On optimal Wasserstein distributionally robust mode estimation

Diego Fonseca DF.FONSECA@UNIANDES.EDU.CO
Universidad de los Andes
Marco Avella Medina MARCO.AVELLA@COLUMBIA.EDU
Columbia University
Mauricio Junca MJ.JUNCA20@UNIANDES.EDU.CO
Universidad de los Andes

Editors: Shipra Agrawal and Aaron Roth

Abstract

We propose a distributionally robust approach to a mode estimation based on a robustification of kernel density estimators. Our method considers the mode of the largest nonparametric density estimator achievable over Wasserstein balls centered around the empirical distribution function. We show that the ∞-Wasserstein case leads to simple tractable estimators that can achieve minimax optimal convergence rates under mild distributional assumptions.

Keywords: Mode estimation, DRO, Wasserstein distance

1. Introduction

In the last few years, there has been a considerable amount of interest in distributionally robust optimization (DRO) formulations of important statistical learning problems. Such formulations provide an interesting framework that explicitly leads to methods that are robust to small changes in the data-generating process. Many recent papers have demonstrated that one can efficiently solve popular estimation problems using DRO approaches based on Wasserstein distances, in some cases recovering some well-known regularized estimators (Blanchet et al., 2019; Gao et al., 2022; Olea et al., 2022; Li et al., 2022; Blanchet et al., 2023).

In this paper, we revisit the problem of mode estimation by considering a Wasserstein DRO formulation of this problem. More specifically, we consider a robustification of the classical mode estimator given by the mode of the kernel density estimator (KDE). Our robustified estimator finds instead the argmax of the largest KDEs that can be achieved over all possible distributions contained in a Wasserstein ball of radius $\varepsilon > 0$ around the empirical distribution. In particular, we show that an ∞-Wasserstein robustification of KDEs defined by kernels supported in compact sets leads to simple mode estimators that roughly correspond to the center of the densest Euclidean ball of radius ε. We show how this general class of mode estimators can be efficiently computed and we provide finite sample estimation error bounds that demonstrate that these mode estimators are minimax optimal for appropriately decreasing sequences of values of ε.

The significance of our study stems from its novel application of concepts from optimal transport theory, specifically the Wasserstein distance, to address a statistical challenge not previously explored through this lens—the estimation of the mode. Beyond simply forging this connection, the methodological framework employed is noteworthy; we leverage stochastic optimization under a Distributionally Robust Optimization (DRO) framework. While the utilization of optimal transport-based distances within DRO methodologies has been documented in various statistical analyses, their application to mode estimation represents a novel endeavor.

© 2024 D. Fonseca, M. Avella Medina & M. Junca.

1.1. Related literature

DISTRIBUTIONALLY ROBUST OPTIMIZATION

Distributionally robust optimization offers an appealing framework for tackling stochastic optimization problems, particularly when the underlying probability distribution is unknown, and only sample data is available. Traditional approaches, like the Sample Average Approximation (SAA) (Shapiro, 2003), directly substitute the expected value with a sample average. While straightforward, the SAA method can be adversely affected by sample contamination and often yields solutions with suboptimal out-of-sample performance, especially with limited data (Esfahani and Kuhn (2018)). The DRO methodology tries to address these limitations by optimizing over a supremum of expected values across a set of probability distributions, termed the ambiguity set. This set is typically conceptualized as a ball centered on the empirical distribution.

In general, DRO problems based on p-Wassertein distances can be hard to solve. Recent work by Esfahani and Kuhn (2018); Gao and Kleywegt (2022); Blanchet and Murthy (2019); Gao et al. (2022) has demonstrated that certain Wasserstein-based DRO problems can often be efficiently solved after some clever reformulations. We will show an analogous result in the context of our robustified mode estimation procedure based on the $p = \infty$ Wasserstein distance.

We note that the robustness notion of DRO is not the same notion usually studied in robust statistics where one is worried about outliers or adversarial contamination. The literature on nonparametric density estimation in this setting is not very extensive. Some representative work in this direction includes (Kim and Scott, 2012; Humbert et al., 2022; Zhang and Ren, 2023). See also Blanchet et al. (2024) for a recent discussion of connections between DRO and robust statistics.

MODE ESTIMATION

The problem of estimating the mode of a probability distribution $\mathbb{P}$ supported on $\mathcal{X}$, equipped with a density function f, finds its origin in the seminal work of Parzen (1962). That work introduced an estimator for the mode as $\arg\max_{x\in\mathcal{X}} f_n(x)$, where f_n is a kernel density estimate of f. This KDE-based mode estimate marks a significant milestone in statistical methodology, offering a framework that has been extensively studied and refined over subsequent decades. Initial proofs of asymptotic normality were introduced by Parzen (1962) for univariate data. This foundational work was later expanded to address multivariate cases, as detailed in Konakov (1974) and Mokkadem and Pelletier (2003). Additional important work studying the asymptotic properties of this estimator include Chernoff (1964); Eddy (1980)

Tsybakov (1990) and Donoho and Liu (1991) independently established that the minimax rate for mode estimation, under the assumption of the density function being β-times differentiable and within a d-dimensional space, is of the order $O(n^{-(\beta-1)/(2\beta+d)})$. This result benchmarks the theoretical best performance achievable by any mode estimator. Klemelä (2005) further advanced the discussion by demonstrating that adaptive KDE-based estimators.

While very intuitive, the finding the mode of the KDE over all $\mathcal{X}$ is not a trivial and motivated the development of alternatives called recursive estimators of the mode Devroye (1979); Tsybakov (1990). Another simple alternative approach is to consider the mode among the points in the observed sample $X_{[n]} = (X_1, \ldots, X_n)$ i.e. $\arg\max_{x\in X_{[n]}} f_n(x)$. This greatly simplifies the optimization to a finite set and in fact preserves the optimality of the mode estimator computed over all of $\mathcal{X}$ Abraham et al. (2003, 2004).

Finally, Dasgupta and Kpotufe (2014) and Arias-Castro et al. (2022) considered very different minimax optimal mode estimators. The former authors established finite sample optimal estimation rates for k-nearest neighbors (k-NN) density estimator and proposed a practical mode estimator based on this approach. The latter considered an adaptive histogram estimator that can also attain optimal finite sample rates.

Our proposed mode estimator leverages the computational framework of the simple estimators of type $\arg\max_{x\in X_{[n]}} f_n(x)$. The key insight is that we view the KDE as an expectation taken over an empirical distribution and view the mode estimation problem as a stochastic optimization problem that we robustify using a Wasserstein approach.

Notation: For any $r > 0$ and $x \in \mathbb{R}^d$, the ball of radius r centered at x with respect to the Euclidean norm is defined as $B_r(x)$. For a natural number $n \in \mathbb{N}$, we use $[n]$ to represent the set $\{1, 2, \ldots, n\}$. Additionally, for any finite set A, the notation $|A|$ represents the cardinality of A. Given a sample $X_1, \ldots, X_n$ from a random variable X, the *empirical distribution* of X with respect to this sample is defined by the probability measure $\widehat{\mathbb{P}}_n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x denotes the Dirac delta function supported at x. We also define $X_{[n]} := \{X_1, \ldots, X_n\}$. For a set $\mathcal{X}$, $\mathcal{P}(\mathcal{X})$ denotes the set of all Borel probability measures on $\mathcal{X}$. Furthermore, for any set $\mathcal{C}$, the indicator function $\mathbf{1}_{\mathcal{C}}$ is defined as $\mathbf{1}_{\mathcal{C}}(y) = 1$ if $y \in \mathcal{C}$, and $\mathbf{1}_{\mathcal{C}}(y) = 0$ if $y \notin \mathcal{C}$.

