Mode estimation via Optimal Transport metric

Diego Fonseca, Mauricio Junca, Marco Avella

Abstract

This study addresses the challenge of estimating the mode of a random vector, particularly when the probability distribution and the density function are unknown. Given a sample of a random vector, our objective is to determine the point in the sample exhibiting the highest density, which can be regarded as an in-sample estimator of the largest mode of the random vector. To tackle this problem, we formulate it as a stochastic optimization problem and employ a Distributionally Robust Optimization (DRO) approach, utilizing Wasserstein metrics. Furthermore, we rigorously establish the consistency of the proposed estimator, demonstrating its validity and reliability in the context of mode estimation for random vectors.

1 Introduction

Mode estimation is a problem with numerous applications, such as computer vision (Chen and Meer, 2002; Vedaldi and Stefano, 2008), econometrics (Kemp and Santos-Silva, 2012; Ho et al., 2017), and clustering (Cheng, 1995; Genovese et al., 2016; Menardi, 2016; Jiang and Kpotufe, 2017; Casa et al., 2020), with the latter being one of the focal points in this work. The significance of these applications has rendered this problem an intriguing research topic. In this work, we investigate the problem of estimating the largest mode of a random vector when the probability distribution and density function are unknown. Nonetheless, when the distribution and density are unknown, we presume that a sample of the random vector is available. Our proposal entails using that sample to estimate the point with the highest density; that is, if all the sample points were evaluated in the density function, we would be interested in the point yielding the highest value in that evaluation.

To provide context, let X be a continuous random vector with support in $\Xi \subseteq \mathbb{R}^d$ and probability distribution $\mathbb{P}$. We assume that X has a density function f. Additionally, let $X_1, X_2, \ldots, X_n$ be an i.i.d sample of X. As mentioned earlier, our proposal aims to find the point in the sample with the highest density when $\mathbb{P}$ and the density function f are unknown. Observe that if f is known, this problem can be formulated as:

$$J_n := \max_{i=1,\ldots,n} f(X_i). \tag{1}$$

Estimating the mode from a sample in the case of discrete probability distributions is typically accomplished through frequentist analysis. However, in the case of continuous probability distributions, this task is more intricate. Some approaches involve estimating the unknown density function f using the sample and then calculating the mode of this estimated density. Nevertheless, determining the estimated density and computing its maximum tend to be computationally complex when the sample size is large, and the random vector has a high dimension. In contrast, this thesis addresses the problem of estimating the mode as a stochastic problem, which is approximated using a Distributionally Robust approach. Our contributions in this regard are as follows:

- We address the mode estimation problem as a stochastic optimization problem, adopting a Distributionally Robust perspective and employing Wasserstein distances to define the ambiguity set.
- The consistency of the proposed mode estimator is demonstrated, and empirical evidence suggests that its convergence rate can be improved.
- Drawing on the approach to the stochastic problem underlying the proposed mode estimator, we propose a clustering method. This method consists of a modification to the well-known spectral clustering technique, resulting in an improvement.

1.1 Related literature

In the *mode estimation* of a probability distribution, the objective is to calculate the global maximum of the associated density function. However, in practical situations, this function is unknown, and only a

sample of this distribution is available. Early approaches were motivated to use the sample to generate an estimate of the unknown density (Parzen, 1962; Chernoff, 1964). For this purpose, estimators based on Kernel Density Estimation (KDE) are utilized, which yield the density function $\frac{1}{h^d n}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$, where K is a fixed density, such as the multivariate standard normal density, and h is a parameter influencing the closeness of the estimator to the true density. These parameters are chosen by minimizing a measure of error. Fitting such parameters constitutes an active research topic, particularly in high-dimensional data, as evidenced in (Wanh and Scott, 2019). Once the h parameter has been fitted, the local maxima of this function are calculated, and the points where these maxima are reached serve as estimators of the unknown density modes. The strategy of using these types of density estimators in statistics was popularized by (Silverman, 1981, 1986).

However, this approach has drawbacks, including the sensitivity of the estimated modes and the number of these modes with respect to the parameter h. Additionally, the performance is unsatisfactory with a moderate sample, and the computation of the local maxima of the estimator can be computationally expensive (Lee et al., 2019). Despite these obstacles, research on this topic has been further motivated, as seen in (Genovese et al., 2016; Ameijeiras-Alonso et al., 2019; Chen, 2018).

In contrast to KDE-based methods, alternative approaches do not rely on Kernel Density Estimation. Among these, we find frequentist analysis-based methods, such as those presented in (Sager, 1978; Bickel, 2002; Burman and Polonik, 2009; Hsu and Wu, 2013; Kirschstein et al., 2016). However, these methods exhibit high sensitivity to contaminations in the sample and can be computationally expensive, depending on the sample size. Another strategy involves approaching the mode-finding problem as an optimization problem, where the objective function is the unknown density function. The Stochastic Gradient Ascent method is used, in which the gradient of the density function is replaced by an estimate (Fukunaga and Hostetler, 1975; Aliyari Ghassabeh, 2015; Kamanchi et al., 2019).

Lastly, there is the k-nearest neighbor approach (Dasgupta and Kpotufe, 2014). In this approach, given a fixed number k much smaller than the sample size, an estimator of the mode is identified as a point for which there exists a ball centered at that point containing at least k sample points, with the ball being as small as possible. In other words, the desired point is the one that concentrates the largest number of sample points within the smallest possible space around it. This method tends to fall into false modes when the sample size is not sufficiently large. However, it converges to the minimax-optimal rate, even though this rate starts to be reached with considerable sample size.

Our proposal has not been presented in previous works. Moreover, the use of Wasserstein distance in methods that aim at estimating modes is not common, even, in this review, we have not found research papers that have this approach. However, in this proposal, we use kernels (Gaussian, uniform spherical) that depend on a parameter h, which could relate our approach to KDE-based techniques. In fact, for certain values of h, our approach has the same convergence rates as estimating the mode using Kernel density estimation, indeed, it could be used for point-to-point density estimation. However, our priority is not to estimate the density, the objective is to estimate the mode. For this purpose, there are values of h that allow a better performance, which makes our method more related to the one proposed in (Dasgupta and Kpotufe, 2014) although they are different. Moreover, in terms of consistency, the optimality rates of our approach and that of (Dasgupta and Kpotufe, 2014) seem to be the same.

1.2 Outline

The outline of the paper is as follows. Section 2 formally introduces/reviews the distributionally robust optimization framework. In Section 3, we approached the problem of mode estimation as a stochastic optimization problem and approach it from a distributionally robust perspective. In section 4 we discuss and demonstrate the reformulations that can be obtained from the distributionally robust approach. Additionally, in section 5, we demonstrate the consistency of the resulting mode estimator. Finally, in Section 5.1 we explore one possible direction to improve the consistency rate. We defer many of the technical proofs to the Appendix.

Notation: For $q \in [1,\infty)\cap\mathbb{N}$, the q-norm in $\mathbb{R}^k$ is is noted as $\|\cdot\|_q$. For $n \in \mathbb{N}$, we let $[n] := \{1, 2, \ldots, n\}$. Additionally, if A is a finite set, the $|A|$ is the number of elements of A. Finally, given a sample $X_1, \ldots, X_n$ of a random variable X, the *empirical distribution* of X with respect to this sample is defined as the probability measure given by $\widehat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$, where δ_x is the Dirac delta function supported at x.

2 Distributionally Robust Optimization

In this section, we present the concept of Distributionally Robust Optimization (DRO). This concept emerged as a means to tackle stochastic optimization problems in the form of

$$\min_{x\in\mathcal{X}} \mathbb{E}_{X\sim\mathbb{P}}[F(x,X)]. \tag{2}$$

where F is a function such that $F:\mathbb{R}^m\times\mathbb{R}^d\to\mathbb{R}$, $X\in\mathbb{R}^d$ is a random vector with (unknown) probability distribution $\mathbb{P}$ supported in $\Xi\subseteq\mathbb{R}^d$, and $\mathcal{X}\subseteq\mathbb{R}^m$ is a set of constraints on the decision vectors. The Distributionally Robust Optimization (DRO) approach for the problem (2) is formulated as

$$J_{\mathcal{D}} := \min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{X\sim\mathbb{Q}}[F(x,X)], \tag{3}$$

where $\mathcal{D}$ is a set of probability distributions, which is known as *ambiguity set*. Note that $J^*\leq J_{\mathcal{D}}$ if $\mathbb{P}\in\mathcal{D}$. The set $\mathcal{D}$ plays a crucial role in the tractability of the problem under consideration. There have been several proposals in the literature on how to define $\mathcal{D}$. For instance, in references (Lagoa and Barmish, 2002; Shapiro, 2006), it is defined as a set of distributions supported at a single point. In contrast, references (Scarf et al., 1958; Shapiro and Kleywegt, 2002; Popescu, 2007; Delage and Ye, 2010) define $\mathcal{D}$ as the set of distributions satisfying certain moment restrictions or belonging to a particular parametric family of distributions.

Another option is to endow the set of probability distributions with a notion of distance, and define $\mathcal{D}$ as a ball in this metric. This ball is often centered on an empirical distribution $\widehat{\mathbb{P}}_n$, computed from a sample $X_1,\ldots,X_N$ of the random vector X, with the radius chosen such that $\mathbb{P}$ belongs to the ball with high probability or such that the out-of-sample performance of the optimal solution is satisfactory. The choice of the distance metric influences the tractability of the resulting DRO. Commonly used metrics include Burg's entropy (Wang et al., 2016), Kullback-Leibler divergence (Jiang and Guan, 2016), and Total Variation distance (Sun and Xu, 2015). In this work, we adopt the Wasserstein distance and define $\mathcal{D}$ as a ball in this metric centered on the empirical distribution and with a properly chosen radius. Note that if the radius is set to 0 in this approach, we recover the SAA strategy.

Definition 2.1 (Wasserstein distance)**.** *The Wasserstein distance $W_p(\mu,\nu)$ between $\mu,\nu\in\mathcal{P}_p(\Xi)$ is defined by*

$$W_p(\mu,\nu) := \left(\inf_{\Pi\in\mathcal{P}(\Xi\times\Xi)} \left\{\int_{\Xi\times\Xi} \mathbf{d}^p(\xi,\zeta)\Pi(d\xi,d\zeta) \;:\; \Pi(\cdot\times\Xi)=\mu(\cdot),\ \Pi(\Xi\times\cdot)=\nu(\cdot)\right\}\right)^{1/p}$$

where

$$\mathcal{P}_p(\Xi) := \left\{\mu\in\mathcal{P}(\Xi) \;:\; \int_{\Xi} \mathbf{d}^p(\xi,\zeta_0)\mu(d\xi)<\infty \text{ for some } \zeta_0\in\Xi\right\}$$

and d is a metric in Ξ.

