Non-Asymptotic Uniform Rates of Consistency for k-NN Regression

Heinrich Jiang
Google Research
Mountain View, CA

Abstract

We derive high-probability finite-sample uniform rates of consistency for k-NN regression that are optimal up to logarithmic factors under mild assumptions. We moreover show that k-NN regression adapts to an unknown lower intrinsic dimension automatically in the sup-norm. We then apply the k-NN regression rates to establish new results about estimating the level sets and global maxima of a function from noisy observations.

Introduction

The k-nearest neighbor (k-NN) regression algorithm is a classical approach to nonparametric regression. The value of the functional is taken to be the unweighted average observation of the k closest samples. Although this procedure has been known for a long time and has a deep practical significance, there is still surprisingly much about its convergence properties yet to be understood.

We derive finite-sample high probability uniform bounds for k-NN regression under a standard additive model $y = f(x)+\xi$ where f is an unknown function, ξ is sub-Gaussian white noise and y is the noisy observation. The samples $\{(x_i, y_i)\}_{i=1}^n$ are drawn i.i.d. as follows: x_i is drawn according to an unknown density p_X, which shares the same support as f, and then observation y_i is generated by the additive model based on x_i.

We then give simple procedures to estimate the level sets and global maximas of a function given noisy observations and apply the k-NN regression bounds to establish new Hausdorff recovery guarantees for these structures. Each of these results are interesting on their own.

The bulk of the work on k-NN regression convergence theory is on its properties under various risk measures or asymptotic convergence. Notions of consistency involving risk measures such as mean squared error are considerably weaker than the sup-norm as the latter imposes a *uniform* guarantee on the error $|f_k(x) - f(x)|$ where f_k is the k-NN regression estimate of function f. Existing work on studying f_k under the sup-norm thus far are asymptotic. We give the first sup-norm *finite-sample* result. This result matches the minimax optimal rate up to logarithmic factors.

We then discuss the setting where the data lies on a lower dimensional manifold. It is already known that k-NN regression is able to automatically adapt to the intrinsic dimension under various risk measures: the rates depend only on the intrinsic dimension and independent of ambient dimension. We show that this is also the case in the sup-norm: we attain finite-sample bounds as if we were operating in the lower intrinsic dimension space without any modifications to the procedure.

We then show the utility of our k-NN regression results in recovering certain structures of an arbitrary function, namely the level-sets and global maximas. The motivation can be traced back to the rich theory of density-based clustering. There, one is given a finite sample from a probability density p. The clusters can then be modeled based on certain structures in the underlying density p. Such structures include the level-sets $\{x : p(x) \geq \lambda\}$ for some density level λ or the local maximas of p. Then to estimate these, one typically uses a plug-in approach using a density estimator $\widehat{p}$ (e.g. for level-sets, $\{x : \widehat{p}(x) \geq \lambda\}$ and for modes, $\operatorname{argmax}_x \widehat{p}(x)$). It turns out that given uniform bounds on $\widehat{p}$, we can estimate these structures with strong guarantees.

In this paper, instead of estimating these structures in a density, we estimate these structures for a general function f. This is possible because of our established finite-sample sup-norm bounds for nonparametric regression. There are however some key differences in our setting. In the density setting, one has access to i.i.d. samples drawn from the density. Here, we have an i.i.d. sample x drawn from some density p_X not necessarily related to f, and then we obtain a noisy observation of the value $f(x)$. This can be viewed as a noisy observation of the *feature* of x. In other words, we estimate the stuctures based on the features of data, while in the density setting, there are no features and the structures are instead based on the dense regions of the dataset.

Related Works and Contributions

k-NN Regression Rates

The consistency properties of k-NN regression have been studied for a long time and we highlight some of the work here. Biau, Cérou, and Guyader (2010) give guarantees under L_2 risk. Devroye et al. (1994) give consistency guarantees under the L_1 risk. Stone (1977) provides results under

L_p for $p \geq 1$. All these notions of consistency so far are under some integrated risk, and thus are weaker than the sup-norm (i.e. L_∞), which imposes a uniform guarantee.