2. Preliminaries on Wassertein distributionally robust optimization

Distributionally robust optimization provides a framework for guaranteeing local uniform protection from model misspecification to stochastic optimization problems in the form of

$$\min_{x\in\mathcal{X}} \mathbb{E}_{X\sim\mathbb{P}}[F(x, X)], \tag{1}$$

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 $\mathcal{X} \subseteq \mathbb{R}^d$, and $\mathbb{X} \subseteq \mathbb{R}^m$ is a set of constraints on the decision vectors. The Distributionally Robust Optimization (DRO) approach for the problem (1) is formulated as

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

where $\mathcal{D}$ is a set of probability distributions known as *ambiguity set*. The set $\mathcal{D}$ plays a crucial role in the tractability of the problem under consideration. It is common to define this set to be a ball that is 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. 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.

Definition 1 (Wasserstein distance) *The p-Wasserstein distance $W_p(\mathbb{P}, \mathbb{Q})$ between $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\mathcal{X})$ is defined by*

$$W_p(\mathbb{P}, \mathbb{Q}) := \begin{cases} \left(\inf_{\Pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})} \left\{ \int_{\mathcal{X}\times\mathcal{X}} \mathbf{d}^p(\xi, \zeta)\Pi(d\xi, d\zeta) \;:\; \Pi \textit{ has marginals } \mathbb{P}, \mathbb{Q} \right\} \right)^{1/p}, & p \in [1, \infty), \\ \inf_{\Pi\in\mathcal{P}(\mathcal{X}\times\mathcal{X})} \left\{ \Pi - \operatorname*{ess\,sup}_{(\xi,\zeta)\in\mathcal{X}\times\mathcal{X}} \mathbf{d}(\xi, \zeta) \;:\; \Pi \textit{ has marginals } \mathbb{P}, \mathbb{Q} \right\}, & p = \infty, \end{cases}$$

where $\mathcal{P}_p(\mathcal{X}) := \left\{\mathbb{Q} \in \mathcal{P}(\mathcal{X}) \,:\, \int_{\mathcal{X}} \mathbf{d}^p(\xi, \zeta_0)\mathbb{Q}(d\xi) < \infty \text{ for some } \zeta_0 \in \mathcal{X}\right\}$ *for each* $p \in [1, \infty)$, $\mathcal{P}_\infty(\mathcal{X}) = \mathcal{P}(\mathcal{X})$, *and* $\mathbf{d}$ *is a metric in* $\mathcal{X}$.

Given the empirical measure $\widehat{\mathbb{P}}_n$ generated by the sample $X_1, \ldots, X_n$ of X, we can use the Wasserstein distance to define the following Wasserstein-based DRO problem

$$\min_{x \in \mathbb{X}} \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon^p(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X \sim \mathbb{Q}}[F(x, X)], \tag{3}$$

where $\mathcal{B}_\varepsilon^p(\mathbb{P}_n) := \{\mathbb{Q} \in \mathcal{P}(\Xi) \mid W_p(\mathbb{P}_n, \mathbb{Q}) \leq \varepsilon\}$ and $\varepsilon > 0$. The following result shows that the supremum in (3) can be conveniently reformulated for upper semicontinuous losses as shown in (Blanchet and Murthy, 2019, Theorem 1) for $p \in [1, \infty)$, and in (Gao et al., 2022, Lemma EC.2) for $p = \infty$. These results will be instrumental in the analysis of our mode estimators.

Theorem 2 *Assume that* F *is upper semicontinuous with respect to* X. *Then, for each* $x \in \mathcal{X}$ *and* $p \in [1, \infty)$, *the following is obtained:*

$$\sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon^p(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X \sim \mathbb{Q}}[F(x, X)] = \begin{cases} \inf\limits_{\lambda, s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{subject to} & \sup\limits_{y \in \Xi} \left(F(x, y) - \lambda \mathbf{d}^p(y, X_i)\right) \leq s_i \quad \forall i \in [n], \\ & \lambda \geq 0. \end{cases} \tag{4}$$

Furthermore, if $p = \infty$, *then*

$$\sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon^\infty(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X \sim \mathbb{Q}}[F(x, X)] = \frac{1}{n}\sum_{i=1}^{n} \sup_{y \in \Xi} \{F(x, y) \,:\, \mathbf{d}(y, X_i) \leq \varepsilon\}. \tag{5}$$

This theorem indicates that reformulation (5) could offer computational advantages under specific circumstances. The choice of $p = \infty$ in our analysis is motivated by such computational expediency. This particular setting not only simplifies the computational process but also streamlines the finite sample analysis of the mode estimator.

3. Mode estimation as a stochastic optimization problem

Assume $X_{[n]} = (X_1, \ldots, X_n)$ is an i.i.d. random sample with common distribution $\mathbb{P}$ and density function $f : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$. We consider kernel density estimators of $f(x)$ given by

$$\hat{f}_h(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right), \tag{6}$$

where $h > 0$ and $K : \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ is an integrable function such that $\int_{\mathbb{R}^d} uK(u)du = 0$ (Tsybakov, 2008). There are three main mode estimators building on the idea KDEs: the natural estimator $\tilde{x} = \arg\max_{x \in \mathbb{R}^d} \hat{f}_h(x)$ (Parzen, 1962; Eddy, 1980; Grund and Hall, 1995), the computationally more efficient recursive estimators of Devroye (1979); Tsybakov (1990) and the simple mode estimator considered in Abraham et al. (2004). The latter takes the form

$$\hat{x}_h = \arg\max_{x \in X_{[n]}} \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{7}$$

Note that the simple KDE mode estimator can be viewed as a sample average approximation of the stochastic optimization problem

$$\arg\max_{x\in X_{[n]}} \mathbb{E}_{X\sim\mathbb{P}}\left[\frac{1}{h^d}K\left(\frac{X-x}{h}\right)\right]. \tag{8}$$

Building on this observation, we propose a DRO approach to the mode estimation problem using the ∞-Wasserstein distance where we the cost function $\mathbf{d} = \|\cdot\|$ to be the Euclidean distance. More specifically, we define the Wasserstein distributionally robust mode estimator as

$$\hat{x}_{h,\varepsilon} := \arg\max_{x\in X_{[n]}} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon^\infty(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X\sim\mathbb{Q}}\left[\frac{1}{h^d}K\left(\frac{X-x}{h}\right)\right] \tag{9}$$

It will be convenient to introduce the notation

$$\hat{\Pi}_{h,\varepsilon}(x) := \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon^\infty(\widehat{\mathbb{P}}_n)} \mathbb{E}_{X\sim\mathbb{Q}}\left[\frac{1}{h^d}K\left(\frac{X-x}{h}\right)\right]. \tag{10}$$

The next two examples show how to compute $\hat{\Pi}_{h,\varepsilon}(x)$ for specific kernel functions.