W_p defines a metric in $\mathcal{P}_p(\Xi)$ for $p\in[1,\infty)$, hence, the ball with respect to some p-Wasserstein distance with radius $\varepsilon>0$ and center $\mu\in\mathcal{P}(\Xi)$ is given by

$$\mathcal{B}_\varepsilon(\mu) := \{\nu\in\mathcal{P}(\Xi) \mid W_p(\mu,\nu)\leq\varepsilon\}. \tag{4}$$

The Wasserstein distance, also referred to as Earth's moving distance in computer science, the Monge-Kantorovich-Rubinstein distance in physics, and the Optimal Transport distance in optimization, was first defined in (Vasershtein, 1969). Although it arose in various fields of science almost simultaneously, it is known by different names depending on the context.

There are numerous theoretical and practical reasons that make the Wasserstein distance particularly appealing, as highlighted in (Villani, 2003). One of its key advantages is its dual representation, which enables a more tractable equivalent formulation of (3). Specifically, using p-Wasserstein distances we obtain the following problem

$$\min_{x\in\mathbb{X}} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X\sim\mathbb{Q}}[F(x,X)], \tag{5}$$

where $\widehat{\mathbb{P}}_n$ is the empirical measure generated by sample $X_1,\ldots,X_n$ of X, and the ball $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$ is defined with respect to the p-Wasserstein distance in $\mathbb{R}^d$ where the cost function used is $\mathbf{d}=\|\cdot\|_q$ (see definition 2.1). Finally, the internal supremum of problem (5) can be reformulated by the following theorem

Theorem 2.1. *Assume that F is upper semicontinuous with respect to X. Then, for each $x \in \mathcal{X}$, the following is obtained:*

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)} \mathbb{E}_{X\sim\mathbb{Q}}[F(x,X)] = \begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{subject to} & \sup\limits_{y\in\Xi}\left(F(x,y) - \lambda d^p(y,X_i)\right) \le s_i \quad \forall i \in [n], \\ & \lambda \ge 0. \end{cases} \tag{6}$$

The previous theorem is formulated and proved in (Blanchet and Murthy, 2019). However, the reformulation (6) has also been obtained under more restrictive assumptions in (Esfahani and Kuhn, 2018; Luo and Mehrotra, 2019). Another compelling aspect concerns identifying the form of the optimal distributions in (5), assuming they exist. The following corollary, proven in (Gao and Kleywegt, 2022), provides a characterization of these distributions.

Corollary 2.1.1 (Optimal distributions)**.** *Let $x \in \mathcal{X}$ and λ^* be an optimal solution of (6). Provided that the set of optimal distributions of the internal supremum of problem (5) is non-empty, there exists a distribution $\mathbb{Q}^* \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$ that is an optimal solution of the internal supremum of (5) which is supported on at most $n+1$ points and has the form*

$$\mathbb{Q}^* = \frac{1}{n}\sum_{i\neq i_o} \delta_{X_*^i} + \frac{p_0}{n}\delta_{\hat{X}_*^{i_0}} + \frac{1-p_0}{n}\delta_{\check{X}_*^{i_0}}$$

where $i_0 \in [n]$, $p_0 \in [0,1]$, and $\hat{X}_*^{i_0}, \check{X}_*^{i_0} \in \operatorname{argmax}_{y\in\Xi}\left(F(x,y) - \lambda^* \|y - X_{i_0}\|^p\right)$, and $X_*^i \in \operatorname{argmax}_{y\in\Xi}\left(F(x,y) - \lambda^* \|y - X_i\|^p\right)$ for each $i \neq i_0$.

In many cases, depending on the form of the function F, the problem (6) may result in a semi-infinite optimization problem with a large number of variables, which can pose a significant challenge. This is because the supremum appearing in the constraints of (6) may not be solvable explicitly, and it is well-known that solving semi-infinite programs is computationally demanding.

3 Mode estimation as a distributionally robust optimization problem

We return to the context established in Section 1, in which the mode estimation problem is formulated as problem (1). Because f is unknown, the proposal is to approximate this problem with the following stochastic problem:

$$J_{n,h} := \max_{i=1,\ldots,n} \mathbb{E}_{X\sim\mathbb{P}}[f_{h,X_i}(X)] \tag{7}$$

for very small values of $h > 0$, where f_{h,X_i} is any uni-modal density function with mode in X_i such that, if $\mathbb{P}_{h,X_i}$ is a probability distribution with density f_{h,X_i}, then $\mathbb{P}_{h,X_i} \to \delta_{X_i}$, when $h \to 0$, in the sense of weak convergence of measures, and δ_{X_i} is the Dirac delta distribution supported in $\{X_i\}$. Note that taking the limit as h approaches zero, (7) becomes (1). Indeed, we have

$$\begin{aligned} \lim_{h\to 0}\max_{i=1,\ldots,n} \mathbb{E}_{X\sim\mathbb{P}}[f_{h,X_i}(X)] &= \max_{i=1,\ldots,n}\lim_{h\to 0} \mathbb{E}_{X\sim\mathbb{P}}[f_{h,X_i}(X)] \\ &= \max_{i=1,\ldots,n}\lim_{h\to 0} \mathbb{E}_{X\sim\mathbb{P}_{h,X_i}}[f(X)] \\ &= \max_{i=1,\ldots,n} \mathbb{E}_{X\sim\delta_{X_i}}[f(X)] \\ &= \max_{i=1,\ldots,n} f(X_i). \end{aligned}$$

The first inequality is achievable because it involves taking the limit of a maximum, where that maximum is determined over a finite set. Nonetheless, the same could still be valid if that maximum is taken over the entire support of X, even if it is not discrete, as long as the function f has bounded derivatives of a certain order. The second inequality is made possible due to a role exchange between the densities appearing in the integral defining this expected value.

Nevertheless, $\mathbb{P}$ is unknown, so a first alternative is to try to solve a sample average approximation SAA of (7), that is,

$$\widehat{J}_{SAA} := \max_{i=1,\dots,n} \frac{1}{n} \sum_{j=1}^{n} f_{h,X_i}(X_j). \tag{8}$$

However, this approach becomes problematic when h is very small. The issue arises from the fact that multiple indices i can be optimal, or even all of them could be optimal. Consequently, we propose addressing (7) using a Distributionally Robust Optimization (DRO) approach with p-Wasserstein distances where the cost function is $\mathbf{d} = \|\cdot\|$ the Euclidean distance (see Section 2), but with a variation that does not involve a min-max problem. In this case, the proposed approach can be formulated as follows:

$$\widehat{J}_{n,h}(\varepsilon) := \max_{i=1,\dots,n} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)} \mathbb{E}_{X\sim\mathbb{Q}}[f_{h,X_i}(X)] = \max_{i=1,\dots,n} \widehat{J}_{n,h}^{(X_i)}(\varepsilon). \tag{9}$$

where

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) := \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)} \mathbb{E}_{X\sim\mathbb{Q}}[f_{h,x}(X)] \tag{10}$$

In this case, the mode estimator is the point in the sample that maximizes the functional $\widehat{J}_{n,h}^{(x)}(\varepsilon)$. Additionally, it can be observed that $\widehat{J}_{n,h}(\varepsilon)$ might be considered an upper estimator of $J_{n,h}$ since, when $\mathbb{P} \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$, $\widehat{J}_{n,h}(\varepsilon) \geq J_{n,h}$ holds true. However, ensuring that $\mathbb{P} \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$ is not the primary concern. Instead, the focus is on obtaining a reasonable mode estimator. To achieve this goal, it is not necessary to guarantee this specific condition.

Based on the aforementioned discussion, the functional $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ holds a significant role in the proposed strategy. Consequently, it becomes necessary to rewrite this functional in a manner that yields a tractable problem. With this in mind, Theorem 2.1 enables us to express $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ as the following optimization problem:

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) := \begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} s_j \\ \text{subject to} & \sup\limits_{y\in\Xi} \left(f_{h,x}(y) - \lambda \left\|X_j - y\right\|_2^p\right) \leq s_j \quad \forall j \in [n]. \\ & \lambda \geq 0. \end{cases} \tag{11}$$

The computational complexity of this optimization problem depends on the form of the $f_{h,x}$ function. In the next Section, we explore various types of functions $f_{h,x}$ and their possible reformulations.

4 Reformulations

In this section, we examine two types of functions $f_{h,x}$, Gaussian and Uniform spherical, and for each, we reformulate the functional $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ defined in (11). This analysis aims to demonstrate that the complexity of computing the functional $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ is dependent on the chosen form of the function $f_{h,x}$. Consequently, we investigate each case individually. Although the functions analyzed in this section are not the only ones to which this analysis can be applied, we focus on these two cases due to their prevalence in the statistical context. Specifically, these types of functions, commonly referred to as kernels, are often encountered in the context of kernel density estimation. However, it should be noted that the use and purpose of these functions in this work differ from their conventional applications in that context.

4.1 Gaussian case

In the first case under consideration, $f_{h,x}(y)$ is assumed to be Gaussian. In this instance, the density function can be expressed as follows:

$$f_{h,x}(y) := \frac{1}{h^d(2\pi)^{d/2}} e^{-\frac{\|x-y\|^2}{2h^2}}.$$

The following theorem establishes a reformulation of problem (11) for this case.

Theorem 4.1. *In the case of the Gaussian densities, we assume $\Xi = \mathbb{R}^d$ and $p \geq 2$. Then the problem (11) is equivalent to*

$$\begin{cases} \inf\limits_{\lambda,s} & \lambda\varepsilon^2 + \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} s_j \\ \text{subject to} & \sup\limits_{t\in[0,1]} \left(\dfrac{1}{h^d(2\pi)^{\frac{d}{2}}} e^{-\frac{(t\|x-X_j\|)^2}{2h^2}} - \lambda \|x - X_j\|^p (1-t)^p \right) \leq s_j \quad \forall j \in [n]. \\ & \lambda \geq 0. \end{cases} \tag{12}$$

Furthermore, if $\lambda^*_{n,h}(\varepsilon)$ is optimal solution in (12), then optimal measure of (10) has n supports $X^*_1(\varepsilon,h), X^*_2(\varepsilon,h), \ldots, X^*_n(\varepsilon,h)$ given by $X^*_j(\varepsilon,h) = (X_j - x)T^*_{n,h}(\varepsilon)_j + x$ where

$$T^*_{n,h}(\varepsilon)_j := \operatorname*{argmax}_{t\in[0,1]} \left(\frac{1}{h^d(2\pi)^{\frac{d}{2}}} e^{-\frac{(t\|x-X_j\|)^2}{2h^2}} - \lambda^*_{n,h}(\varepsilon) \|x - X_j\|^p (1-t)^p \right)$$

for each $j \in [n]$.

