\pi}\int h(q'_{j,h}(u))^2du \\ &\leq C\ \left(\int_{-1/h}^{1/h}(r'_j(u))^2du + h^2\int_{-1/h}^{1/h} r_j^2(u)du\right),\end{aligned}$$

the last inequality being true because k^* and $(k^*)'$ are compactly supported over $[-1,1]$. Consequently

$$\sqrt{\int(1+u_j^2h_j^2)\frac{1}{h_j}(\tilde{k}_{j,h_j}(u_j))^2du_j} \leq CI_j(h_j)\,.$$

In the same way

$$(-iu)^{\lceil p\rceil+1}2\pi\tilde{k}_{j,h}(u) = (q_{j,h}^{(\lceil p\rceil+1)})^*(u)$$

and

$$\int h^{2\lceil p\rceil+1}u^{2\lceil p\rceil+2}(\tilde{k}_{j,h}(u))^2du = \frac{1}{2\pi}\int h^{2\lceil p\rceil+1}(q_{j,h}^{(\lceil p\rceil+1)}(u))^2du.$$

Now, since $k^*, (k^*)', \ldots, (k^*)^{(\lceil p\rceil+1)}$ are compactly supported over $[-1,1]$,

$$\int h^{2\lceil p\rceil+1}(q_{j,h}^{(\lceil p\rceil+1)}(u))^2du \leq C\sum_{k=0}^{\lceil p\rceil+1} h^{2(\lceil p\rceil+1-k)}\int_{-1/h}^{1/h}(r_j^{(k)}(u))^2du\,.$$

Consequently

$$\begin{aligned}&\int(1+|u_\ell|^{2p+2}h_\ell^{2p+2})\frac{1}{h_\ell}(\tilde{k}_{\ell,h_\ell}(u_\ell))^2du_\ell \\ &\qquad\leq 2\int(1+|u_\ell|^{2\lceil p\rceil+2}h_\ell^{2\lceil p\rceil+2})\frac{1}{h_\ell}(\tilde{k}_{\ell,h_\ell}(u_\ell))^2du_\ell \leq CJ_\ell(h_\ell)\,.\end{aligned}$$

The results follows. □

3.2 The general case

Here, as in the introduction, we shall assume that there exists an invertible matrix A such that the coordinates of the vector $A\varepsilon_1$ are independent. Applying A to the random variables Y_i in (1), we obtain the new model

$$AY_i = AX_i + A\varepsilon_i\,,$$

that is: a convolution model in which each error vector $\eta_i = A\varepsilon_i$ has independent coordinates.

To estimate the image measure μ^A of μ by A, we use the preliminary estimator (21), that is

$$\hat{f}_{n,A}(x_1,\ldots,x_d) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1\ldots d}\frac{1}{h_j}\tilde{k}_{j,h_j}\Big(\frac{x_j-(AY_i)_j}{h_j}\Big)\,,$$

and the estimator $\hat{\mu}_{n,A}$ of μ^A is deduced from $\hat{f}_{n,A}$ as in Section 3.1. This estimator $\hat{\mu}_{n,A}$ has the density $\hat{g}_{n,A}$ with respect to the Lebesgue measure.

To estimate μ, we define $\hat{\mu}_n = \hat{\mu}_{n,A}^{A^{-1}}$ as the image measure of $\hat{\mu}_{n,A}$ by A^{-1}. This estimator has the density $\hat{g}_n = |A|\hat{g}_{n,A} \circ A$ with respect to the Lebesgue measure. It can be deduced from the preliminary estimator $\hat{f}_n = |A|\hat{f}_{n,A} \circ A$ as in Section 3.1. Now

$$\begin{aligned} W_p^p(\hat{\mu}_n, \mu) &= \min_{\lambda \in \Pi(\hat{\mu}_n, \mu)} \int \|x - y\|^p \lambda(dx, dy) \\ &= \min_{\pi \in \Pi(\hat{\mu}_{n,A}, \mu^A)} \int \|A^{-1}(x - y)\|^p \pi(dx, dy) . \end{aligned}$$

Consequently, if $\|A^{-1}\| = \sup_{\|x\|=1} \|A^{-1}x\|$, we obtain that

$$W_p^p(\hat{\mu}_n, \mu) \le \|A^{-1}\|^p W_p^p(\hat{\mu}_{n,A}, \mu^A) , \tag{24}$$

which is an equality if A is an unitary matrix. Note also that $\mu \in \mathcal{D}_A(M, p)$ if and only if $\mu^A \in \mathcal{D}(M, p)$.