Example 1 (Uniform kernel) Consider the kernel function

$$K\left(\frac{y-x}{h}\right) := \begin{cases} \frac{1}{v_d} & \text{if } \|x-y\|\le h \\ 0 & \text{otherwise,}\end{cases} \tag{11}$$

where v_d is the volume of a unit radius d-dimensional sphere. Note that

$$\begin{aligned}\sup_{y\in\mathcal{X}}\left\{K\left(\frac{y-x}{h}\right) \,:\, \|X_i-y\|\le\varepsilon\right\} &= \begin{cases} \frac{1}{v_d} & \text{if } B_h(x)\cap B_\varepsilon(X_i)\neq\emptyset, \\ 0 & \text{if } B_h(x)\cap B_\varepsilon(X_i)=\emptyset.\end{cases} \\ &= \begin{cases} \frac{1}{v_d} & \text{if } \|X_i-x\|\le h+\varepsilon, \\ 0 & \text{if } \|X_i-x\|> h+\varepsilon.\end{cases}\end{aligned}$$

Now, Theorem 2 shows that

$$\widehat{\Pi}_{h,\varepsilon}(x) = \frac{1}{nh^d v_d}\left|\{j\in\{1,2,\ldots,n\} \,:\, \|x-X_j\|\le h+\varepsilon\}\right| = \frac{1}{nh^d v_d}\sum_{i=1}^n \mathbf{1}_{B_{h+\varepsilon}(x)}(X_i).$$

It follows that the proposed Wasserstein distributionally robust mode estimator for the uniform kernel is

$$\hat{x}_{h,\varepsilon} := \arg\max_{x\in X_{[n]}} \sum_{i=1}^n \mathbf{1}_{B_{h+\varepsilon}(x)}(X_i).$$

This estimator finds the point in $X_{[n]}$ that leads to the densest ball of radius $h+\varepsilon$.

Example 2 (Quadratic kernel) Consider the kernel

$$\frac{1}{h^d}K\left(\frac{y-x}{h}\right) := C_{h,d}\max\left\{\left(1-\frac{\|x-y\|^2}{h^2}\right), 0\right\},$$

where

$$C_{h,d} := \begin{cases} \frac{d(d+2)\left(\frac{d}{2}-1\right)!}{4\pi^{d/2}h^d} & \text{if } d \text{ is even.} \\ \frac{d(d+2)(d-2)!}{2^d\pi^{\frac{d-1}{2}}\left(\frac{d-1}{2}-1\right)!h^d} & \text{if } d \text{ is odd.} \end{cases}$$

for $d \geq 2$. When $d = 1$, $C_{h,d}$ is given by $C_{h,d} := \frac{3}{4h}$. We note that for this kernel function

$$\sup_{y\in\mathcal{X}}\left\{\frac{1}{h^d}K\left(\frac{y-x}{h}\right) \,:\, \|y-X_i\| \leq \varepsilon\right\} = \begin{cases} C_{h,d} & \text{if } \|x-X_i\| \leq \max\{\varepsilon,h\}, \\ C_{h,d}\left(1-\frac{(\|x-X_i\|-\varepsilon)^2}{h^2}\right) & \text{if } \max\{\varepsilon,h\} < \|x-X_i\| \leq \varepsilon+h, \\ 0 & \text{if } \|x-X_i\| > \varepsilon+h. \end{cases}$$

It follows from this supremum and Theorem 2 that

$$\widehat{\Pi}_{h,\varepsilon}(x) = \frac{C_{h,d}}{n}\sum_{i=1}^{n}\mathbf{1}_{B_{\max\{\varepsilon,h\}}(x)}(X_i) + \frac{C_{h,d}}{n}\sum_{i=1}^{n}\left(1-\frac{(\|x-X_i\|-\varepsilon)^2}{h^2}\right)\mathbf{1}_{B_{h+\varepsilon}(x)\setminus B_{\max\{\varepsilon,h\}}(x)}(X_i).$$

We note that in this case, the mode estimator will approximately find the point in the sample $X_{[n]}$ that defines the densest ball of radius $\max\{\varepsilon, h\}$.

4. Optimal mode estimation

We will need a few definitions and assumptions in order to provide finite sample rates of convergence for the proposed mode estimator.

Definition 3 *A point $y \in \mathcal{X}$ is designated as a mode of the function f if there exists a radius $\gamma > 0$ such that for all $x \in B_\gamma(y)$, the inequality $f(x) < f(y)$ holds.*

Furthermore, it is assumed that if f possesses a mode, this mode resides in the interior of the support set $\mathcal{X}$. The above definition allows us to define the (global) mode of f as

$$x^* = \arg\max_{x\in\mathcal{X}} f(x)$$

Our analysis also relies on a local regularity condition similar to the ones considered in (Dasgupta and Kpotufe (2014); Arias-Castro et al. (2022)).

Assumption 4.1 *We posit that f has a unique mode x^* and that there exists $r_{x^*} > 0$ and $\lambda > 0$ such that the ball $B_{r_{x^*}}(x^*)$ is within $\mathcal{A}_{x^*} = \{x \,:\, f(x) \geq \lambda\}$ and that $\mathcal{A}_{x^*}$ is a connected set. Furthermore, it is assumed that for any $x \in \mathcal{A}_{x^*}$, the inequality*

$$\check{C}_{x^*}\|x-x^*\|^2 \leq f(x^*) - f(x) \leq \hat{C}_{x^*}\|x-x^*\|^2 \tag{12}$$

holds true for certain constants $\hat{C}_{x^*}, \check{C}_{x^*} > 0$.

It is noteworthy that if the function f is s twice differentiable at x^* and its Hessian matrix at x^*, denoted as $\nabla^2 f(x^*)$, is negative definite, then Assumption 4.1 is satisfied.

In the rest of this section, we will show that any Wasserstein distributionally robust mode estimator of the form (9), computed with a kernel function with bounded support, can be minimax optimal under the above regularity conditions. We will start by establishing this result for the densest ball mode estimator of Example 1. The simple form of this estimator will enable us to highlight the core proof ideas. We then generalize the result to general kernels supported in compact sets by showing that for large n they all behave essentially like the uniform kernel.

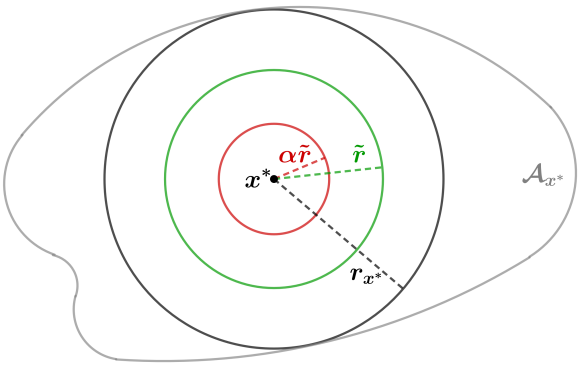

Figure 1: Different regions around a mode x^*.

4.1. Analysis of the robustified uniform kernel mode estimator

A direct consequence of Example 1 is that in the case of the uniform kernel, the robustified mode estimator (9) finds the point in $X_{[n]}$ that minimizes a function of the form

$$\widehat{\Pi}'_{\bar{h}}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{B_{\bar{h}}(x)}(X_i). \tag{13}$$

Note the slight change in notation from Example 1 as it will be convenient for our intuition to analyze $\widehat{\Pi}'_{\bar{h}}(x)$ instead of $\widehat{\Pi}_{h,\varepsilon}$ and consider $\bar{h} := h + \varepsilon$. Therefore, the primary objective of this section is to establish high probability bounds on $\|\hat{x}_{\bar{h}} - x^*\|$, where $\hat{x}_{\bar{h}} = \arg\max_{x \in X_{[n]}} \widehat{\Pi}'_{\bar{h}}(x)$.

To prove the finite sample optimal rates of convergence of our estimator we follow a strategy inspired in Dasgupta and Kpotufe (2014), which consists of showing that with high probability $\widehat{\Pi}'_{\bar{h}}(x)$ is maximized around the mode. Specifically, for any $\delta > 0$, there exist $0 < \alpha < 1$ and $\tilde{r} < r_{x^*}$ such that with probability at least $1 - \delta$

$$\inf_{x \in B_{\alpha\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x) \geq \sup_{x \in \mathcal{X} \setminus B_{\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x). \tag{14}$$

In fact, we show that $\tilde{r} = O\left(\frac{\log(n)\log(2/\delta)}{n}\right)^{\frac{1}{d+4}}$, determining the correct size of $\bar{h}$.