A number of works such as Mack and Silverman (1982), Cheng (1984), Devroye (1978), Lian (2011), Kudraszow and Vieu (2013) give strong uniform convergence rates. However, these results are asymptotic. Our bounds explore the *finite-sample* consistency properties of k-NN regression, which we will demonstrate later can show strong results about k-NN based learning algorithms which were not possible with existing results. To the best of our knowledge, this is the first such finite-sample uniform consistency result for this procedure, which matches the minimax rate up to logarithmic factors.

We then extend our results to the setting where the data lies on a lower dimensional manifold. This is of practical interest because the curse of dimensionality forces nonparametric methods such as k-NN to require an exponential-in-dimension sample complexity; however as a concession, we can show that many of these methods can have sample complexity depending on the intrinsic dimension (e.g. doubling dimension, manifold dimension, covering number) and independent of the ambient dimension. In modern data applications where the dimension can be arbitrarily high, oftentimes the number of degrees of freedom remains much lower. It thus becomes important to understand these methods under this setting.

Kulkarni and Posner (1995) give results for k-NN regression based on the covering numbers of the support of the distribution. Kpotufe (2011) shows that k-NN regression actually adapts to the local intrinsic dimension without any modifications to the procedure or data in the L_2 norm. In this paper, we show that this holds in the sup-norm as well for a global intrinsic dimension.

Level Set Estimation

Density level-set estimation has been extensively studied and has significant implications to density-based clustering. Some works include Tsybakov (1997) and Singh, Scott, and Nowak (2009). It involves estimating $L_p(\lambda) := \{x : p(x) \geq \lambda\}$ given a finite i.i.d. sample X from p, where λ is some known density level and p is the unknown density. $L_p(\lambda)$ can be seen as the high density regions of the data and thus the connected components can be used as the core-sets in clustering. It can be shown that given a density estimator $\widehat{p}_n$ with guarantees on $|\widehat{p}_n - p|_\infty$, then taking $\widehat{L}_p(\lambda) := \{x \in X : \widehat{p}_n(x) \geq \lambda\}$, the Hausdorff distance between $L_p(\lambda)$ and $\widehat{L}_p(\lambda)$ can also be bounded.

In this paper, we extend this idea to functions f which are not necessarily densities given noisy observations of f. We obtain similar results to those familiar in the density setting, which are made possible by our established bounds for estimating f. An advantage of this approach is that it can be applied to clustering where there are features where clusters are defined as regions of similar feature value rather than similar density. In density-based clustering, it is typical that one does not assume access to the features and thus such procedures fail to readily take advantage of the features when performing clustering. A similar approach was taken by Willett and Nowak (2007) by using nonparametric regression to estimate the level sets of a function; however our consistency results are instead under the Hausdorff metric.

Global Maxima Estimation

We next give an interesting result for estimating the global maxima of a function. Given n i.i.d. samples from some distribution on the input space and seeing a noisy observations of f at the samples, we show a guarantee on the distance between the sample point with the highest k-NN regression value and the (unique) point which maximizes f. This gives us insight into how well a grid search or randomized search can estimate the maximum of a function.

This result can be compared to mode estimation in the density setting where the object is to find the point which maximizes the density function (Tsybakov 1990). Dasgupta and Kpotufe (2014) show that given n draws from a density, the sample point which maximizes the k-NN density estimator is close to the true maximizer of the density; moreover they give finite-sample rates. Earlier works such as Romano (1988) provide asymptotic rates.

k-NN Regression

Throughout the paper, we assume a function f with compact support $\mathcal{X} \subseteq \mathbb{R}^D$ and that we have datapoints $(x_1, y_1), ..., (x_n, y_n)$ drawn follows. The x_i's are drawn i.i.d. from density p_X with support $\mathcal{X}$. Then $y_i = f(x_i) + \xi_{x_i}$ where ξ_{x_i} are i.i.d. drawn according to random variable ξ.

Definition 1. *$f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subseteq \mathbb{R}^D$ is compact.*

The first regularity assumption ensures that the support $\mathcal{X}$ does not become arbitrarily thin anywhere. Otherwise, it becomes impossible to estimate the function in such areas from a random sample.

Assumption 1 (Support Regularity)**.** *There exists $\gamma > 0$ and $r_0 > 0$ such that Vol$(\mathcal{X} \cap B(x, r)) \geq \gamma \cdot$ Vol$(B(x, r))$ for all $x \in \mathcal{X}$ and $0 < r < r_0$.*

The next assumption ensures that with a sufficiently large sample, we will obtain a good covering of the input space.