Let μ_η be the distribution of the η_i's. Since the coordinates of the η_i's are independent, μ_η can be written as $\mu_\eta = \mu_{\eta,1} \otimes \cdots \otimes \mu_{\eta,d}$. As in Section 3.1, let $r_j := 1/\mu^*_{\eta,j}$. Assume that the r_j's satisfy the condition (22). It follows from (24) and Theorem 4 that, taking $h_1 = \cdots = h_d = (4d/(\gamma_2 \log(n))^{1/\beta}$, there exists a positive constant C such that

$$\sup_{\mu \in \mathcal{D}_A(M,p)} \mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}} \left(W_p^p \left(\mu, \hat{\mu}_n \right) \right) \le C \left(\log n \right)^{-\frac{p}{\beta}} .$$

3.3 Examples of rates of convergence

Gaussian noise. Assume that we observe $Y_1, \ldots, Y_n$ in the multivariate convolution model (1), where ε is a centered non degenerate Gaussian random vector. In that case, there always exists an invertible matrix A such that the coordinates of $A\varepsilon_1$ are independent. The distribution of $(A\varepsilon_1)_j$ is a either a Dirac mass at zero or a centered Gaussian random variable with positive variance. Since ε is non degenerate, there exists at least one index j_0 for which $(A\varepsilon_1)_{j_0}$ is non zero.

Now, the distribution of $(A\varepsilon_1)_{j_0}$ satisfies the assumptions of Theorem 2, for any $p \ge 1$ and $\beta = 2$ (Conditions (3) and (4) follow from Lemma 1). Moreover, denoting by $\mu_{\eta,j}$ the distribution of $\eta_{1,j} = (A\varepsilon_1)_j$, then the quantity $r_j^* = 1/\mu^*_{\eta,j}$ satisfies (22) for any $p \ge 1$ and $\beta = 2$. Theorem 1 follows then from Theorems 2 and 4 (more precisely, the estimator $\hat{\mu}_n$ of Theorem 1 is constructed as in Section 3.2).

Other supersmooth distributions. For $\alpha \in]0, 2[$, we denote by s_α the symmetric α-stable density, whose Fourier transform q_α is given by

$$s_\alpha^*(x) = q_\alpha(x) = \exp\left(-|x|^\alpha\right) .$$

Let $q_{\alpha,1} = q_\alpha$ and $q_{\alpha,2} = q_\alpha \star q_\alpha$. For any positive integer $k > 2$, define by induction $q_{\alpha,k} = q_{\alpha,k-1} \star q_\alpha$.

Lemma 2. *Let k be a positive integer. The function $q_{\alpha,k}$ satisfies the following properties*

1. *$q_{\alpha,k}$ is $k-1$ times differentiable, and $q_{\alpha,k}^{(k-1)}$ is absolutely continuous, with almost sure derivative $q_{\alpha,k}^{(k)}$. Moreover, if $\alpha \in]0, 1[$ then $q_{\alpha,k}^{(k-1)}$ is bounded, and if $\alpha \in [1, 2[$ then $q_{\alpha,k}^{(k)}$ is bounded.*

2. There exists two positive constants $a_{\alpha,k}$ and $b_{\alpha,k}$ such that for any $x \in \mathbb{R}$,

$$a_{\alpha,k} \exp\left(-|x|^\alpha\right) \le q_{\alpha,k}(x) \le b_{\alpha,k} \exp\left(-\frac{|x|^\alpha}{2^{(k-1)\alpha}}\right).$$

The proof of Lemma 2 is given in Appendix C. Next, for any integer $k \ge 2$, we introduce the supersmooth density

$$f_{\alpha,k}(x) = \frac{(s_\alpha(x))^k}{\int (s_\alpha(x))^k dx},$$

and we note that $f^*_{\alpha,k} = q_{\alpha,k}/q_{\alpha,k}(0)$ and $f_{\alpha,k}(x) = O(|x|^{-k(\alpha+1)})$. Let $r_{\alpha,k} = 1/f^*_{\alpha,k}$. According to Lemma 2, it follows that r is k times differentiable with

$$|r^{(\ell)}_{\alpha,k}(x)| \le C_\ell \sum_{i=1}^{\ell} \frac{|(f^*_{\alpha,k})^{(i)}(x)|}{(f^*_{\alpha,k}(x))^{\ell+2-i}} \qquad \text{for } 1 \le \ell \le k.$$