To establish (14) we compare $\widehat{\Pi}'_{\bar{h}}(x)$ with the function:

$$\Pi'_{\bar{h}}(x) := \mathbb{E}[\widehat{\Pi}'_{\bar{h}}(x)] = \mathbb{P}\left(\|X_i - x\| \leq \bar{h}\right) = \int_{B_{\bar{h}}(x)} f(u)du. \tag{15}$$

This is facilitated by Lemma 8, which leverages relative Vapnick-Chervonenkis bounds and is a direct consequence of the results in (Dasgupta and Kpotufe (2014)), though stated in (Bousquet et al. (2004)). This lemma leads to Lemma 4, elucidating the relationship between $\Pi_{\bar{h}}$ and $\widehat{\Pi}_{\bar{h}}$.

Lemma 4 *Given* $0 < \delta < 1$ *and defining* $C_{\delta,n} := \sqrt{\frac{16d\log(n)\log(2/\delta)}{n}}$, *there exists* $n_\delta > 0$ *such that for* $n \geq n_\delta$, *with probability at least* $1 - \delta$, *for each* $x \in \mathcal{X}$, *the following inequalities hold:*

$$\widehat{\Pi}'_{\bar{h}}(x) \leq \Pi'_{\bar{h}}(x) + C^2_{\delta,n} + C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}. \tag{16}$$

$$\widehat{\Pi}'_{\bar{h}}(x) \geq \Pi'_{\bar{h}}(x) - C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}. \tag{17}$$

Next, using Assumption 4.1 we prove the following inequalities with probability at least $1-\delta$

$$\sup_{x\in\mathcal{X}\setminus B_{\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x) \leq \upsilon_d \bar{h}^d \left(f(x^*) - \check{C}_{x^*}(\tilde{r}-\bar{h})^2\right) + C_{\delta,n}^2 + C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}.$$
$$\inf_{x\in B_{\alpha\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x) \geq \upsilon_d \bar{h}^d \left(f(x^*) - \hat{C}_{x^*}(\alpha\tilde{r}+\bar{h})^2\right) - C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}.$$

Finally, after optimizing over the parameter $\bar{h}$ we can obtain inequality (14) and the minimax rate of convergence. The last part is to show that one point in $X_{[n]}$ achieves the desired error, and to this end, we show the following lemma derived also from Lemma 8:

Lemma 5 *Let $0<\delta<1$ and $C_{\delta,n}$ be as defined in Lemma 4. Consider $\gamma>0$ such that $B_\gamma(x^*)\subset\mathcal{X}$, and X a random variable with density f. If $\mathbb{P}\left(X\in B_\gamma(x^*)\right) > C_{\delta,n}^2$, then $\sum_{i=1}^n \mathbf{1}_{B_\gamma(x^*)}(X_i) > 0$ holds with probability at least $1-\delta$.*

Now, we can formally state the main result of our work.

Theorem 6 *Let $\delta>0$. Assuming f has a unique mode denoted x^*, and f satisfies Assumption 4.1, for ease of notation, let $0<\alpha<\sqrt{\frac{\check{C}_{x^*}}{\hat{C}_{x^*}}}$, $A := \check{C}_{x^*} - \alpha^2\hat{C}_{x^*}$, $B = \check{C}_{x^*} + \alpha\hat{C}_{x^*}$, and $D = (\alpha^2+1)^2\check{C}_{x^*}\hat{C}_{x^*}$. Define $\Psi_{d,x^*} := \frac{(4+8\sqrt{f(x^*)})d^2}{\upsilon_d^{1/2}}$ and $\Phi_{d,x^*} := 4\left(dD+2B^2+B\sqrt{4B^2+dD(4+d)}\right)$. Then, there exists $n_\delta>0$ such that for $n\geq n_\delta$ and $\bar{h} = \frac{(d\log(n)\log(2/\delta))^{\frac{1}{d+4}}}{n^{\frac{1}{d+4}}}\left(\frac{\Psi_{d,x^*}}{\Phi_{d,x^*}}\right)^{\frac{2}{d+4}}$, we have:*

$$\mathbb{P}\left(\|\hat{x}_{\bar{h}} - x^*\| \leq \left(\frac{d\log(n)\log(2/\delta)}{n}\right)^{\frac{1}{d+4}} \Lambda_{d,x^*}\right) \geq 1-2\delta \tag{18}$$

where $\Lambda_{d,x^*} := \frac{\Psi_{d,x^*}}{A\Phi_{d,x^*}}\left(B+\sqrt{D+\left(\frac{4+8\sqrt{f(x^*)}}{\upsilon_d^{1/2}}\right)\frac{\Phi_{d,x^*}}{\Psi_{d,x^*}}}\right)$.

The theorem ensures that the consistency rate of our estimator is $O\left((\log(n)/n)^{\frac{1}{d+4}}\right)$ when expressed in terms of the expected value.

4.2. General kernels with bounded support

We now extend the result of the previous subsection to more general bounded-support kernels. We will argue that the similarity between the robust quadratic kernel estimator of Example 2 and that of its uniform counterpart is in fact more general and implies that a wider class of estimators is equivalent from a finite sample error rates perspective.

Assumption 4.2 *Let $K:\mathbb{R}^d\mapsto\mathbb{R}_{\geq 0}$ unimodal such that $\int K(u)du = 1$ and $supp(K)=\mathcal{C}$ is a bounded set of unit diameter. Furthermore, assume that $\sup_u K(u) = K(0)$.*

It follows from Assumption 4.2 and Theorem 2 that for $h < \varepsilon$

$$\widehat{\Pi}_{h,\varepsilon}(x) \leq \frac{K(0)}{n} \sum_{i=1}^{n} \mathbf{1}_{B_{h+\varepsilon}(x)}(X_i) \tag{19}$$

and

$$\widehat{\Pi}_{h,\varepsilon}(x) \geq \frac{K(0)}{n} \sum_{i=1}^{n} \mathbf{1}_{B_{\varepsilon}(x)}(X_i). \tag{20}$$

We denote $\widehat{\Pi}''_{h,\varepsilon}(x) := \frac{\widehat{\Pi}_{h,\varepsilon}(x)}{K(0)}$, so

$$\hat{x}_{h,\varepsilon} = \arg\max_{x \in X_{[n]}} \widehat{\Pi}''_{\bar{h}}(x).$$

Note that $\widehat{\Pi}''_{h,\varepsilon}(x)$ is bounded by $\widehat{\Pi}'_{h+\varepsilon}(x)$ and $\widehat{\Pi}'_{\varepsilon}(x)$ respectively, suggesting that the minimax consistency rate is equally applicable to this type of kernel. This transition hinges on showing a result analogous to Lemma 4.

Lemma 7 *Given $0 < \delta < 1$, defining $C_{\delta,n}$ as in Lemma 4, and considering $\bar{h} = h + \varepsilon$ with $h < \varepsilon$, there exists $n_\delta > 0$ such that for all $n \geq n_\delta$, with probability at least $1 - \delta$, for each $x \in \mathcal{X}$, the following inequalities hold:*

$$\widehat{\Pi}''_{h,\varepsilon}(x) \leq \Pi'_{\bar{h}}(x) + C^2_{\delta,n} + C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}, \tag{21}$$

$$\widehat{\Pi}''_{h,\varepsilon}(x) \geq \Pi'_{\bar{h}}(x) - 3C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}, \tag{22}$$

where $\Pi'_{\bar{h}}$ is defined in (15).