Assumption 2 (p_X bounded from below)**.** *$p_{X,0} := \inf_{x \in \mathcal{X}} p_X(x) > 0$.*

Finally, we have a standard sub-Gaussian white noise assumption in our additive model.

Assumption 3 (Sub-Gaussian White noise)**.** *ξ satisfies $E[\xi] = 0$ and sub-Gaussian with parameter σ^2 (i.e. $E[\exp(\lambda\xi)] \leq \exp(\sigma^2\lambda^2/2)$ for all $\lambda \in \mathbb{R}$).*

Then define k-NN regression as follows.

Definition 2 (k-NN)**.** *Let the k-NN radius of $x \in \mathcal{X}$ be $r_k(x) := \inf\{r : |B(x, r) \cap X| \geq k\}$ where $B(x, r) := \{x' \in \mathcal{X} : |x - x'| \leq r\}$ and the k-NN set of $x \in \mathcal{X}$ be $N_k(x) := B(x, r_k(x)) \cap X$. Then for all $x \in \mathcal{X}$, the k-NN regression function with respect to the samples is defined as*

$$f_k(x) := \frac{1}{|N_k(x)|} \sum_{i=1}^{n} y_i \cdot 1\left[x_i \in N_k(x)\right].$$

Next, we define the following pointwise modulus of continuity, which will be used to express the bias for an arbitrary function in later result.

Definition 3 (Modulus of continuity)**.** $u_f(x,r) := \sup_{x' \in B(x,r)} |f(x) - f(x')|$.

We now state our main result about k-NN regression. Informally, it says that under the mild assumptions described above, for $k \gtrsim \log n$, $|f_k(x) - f(x)| \lesssim u_f(x, (k/n)^{1/D}) + \sqrt{(\log n)/k}$ uniformly in $x \in \mathcal{X}$ with high probability.

The first term correponds to the bias term. Using uniform VC-type concentration bounds, it can be shown that the k-NN radius can be uniformly bounded by approximately distance $(k/n)^{1/D}$ and hence no point in the k-NN set will be that far. The bias can then be expressed in terms of that distance and u_f.

The second term corresponds to the variance. The $1/\sqrt{k}$ factor is not surprising since the noise terms are averaged over k observations and the extra $\sqrt{\log n}$ factor comes from the cost of obtaining a uniform bound.

Definition 4. *Let v_D be the volume of a D-dimensional unit ball.*

Theorem 1 (k-NN Regression Rate)**.** *Suppose that Assumptions 1, 2, and 3 hold and that*

$$2^8 \cdot D \log^2(4/\delta) \cdot \log n \leq k \leq \frac{1}{2} \cdot \gamma \cdot p_{X,0} \cdot v_D \cdot r_0^D \cdot n.$$

Then probability at least $1-\delta$, the following holds uniformly in $x \in \mathcal{X}$.

$$|f(x) - f_k(x)| \leq u_f \left(x, \left(\frac{2k}{\gamma \cdot p_{X,0} \cdot v_D \cdot n} \right)^{1/D} \right) + 2\sigma \sqrt{\frac{D \log n + \log(2/\delta)}{k}}.$$

Note that the above result is fairly general and makes no smoothness assumptions. In particular, f need not even be continuous. It is also important to point out that n must be sufficiently large in order for there to exist a k that satisfies the conditions. We can then apply this to the class of Hölder continuous functions to obtain the following result.

Corollary 1 (Rate for α-Hölder continuous functions)**.** *Let $0 < \alpha \leq 1$. Suppose that Assumptions 1, 2, and 3 hold and*

$$2^8 \cdot D \log^2(4/\delta) \cdot \log n \leq k \leq \frac{1}{2} \cdot \gamma \cdot v_D \cdot p_{X,0} \cdot r_0^D \cdot n.$$

If f is Hölder continuous (i.e. $|f(x) - f(x')| \leq C_\alpha |x - x'|^\alpha$), then the following holds:

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |f(x) - f_k(x)| \leq C_\alpha \left(\frac{2k}{\gamma \cdot p_{X,0} \cdot v_D \cdot n} \right)^{\alpha/D} + 2\sigma \sqrt{\frac{D \log n + \log(2/\delta)}{k}} \right) \geq 1 - \delta.$$

Remark 1. *Taking $k = O(n^{2\alpha/(2\alpha+D)})$ gives us a rate of*

$$\sup_{x \in \mathcal{X}} |f(x) - f_k(x)|_\infty \lesssim \widetilde{O}(n^{-\alpha/(2\alpha+D)}),$$

which is the minimax optimal rate for estimating a Hölder function, up to logarithmic factors.