Applying Lemma 2 we see that: if $\alpha \in]0,1[$, then for any $\ell \in \{0, \dots, k-1\}$

$$|r^{(\ell)}_{\alpha,k}(x)| \le K_{\alpha,\ell} \exp(|x|^\alpha)\,, \tag{25}$$

and the same holds for any $\ell \in \{0, \dots, k\}$ if $\alpha \in [1,2[$. Moreover, we also have the lower bound

$$|r_{\alpha,k}(x)| \ge c_{\alpha,k} \exp\left(\frac{|x|^\alpha}{2^{(k-1)\alpha}}\right). \tag{26}$$

Now, assume that we observe $Y_1, \dots, Y_n$ in the multivariate convolution model (1). Let $p \ge 1$, assume that there exists an invertible matrix A such that, for any $j \in \{1, \dots, d\}$, $(A\varepsilon_1)_j$ has the distribution f_{α_j,k_j} for some $\alpha_j \in]0,2[$ and such that $k_j \ge \lceil p \rceil + 1 + 2\mathbf{1}_{\alpha_j \in]0,1[}$. Let $\alpha = \max_{1 \le j \le d} \alpha_j$.

Inequality (26) gives Condition (2) in Theorem 4 for $\beta = \alpha$. Lemma 1 can be applied with $a = (\lceil p \rceil + 1 + 2\mathbf{1}_{\alpha \in]0,1[})(\alpha+1)$ and then Conditions (3) and (4) of Theorem 2 are also satisfied. Next, according to (25), Condition (22) in Theorem 4 is satisfied for $\beta = \alpha$. Theorems 2 and 4 finally give the following result:

1. There exists a constant $C > 0$ such that for all estimator $\tilde{\mu}_n$ of the measure μ:

$$\liminf_{n \to \infty} (\log n)^{p/\alpha} \sup_{\mu \in \mathcal{D}_A(M,p)} \mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}}(W_p^p(\tilde{\mu}_n, \mu)) \ge C.$$

2. The estimator $\hat{\mu}_n$ of μ constructed in Section 3.2 is such that

$$\sup_{n \ge 1} \sup_{\mu \in \mathcal{D}_A(M,p)} (\log n)^{p/\alpha}\, \mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n, \mu)) \le K\,,$$

   for some positive constant K.

Mixtures of distributions. Of course, the independent coordinates of $A\varepsilon_1$ need not all be Gaussian or even supersmooth.

For instance if there exists j_0 such that $(A\varepsilon_1)_{j_0}$ is a non degenerate Gaussian random variable, and the other coordinates have distribution which is either a Dirac mass at 0 or a Laplace distribution, or a supersmooth distribution $f_{\alpha,k}$ for some $\alpha \in]0,2[$ and $k \ge \lceil p \rceil + 1 + 2\mathbf{1}_{\alpha \in]0,1[}$ (this list in non exhaustive), then the estimator $\hat{\mu}_n$ of μ constructed in Section 3.2 is such that

$$\sup_{n \ge 1} \sup_{\mu \in \mathcal{D}_A(M,p)} (\log n)^{p/2}\, \mathbb{E}_{(\mu \star \mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n, \mu)) \le K\,,$$

and this rate is minimax.

In the same way if there exists j_0 such that $(A\varepsilon_1)_{j_0}$ is supersmooth with density $f_{\alpha,k}$ for some $\alpha \in]0,2[$ and $k \geq \lceil p \rceil + 1 + 2\mathbf{1}_{\alpha \in]0,1[}$, and the other coordinates have distribution which is either a Dirac mass at 0 or a Laplace distribution, or a supersmooth distribution $f_{\beta,m}$ for some $\beta \in]0,\alpha]$ and $m \geq \lceil p \rceil + 1 + 2\mathbf{1}_{\beta \in]0,1[}$, then the estimator $\hat{\mu}_n$ of μ constructed in Section 3.2 is such that

$$\sup_{n\geq 1} \sup_{\mu \in \mathcal{D}_A(M,p)} (\log n)^{p/\alpha} \, \mathbb{E}_{(\mu\star\mu_\varepsilon)^{\otimes n}}(W_p^p(\hat{\mu}_n, \mu)) \leq K \,,$$

and this rate is minimax.

4 Discussion

In the supersmooth case, we have seen that lower bounds for the Wasserstein deconvolution problem in any dimension can be deduced from lower bounds for the deconvolution of the c.d.f in dimension one. But this method cannot work in the ordinary smooth case for $d > 1$, because, contrary to the supersmooth case, the rates of convergence depends on the dimension.