The implications of this lemma are profound in that they facilitate a proof of consistency that parallels the argumentation and methodology employed in Theorem 6. The adaptation required for this proof involves substituting $\widehat{\Pi}'_{\bar{h}}(x)$ with $\widehat{\Pi}''_{h,\varepsilon}(x)$ and adjusting certain constants to accommodate the modification introduced in inequality (22). Consequently, it is established that the mode estimators derived from this generalized kernel formulation inherit the consistency rate ascertained in Theorem 6. It is worth pointing out that even if the estimator obtained with a general kernel has optimal rates, its actual computation will depend heavily on the specific form of K.

Acknowledgments

Mauricio Junca was supported by the Research Fund of the Facultad de Ciencias, Universidad de los Andes INV-2021-128-2307.

References

C. Abraham, G. Biau, and B. Cadre. On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics*, 8(1):1–11, 2004.

Christophe Abraham, Gérard Biau, and Benoit Cadre. Simple estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, 31(1):23–34, 2003.

E. Arias-Castro, W. Qiao, and L. Zheng. Estimation of the global mode of a density: Minimaxity, adaptation, and computational complexity. *Electronic Journal of Statistics*, 16(1):2774 – 2795, 2022.

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

J. Blanchet, Y. Kang, and K. Murthy. Robust Wasserstein profile inference and applications to Machine Learning . *Journal of Applied Probability*, 56(3):830–857, 2019.

J. Blanchet, Y. Kang, J. L. Montiel Olea, V. A. Nguyen, and X. Zhang. Dropout training is distributionally robust optimal. *Journal of Machine Learning Research*, 24(180):1–60, 2023.

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. *Introduction to Statistical Learning Theory*, pages 169–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

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.

Luc Devroye. Recursive estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, pages 159–167, 1979.

D. L. Donoho and R. C. Liu. Geometrizing Rates of Convergence, III. *The Annals of Statistics*, 19 (2):668 – 701, 1991.

W. F. Eddy. Optimum Kernel Estimators of the Mode. *The Annals of Statistics*, 8(4):870 – 882, 1980.

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

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

R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 0(0), 2022.

Birgit Grund and Peter Hall. On the minimisation of L^p error in mode estimation. *Annals of Statistics*, 23(6):2264–2284, 1995.

P. Humbert, B. L. Bars, and L. Minvielle. Robust kernel density estimation with median-of-means principle. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9444–9465. PMLR, 17–23 Jul 2022.

J. Kim and C. D. Scott. Robust kernel density estimation. *Journal of Machine Learning Research*, 13(82):2529–2565, 2012.

J. Klemelä. Adaptive estimation of the mode of a multivariate density. *Journal of Nonparametric Statistics*, 17(1):83–105, 2005.

V. D. Konakov. On the asymptotic normality of the mode of multidimensional distributions. *Theory of Probability & Its Applications*, 18(4):794–799, 1974.

Jiajin Li, Sirui Lin, José Blanchet, and Viet Anh Nguyen. Tikhonov regularization is optimal transport robust under martingale constraints. *Advances in Neural Information Processing Systems*, 35:17677–17689, 2022.

A. Mokkadem and M. Pelletier. The law of the iterated logarithm for the multivariate kernel mode estimator. *ESAIM: Probability and Statistics*, 7:1–21, 2003.

José Luis Montiel Olea, Cynthia Rush, Amilcar Velez, and Johannes Wiesel. On the generalization error of norm penalty linear regression models. *arXiv preprint arXiv:2211.07608*, 2022.

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

A. Shapiro. Monte carlo sampling methods. In *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, pages 353–425. Elsevier, 2003.

Aleksandr Borisovich Tsybakov. Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii*, 26(1):38–45, 1990.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2008.

Peiliang Zhang and Zhao Ren. Adaptive minimax density estimation on $\mathbb{R}^d$ for Huber's contamination model. *Information and Inference: A Journal of the IMA*, 12(4):iaad045, 2023.

Appendix A. Auxiliary Lemmas and proofs

In this part, we present some results that are necessary to prove the Theorem presented in this work.

Lemma 8 *Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ to $\{0,1\}$ with* VC *dimension $d < \infty$, and $\mathbb{P}$ a probability distribution on $\mathcal{X}$. Let $\mathbb{E}$ denote expectation with respect to $\mathbb{P}$. Suppose n points are drawn independently at random from $\mathbb{P}$; let $\mathbb{E}_n$ denote expectation with respect to this sample. Then for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $g \in \mathcal{G}$:*

$$-\min\left(\beta_n\sqrt{\mathbb{E}_n[g]}, \beta_n^2 + \beta_n\sqrt{\mathbb{E}[g]}\right) \leq \mathbb{E}[g] - \mathbb{E}_n[g] \leq \min\left(\beta_n^2 + \beta_n\sqrt{\mathbb{E}_n[g]}, \beta_n\sqrt{\mathbb{E}[g]}\right),$$

where $\beta_n = \sqrt{(4/n)(d\ln 2n + \ln(8/\delta))}$.

Proof of Lemma 4 Applying Lemma 8 with $g = \mathbf{1}_{B_{\bar{h}}(x)}$, we deduce $\mathbb{E}[g] = \Pi'_{\bar{h}}(x)$ and $\mathbb{E}_n[g] = \widehat{\Pi}'_{\bar{h}}(x)$. Utilizing the inequalities $\mathbb{E}[g] - \mathbb{E}_n[g] \leq \beta_n\sqrt{\mathbb{E}[g]}$ and $-(\beta_n^2 + \beta_n\sqrt{\mathbb{E}[g]}) \leq \mathbb{E}[g] - \mathbb{E}_n[g]$ yields:

$$\widehat{\Pi}'_{\bar{h}}(x) \leq \Pi'_{\bar{h}}(x) + \beta_n^2 + \beta_n\sqrt{\Pi_{\bar{h}}(x)}, \tag{23}$$

$$\widehat{\Pi}'_{\bar{h}}(x) \geq \Pi'_{\bar{h}}(x) - \beta_n\sqrt{\Pi_{\bar{h}}(x)}. \tag{24}$$

Observing that $\sqrt{\Pi'_{\bar{h}}(x)} \leq \sqrt{\upsilon_d \bar{d}^d f(x^*)}$, (23) and (24) transform into:

$$\widehat{\Pi}'_{\bar{h}}(x) \leq \Pi'_{\bar{h}}(x) + \beta_n^2 + \beta_n\sqrt{\upsilon_d \bar{d}^d f(x^*)}, \tag{25}$$

$$\widehat{\Pi}_{\bar{h}}(x) \geq \Pi'_{\bar{h}}(x) - \beta_n\sqrt{\upsilon_d \bar{d}^d f(x^*)}. \tag{26}$$

Note that if there exists C such that $\beta_n \leq C$, this ensures (25) and (26) remain valid with β_n replaced by C. Thus, proving the lemma requires demonstrating the existence of n_δ such that if $n \geq n_\delta$ then $\beta_n \leq C_{\delta,n}$. This holds true as:

$$\begin{aligned}
\beta_n \leq C_{\delta,n} &\iff \beta_n^2 \leq C_{\delta,n}^2. \\
&\iff \frac{4}{n}\left(d\log(2n) + \log(8\delta)\right) \leq \frac{16 d \log(n)\log(2/\delta)}{n}. \\
&\iff \log\left(\frac{2^{d+3}}{\delta}\right) \leq \log\left(\frac{2^4}{\delta^4 e}\right)\log(n).
\end{aligned} \tag{27}$$