This theorem could potentially be extended to accommodate cases where Ξ solely fulfills the convexity condition. However, such an extension would necessitate incorporating additional technical details into the proof. The proof of this Theorem can be consulted in Section A.1.1.

Based on this theorem, it can be deduced that working with Gaussian densities entails addressing a semi-infinite optimization problem, as shown in (12). The problem is considered semi-infinite since obtaining an analytical solution for the supremum appearing in the constraint is not possible. In addition, in this reformulation, compared to (11), the nature of the semi-infinite constraint, represented by a supremum, is distinct due to its reliance on a one-dimensional set. This difference carries importance because employing cutting plane algorithms becomes the most viable option in this case, (Luo and Mehrotra, 2019). These algorithms require estimating the supremum in each iteration, and given that the supremum is taken over a one-dimensional set, the calculation has the potential to be less complex.

4.2 Uniform spherical case

In this case, the density under consideration is defined by $f_{h,x}(y) := \max\left\{C_{h,d}\mathcal{X}_{\mathcal{B}_h(x)}(y), 0\right\}$, where

$$\mathcal{X}_{\mathcal{B}_h(x)}(y) := \begin{cases} 1 & \text{if } \|x-y\| \leq h, \\ -\infty & \text{if } \|x-y\| > h. \end{cases}$$

For $d = 1$, $C_{h,d} = \frac{1}{2h}$, and for $d \geq 2$, $C_{h,d} = \frac{1}{h^d V_d}$, with V_d representing the volume of a unit $d-$dimensional sphere. It is important to note that the support of this density is the spherical region $\|x - y\| \leq h$, which is the basis for its name.

Before proceeding with the reformulation of (11) for this case, some concepts must be defined. First, the random variables $X^{(x)}_1, \ldots, X^{(x)}_n$ are defined by $X^{(x)}_j := \|x - X_j\|$, representing the distance from point x to the j-th point of the sample. Similarly, we consider the random variables $X^{(x)}_{(1)}, \ldots, X^{(x)}_{(n)}$, which are the order statistics of $X^{(x)}_1, \ldots, X^{(x)}_n$, satisfying $X^{(x)}_{(1)} < X^{(x)}_{(2)} < \cdots < X^{(x)}_{(n)}$. From these variables, we define the expression $k^{(x)}_h$ given by $k^{(x)}_h := \min\left\{k \in [n] \mid h < X^{(x)}_k\right\}$. It is important to note that $k^{(x)}_h$ approaches one as h approaches zero. Furthermore, for each $i = k^{(x)}_h, k^{(x)}_h + 1, \ldots, n$, we define the random variable $Z^{(x)}_i := \frac{1}{n} \sum\limits_{\substack{j=1 \\ X^{(x)}_j \leq X^{(x)}_{(i)}}}^{n} (X^{(x)}_j - h)^p$. To facilitate future calculations, we define $Z^{(x)}_{n+1} := \infty$. The variable $Z^{(x)}_i$ can be interpreted as a truncated average since it is averaged over a subset of the sample.

With these considerations in mind, the theorem enabling the reformulation is as follows.

Theorem 4.2. *Assume $\Xi = \mathbb{R}^d$ and $p \geq 1$. Then the optimal value $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ of (11) is given by*

$$\widehat{J}^{(x)}_{n,h}(\varepsilon) = \frac{C_{h,d}}{n} \left| \left\{ j \in \{k^{(x)}_h, \ldots, n\} \,:\, Z^{(x)}_j \leq \varepsilon^p \right\} \right| + \frac{C_{h,d}}{n} \left(k^{(x)}_h - 1 + \frac{\varepsilon^p - Z^{(x)}_{i^*_-}(\varepsilon,h)}{Z^{(x)}_{i^*_+}(\varepsilon,h) - Z^{(x)}_{i^*_-}(\varepsilon,h)} \right) \tag{13}$$

where $Z^{(x)}_{i^*_+}(\varepsilon,h) := \min\left\{ Z^{(x)}_j \,:\, Z^{(x)}_j > \varepsilon^p, j = k^{(x)}_h, \ldots, n+1 \right\}$ *and* $Z^{(x)}_{i^*_-}(\varepsilon,h) := \max\left\{ Z^{(x)}_j \,:\, Z^{(x)}_j \leq \varepsilon^p, j = k^{(x)}_h, \ldots, n+1 \right\}$.

Furthermore, let $i^*(\varepsilon,h) := \operatorname{argmin}\left\{j \in \{k_h^{(x)},\dots,n+1\} : Z_j^{(x)} > \varepsilon^p\right\}$. *Then the optimal measure of (10) has at least n supports* $X_1^*(\varepsilon,h), X_2^*(\varepsilon,h),\dots,X_n^*(\varepsilon,h)$ *given by* $X_i^*(\varepsilon,h) = \widehat{\xi}_i$ *if* $\|x-X_i\| \leq h$ *or* $\|x-X_i\| \geq \left\|x-X_{i^*(\varepsilon,h)}\right\|$, *and* $X_i^*(\varepsilon,h) = x + \frac{h(x-X_i)}{\|x-X_i\|}$ *if* $h \leq \|x-X_i\| \leq \left\|x-X_{i^*(\varepsilon,h)}\right\|$.

The proof of Theorem 4.2 is provided in Subsubsection A.1.2.

From this theorem, it is observed that $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ is expressed as the sum of two expressions. The first expression tends to be less computationally demanding, while the second expression might initially appear intricate. However, in reality, the impact of the second expression is not significant because it can be demonstrated that when h approaches 0, this expression vanishes. Consequently, the importance of the reformulation lies in the first expression. Taking these aspects into consideration, the uniform spherical case presents a more computationally amenable reformulation when compared to the one obtained in the Gaussian case.

A final aspect to address in this section is that other types of densities $f_{h,x}$ can also achieve formulations that are nearly as tractable as the one derived in this case. For instance, the Epanechikov density supported on the d-dimensional ball of radius h induces a reformulation that involves a similar number of calculations compared to those required in the formulation obtained herein. Nevertheless, the resulting form in the Epanechikov case is not as elegant from an algebraic standpoint.

5 Consistency

In this section, we determine the conditions under which the proposed mode estimator exhibits consistency. Early in this work, it was noted that the density f can possess multiple modes. Our focus lies on the mode with the highest value in the density, and we seek to establish an estimate of this specific mode. To demonstrate the consistency of the proposed estimator, the density f must meet the following conditions.

Assumption 5.1. *Without loss of generality, suppose that* f *has* m *modes. Additionally, assume that* f *satisfies the following conditions:*

1. f *is twice differentiable with these derivatives bounded, that is, there exists* M_f *such that* $|\frac{\partial^2 f(x)}{\partial x_i \partial x_j}| \leq M_f$ *for all* x *and* $i,j = 1,\dots,d$.
2. *Let* x^* *be the mode whose value in density* f, J^*, *is the highest. There exists* $K^* > 0$ *such that the level set* $\mathcal{C}_{K^*} := \{x : f(x) \geq J^* - K^*\}$ *is a connected set, and it satisfies* $\mathcal{C}_{K^*} \cap \mathcal{M} = \{x^*\}$ *and* $\{x \in \mathcal{C}_{K^*} \setminus \{x^*\} : \nabla f(x) = 0\} = \emptyset$ *where* $\mathcal{M}$ *is the set of modes of* f. *This means that* f *is a density such that there exists a connected level set containing* x^* *in which* x^* *is the only critical point of* f.

Let $i_{n,h}^*(\varepsilon)$ be the index that attains the maximum in (9). We propose $X_{i_{n,h}^*(\varepsilon)}$ as the estimator of the sample point with the highest density. In other words, $X_{i_{n,h}^*(\varepsilon)}$ serves as an estimator of X_{i^*}, where $i^* := \underset{i=1,\dots,n}{\operatorname{argmax}} f(X_i)$.

The subsequent theorem establishes bounds for the probabilities that delineate consistency. In this result, we denote $\mathbb{P}^n$ and $\mathbb{E}^n$ as the probability and the expected value taken with respect to the sample of size n.

Theorem 5.1. *Assuming* $p = 2$, *suppose that* $f_{h,x}(y) = \frac{1}{h^d} K\left(\frac{y-x}{h}\right)$ *where* K *is a bounden density function such that* $K(z) \geq 0$ *for all* z, $\int z_i z_j K(z)dz = 0$, $\int z_i K(z)dz = 0$, $\int z_i^2 K(z)dz = 1$ *for all* $i,j = 1,\dots,d$ *with* $i \neq j$, *and* $f_{h,x}$ *is a Lipschitz function with Lipschitz constant of the form* $\frac{L_d}{h^{d+1}}$ *where* L_d *is a constant that depends on the dimension* d. *Then, the following is obtained:*

1. *Suppose that condition 1 of Assumption 5.1 hold, then*

$$\mathbb{P}^n\left(\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \frac{4h^4 M_f^2}{9k^2} + \frac{4\mathcal{T}}{nh^d k^2} + \frac{4L_d^2\varepsilon^2}{h^{2d+2}k^2}. \tag{14}$$

Furthermore, considering $J^* := \max_{x\in\mathbb{R}^d} f(x)$, *and* $\mathcal{Q}_k(\mathbb{P}) := \mathbb{P}\left(J^* - f(\xi) > k\right)$, *then*

$$\mathbb{P}^n\left(\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \mathcal{Q}_{\frac{k}{2}}^n(\mathbb{P}) + \frac{16h^4 M_f^2}{9k^2} + \frac{16\mathcal{T}}{nh^d k^2} + \frac{16L_d^2\varepsilon^2}{h^{2d+2}k^2}. \tag{15}$$

2. *Suppose that all conditions of Assumption 5.1 hold, then there exists a constant R_t that depends on t such that*

$$\mathbb{P}^n\left(\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| > t\right) \leq \mathcal{Q}^n_{R_t}(\mathbb{P}) + \mathcal{Q}^n_{\frac{R_t}{4}}(\mathbb{P}) + \frac{36h^4M_f^2}{R_t^2} + \frac{72\mathcal{T}}{nh^dR_t^2} + \frac{72L_d^2\varepsilon^2}{h^{2d+2}R_t^2}. \tag{16}$$

Furthermore, there exists a constant D_t that depends only on t such that

$$\mathbb{P}^n\left(\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| > t\right) \leq \mathcal{Q}^n_{\frac{D_t}{2}}(\mathbb{P}) + \frac{9h^4M_f^2}{D_t^2} + \frac{18\mathcal{T}}{nh^dD_t^2} + \frac{18L_d^2\varepsilon^2}{h^{2d+2}D_t^2}. \tag{17}$$

All previous inequalities are valid for any $k, t > 0$. Additionally, $\mathcal{T}$ is a constant that depends only on the kernel K and J^*.