Let us briefly discuss the case where $d = 1$ and the error distribution is ordinary smooth. It is actually well known that establishing optimal rates of convergence in the ordinary smooth case is more difficult than in the supersmooth case, even for pointwise estimation, as noticed by Fan in [Fan91b]. When the density is m times differentiable, Fan gives in this paper pointwise lower and upper bounds for the estimation of the c.d.f. in both the supersmooth case and the ordinary smooth case. He finds the optimal rates in the supersmooth case and he conjectures that his upper bound is actually optimal in the ordinary smooth case (see his Remark 3). Optimal pointwise rates for the deconvolution of the c.d.f. in the ordinary smooth case was an open question until recently. This problem has been solved in [DGJ11] when the density belongs to a Sobolev class.

When $d = 1$ and the error distribution is ordinary smooth, some results about integrated rates of convergence for the density (and its derivatives) can be found in [Fan93, Fan91a] but the case of the c.d.f. (for the integrated risk) is not studied in these papers. However, some lower bounds can be easily computed by following the method of [Fan93] and using the pointwise rates of [Fan91b] : for a class of ordinary smooth noise densities of order β and assuming only that the unknown distribution μ has a moment of order 4, we find that the minimax integrated risk is lower bounded by $n^{-1/(2\beta+1)}$ and we then obtain the same lower bound for W_1. As for the pointwise estimation described in [Fan91b], these rates do not match with the upper bounds given by Proposition 1 for W_1. For instance, for Laplace errors ($\beta = 2$), the rate of convergence of the kernel estimator under W_1 is upper bounded by $n^{-1/7}$. We are currently working on this issue, and we conjecture that the minimax rates of convergence for W_1 when $d = 1$ is of order $n^{-1/(2\beta+2)}$ for a class of ordinary smooth errors distributions of order β. If this conjecture is correct, it means that the existing lower and upper bounds have to be improved.

A Some known lemmas

The following lemma is given in [FT93] (Lemma 1):

Lemma 3. *Let H be a function such that*

$$|H(t)| \leq C(1+t^2)^{-r}$$

for some $C > 0$ and some $r > 0.5$. Then there exists a positive constant $\tilde{C}$ such that for any sequence $b_n \to \infty$,

$$\sum_{s=1}^{b_n} |H\left(b_n(t-s/n)\right)| \leq \tilde{C}(1+t^2)^{-r}.$$

Let $f_{0,r}$ be the function defined in (5). The following lemma can be found in [Fan91b] (Lemma 5.1):

Lemma 4. *For any probability measure μ, there exists a constant $C_r > 0$ such that*

$$f_{0,r} \star \mu(t) \geq C_r t^{-2r} \quad \text{as } |t| \text{ tends to infinity.}$$

The following lemma is rewritten from [Fan91b] (Lemma 5.2):

Lemma 5. *Let $r > 0$. Suppose that $P(|\varepsilon_1' - t| \leq |t|^{\kappa_1}) = O(|t|^{-\kappa_2})$ as $|t|$ tends to infinity for some $0 < \kappa_1 < 1$ and $\kappa_2 > 1$. Let H be a bounded function such that $|H(t)| \leq O(|t|^{-2r})$ for some $r > \kappa_2/(2\kappa_1)$. Then there exists a large T and a constant C such that when $|v|/b_n \geq T$:*

$$\int_{-\infty}^{+\infty} H(v-y)g(y/b_n)\, dy/b_n \leq C(|v|/b_n)^{-\kappa_2}.$$

B Distances between probability measures

The first lemma follows straightforwardly from the definition of W_1.

Lemma 6. *Let μ and $\tilde{\mu}$ be two measures on $\mathbb{R}^d$ with finite first moments, and let μ_1 and $\tilde{\mu}_1$ be their first marginals. Then $W_1(\mu, \tilde{\mu}) \geq W_1(\mu_1, \tilde{\mu}_1)$.*

The following Lemma is a particular case of the famous Le Cam's inequalities. See for instance Section 2.4 in [Tsy09] for more details.

Lemma 7. *Let h and $\tilde{h}$ be two densities on $\mathbb{R}^n$, then*

$$\int_{\mathbb{R}^n} \min\left(h(x), \tilde{h}(x)\right) dx \geq \frac{1}{2}\left\{\int_{\mathbb{R}^n} \sqrt{h(x)\tilde{h}(x)}dx\right\}^2.$$

The next lemma can be found for instance in Section 2.4 of [Tsy09].