Given that $0 < \delta < 1$, we can deduce $\log\left(\frac{2^{d+3}}{\delta}\right) > 0$ and $\log\left(\frac{2^4}{\delta^4 e}\right) > 0$, which validates (27) as:

$$\frac{\log\left(\frac{2^{d+3}}{\delta}\right)}{\log\left(\frac{2^4}{\delta^4 e}\right)} \leq \log(n). \tag{28}$$

Consequently, we set $n_\delta := \exp\left(\frac{\log\left(\frac{2^{d+3}}{\delta}\right)}{\log\left(\frac{2^4}{\delta^4 e}\right)}\right)$, completing the proof. ■

Proof of Lemma 5 Employing Lemma 8 with $g = \mathbf{1}_{B_\gamma(x^*)}$, we deduce that $\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}] - \mathbb{E}_n[\mathbf{1}_{B_\gamma(x^*)}] \leq \beta_n\sqrt{\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}]}$. This relation can be reformulated as $\sqrt{\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}]}\left(\sqrt{\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}]} - \beta_n\right) \leq \mathbb{E}_n[\mathbf{1}_{B_\gamma(x^*)}]$. Thus, if $\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}] > \beta_n^2$, it implies that $\mathbb{E}_n[\mathbf{1}_{B_\gamma(x^*)}] > 0$. Given that $C_{\delta,n}^2 \geq \beta_n^2$, it follows that if $\mathbb{E}[\mathbf{1}_{B_\gamma(x^*)}] > C_{\delta,n}^2$, then $\mathbb{E}_n[\mathbf{1}_{B_\gamma(x^*)}] > 0$. ■

Lemma 9 *Let f be a unimodal density function with mode denoted by x^* and $\gamma > 0$ such that $B_\gamma(x^*) \subseteq \mathcal{X}$. Then,*

$$\sup_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x) = \sup_{x\in\partial B_\gamma(x^*)} f(x),$$

where $\partial B_\gamma(x^*)$ denotes the boundary of $B_\gamma(x^*)$.

Proof We argue by contradiction. Assume that the equality does not hold. Then, we have

$$\sup_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x) = \sup_{x\in\mathbf{cl}(\mathcal{X}\setminus B_\gamma(x^*))} f(x) > \sup_{x\in\partial B_\gamma(x^*)} f(x),$$

where $\mathbf{cl}(\cdot)$ denotes the closure. This equality stems from the continuity of f. As a result,

$$\operatorname*{arg\,max}_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x) \subsetneq \mathbf{int}\left(\mathcal{X}\setminus B_\gamma(x^*)\right).$$

where $\mathbf{int}(\cdot)$ denote the interior. Given our assumption that if f has more than one mode, the set of all modes is discrete, it follows that $\operatorname*{arg\,max}_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x)$ must be discrete. This contradicts the unimodality of f, as each element in $\operatorname*{arg\,max}_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x)$ would constitute a local mode.

It is also worth noting that a similar contradiction would arise if $\operatorname*{arg\,max}_{x\in\mathcal{X}\setminus B_\gamma(x^*)} f(x)$ were not discrete. Addressing this scenario would require some additional technical details, which, while not overly complex, would extend this proof. ■

Proof of Theorem 6 We consider $\tilde{r} > 0$ such that $\bar{h} < r_{x^*} - \tilde{r}$ and $\tilde{r} < r_{x^*}$, theses $\tilde{r}$ and $\bar{h}$ depends on n but we will see that dependence later (see Figure 1). Thus, we analyze $\Pi_{\bar{h}}(x)$ for x in subsets

$\mathcal{X} \setminus B_{\tilde{r}}(x^*)$ and $B_{\alpha\tilde{r}}(x^*)$ where $0 < \alpha < 1$. In fact, for each $x \in \mathcal{X} \setminus B_{\tilde{r}}(x^*)$:

$$\begin{aligned}
\Pi'_{\bar{h}}(x) &= \int_{B_{\bar{h}}(x)} f(u)du \\
&\leq \int_{B_{\bar{h}}(x)} \left(\sup_{B_{\bar{h}}(x)} f(w) \right) du \\
&\leq \int_{B_{\bar{h}}(x)} \left(\sup_{\mathcal{X} \setminus B_{\tilde{r}-\bar{h}}(x^*)} f(w) \right) du \\
&= \int_{B_{\bar{h}}(x)} \left(\sup_{\partial B_{\tilde{r}-\bar{h}}(x^*)} f(w) \right) du \qquad (29) \\
&\leq \int_{B_{\bar{h}}(x)} \left(f(x^*) - \check{C}_{x^*}(\tilde{r} - \bar{h})^2 \right) du \\
&\leq \upsilon_d \bar{h}^d f(x^*) - \check{C}_{x^*} \upsilon_d \bar{h}^d (\tilde{r} - \bar{h})^2 \\
&= \upsilon_d \bar{h}^d \left(f(x^*) - \check{C}_{x^*}(\tilde{r} - \bar{h})^2 \right)
\end{aligned}$$

provided that $\tilde{r} - \bar{h} > 0$, where υ_d is the volume of the unit ball in $\mathbb{R}^d$. Inequality (29) is the result of applying Lemma 9. In the same sense, for each $x \in B_{\alpha\tilde{r}}(x^*)$ we have

$$\begin{aligned}
\Pi'_{\bar{h}}(x) &= \int_{B_{\bar{h}}(x)} f(u)du \geq \int_{B_{\bar{h}}(x)} \left(f(x^*) - \hat{C}_{x^*} \|u - x^*\|^2 \right) du \\
&= \upsilon_d \bar{h}^d f(x^*) - \hat{C}_{x^*} \int_{B_{\bar{h}}(x)} \|u - x^*\|^2 du \\
&\geq \upsilon_d \bar{h}^d f(x^*) - \hat{C}_{x^*} \upsilon_d \bar{h}^d (\alpha\tilde{r} + \bar{h})^2 \\
&= \upsilon_d \bar{h}^d \left(f(x^*) - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 \right)
\end{aligned}$$

Therefore, we can infer

$$\sup_{x \in \mathcal{X} \setminus B_{\tilde{r}}(x^*)} \Pi'_{\bar{h}}(x) \leq \upsilon_d \bar{h}^d \left(f(x^*) - \check{C}_{x^*}(\tilde{r} - \bar{h})^2 \right). \qquad (30)$$

$$\inf_{x \in B_{\alpha\tilde{r}}(x^*)} \Pi'_{\bar{h}}(x) \geq \upsilon_d \bar{h}^d \left(f(x^*) - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 \right). \qquad (31)$$

According to inequalities (16) and (17) in Lemma 4 and this last one, the following is obtained

$$\sup_{x \in \mathcal{X} \setminus B_{\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x) \leq \upsilon_d \bar{h}^d \left(f(x^*) - \check{C}_{x^*}(\tilde{r} - \bar{h})^2 \right) + C_{\delta,n}^2 + C_{\delta,n} \sqrt{\upsilon_d \bar{h}^d f(x^*)}. \qquad (32)$$

$$\inf_{x \in B_{\alpha\tilde{r}}(x^*)} \widehat{\Pi}'_{\bar{h}}(x) \geq \upsilon_d \bar{h}^d \left(f(x^*) - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 \right) - C_{\delta,n} \sqrt{\upsilon_d \bar{h}^d f(x^*)}. \qquad (33)$$

each with a probability at least $1 - \delta$. The crux of the proof lies in finding $\tilde{r}$ such that $\tilde{r} > \bar{h}$ and that satisfies the inequality:

$$\upsilon_d \bar{h}^d \left(f(x^*) - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 \right) - C_{\delta,n} \sqrt{\upsilon_d \bar{h}^d f(x^*)} \geq \upsilon_d \bar{h}^d \left(f(x^*) - \check{C}_{x^*}(\tilde{r} - \bar{h})^2 \right) + C_{\delta,n}^2 + C_{\delta,n} \sqrt{\upsilon_d \bar{h}^d f(x^*)}.$$