Although this theorem provides bounds for the probabilities, it is crucial to understand the rate of convergence concerning n and to establish h, and ε as functions of n. To this end, and simplifying the notation, note that the upper bounds on (14), (15), (16), and (17) are in the form

$$A^n + B^n + Ch^4 + \frac{D}{nh^d} + \frac{E\varepsilon^2}{h^{2d+2}} \tag{18}$$

where A, B, C, D, and E are positive constants, $0 \leq A < 1$, and $0 \leq B < 1$. For instance, the upper bound on (14) can be obtained by considering $A = 0$, $B = 0$, $C = \frac{4M_f^2}{9k^2}$, $D = \frac{4\mathcal{T}}{k^2}$, and $E = \frac{4L_d^2}{k^2}$. Comprehending this is necessary for the following corollary, in which the rate of convergence is determined.

Corollary 5.1.1. *Assuming the same setup as Theorem 5.1, each of the probabilities in (14), (15), (16), and (17) has an order of convergence of $O\left(\frac{1}{n^{\frac{4}{d+4}}}\right)$ when $h = \left(\frac{(D+E\tau)d}{4Cn}\right)^{\frac{1}{d+4}}$ where τ is a positive constant and $\varepsilon = \frac{\varrho_{d,\tau}}{n^{1-\frac{1}{d+4}}}$ with $\varrho_{d,\tau} = \left(\frac{(D+E\tau)d}{4C}\right)^{\frac{d+2}{2d+8}}$.*

The proofs of Theorem 5.1 and Corollary 5.1.1 are relegated to Subsection A.2.

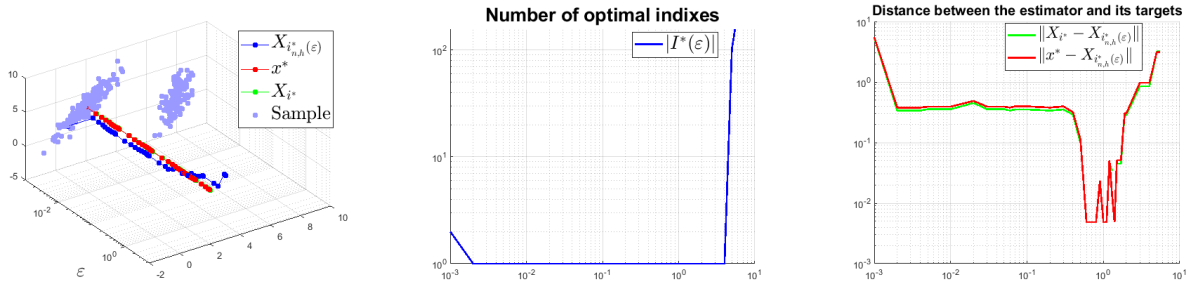
The hypotheses imposed to achieve this consistency result appear to exclude the case where $f_{h,x}$ is a Uniform Spherical density, given that this density is not Lipschitz. However, the same convergence rate can be obtained for this case as well. To accomplish this, it is sufficient to approximate the uniform spherical density by densities that satisfy the hypotheses. In this context, a family of densities that meet these requirements is the Subbotin densities[1].

5.1 One possible direction to improve the consistency rate

In this part, our objective is to propose a potential direction for enhancing the convergence rate demonstrated in Corollary 5.1.1. Specifically, we intend to improve the convergence rate in probability from $X_{i^*_{n,h}(\varepsilon)}$ to x^* and X_{i^*}. To comprehend this, it is essential to recall that the method of obtaining the rate stated in Corollary 5.1.1 involved ensuring that the optimal value $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ of (10) also functions as an estimator of $f(x)$ for any x, as indicated in Theorem A.1. This fact serves as the foundation for proving Theorem 5.1, and subsequently, Corollary 5.1.1. However, this results in the values of h, which guarantee convergence in probability from $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ to $f(x)$, being identical to those ensuring the same convergence from $X_{i^*_{n,h}(\varepsilon)}$ to x^* and X_{i^*}. Moreover, the notion that $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ can be considered an estimator of $f(x)$ implies that the values of h cannot be arbitrarily small. We conjecture that a superior rate of convergence in probability from $X_{i^*_{n,h}(\varepsilon)}$ to x^* and X_{i^*} compared to that stated in Corollary 5.1.1 can be achieved, and the strategy for accomplishing this involves addressing the problem directly without ensuring that $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ is an estimator of $f(x)$. The following example may serve as an illustration of this situation.

Let us consider that the sample $X_1, \cdots, X_n$ is drawn from a mixture of two multivariate normal distributions with density f given by $f := \frac{1}{2}f_1 + \frac{1}{2}f_2$, where f_1 and f_2 are multivariate normal with mean $\mu_1 = [2, 3]$ and covariance matrix $\Sigma^{(1)}$ defined by $\Sigma^{(1)}_{1,1} = 1$, $\Sigma^{(1)}_{1,2} = \Sigma^{(1)}_{2,1} = 1.5$, and $\Sigma^{(1)}_{2,2} = 3$ for f_1, and mean $\mu_2 = [7, 0]$ and covariance matrix $\Sigma^{(2)}$ defined by $\Sigma^{(2)}_{1,1} = 0.3$, $\Sigma^{(2)}_{1,2} = \Sigma^{(2)}_{2,1} = 0.6$, and $\Sigma^{(2)}_{2,2} = 4$ for f_2. In this case, the true mode of f is $[2, 3]$. Additionally, we consider $h = 0.0001$.

[1] Subbotin densities are characterized by functions of the form $f_{h,x}(y) = Ce^{-\frac{\|x-y\|^r}{2h^r}}$, where $r > 0$ and C is a constant that ensures this function is a density. This family of densities bears a resemblance to normal densities and may be referred to by alternative names that associate it with the normal distribution. However, it derives its name from the author who first introduced it in the statistical context Subbotin (1923).



(a) (b) (c)

Figure 5.1: (a) Three-dimensional plot of the performance of the proposed estimator with respect to ε. (b) Number of optimal indexes $|I^*(\varepsilon)|$. (c) Distance between the estimator and the true mode, and the estimator and the highest density point. In this example, $h = 0.0001$, $n = 300$, and Uniform spherical density were used.

Figure 5.1 displays the performance of the estimator $X_{i^*_{n,h}(\varepsilon)}$ as ε varies for the sample described above. To interpret this figure, we denote by $I^*(\varepsilon)$ the set of indices that are optimal indices in (9), so $|I^*(\varepsilon)|$ is the number of indices that are optimal in (9).

Figure 5.1(a) indicates that, for an interval of ε values, the estimator $X_{i^*_{n,h}(\varepsilon)}$ is close to its targets, which are X_{i^*} and x^*, although the latter is not the primary target. However, for the initial ε values, it is observed that $X_{i^*_{n,h}(\varepsilon)}$ is far from X_{i^*} because, for those values, the optimal number of indices $|I^*(\varepsilon)|$ is greater than 1, as shown in Figure 5.1(b). Consequently, the estimators for those values are not reliable. The same occurs for $\varepsilon > 4$, for the same reason. Additionally, Figure 5.1(c) attempts to quantify the proximity of the estimator with respect to X_{i^*} and x^*, demonstrating that for an interval similar to the one mentioned earlier, this distance is small with some distances reaching around 10^{-2}. Some of the results shown in this example can be generalized, which is the objective of the following proposition.

Proposition 5.1. *The problem (9) satisfy the following properties:*

1. *If $\varepsilon \to \infty$ then $|I^*(\varepsilon)| \to n$ for all $h > 0$.*
2. *If $h \to 0$, then $|I^*(0)| \to n$.*

Proof. 1. Note that there exist $\varepsilon^* > 0$ such that all Dirac measures supported at sample points are contained in $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$. That is, there exists $\varepsilon^* > 0$ such that $\delta_{X_i} \in \mathcal{B}_{\varepsilon^*}\left(\widehat{\mathbb{P}}_n\right)$ for all $i = 1, \ldots, n$. However, if δ_{X_i} is on $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)$, that measure will be an optimal measure of the internal maximization problem $\sup_{\mathbb{Q}\in\mathcal{D}} \mathbb{E}_{X\sim\mathbb{Q}}[f_{h,X_i}(X)]$ of (9) for each $i = 1, \ldots, n$, with optimal value $C_{h,d}$ for Uniform spherical case, and $\frac{1}{h^d(2\pi)^{d/2}}$ for Gaussian density case. Therefore, $|I^*(\varepsilon)| = n$ for all $\varepsilon > \varepsilon^*$.

2. Note that $\widehat{J}_n(0) = \widehat{J}_{SAA}$. If $h \to 0$ then $f_{h,X_i}(X_i) \to \infty$ for $i = 1, \ldots, n$. Therefore, $\widehat{J}_{SAA} \to \infty$ and $\underset{i=1,\ldots,n}{\operatorname{argmax}} \frac{1}{n}\sum_{j=1}^n f_{h,X_i}(X_j) \to \{1, \ldots, n\}$, which means that $|I^*(0)| \to n$.
   Furthermore, in the case of Uniform spherical density, if $h < \overline{h}$ where

$$\overline{h} := \min\{\|X_i - X_j\| \mid i, j = 1 \ldots, n \text{ with } i \neq j\},$$

   then $|I^*(0)| = n$.

□

Note that in this example, $h = 0.0001$ was used. However, this value of h tends to result in $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ not being very close to $f(x)$. Indeed, if we focus on x^*, Figure 5.2 shows that the smallest absolute distance between $\widehat{J}^{(x^*)}_{n,h}(\varepsilon)$ and $f(x^*)$ is around 10^4, which is considerably large. The same situation occurs for values of x distinct from x^*. This allows us to infer that this value of h does not make $\widehat{J}^{(x)}_{n,h}(\varepsilon)$ a good estimator of $f(x)$, but it does make $X_{i^*_{n,h}(\varepsilon)}$ an acceptable estimator of x^*.

Considering the discussion thus far, it is natural to inquire how much the convergence rate for the mode estimator can be enhanced. One approach to address this question is to compare the proposed estimator with an estimator that is proven to exhibit a superior convergence rate than the one previously obtained. The objective is to illustrate that the proposed estimator can emulate the performance of an

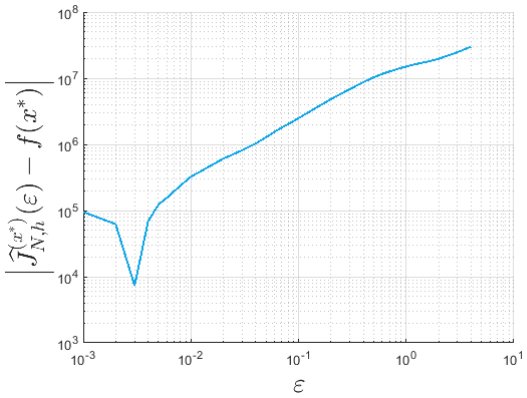


Figure 5.2: Absolute distance between $\widehat{J}_{n,h}^{(x^*)}(\varepsilon)$ and $f(x^*)$ as function of ε.

estimator with a better convergence rate. In this context, the mode estimator investigated in Dasgupta and Kpotufe (2014) emerges as a potential candidate for the estimator with the optimal convergence rate. This estimator is referred to as the k-nearest neighbors density-based estimator.