Remark 2. *It is understood that all our results will also hold under the assumption that the x_i's are fixed and deterministic (e.g. on a grid) as long as there is a sufficient covering of the space.*

Regression On Manifolds

In this section, we show that if the data has a lower intrinsic dimension, then k-NN will automatically attain rates as if it were in the lower dimensional space and independent of the ambient dimension.

We make the following regularity assumptions which are standard among works in manifold learning e.g. (Genovese et al. 2012) and (Balakrishnan et al. 2013).

Assumption 4. *$\mathcal{P}$ is supported on M where:*

- *M is a d-dimensional smooth compact Riemannian manifold without boundary embedded in compact subset $\mathcal{X} \subseteq \mathbb{R}^D$.*
- *The volume of M is bounded above by a constant.*
- *M has condition number $1/\tau$, which controls the curvature and prevents self-intersection.*

Let p_X be the density of $\mathcal{P}$ with respect to the uniform measure on M.

We now give the manifold analogues of Theorem 1 and Corollary 1.

Theorem 2 (k-NN Regression Rate)**.** *Suppose that Assumptions 2, 3, and 4 hold and that*

$$k \geq 2^8 \cdot D \log^2(4/\delta) \cdot \log n$$
$$k \leq \frac{1}{4} \left(\min \left\{ \frac{\tau}{4d}, \frac{1}{\tau} \right\} \right)^d p_{X,0} \cdot v_d \cdot n.$$

Then with probability at least $1 - \delta$, the following holds uniformly in $x \in \mathcal{X}$.

$$|f(x) - f_k(x)| \leq u_f \left(x, \left(\frac{4k}{v_d \cdot n \cdot p_{X,0}} \right)^{1/d} \right) + 2\sigma \sqrt{\frac{D \log n + \log(2/\delta)}{k}}.$$

Similar to the full dimensional case, we can then apply this to the class of Hölder continuous functions.

Corollary 2 (Rate for α-Hölder continuous functions)**.** *Let $0 < \alpha \leq 1$. Suppose that Assumptions 2, 3, and 4 hold and*

$$k \geq 2^8 \cdot D \log^2(4/\delta) \cdot \log n$$
$$k \leq \frac{1}{4} \left(\min \left\{ \frac{\tau}{4d}, \frac{1}{\tau} \right\} \right)^d p_{X,0} \cdot v_d \cdot n.$$

If f is Hölder continuous (i.e. $|f(x) - f(x')| \leq C_\alpha |x - x'|^\alpha$), then the following holds

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |f(x) - f_k(x)| \leq C_\alpha \left(\frac{4k}{v_d \cdot n \cdot p_{X,0}} \right)^{\alpha/d} + 2\sigma \sqrt{\frac{D \log n + \log(2/\delta)}{k}} \right) \geq 1 - \delta.$$

Remark 3. *Taking $k = O(n^{2\alpha/(2\alpha+d)})$ gives us a rate of $\widetilde{O}(n^{-\alpha/(2\alpha+d)})$, which is more attractive than the full dimensional version $\widetilde{O}(n^{-\alpha/(2\alpha+D)})$ when intrinsic dimension d is lower than ambient dimension D. We note that the bound contains a constant factor depending on D but the rate at which it decreases as n grows does not.*

Level Set Estimation

The level set is the region of the input space that have value greater than a fixed threshold.

Definition 5 (Level-Set)**.**

$$L_f(\lambda) := \{x \in \mathcal{X} : f(x) \geq \lambda\}.$$

In order to estimate the level-sets, we require the following regularity assumption. It states that for each maximal connected component of the level-set, the change in the function around the boundary has a Lipschitz form with smoothness and curvature $\beta > 0$ around some neighborhood of the boundary. This notion of regularity at the boundaries of the level-sets is a standard one in density level-set estimation e.g. Tsybakov; Singh, Scott, and Nowak (2009).