This is equivalent to proving the inequality

$$\upsilon_d \bar{h}^d \left(\check{C}_{x^*}(\tilde{r} - \bar{h})^2 - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 \right) \geq C_{\delta,n}^2 + 2C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}. \tag{34}$$

Our aim is to deduce conditions on $\bar{r}$ and α that ensure the validity of inequality (34). A preliminary observation reveals that $\check{C}_{x^*}(\tilde{r} - \bar{h})^2 - \hat{C}_{x^*}(\alpha\tilde{r} + \bar{h})^2 > 0$ is a requisite for this inequality, which implies:

$$\alpha < \sqrt{\frac{\check{C}_{x^*}}{\hat{C}_{x^*}}} \quad \text{and} \quad \tilde{r} > \frac{\left(\sqrt{\hat{C}_{x^*}} - \sqrt{\check{C}_{x^*}}\right)}{\left(\sqrt{\check{C}_{x^*}} - \alpha\sqrt{\hat{C}_{x^*}}\right)}\bar{h}.$$

Of these, the constraint $\alpha < \sqrt{\frac{\check{C}_{x^*}}{\hat{C}_{x^*}}}$ is crucial. It leads to $\check{C}_{x^*} - \alpha^2\hat{C}_{x^*} > 0$, a condition pivotal for recasting inequality (34) as an inequality in which we look for the values at which a concave upward polynomial is greater than a given level. Specifically, inequality (34) can be reframed as:

$$\upsilon_d \bar{h}^d \left((\check{C}_{x^*} - \alpha^2\hat{C}_{x^*})\tilde{r}^2 - 2\bar{h}(\check{C}_{x^*} + \alpha\hat{C}_{x^*})\tilde{r} + (\check{C}_{x^*} - \hat{C}_{x^*})\bar{h}^2 \right) \geq C_{\delta,n}^2 + 2C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}. \tag{35}$$

Given that $\bar{h} > 0$, inequality (35) can be equivalently expressed as:

$$(\check{C}_{x^*} - \alpha^2\hat{C}_{x^*})\tilde{r}^2 - 2\bar{h}(\check{C}_{x^*} + \alpha\hat{C}_{x^*})\tilde{r} + (\check{C}_{x^*} - \hat{C}_{x^*})\bar{h}^2 - \frac{1}{\upsilon_d \bar{h}^d}\left(C_{\delta,n}^2 + 2C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)} \right) \geq 0. \tag{36}$$

Inequality (36) brings into focus the aforementioned upward concave polynomial. Subsequent algebraic manipulations yield the condition for the searched $\tilde{r}$:

$$\tilde{r} \geq \frac{1}{(\check{C}_{x^*} - \alpha^2\hat{C}_{x^*})}\left((\check{C}_{x^*} + \alpha\hat{C}_{x^*})\bar{h} + \sqrt{(\alpha^2+1)^2\check{C}_{x^*}\hat{C}_{x^*}\bar{h}^2 + \frac{1}{\upsilon_d \bar{h}^d}\left(C_{\delta,n}^2 + 2C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)} \right)} \right) \tag{37}$$

$$= \frac{1}{A}\left(B\bar{h} + \sqrt{D\bar{h}^2 + \frac{1}{\upsilon_d \bar{h}^d}\left(C_{\delta,n}^2 + 2C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)} \right)} \right)$$

$$= \frac{1}{A}\left(B\bar{h} + \sqrt{D\bar{h}^2 + \frac{16d\log(n)\log(2/\delta)}{\upsilon_d \bar{h}^d n} + \frac{8\sqrt{d\log(n)\log(2/\delta)}}{\bar{h}^{d/2} n^{1/2} \upsilon_d^{1/2}}\sqrt{f(x^*)}} \right) \tag{38}$$

Since the right-hand side of (37) is greater than $\bar{h}$ because $\frac{B}{A} \geq 1$, selecting $\tilde{r}$ in this manner also ensures the satisfaction of the condition $\tilde{r} > \bar{h}$.

Our next endeavor involves optimizing the expression delineated in equation (38), focusing on minimizing it with respect to $\bar{h}$. However, optimizing (38) analytically is not feasible; hence, our strategy pivots to optimizing an upper bound of (38). This approach necessitates an assumption: the existence of $n_{\delta,1}$ such that for all $n \geq n_{\delta,1}$, the following condition is met:

$$\frac{16d\log(n)\log(2/\delta)}{\upsilon_d n} \leq \bar{h}^d. \tag{39}$$

We will later demonstrate that the forthcoming characterization of $\bar{h}$ aligns with this assumption. Under this premise, we can infer:

$$\begin{aligned}
&\frac{1}{A}\left(B\bar{h}+\sqrt{D\bar{h}^2+\frac{16d\log(n)\log(2/\delta)}{\upsilon_d\bar{h}^d n}+\frac{8\sqrt{d\log(n)\log(2/\delta)}}{\bar{h}^{d/2}n^{1/2}\upsilon_d^{1/2}}\sqrt{f(x^*)}}\right)\\
\leq&\frac{1}{A}\left(B\bar{h}+\sqrt{D\bar{h}^2+\frac{\sqrt{16d\log(n)\log(2/\delta)}}{\upsilon_d^{1/2}\bar{h}^{d/2}n^{1/2}}+\frac{8\sqrt{d\log(n)\log(2/\delta)}}{\bar{h}^{d/2}n^{1/2}\upsilon_d^{1/2}}\sqrt{f(x^*)}}\right)\\
=&\frac{1}{A}\left(B\bar{h}+\sqrt{D\bar{h}^2+\frac{\sqrt{d\log(n)\log(2/\delta)}}{\bar{h}^{d/2}n^{1/2}}\left(\frac{4+8\sqrt{f(x^*)}}{\upsilon_d^{1/2}}\right)}\right).
\end{aligned} \tag{40}$$

Observing that (37) is satisfied if

$$\tilde{r}\geq\frac{1}{A}\left(B\bar{h}+\sqrt{D\bar{h}^2+\frac{\sqrt{d\log(n)\log(2/\delta)}}{\bar{h}^{d/2}n^{1/2}}\left(\frac{4+8\sqrt{f(x^*)}}{\upsilon_d^{1/2}}\right)}\right). \tag{41}$$

it becomes evident that our primary objective is to minimize the right-hand side of this inequality with respect to $\bar{h}$. Application of elementary optimization techniques, particularly differentiation followed by setting the derivative to zero, facilitates this process. These steps lead us to conclude that the optimal value of $\bar{h}$, which minimizes the expression in (41), is:

$$\bar{h}=\frac{(d\log(n)\log(2/\delta))^{\frac{1}{d+4}}}{n^{\frac{1}{d+4}}}\left(\frac{\Psi_{d,x^*}}{\Phi_{d,x^*}}\right)^{\frac{2}{d+4}}. \tag{42}$$