Definition 5.1 (k-nn density-based estimator)**.** *For every $x \in \mathbb{R}^d$, let $r_{k,n}(x)$ denote the distance from x to its k-th nearest neighbor in $\widehat{\Xi}_n := \{X_1, \ldots, X_n\}$. The density estimate is given as:*

$$f_k(x) := \frac{k}{n \cdot V_d \cdot r_{k,n}(x)^d},$$

where V_d denotes the volume of the unit sphere in $\mathbb{R}^d$. The k-nn density-based mode estimator is defined as $X_{i^*_{n,k_{nn}}} := \underset{x \in X_n}{\operatorname{argmax}} f_k(x)$.

The k-nearest neighbors density-based estimator achieves the minimax rate with $k = O\left(n^{\frac{4}{4+d}}\right)$, where the minimax rate represents the lowest achievable rate. For further discussion on this topic, refer to Dasgupta and Kpotufe (2014). To compare this estimator with the one proposed in this study, a correspondence between k and ε must be established. Specifically, for the case where $f_{h,x}$ is the uniform spherical density, given $k \in \ltimes$, we define ε induced by k as the expression

$$\varepsilon_k := \frac{1}{n} \sum_{j \in I_{k,n}}^{n} \left(\left\| X_j - X_{i^*_{n,k_{nn}}} \right\| - h \right)^p$$

where $I_{k,n} := \left\{ j = 1, \ldots, n \ : \ h \leq \left\| X_j - X_{i^*_{n,k_{nn}}} \right\| \leq r_{k,n}\left(X_{i^*_{n,k_{nn}}} \right) \right\}$.

Consequently, the comparison is made between the estimators $X_{i^*_{n,k_{nn}}}$ and $X_{i^*_{n,h}(\varepsilon_k)}$. Figure 5.3 illustrates the behavior of these estimators. In particular, Figure 5.3(a) displays the expressions $\left\| X_{i^*_{n,k_{nn}}} - x^* \right\|$ (blue) and $\left\| X_{i^*_{n,h}(\varepsilon_k)} - x^* \right\|$ (orange), while Figure 5.3(b) exhibits ε_k. All of these vary as a function of n. In both instances, the shaded regions represent the tube between the 20% and 80% quantiles for 200 simulations. Based on the observations in these figures, it is evident that there exist ε values capable of achieving the same convergence rate as the k-nearest neighbor density-based estimator. The emerging task is to substantiate this finding and identify the differences between the estimators. Indeed, this is one of the issues we intend to investigate further in future research.

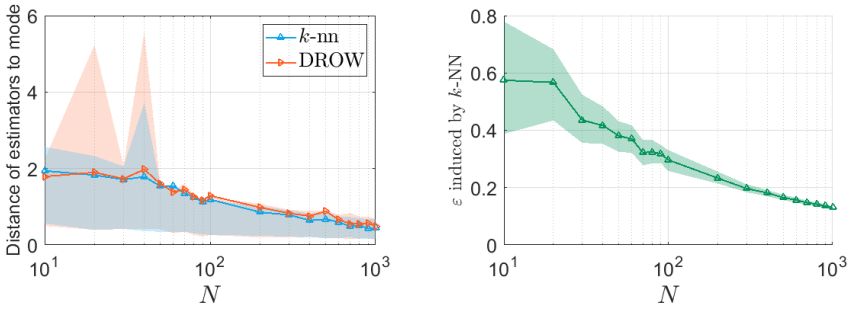
A Proofs of Lemmas and Theorems

A.1 Proofs of the reformulations

A.1.1 Proof of Theorem 4.1

Proof of Theorem 4.1. The strategy of this proof is to reformulate the supremum of the constraints of (11) for this case, this is, for each $j = 1, 2, \ldots, n$, the idea is to reformulate the problem

$$\sup_{y \in \mathbb{R}^d} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{\|x-y\|^2}{2h^2}} - \lambda \|X_j - y\|^p \right). \tag{19}$$



(a) (b)

Figure 5.3: (a) Distance of the estimators to the mode x^*. (b) Value of ε_k. In this case, $k = n^{\frac{4}{4+d}}$.

In fact, we want to show that any solution of (19) belongs to the line of equation $y_t := (X_j - x)t + x$ where $t \in \mathbb{R}$ is a variable. In addition, we also want to show that we can assume that $t \in [0, 1]$. To achieve the former, for each t, we consider the set

$$\begin{aligned}\Lambda_t &:= \left\{ y \in \mathbb{R}^d \;\middle|\; y = rs + y_t \text{ where } s \in \mathbb{R} \text{ y } r \in \mathbb{R}^d \text{ such that } r \perp (X_j - x) \text{ with } \|r\| = 1 \right\} \\ &= \bigcup_{r \in R} \left\{ y \in \mathbb{R}^d \;\middle|\; y = rs + y_t \text{ where } s \in \mathbb{R} \right\}\end{aligned}$$

where $R = \left\{ r \in \mathbb{R}^d \mid r \perp (X_j - x) \text{ with } \|r\| = 1 \right\}$ and $r \perp (X_j - x)$ means that r is perpendicular to $(X_j - x)$.

Note that $\mathbb{R}^d = \bigcup_{t \in \mathbb{R}} \Lambda_t$, and this union is disjoint, so (19) can be rewritten as

$$\begin{aligned}&\sup_{t \in \mathbb{R}} \sup_{y \in \Lambda_t} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{\left((x-y)^T (x-y)\right)}{2h^2}} - \lambda \left((y - X_j)^T (y - X_j) \right)^{p/2} \right) \\ &= \sup_{t \in \mathbb{R}} \sup_{r \in R} \sup_{s \in \mathbb{R}} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{\left((x - rs - y_t)^T (x - rs - y_t)\right)}{2h^2}} - \lambda \left((rs + y_t - X_j)^T (rs + y_t - X_j) \right)^{p/2} \right) \\ &= \sup_{t \in \mathbb{R}} \sup_{r \in R} \sup_{s \in \mathbb{R}} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{\left(s^2 - 2x y_t + \|y_t\|^2 + \|x\|^2\right)}{2h^2}} - \lambda \left(s^2 - 2X_j y_t + \|y_t\|^2 + \|X_j\|^2 \right)^{p/2} \right).\end{aligned}$$

In the innermost supremum of the above expression, the objective function is of one variable, so, for any $t \in \mathbb{R}$ and $r \in R$, we have

$$0 = \operatorname*{argmax}_{s \in \mathbb{R}} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{(s^2 - 2x y_t + \|y_t\|^2 + \|x\|^2)}{2h^2}} - \lambda \left(s^2 - 2X_j y_t + \|y_t\|^2 + \|X_j\|^2 \right)^{p/2} \right)$$

Therefore, (19) can be rewritten as

$$\begin{aligned}&\sup_{t \in \mathbb{R}} \left(\frac{1}{h^d (2\pi)^{d/2}} e^{-\frac{(x - y_t)^T (x - y_t)}{2h^2}} - \lambda \left(y_t - X_j \right)^T (y_t - X_j) \right)^{p/2} \Bigg) \\ &= \sup_{t \in \mathbb{R}} \left(\frac{1}{h^d (2\pi)^{\frac{d}{2}}} e^{-\frac{(t \|x - X_j\|)^2}{2h^2}} - \lambda |t - 1|^p \|x - X_j\|^p \right).\end{aligned}$$

It only remains to be shown that it is correct to change $\mathbb{R}$ to $[0, 1]$. In fact, it is sufficient to show that the function

$$\ell(t) := \frac{1}{h^d (2\pi)^{\frac{d}{2}}} e^{-\frac{(t \|x - X_j\|)^2}{2h^2}} - \lambda |t - 1|^p \|x - X_j\|^p$$

is increasing in the interval $(-\infty, 0)$ and decreasing in the interval $(1, \infty)$. This is an immediate consequence of analyzing the derivative ℓ'. In addition, the final part of this theorem is an immediate consequence of Corollary 2.1.1. □

A.1.2 Proof of Theorem 4.2

Proof of Theorem 4.2. In this case, (11) can be expressed as:

$$\begin{aligned}
\widehat{J}_{n,h}^{(x)}(\varepsilon) &= \inf_{\lambda\geq 0} \lambda\varepsilon^p + \frac{1}{n}\sum_{j=1}^{n} \sup_{y\in\mathbb{R}^d} \left(\max\left\{C_{h,d}\mathcal{X}_{\mathcal{B}_h(x)}(y), 0\right\} - \lambda\left\|z - X_j\right\|^p\right) \\
&= \inf_{\lambda\geq 0} \lambda\varepsilon^p + \frac{1}{n}\sum_{j=1}^{n} \max\left\{\sup_{y\in\mathbb{R}^d}\left(C_{h,d}\mathcal{X}_{\mathcal{B}_h(x)}(y) - \lambda\left\|z - X_j\right\|^p\right), 0\right\} \\
&= \inf_{\lambda\geq 0} \lambda\varepsilon^p + \frac{1}{n}\sum_{\substack{j=1\\ X_j^{(x)}\leq h}}^{n} C_{h,d} + \frac{1}{n}\sum_{\substack{j=1\\ X_j^{(x)}> h}}^{n} \max\left\{C_{h,d} - \lambda\left(X_j^{(x)} - h\right)^p, 0\right\} \\
&= \inf_{\lambda\geq 0} \lambda\varepsilon^p + C_{h,d} - \frac{\lambda}{n}\sum_{\substack{j=1\\ X_j^{(x)}> h}}^{n} \min\left\{\left(X_j^{(x)} - h\right)^p, \frac{C_{h,d}}{\lambda}\right\} \\
&= \inf_{\lambda\geq 0} \lambda\varepsilon^p + C_{h,d} - \frac{\lambda}{n}\sum_{j=k_h^{(x)}}^{n} \min\left\{\left(X_{(j)}^{(x)} - h\right)^p, \frac{C_{h,d}}{\lambda}\right\}. \qquad (20)
\end{aligned}$$

Now, we define $\lambda_{(j)} := \frac{C_{h,d}}{\left(X_{(j)}^{(x)} - h\right)^p}$ and $M_{(j)} := \left|\left\{j \in \{1,\ldots,n\} : X_j^{(x)} \leq X_{(k_h^{(x)}-1)}^{(x)}\right\}\right|$ for each $j = k_h^{(x)}, k_h^{(x)}+1, \ldots, n$. With this consideration and using (20), $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ can be expressed as follows:

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) = \inf_{\lambda\geq 0}\begin{cases}
\lambda\varepsilon^p + \frac{C_{h,d}}{n}M_{(k_h^{(x)}-1)} & \text{if } \lambda_{(k_h^{(x)})} \leq \lambda, \\
\lambda\left(\varepsilon^p - Z_{k_h^{(x)}}^{(x)}\right) + \frac{C_{h,d}}{n}M_{(k_h^{(x)})} & \text{if } \lambda_{(k_h^{(x)}+1)} \leq \lambda < \lambda_{(k_h^{(x)})}, \\
\quad\vdots & \quad\vdots \\
\lambda\left(\varepsilon^p - Z_j^{(x)}\right) + \frac{C_{h,d}}{n}M_{(j)} & \text{if } \lambda_{(j+1)} \leq \lambda < \lambda_{(j)}, \\
\lambda\left(\varepsilon^p - Z_{j+1}^{(x)}\right) + \frac{C_{h,d}}{n}M_{(j+1)} & \text{if } \lambda_{(j+2)} \leq \lambda < \lambda_{(j+1)}, \\
\quad\vdots & \quad\vdots \\
\lambda\left(\varepsilon^p - Z_{n-1}^{(x)}\right) + \frac{C_{h,d}}{n}M_{(n-1)} & \text{if } \lambda_{(n)} \leq \lambda < \lambda_{(n-1)}, \\
\lambda\left(\varepsilon^p - Z_n^{(x)}\right) + C_{h,d} & \text{if } 0 \leq \lambda < \lambda_{(n)}.
\end{cases}$$

From this, it is clear that $\widehat{J}_{n,h}^{(x)}(\varepsilon)$ represents the infimum of a piecewise function, where each component function is linear with respect to λ. Furthermore, the derivative of each component function is given by $\varepsilon^p - Z_j^{(x)}$. Consequently, the infimum occurs at the extreme point where a change of sign appears between the derivatives of two consecutive component functions sharing that extreme. This change of sign depends on the value of ε. In order to find the interval containing the value of λ that achieves the infimum, we must consider that the set of $Z_j^{(x)}$ forms a partition of $\mathbb{R}_{\geq 0}$, as it satisfies

$$0 < Z_{k_h^{(x)}}^{(x)} < Z_{k_h^{(x)}+1}^{(x)} < \cdots < Z_j^{(x)} < Z_{j+1}^{(x)} < \cdots < Z_m^{(x)}.$$

Thus, ε can only reside in one of the intervals of this partition, with each interval associated with one of the component functions.

In fact, if $\varepsilon^p \in \left[0, Z_{k_h^{(x)}}^{(x)}\right)$, then the change of sign in the derivative occurs between the first and second component functions at the extreme point $\lambda^* = C_{h,d} \Big/ \left(X_{k_h^{(x)}}^{(x)} - h\right)^p$. Therefore, the infimum is attained by evaluating λ^* in either of these two component functions, leading to the conclusion that

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) = \frac{C_{h,d}\varepsilon^p}{\left(X_{k_h^{(x)}}^{(x)} - h\right)^p} + \frac{C_{h,d}}{n}M_{(k_h^{(x)}-1)}.$$

Similarly, if $\varepsilon^p \in \left[Z_j^{(x)}, Z_{j+1}^{(x)}\right)$, then the infimum is reached at $\lambda^* = C_{h,d} \Big/ \left(X_{j+1}^{(x)} - h\right)^p$, resulting in

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) = \frac{C_{h,d}\left(\varepsilon^p - Z_j^{(x)}\right)}{\left(X_{j+1}^{(x)} - h\right)^p} + \frac{C_{h,d}}{n} M_{(j)}.$$

Analogously, if $\varepsilon^p \in \left[Z_n^{(x)}, \infty\right)$, then the infimum is reached at $\lambda^* = 0$, resulting in $\widehat{J}_{n,h}^{(x)}(\varepsilon) = C_{h,d}$.

To conclude (13), we perform a small convention consisting in defining $Z_{k_h^{(x)}-1}^{(x)} := 0$ and $Z_{n+1}^{(x)} = \infty$. In that sense, let $i^* \in \{k_h^{(x)} - 1, \ldots, n\}$ be such that $\varepsilon^p \in \left[Z_{i^*}^{(x)}, Z_{i^*+1}^{(x)}\right)$, then we have that

$$M_{(i^*)} = \left|\left\{j \in \{k_h^{(x)}, \ldots, n\} \ : \ Z_j^{(x)} \leq \varepsilon^p\right\}\right| + \left(k_h^{(x)} - 1\right).$$

And additionally, according to definition of $Z_{i_+^*}^{(x)}(\varepsilon, h)$ and $Z_{i_-^*}^{(x)}(\varepsilon, h)$, we have $Z_{i^*}^{(x)} = Z_{i_-^*}^{(x)}(\varepsilon, h)$ and $\left(X_{i^*+1}^{(x)} - h\right)^p = Z_{i_+^*}^{(x)}(\varepsilon, h) - Z_{i_-^*}^{(x)}(\varepsilon, h)$. This allow infer that:

$$\frac{C_{h,d}\left(\varepsilon^p - Z_{i^*}^{(x)}\right)}{\left(X_{i^*+1}^{(x)} - h\right)^p} = \frac{C_{h,d}}{n}\left(\frac{\varepsilon^p - Z_{i_-^*}^{(x)}(\varepsilon, h)}{Z_{i_+^*}^{(x)}(\varepsilon, h) - Z_{i_-^*}^{(x)}(\varepsilon, h)}\right).$$

This results in the conclusion presented in (13). The only remaining aspect to analyze is the support. Following the same line of reasoning, from Corollary 2.1.1, it is deduced that

$$X_j^*(\varepsilon, h) = \sup_{y \in \mathbb{R}^d} \left(\max\left\{C_{h,d}\mathcal{X}_{\mathcal{B}_h(x)}(y), 0\right\} - \lambda \left\|z - X_j\right\|^p\right) \tag{21}$$

$$= \begin{cases} X_j & \text{if } \|x - X_j\| \leq h \text{ or } \|x - X_j\| > \left\|x - Z_{i^*+1}^{(x)}\right\|, \\ x + \frac{h(X_j - x)}{\|x - X_j\|} & \text{if } h < \|x - X_j\| \leq \left\|x - Z_{i^*+1}^{(x)}\right\|. \end{cases} \tag{22}$$

The proof is completed with the observation that $i^* + 1 = i^*(\varepsilon, h)$. □

A.2 Proof of estimator consistency

In this part of this work, we establish the conditions under which the proposed mode estimator is consistent. This is equivalent to showing Theorem 5.1 and Corollary 5.1.1. To this end, we analyze the internal problem

$$\widehat{J}_{n,h}^{(x)}(\varepsilon) := \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_n\right)} \mathbb{E}_{X \sim \mathbb{Q}}[f_{h,x}(X)]$$

for any x. The first objective is to show that the optimal value of this problem converges in probability to $f(x)$. To achieve this goal the following lemma is a necessary result.

Lemma A.1. *Assuming $p = 2$, suppose that condition 1 of Assumption 5.1 hold, and $f_{h,x}(y) = \frac{1}{h^d} K\left(\frac{y-x}{h}\right)$ where $K(z) \geq 0$ or all z, $\int z_i z_j K(z) dz = 0$, $\int z_i K(z) dz = 0$, and $\int z_i^2 K(z) dz = 1$ for all $i, j = 1, \ldots, d$ with $i \neq j$. Then*

$$\mathbb{E}^n\left[\mathbb{E}_{X \sim \widehat{\mathbb{P}}_n}[f_{h,x}(X)] - f(x)\right] \leq \frac{h^2 M_f}{3}.$$

Proof.

$$\begin{aligned}
\mathbb{E}^n\left[\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - f(x)\right] &= \mathbb{E}_{X\sim\mathbb{P}}\left[f_{h,x}(X) - f(x)\right] \\
&= \int f_{h,x}(y)f(y)dy - f(x) \\
&= \int \frac{1}{h^d}K\left(\frac{y-x}{h}\right)f(y)dy - f(x) \\
&= \int K(z)\, f(x+hz)dz - f(x) \\
&= \int K(z)\left(f(x) + h\nabla f(x)z + \right. \\
&\qquad \left. h^2\int_0^1 (1-s)^2 z^T H_f(x+zsh)zds\right)dz - f(x)
\end{aligned}$$

where $H_f(x+zsh)$ is the Hessian matrix of f evaluated at $x+zsh$. Due to the hypotheses, we obtain

$$\begin{aligned}
\mathbb{E}_{X\sim\mathbb{P}}\left[f_{h,x}(X) - f(x)\right] &= h^2\int K(z)\int_0^1 (1-s)^2 z^T H_f(x+zsh)zdsdz \\
&\leq \frac{h^2 M_f}{3}
\end{aligned} \tag{23}$$

□

The following theorem establishes conditions for finding an upper bound of the probability that defines convergence in probability.

Theorem A.1. *Assuming the same hypotheses of Lemma A.1, suppose that $f_{h,x}$ is a Lipschitz function with Lipschitz constant of the form $\frac{L_d}{h^{d+1}}$ where L_d is a constant that depends on the dimension d. Additionally, assume that $\int K^2(u)du < \infty$. Then*

$$\mathbb{P}^n\left(\left|f(x) - \widehat{J}^{(x)}_{n,h}(\varepsilon)\right| > k\right) \leq \frac{2h^4M_f^2}{9k^2} + \frac{2\mathcal{T}}{nh^dk^2} + \frac{2L_d^2\varepsilon^2}{h^{2d+2}k^2}.$$

where $\mathcal{T} = J^*\int K^2(u)du$. Note that the above bound does not depend on x.

Proof. Let $\mathbb{Q}^*_{n,h}$ be an optimal measure of $\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_n)}\mathbb{E}_{X\sim\mathbb{Q}}[f_{h,x}(X)]$. Note that $\mathbb{Q}^*_{n,h}$ depends on the sample. Therefore, considering the following probability and expected values with respect to the random in $\mathbb{Q}^*_{n,h}$, and using Markov's inequality, we obtain

$$\begin{aligned}
&\mathbb{P}^n\left(\left|f(x) - \widehat{J}^{(x)}_{n,h}(\varepsilon)\right| > k\right) \\
&= \mathbb{P}^n\left(\left|f(x) - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right| > k\right) \\
&\leq \frac{1}{k^2}\mathbb{E}^n\left[\left(f(x) - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right] \\
&\leq \frac{2}{k^2}\left(\mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - f(x)\right)^2\right] + \mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right]\right) \\
&= \frac{2}{k^2}\left(\left(\mathbb{E}^n\left[\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - f(x)\right]\right)^2 + \mathrm{Var}\left[\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)]\right]\right. \\
&\qquad \left. + \mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right]\right) \\
&\leq \frac{2}{k^2}\left(\frac{h^4M_f^2}{9} + \mathrm{Var}\left[\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)]\right] + \mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right]\right) \qquad (24) \\
&\leq \frac{2}{k^2}\left(\frac{h^4M_f^2}{9} + \frac{\mathcal{T}}{nh^d} + \mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right]\right) \qquad (25)
\end{aligned}$$

The inequality (24) is due to Lemma A.1, and (25) is inferred from the fact that $\text{Var}\left[\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)]\right] \leq \frac{\mathcal{T}}{nh^d}$ where $\mathcal{T} = J^* \int K^2(u)du$ assuming $\int K^2(u)du < \infty$ which is satisfied by the types of kernels proposed in this work.