Lemma 8. *Let h and $\tilde{h}$ be two densities for the Lebesgue measure on $\mathbb{R}$, then*

$$\int_{\mathbb{R}} \sqrt{h(y)\,\tilde{h}(y)}\, dy \geq 1 - \frac{1}{2}\chi^2(h, \tilde{h}).$$

C Auxiliary results

Proof of Lemma 2 The proof of Item 1 is standard. Note first that q_α is bounded and absolutely continuous with almost sure derivative q_α', which is bounded as soon as $\alpha \in [1, 2[$. This proves the result for $k = 1$. It follows that $q_\alpha \star q_\alpha$ is differentiable with derivative $q_\alpha \star q_\alpha'$. This derivative is absolutely continuous with almost sure derivative $q_\alpha' \star q_\alpha'$. Moreover $q_\alpha \star q_\alpha'$ is bounded (because q_α' is integrable and q_α is bounded), and if $\alpha \in [1, 2[$ then $q_\alpha' \star q_\alpha'$ is bounded (because in that case q_α' is bounded). This proves Item 1 for $k = 2$. The general case follows by induction.

In the same way, it suffices to prove Item 2 for $k = 2$, and the general case follows by induction. Since $q_\alpha \star q_\alpha$ is symmetric, it suffices to prove the result for $x > 0$. Now, for any $x > 0$,

$$\begin{aligned} q_\alpha \star q_\alpha(x) &= 2\int_{x/2}^{+\infty} \exp\left(-|x-t|^\alpha - t^\alpha\right) dt \\ &\leq a_{\alpha,2} \exp\left(-(x/2)^\alpha\right). \end{aligned}$$

On the other hand, for any $x > 1$, there exist a positive constant c_α such that

$$\begin{aligned} q_\alpha \star q_\alpha(x) &\geq \int_{x/2}^{x} \exp\left(-|x-t|^\alpha - t^\alpha\right) dt \\ &\geq \exp\left(-x^\alpha\right) \int_0^{x/2} \exp\left(-u^\alpha\right) du \\ &\geq c_\alpha \exp\left(-x^\alpha\right). \end{aligned} \tag{27}$$

The function $x \mapsto q_\alpha \star q_\alpha(x) \exp(x^\alpha)$ is continuous and positive on $[0, 1]$ and thus (27) is also true on $[0, 1]$ for some other positive constant c'_α. The lower bound follows by taking $b_{\alpha,2} = \min\{c_\alpha, c'_\alpha\}$.

References

[BT08a] C. Butucea and A. B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. I. *Theory Probab. Appl.*, 52:24–39, 2008.

[BT08b] C. Butucea and A. B. Tsybakov. Sharp optimality in density deconvolution with dominating bias. II. *Theory Probab. Appl.*, 52:237–249, 2008.

[CCDM11] Claire Caillerie, Frédéric Chazal, Jérôme Dedecker, and Bertrand Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Stat.*, 5:1394–1423, 2011.

[CCSM11] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Found. Comput. Math.*, 11:733–751, 2011.

[CH88] Raymond J. Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186, 1988.

[CL11] F. Comte and C. Lacour. Data driven density estimation in presence of unknown convolution operator. *J. Royal Stat. Soc., Ser B*, 73:601–627, 2011.

[DGJ11] I. Dattner, A. Goldenshluger, and A. Juditsky. On deconvolution of distribution functions. *Ann. Statist.*, 39(5):2477–2501, 2011.

[Fan91a] Jianqing Fan. Global behavior of deconvolution kernel estimates. *Statist. Sinica*, 1(2):541–551, 1991.

[Fan91b] Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991.

[Fan92] Jianqing Fan. Deconvolution with supersmooth distributions. *Canad. J. Statist.*, 20(2):155–169, 1992.

[Fan93] Jianqing Fan. Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, 21(2):600–610, 1993.

[FT93] Jianqing Fan and Young K. Truong. Nonparametric regression with errors in variables. *Ann. Statist.*, 21(4):1900–1925, 1993.

[GPPVW12] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.*, 40:941–963, 2012.

[Mei09] Alexander Meister. *Deconvolution problems in nonparametric statistics*, volume 193 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 2009.

[RR98] Svetlozar T. Rachev and Ludger Rüschendorf. *Mass transportation problems. Vol. II.* Probability and its Applications. Springer-Verlag, 1998.

[Tan94] X. Y. Tang. Effect of dimension in multivariate deconvolution problems. 1994.

[Tsy09] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag, 2008.