Note that this $\bar{h}$ satisfies the assumed condition represented in (39). This is a direct consequence of the fact that this $\bar{h}$ satisfies

$$\lim_{n\to\infty}\frac{16d\log(n)\log(2/\delta)}{\upsilon_d\bar{h}^d n}=\lim_{n\to\infty}\frac{1}{\upsilon_d}\left(\frac{16d\log(n)\log(2/\delta)}{n}\right)^{\frac{4}{d+4}}\left(\frac{\Psi_{d,x^*}}{\Phi_{d,x^*}}\right)^{\frac{-2d}{d+4}}=0.$$

In addition, the minimum value attained by the expression in (41) can be expressed as $\left(\frac{d\log(n)\log(2/\delta)}{n}\right)^{\frac{1}{d+4}}\Lambda_{d,x^*}$. Consequently, this analysis indicates that $\tilde{r}$ must adhere to the following condition:

$$\tilde{r}\geq\left(\frac{d\log(n)\log(2/\delta)}{n}\right)^{\frac{1}{d+4}}\Lambda_{d,x^*}. \tag{43}$$

The inequality (43) could be changed to equality because $\tilde{r}$ should be as small as possible. Moreover, under this setting the condition $\tilde{r}<r_{x^*}$ is satisfied for $n\geq n_{\delta,2}$ where $n_{\delta,2}$ is sufficiently large. Hence, by selecting $\tilde{r}$ in the prescribed manner, we can deduce that

$$\inf_{x\in B_{\alpha\tilde{r}}(x^*)}\widehat{\Pi}'_{\bar{h}}(x)\geq\sup_{x\in\mathcal{X}\setminus B_{\tilde{r}}(x^*)}\widehat{\Pi}'_{\bar{h}}(x)$$

with probability at least $1-\delta$. This inference subsequently implies that $x_{\bar{h}}\in B_{\tilde{r}}(x^*)$ with a probability of at least $1-\delta$, where $x_{\bar{h}}$ is defined as $x_{\bar{h}}:=\arg\max_{x\in\mathcal{X}}\widehat{\Pi}'_{\bar{h}}(x)$.

The proof concludes upon demonstrating that at least one sample element from $X_{[n]}$ belongs to $B_{\alpha\tilde{r}}(x^*)$ with probability at least $1-\delta$. Lemma 5 facilitates this, asserting that it suffices to show that $\mathbb{P}\left(X \in B_{\alpha\tilde{r}}(x^*)\right) > C_{\delta,n}^2$. Since $\mathbb{P}\left(X \in B_{\alpha\tilde{r}}(x^*)\right) \geq \upsilon_d \alpha^d \tilde{r}^d f(x^*)$, we need only to ensure that $\upsilon_d \alpha^d \tilde{r}^d f(x^*) > C_{\delta,n}^2$. This can be formalized as follows:

$$\begin{aligned}\upsilon_d \alpha^d \tilde{r}^d f(x^*) > C_{\delta,n}^2 &\iff \upsilon_d \alpha^d f(x^*) \Lambda_{d,x^*}^d \left(\frac{d\log(n)\log(2/\delta)}{n}\right)^{\frac{d}{d+4}} > \frac{16d\log(n)\log(2/\delta)}{n} \\ &\iff \frac{\upsilon_d \alpha^d f(x^*)\Lambda_{d,x^*}^d}{16 d^{\frac{d}{d+4}}} > \left(\frac{\log(n)\log(2/\delta)}{n}\right)^{\frac{4}{d+4}} \end{aligned} \tag{44}$$

As $\left(\frac{\log(n)\log(2/\delta)}{n}\right)^{\frac{4}{d+4}} \to 0$ when $n \to \infty$, there exists $n_{\delta,3}$ such that for $n \geq n_{\delta,3}$, condition (44) is satisfied. Thus, for these n values, Lemma 5 implies $\sum_{i=1}^n \mathbf{1}_{B_{\alpha\tilde{r}}(x^*)}(X_i) > 0$ with probability at least $1-\delta$, which aligns with our objective in this segment.

Finally, with these facts established, we can infer that

$$\inf_{x \in B_{\alpha\tilde{r}}(x^*) \cap X_{[n]}} \widehat{\Pi}'_{\bar{h}}(x) \geq \sup_{x \in (\mathcal{X} \setminus B_{\tilde{r}}(x^*)) \cap X_{[n]}} \widehat{\Pi}'_{\bar{h}}(x)$$

with probability at least $1-2\delta$, implying that $\hat{x}_{\bar{h}} \in B_{\tilde{r}}(x^*)$ with probability at least $1-2\delta$ for $n \geq n_\delta := \max\{n_{\delta,1}, n_{\delta,2}, n_{\delta,3}\}$, thereby completing the proof. ■

Proof of Lemma 7 Inequality (21) follows from (19) and Lemma 4. To corroborate inequality (22), we reference a preliminary outcome delineated henceforth. Drawing upon Lemma 8, and employing it analogous to its application in the proof of Lemma 4, we deduce the existence of n_{δ_1} for which $n \geq n_{\delta_1}$ ensures:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{B_{h+\varepsilon}(x)\setminus B_\varepsilon(x)}(X_i) &\leq \mathbb{P}\left(X \in B_{h+\varepsilon}(x) \setminus B_\varepsilon(x)\right) + C_{\delta,n}^2 + C_{\delta,n}\sqrt{\upsilon_d((h+\varepsilon)^d - \varepsilon^d) f(x^*)} \\ &\leq \mathbb{P}\left(X \in B_{h+\varepsilon}(x) \setminus B_\varepsilon(x)\right) + C_{\delta,n}^2 + C_{\delta,n}\sqrt{\upsilon_d (h+\varepsilon)^d f(x^*)} \\ &\leq \upsilon_d f(x^*) d(h+\varepsilon)^{d-1} h^d + C_{\delta,n}^2 + C_{\delta,n}\sqrt{\upsilon_d (h+\varepsilon)^d f(x^*)}.\end{aligned}$$

Hence,

$$\widehat{\Pi}''_{h,\varepsilon}(x) \geq \Pi'_{\bar{h}}(x) - \upsilon_d f(x^*) d(h+\varepsilon)^{d-1} h^d - C_{\delta,n}^2 - C_{\delta,n}\sqrt{\upsilon_d (h+\varepsilon)^d f(x^*)} \tag{45}$$

with a probability at least $1-\delta$. Given this framework and considering that $h+\varepsilon$ is delineated as $\mathcal{O}((\log(n)/n)^{1/(d+4)})$ in line with Theorem 6, we note that the term $C_{\delta,n}\sqrt{\upsilon_d \bar{h}^d f(x^*)}$ is $\mathcal{O}((\log(n)/n)^{\frac{d+4}{2d+8}})$. The proof reaches fruition upon demonstrating that the trio of terms bearing negative signs, upon their sign removal, exhibit asymptotic upper bounds that diminish more rapidly or equal than $\mathcal{O}((\log(n)/n)^{\frac{d+4}{2d+8}})$. In fact, this is attainable since $C_{\delta,n}^2$ diminishes at $\mathcal{O}(\log(n)/n)$, and the magnitude of $\upsilon_d f(x^*) d(h+\varepsilon)^{d-1} h^d$ can be modulated by adjusting h, to ensure the aggregate expression wanes quicker than $\mathcal{O}((\log(n)/n)^{\frac{d+4}{2d+8}})$, that adjustment is to consider h of order $O((\log(n)/n)^m)$ with $m > \frac{1}{d+4}$. Consequently, acknowledging these rates of decay surpass

$\mathcal{O}((\log(n)/n)^{\frac{d+4}{2d+8}})$, it is established that there exists for $n_\delta > n_{\delta_1}$ such that for $n \geq n_\delta$ inequality (22) is upheld with a probability surpassing $1 - \delta$. ■