To upper bound the term $\mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right]$ we write it in terms of the 2-Wasserstein distance. In that sense, from Theorem 4.1 of Villani (2003) it follows that there exists a coupling Π between $\widehat{\mathbb{P}}_n$ and $\mathbb{Q}^*_{n,h}$ such that

$$W_2^2(\mathbb{Q}^*_{n,h}, \widehat{\mathbb{P}}_n) = \mathbb{E}_{(X,Z)\sim\Pi}\left[\|X - Z\|^2\right] \leq \varepsilon^2$$

With this in mind, the following follows

$$\begin{aligned}
\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2 &= \left(\mathbb{E}_{(X,Z)\sim\Pi}\left[f_{h,x}(X) - f_{h,x}(Z)\right]\right)^2 \\
&\leq \mathbb{E}_{(X,Z)\sim\Pi}\left[\left(f_{h,x}(X) - f_{h,x}(Z)\right)^2\right] \\
&\leq \frac{L_d^2}{h^{2d+2}}\mathbb{E}_{(X,Z)\sim\Pi}\left[\|X - Z\|^2\right] \\
&\leq \frac{L_d^2}{h^{2d+2}}\varepsilon^2
\end{aligned}$$

Therefore, calculating the expected value the following is obtained

$$\mathbb{E}^n\left[\left(\mathbb{E}_{X\sim\widehat{\mathbb{P}}_n}[f_{h,x}(X)] - \mathbb{E}_{X\sim\mathbb{Q}^*_{n,h}}[f_{h,x}(X)]\right)^2\right] \leq \frac{L_d^2}{h^{2d+2}}\varepsilon^2.$$

□

The above result holds when the supreme is taken. This is what the following corollary establishes.

Corollary A.1.1. *Assuming the same setup as Theorem A.1, then*

$$\mathbb{P}^n\left(\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \frac{4h^4M_f^2}{9k^2} + \frac{4\mathcal{T}}{nh^dk^2} + \frac{4L_d^2\varepsilon^2}{h^{2d+2}k^2}. \tag{26}$$

Furthermore, considering $J^* := \max_{x\in\mathbb{R}^d} f(X)$, *and* $\mathcal{Q}_k(\mathbb{P}) := \mathbb{P}\left(J^* - f(X) > k\right)$, *then*

$$\mathbb{P}^n\left(\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \mathcal{Q}^n_{\frac{k}{2}}(\mathbb{P}) + \frac{16h^4M_f^2}{9k^2} + \frac{16\mathcal{T}}{nh^dk^2} + \frac{16L_d^2\varepsilon^2}{h^{2d+2}k^2}.$$

Proof. Taking into account the conventions established after Assumption 5.1, we have that $\widehat{J}^{(X_{i^*_{n,h}(\varepsilon)})}_{n,h}(\varepsilon) = \widehat{J}_{n,h}(\varepsilon)$ and $f(X_{i^*}) = J_n$ where $\widehat{J}_{n,h}(\varepsilon)$ was defined in (9). This helps to simplify the notation. In that sense, we have

$$\begin{aligned}
\left\{\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| \leq k\right\} &= \left\{J_n \leq \widehat{J}_{n,h}(\varepsilon) + k\right\} \cap \left\{\widehat{J}_{n,h}(\varepsilon) \leq J_n + k\right\} \\
&\supseteq \left\{f(X_{i^*}) \leq \widehat{J}^{(X_{i^*})}_{n,h}(\varepsilon) + k\right\} \cap \left\{\widehat{J}^{(X_{i^*_{n,h}(\varepsilon)})}_{n,h}(\varepsilon) \leq f(X_{i^*_{n,h}(\varepsilon)}) + k\right\} \\
&\supseteq \left\{\left|f(X_{i^*}) - \widehat{J}^{(X_{i^*})}_{n,h}(\varepsilon)\right| \leq k\right\} \cap \left\{\left|\widehat{J}^{(X_{i^*_{n,h}(\varepsilon)})}_{n,h}(\varepsilon) - f(X_{i^*_{n,h}(\varepsilon)})\right| \leq k\right\}.
\end{aligned}$$

Therefore, by Theorem A.1, the following is concluded:

$$\begin{aligned}
\mathbb{P}^n\left(\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) &\leq \mathbb{P}^n\left(\left|f(X_{i^*}) - \widehat{J}^{(X_{i^*})}_{n,h}(\varepsilon)\right| > k\right) + \mathbb{P}^n\left(\left|\widehat{J}^{(X_{i^*_{n,h}(\varepsilon)})}_{n,h}(\varepsilon) - f(X_{i^*_{n,h}(\varepsilon)})\right| > k\right) \\
&\leq \frac{4h^4M_f^2}{9k^2} + \frac{4\mathcal{T}}{nh^dk^2} + \frac{4L_d^2\varepsilon^2}{h^{2d+2}k^2}.
\end{aligned}$$

This shows the first part of this corollary. To show the final part, note that

$$\begin{aligned}\left\{\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| \leq k\right\} &\supseteq \left\{|J^* - J_n| + \left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| \leq k\right\} \\ &\supseteq \left\{|J^* - J_n| \leq \frac{k}{2}\right\} \cap \left\{\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| \leq \frac{k}{2}\right\}.\end{aligned}$$

Therefore, we have

$$\mathbb{P}^n\left(\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \mathbb{P}^n\left(|J^* - J_n| > \frac{k}{2}\right) + \mathbb{P}^n\left(\left|J_n - \widehat{J}_{n,h}(\varepsilon)\right| > \frac{k}{2}\right).$$

The second expression on the right-hand side of this inequality can be bounded by the first part of this corollary. To bound the first expression, we have to take into account that $J_n \leq J^*$, so we have the following

$$\begin{aligned}\mathbb{P}^n\left(|J^* - J_n| > \frac{k}{2}\right) &= \mathbb{P}^n\left(J^* - J_n > \frac{k}{2}\right) \\ &= \mathbb{P}^n\left(\min_{i=1,\ldots,n}\{J^* - f(X_i)\} > \frac{k}{2}\right) \\ &= \left(\mathbb{P}\left(J^* - f(X) > \frac{k}{2}\right)\right)^n =: \mathcal{Q}^n_{\frac{k}{2}}(\mathbb{P}). \end{aligned} \tag{27}$$

This allows us to conclude the following

$$\mathbb{P}^n\left(\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > k\right) \leq \mathcal{Q}^n_{\frac{k}{2}}(\mathbb{P}) + \frac{16h^4M_f^2}{9k^2} + \frac{16\mathcal{T}}{nh^dk^2} + \frac{16L_d^2\varepsilon^2}{h^{2d+2}k^2}.$$

□

Up to this point, we have only obtained values for the probability that the distance between $\widehat{J}_{n,h}(\varepsilon)$ and J^* surpasses a specific value k. Nevertheless, we also aim to achieve a similar result for the distance between $X_{i^*_{n,h}(\varepsilon)}$ and x^*. To accomplish this, the following function is crucial. Let $\gamma : [0, K^*] \to \mathbb{R}$ be the function defined by

$$\gamma(k) := \sup\{\|x^* - x\| \ : \ x \in \mathcal{C}_k\}$$

for each $k \in [0, K^*]$ where $\mathcal{C}_k := \{x : f(x) \geq J^* - k\}$ and K^* is given by Assumption 5.1. This function is well defined because for each k the connected level set $\mathcal{C}_k$ is unique. Furthermore, because f is a density, $\mathcal{C}_k$ is a bounded set.

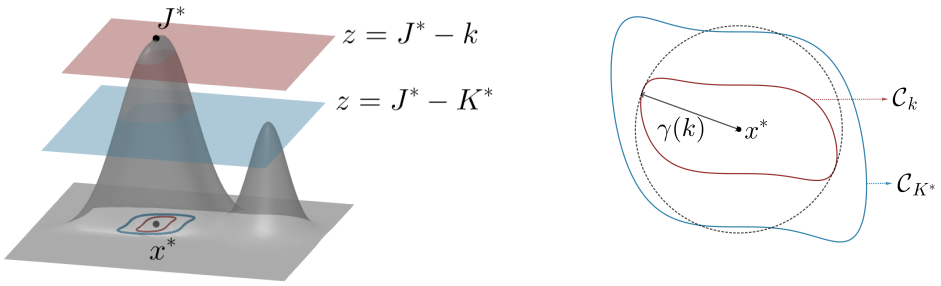
Lemma A.2. *Under Assumption 5.1, the function γ satisfies the following properties:*

1. *γ is increasing in $[0, K^*]$.*
2. *$\lim_{k\to 0^+} \gamma(k) = 0$.*
3. *γ is continuous in $(0, K^*)$.*

Although γ may not be invertible, we define $\gamma^{-1}(t) := \min\{k \in [0, K^*] : \gamma(k) = t\}$. This last function is well-defined due to the previous lemma. Taking all the above into account, we proceed with the proof of Theorem 5.1.

Proof of Theorem 5.1. Item 1 of this theorem follows from Corollary A.1.1. The only remaining task is to demonstrate item 2. Initially, we concentrate on proving (17). In this regard, for $0 \leq t \leq \gamma(K^*)$, note that

$$\begin{aligned}\left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \leq t\right\} &\supseteq \left\{X_{i^*_n(\varepsilon)} \in \mathcal{C}_{\gamma^{-1}(t)}\right\} \\ &= \left\{\left|J^* - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| \leq \gamma^{-1}(t)\right\}. \\ &\supseteq \left\{\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| + \left|\widehat{J}_{n,h}(\varepsilon) - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| \leq \gamma^{-1}(t)\right\}. \\ &\supseteq \left\{\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| \leq \frac{\gamma^{-1}(t)}{2}\right\} \cap \left\{\left|\widehat{J}_{n,h}(\varepsilon) - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| \leq \frac{\gamma^{-1}(t)}{2}\right\}.\end{aligned}$$



(a) (b)

Figure A.1: (a) Example of a density satisfying condition 2 of Assumption 5.1. Graphical intuition of the function γ.