Definition 6 (Level-Set Regularity)**.** *Let $d(x, C) := \inf_{x' \in C} |x - x'|$, ∂C be the boundary of C, and $C \oplus r := \{x' : d(x', C) \leq r\}$. A function f satisfies β-regularity at level λ if the following holds. There exists $r_M, \check{C}, \hat{C} > 0$ such that for each maximal connected subset $C \subseteq L_f(\lambda)$, we have*

$$\check{C} \cdot d(x, \partial C)^\beta \leq |\lambda - f(x)| \leq \hat{C} \cdot d(x, \partial C)^\beta,$$

for all $x \in \partial C \oplus r_M$.

Remark 4. *The upper bound on $|\lambda - f(x)|$ ensures that f is sufficiently smooth so that k-NN regression will give us sufficiently accurate estimates near the boundaries. The lower bound on $|\lambda - f(x)|$ ensures that the level-set is salient enough to be detected.*

To recover $L_f(\lambda)$ based on the samples, we use the following estimator, where $X := \{x_1, ..., x_n\}$.

$$\widehat{L}(\lambda) := \{x \in X : f_k(x) \geq \lambda - \epsilon\},$$

where $\epsilon := 4\hat{\sigma}\sqrt{\frac{D \log n + \log(2/\delta)}{k}}$ and $\hat{\sigma} := \sqrt{\frac{2}{n}\sum_{i=1}^m y_i^2}$. It will become clear later in the proofs that $\hat{\sigma}$ is meant to be an upper bound on σ and thus ϵ is an upper bound on twice the variance of term of the k-NN bound.

There are three simple but key differences of our estimator when compared to $L_f(\lambda)$. The first is that since we don't have access to the true function f, we use the k-NN regression estimate f_k. Next, instead of taking $x \in \mathcal{X}$, we instead restrict to the samples X. This makes our estimator feasible to compute since it will be a subset of the sample points. Finally, we have the ϵ to bound the uniform deviation of $|f_k - f|$ near the boundary of the level-set (as will be apparent in the proof). The main difficulty is choosing ϵ large enough to bound this uniform deviation, but not too large to overestimate the level-set and finally ensuring that ϵ can be computed without knowledge of f or any unknown constants (we only need confidence parameter δ and the dimension, as well as k). Thus, our estimator is practical.

We provide consistency result under the Hausdorff metric. We note that this is a strong notion of consistency since it a uniform guarantee on the constituents of our estimator.

Definition 7 (Hausdorff Distance)**.**

$$d_H(X, Y) = \inf\{\epsilon \geq 0 : X \subseteq Y \oplus \epsilon, Y \subseteq X \oplus \epsilon\}.$$

The next result gives us finite-sample consistency rates for our estimator.

Theorem 3 (Level Set Recovery)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let f be continuous and satisfy β-regularity at level λ. Define $M := \sqrt{\mathbb{E}[y_1^2]}$ where the expectation is taken over p_X and ξ, and suppose that n is sufficiently large depending on ξ, f and δ. If k satisfies*

$$k \geq 8 \max\left\{1, \frac{40M^2}{(2\min\{r_M, r_0\})^{2\beta}\check{C}^2}\right\} \log(4/\delta) D \cdot \log n,$$

$$k \leq (4\sigma^2/\hat{C})^{2D/(2\beta+D)} \cdot (D \log n + \log(4/\delta))^{\beta/(2\beta+D)} \cdot (2\gamma \cdot p_{X,0} \cdot v_D)^{2\beta/(2\beta+D)} \cdot n^{2\beta/(2\beta+D)},$$

then with probability at least $1 - 2\delta$,

$$d_H(L_f(\lambda), \widehat{L}_f(\lambda)) \leq 2 \cdot \left(\frac{24M}{\check{C}}\right)^{1/\beta} \cdot (D \log n \cdot \log(2/\delta))^{1/2\beta} \cdot k^{-1/2\beta}.$$