Similarly, for $t > \gamma(K^*)$ we have

$$\begin{aligned}\left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le t\right\} &\supseteq \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \gamma(K^*)\right\} \\ &\supseteq \left\{\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| \le \frac{K^*}{2}\right\} \cap \left\{\left|\widehat{J}_{n,h}(\varepsilon) - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| \le \frac{K^*}{2}\right\}.\end{aligned}$$

Therefore, for all $t > 0$, defining $D_t := \frac{\gamma^{-1}(\min\{t,\gamma(K^*)\})}{2}$, we have

$$\begin{aligned}\left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| > t\right\} &\subseteq \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| > \min\{t,\gamma(K^*)\}\right\} \\ &\subseteq \left\{\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > D_t\right\} \cup \left\{\left|\widehat{J}_{n,h}(\varepsilon) - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| > D_t\right\}.\end{aligned}$$

This allows us to conclude that

$$\begin{aligned}\mathbb{P}^n\left(\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| > t\right) &\le \mathbb{P}^n\left(\left|J^* - \widehat{J}_{n,h}(\varepsilon)\right| > D_t\right) + \mathbb{P}^n\left(\left|\widehat{J}_{n,h}(\varepsilon) - f\left(X_{i^*_{n,h}(\varepsilon)}\right)\right| > D_t\right) \\ &\le \mathcal{Q}^n_{\frac{D_t}{2}}(\mathbb{P}) + \frac{9h^4M_f^2}{D_t^2} + \frac{18\mathcal{T}}{nh^dD_t^2} + \frac{18L_d^2\varepsilon^2}{h^{2d+2}D_t^2}.\end{aligned} \tag{28}$$

The first probability on the right-hand side of the inequality (28) is bounded using Corollary A.1.1, the second is bounded using Theorem A.1.

Finally, it only remains to prove (16). In that sense, for $0 \le t \le 2\gamma(K^*)$ note that

$$\begin{aligned}\left\{\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| \le t\right\} &\supseteq \left\{\|X_{i^*} - x^*\| + \left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le t\right\} \\ &\supseteq \left\{\|X_{i^*} - x^*\| \le \frac{t}{2}\right\} \cap \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \frac{t}{2}\right\}\end{aligned}$$

Similarly, for $t > 2\gamma(K^*)$ we have

$$\begin{aligned}\left\{\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| \le t\right\} &\supseteq \left\{\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| \le 2\gamma(K^*)\right\} \\ &\supseteq \{\|X_{i^*} - x^*\| \le \gamma(K^*)\} \cap \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \gamma(K^*)\right\}.\end{aligned}$$

Therefore, for all $t > 0$, defining $R_t := \gamma^{-1}\left(\min\{t/2, \gamma(k^*)\}\right)$, we have

$$\begin{aligned}\left\{\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| \le t\right\} &\supseteq \left\{\left\|X_{i^*} - X_{i^*_{n,h}(\varepsilon)}\right\| \le \min\{t, 2\gamma(k^*)\}\right\} \\ &\supseteq \{\|X_{i^*} - x^*\| \le \min\{t/2, \gamma(K^*)\}\} \cap \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \min\{t/2, \gamma(K^*)\}\right\} \\ &\supseteq \{X_{i^*} \in \mathcal{C}_{R_t}\} \cap \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \min\{t/2, \gamma(K^*)\}\right\} \\ &\supseteq \{|J^* - J_n| \le R_t\} \cap \left\{\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| \le \min\{t/2, \gamma(K^*)\}\right\}.\end{aligned} \tag{29}$$

This allows us to conclude that

$$\begin{aligned}\mathbb{P}^n\left(\left\|X_{i^*} - X_{i^*_{n}(\varepsilon)}\right\| > t\right) &\le \mathbb{P}^n\left(|J^* - J_n| > R_t\right) + \mathbb{P}^n\left(\left\|x^* - X_{i^*_{n,h}(\varepsilon)}\right\| > \min\{t/2, \gamma(K^*)\}\right) \\ &\le \mathcal{Q}^n_{R_t}(\mathbb{P}) + \mathcal{Q}^n_{\frac{R_t}{4}}(\mathbb{P}) + \frac{36h^4M_f^2}{R_t^2} + \frac{72\mathcal{T}}{nh^dR_t^2} + \frac{72L_d^2\varepsilon^2}{h^{2d+2}R_t^2}.\end{aligned} \tag{30}$$

The first probability on the right-hand side of the inequality (30) is bounded by $\mathcal{Q}^n_{R_t}(\mathbb{P})$, the second is bounded using the bound found in (28), and taking into account that $D_{\min\{t/2,\gamma(K^*)\}} = \frac{R_t}{2}$.

□

The following is the proof of the Corollary 5.1.1.

Proof of Corollary 5.1.1. As said before, simplifying the notation, note that the upper bounds on (14), (15), (16), and (17) are of the form

$$A^n + B^n + Ch^4 + \frac{D}{Nh^d} + \frac{E\varepsilon^2}{h^{2d+2}}$$

where A, B, C, D, and E are positive constants, $0 \leq A < 1$,and $0 \leq B < 1$. Determining the h that minimizes this expression algebraically is not achievable. Consequently, one approach is to attain a convergence rate similar to that of an estimator based solely on Kernel Density Estimation. This can be achieved by taking $\varepsilon^2 = \tau\frac{h^{d+2}}{n}$, where τ represents a positive constant. This results in the following expression:

$$A^n + B^n + Ch^4 + \frac{D + E\tau}{Nh^d}.$$

The minimum value that this expression (18) can take is achieved when $h = \left(\frac{(D+E\tau)d}{4Cn}\right)^{\frac{1}{d+4}}$. Therefore, $\varepsilon = \frac{\varrho_{d,\tau}}{n^{1-\frac{1}{d+4}}}$ where $\varrho_{d,\tau} = \left(\frac{(D+E\tau)d}{4C}\right)^{\frac{d+2}{2d+8}}$, and the smallest upper bound in terms of n that can be aspired is one of the form

$$A^n + B^n + \left(C + \frac{4C}{d}\right)\left(\frac{(D + E\tau)d}{4Cn})\right)^{\frac{4}{d+4}}.$$

Therefore, the fastest convergence that can be obtained is of order $O\left(\frac{1}{n^{\frac{4}{d+4}}}\right)$.

□

References

Y. Aliyari Ghassabeh. A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. *Journal of Multivariate Analysis*, 135:1–10, 2015.

J. Ameijeiras-Alonso, R. Crujeiras, and A. Rodríguez-Casal. Mode testing, critical bandwidth, and excess mass. *TEST*, 28:900–919, 2019.

D. R. Bickel. Robust estimators of the mode and skewness of continuous data. *Computational Statistics & Data Analysis*, 39(2):153–163, 2002.

J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

P. Burman and P. Polonik. Multivariate mode hunting: Data analytic tools with measures of significance. *Journal of Multivariate Analysis*, 100(6):1198–1218, 2009.

A. Casa, J. Chacón, and G. Menardi. Modal clustering asymptotics with applications to bandwidth selection. *Electronic Journal of Statistics*, 14(1):835 – 856, 2020.

H. Chen and P. Meer. Robust computer vision through kernel density. *In Proceedings of the European Conference on Computer Vision*, pages 236–250, 2002.

Y. Chen. Modal regression using kernel density estimation: A review. *WIREs Computational Statistics*, 10(4):e1431, 2018.

Y. Cheng. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

H. Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(3):31–41, 1964.

S. Dasgupta and S. Kpotufe. Optimal rates for k-nn density and mode estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'14*, page 2555–2563. MIT Press, 2014.

E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2010.

P. Esfahani and D. Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2018.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

R. Gao and A. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *Mathematics of Operations Research*, 0(0), 2022.

C. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Non-parametric inference for density modes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 78(1):99–126, 2016.

C. Ho, C. Damien, and S. Walker. Bayesian mode regression using mixtures of triangular densities. *Journal of Econometrics*, 197(2):273–283, 2017.

C. Hsu and T. Wu. Efficient estimation of the mode of continuous multivariate data. *Computational Statistics & Data Analysis*, 63:148 – 159, 2013.

H. Jiang and S. Kpotufe. Modal-set estimation with an application to clustering. *In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:1197–1206, 2017.

R. Jiang and Y. Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158:291–327, 2016.

C. Kamanchi, R. B. Diddigi, R. B. Prabuchandran, and S. Bhatnagar. An online sample-based method for mode estimation using ode analysis of stochastic approximation algorithms. *IEEE Control Systems Letters*, 3(3):697–702, 2019.

G. Kemp and J. Santos-Silva. Regression towards the mode. *Journal of Econometrics*, 170(1):92–101, 2012.

T. Kirschstein, S. Liebscher, G. Porzio, and G. Ragozini. Minimum volume peeling: A robust nonparametric estimator of the multivariate mode. *Computational Statistics & Data Analysis*, 93:456–468, 2016.

C. M. Lagoa and R. B. Barmish. Distributionally robust Monte Carlo simulation. *In Proceedings of the International Federation of Automatic Control World Congress*, pages 1–12, 2002.

J. Lee, J. Li, C. Musco, J. Phillips, and W. Tai. Finding the mode of a kernel density estimate. *arXiv1912.07673*, 2019.

F. Luo and S. Mehrotra. Decomposition algorithm for distributionally robust optimization using Wasserstein metric with an application to a class of regression models. *European Journal of Operational Research*, 278(1):20–35, 2019.

G. Menardi. A review on modal clustering. *International Statistical Review*, 84(3):413–433, 2016.

E. Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962.

I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1): 98–112, 2007.

T. Sager. Estimation of a Multivariate Mode. *The Annals of Statistics*, 6(4):802–812, 1978.

H. Scarf, K. Arrow, and S. Karlin. A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 10:201–209, 1958.

A. Shapiro. Worst-case distribution analysis of stochastic programs. *Mathematical Programming*, 107(1): 91–96, 2006.

A. Shapiro and A. Kleywegt. Minimax analysis of stochastic problems. *Optimizations Methods and Software*, 17(3):523–542, 2002.

B. Silverman. Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society*, 43(1):97–99, 1981.

B. W. Silverman. Density Estimation for Statistics and Data Analysis. *Chapman & Hall.*, 1986.

M. T. Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31:296–301, 1923.

H. Sun and H. Xu. Convergence analysis for distributionally robust optimization and equilibrium problems. *Mathematics of Operations Research*, 41(2):377–401, 2015.

L. N. Vasershtein. Markov processes over denumerable products of spaces describing large system of automata. *Probl. Peredachi Inf.*, 5(3):64–72, 1969.

A. Vedaldi and S. Stefano. Quick shift and kernel methods for mode seeking. *European Conference on Computer Vision*, pages 705–718, 2008.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2003.

Z. Wang, P. Glynn, and Y. Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13:241–261, 2016.

Z. Wanh and D. Scott. Nonparametric density estimation for high-dimensional data—algorithms and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(4), 2019.