Remark 5. *Although the statement may appear obfuscated, it essentially says that as long as f is a continuous function satisfying β-regularity at level λ, then if k lies within the following range:*

$$\log n \lesssim k \lesssim n^{2\beta/(2\beta+D)},$$

then with high probability,

$$d_H(L_f(\lambda), \hat{L}_f(\lambda)) \lesssim k^{-1/(2\beta)}.$$

Remark 6. *Choosing k at the optimal setting $k \approx n^{2\beta/(2\beta+D)}$, we have $\epsilon = \widetilde{O}(n^{-\beta/(2\beta+D)})$. Then it follows that we recover the level sets at a Hausdorff rate of $\widetilde{O}(n^{-1/(2\beta+D)})$. This can be compared to the lower bound $O(n^{-1/(2\beta+D)})$ established by Tsybakov (1997) for estimating the level sets of an unknown density.*

We can give a similar result when the data lies on a lower dimensional manifold. Interestingly, we can use the exact same estimator as before as if we were operating in the full dimensional space.

Theorem 4 (Level Set Recovery on Manifolds)**.** *Suppose that Assumptions 1, 2, 3, and 4 hold. Let f be continuous and satisfy β-regularity at level λ. Define $M := \sqrt{\mathbb{E}[y_1^2]}$ where the expectation is taken over p_X and ξ, and suppose that n is sufficiently large depending on ξ, f, τ, and δ. If k satisfies*

$$k \geq 8 \max\left\{1, \frac{40M^2}{(2\min\{r_M, r_0\})^{2\beta}\check{C}^2}\right\} \log(4/\delta) D \cdot \log n,$$

$$k \leq (4\sigma^2/\hat{C})^{2d/(2\beta+d)} \cdot (D \log n + \log(4/\delta))^{\beta/(2\beta+d)} \cdot (p_{X,0} \cdot v_D)^{2\beta/(2\beta+d)} \cdot n^{2\beta/(2\beta+d)},$$

then with probability at least $1-2\delta$,

$$d_H(L_f(\lambda), \widehat{L}_f(\lambda)) \leq 2 \cdot \left(\frac{24M}{\check{C}}\right)^{1/\beta} \cdot (D \log n \cdot \log(2/\delta))^{1/2\beta} \cdot k^{-1/2\beta}.$$

Remark 7. *The main difference from the full-dimensional version is that we need* k *to satisfy*

$$\log n \lesssim k \lesssim n^{2\beta/(2\beta+d)}.$$

Choosing k *at the optimal setting* $k \approx n^{2\beta/(2\beta+d)}$*, we recover the level sets at a rate of* $\widetilde{O}(n^{-1/(2\beta+d)})$.

Remarkably, we obtain the rate as if we were operating on the lower dimensional space. This has not been shown for level-set estimation on manifolds for density functions (which is a different problem).

The rate for density functions under similar regularity assumptions is $\widetilde{O}(n^{-1/(2\beta+d\cdot\max\{1,\beta\})})$ (Jiang 2017), which is slower. In other words, we escape the curse of dimensionality with regression level-set estimation but do not escape it for density level-set estimation.

Global Maxima Estimation

In this section, we give guarantees on estimating the global maxima of f.

Definition 8. x_0 *is a maxima of* f *if* $f(x) < f(x_0)$ *for all* $x \in B(x_0, r) \backslash \{x_0\}$ *for some* $r > 0$.

We then make the following assumptions, which states that f has a unique maxima, where it has a negative-definite Hessian.

Assumption 5. f *has a unique maxima* $x_0 := \text{argmax}_{x\in\mathcal{X}} f(x)$ *and* f *has a negative-definite Hessian at* x_0.

These assumptions lead to the following, which states that f has quadratic smoothness and decay around x_0.

Lemma 1 (Dasgupta and Kpotufe (2014))**.** *Let* f *satisfy Assumption 5. Then there exists* $\hat{C}, \check{C}, r_M, \lambda > 0$ *such that the following holds.*

$$\check{C} \cdot |x_0 - x|^2 \leq f(x_0) - f(x) \leq \hat{C} \cdot |x_0 - x|^2$$

for all $x \in A_0$ *where* A_0 *is a connected component of* $\{x : f(x) \geq \lambda\}$ *and* A_0 *contains* $B(x_0, r_M)$.

We utilize the following estimator, which is the maximizer of f_k amongst sample points $X = \{x_1, ..., x_n\}$.

$$\widehat{x} := \underset{x\in X}{\text{argmax}}\, f_k(x).$$

We next give the result of the accuracy of $\widehat{x}$ in estimating x_0.

Theorem 5. *Suppose that* f *is continuous and that Assumptions 1, 2, 3, and 5 hold. Let* k *satisfy*

$$k \geq \frac{2^{10} \cdot D \log^2(4/\delta) \cdot \log n}{\min\{1, \check{C}^2 \cdot r_M^4/\sigma^2\}}$$

$$k \leq \frac{1}{2} \cdot \gamma \cdot p_{X,0} \cdot v_D \cdot \min\left\{r_0^D, \left(\frac{\check{C} \cdot r_M^2}{32 \cdot \hat{C}}\right)^{D/2}\right\} \cdot n.$$

Then the following holds with probability at least $1-\delta$.

$$|\hat{x} - x_0|^2 \leq \max\left\{\frac{32\sigma}{\check{C}}\sqrt{\frac{D\log n + \log(2/\delta)}{k}}, \frac{32\hat{C}}{\check{C}}\left(\frac{2k}{\gamma \cdot p_{X,0} \cdot v_D \cdot n}\right)^{2/D}\right\}.$$

Remark 8. *Taking* $k \approx n^{4/(4+D)}$ *optimizes the above expression so that* $|\hat{x} - x_0| \lesssim \widetilde{O}(n^{-1/(4+D)})$*. This can be compared to the minimax rate for mode estimation* $O(n^{-1/(4+D)})$ *established by Tsybakov (1990). We stress however that estimating the mode of density function is a different problem.*

Remark 9. *An analogue for global minima also holds. Moreover, in the manifold setting, we can obtain a rate of* $\widetilde{O}(n^{-1/(4+d)})$*, which has not been shown for mode estimation in densities.*

Proofs

Proof of Theorem 1

The follow bounds $r_k(x)$ uniformly in $x \in \mathcal{X}$.

Lemma 2. *The following holds with probability at least* $1 - \delta/2$*. If*

$$2^8 \cdot D \log^2(4/\delta) \cdot \log n \leq k \leq \frac{1}{2} \cdot \gamma \cdot p_{X,0} \cdot v_D \cdot r_0^D \cdot n,$$

then $\sup_{x\in\mathcal{X}} r_k(x) \leq \left(\frac{2k}{\gamma \cdot v_D \cdot n \cdot p_{X,0}}\right)^{1/D}$.

Proof. Let $r = \left(\frac{2k}{\gamma \cdot v_D \cdot n \cdot p_{X,0}}\right)^{1/D}$. We have $\mathcal{P}(B(x,r)) \geq \gamma \inf_{x' \in B(x,r)\cap\mathcal{X}} p_X(x') \cdot v_D r^D \geq \gamma p_{X,0} v_D r^D = \frac{2k}{n}$. By Lemma 7 of (Chaudhuri and Dasgupta 2010) and the condition on k, it follows that with probability $1 - \delta/2$, uniformly in $x \in \mathcal{X}$, $\mathcal{P}_n(B(x,r)) \geq \frac{k}{n}$. Hence, $r_k(x) < r$ and the result follows immediately. □

The next result bounds the number of distinct k-NN sets over $\mathcal{X}$.

Lemma 3. *Let* M *be the number of distinct* k*-NN sets over* $\mathcal{X}$*, that is,* $M := |\{N_k(x) : x \in \mathcal{X}\}|$*. Then* $M \leq D \cdot n^D$.

Proof. First, let $\mathcal{A}$ be the partitioning of $\mathcal{X}$ induced by the $\binom{n}{2}$ hyperplanes defined as the perpendicular bisectors of each pair of points x_i, x_j for $i \neq j$. Let us denote this set of hyperplanes as $\mathcal{H}$. We have that if x, x' are in the same partition of $\mathcal{A}$, then $N_k(x) = N_k(x')$. If not, then any path from x to x' must cross some perpendicular bisector in $N_k(x') - N_k(x)$, which would be a contradiction. Thus, $M \leq |\mathcal{A}|$.

Now we will bound $|\mathcal{A}|$. Since $\mathcal{H}$ is finite, choose vectors $e_1, ..., e_D$ such that they form an orthogonal basis of $\mathbb{R}^D$ and none of these vectors are perpendicular to any $H \in \mathcal{H}$. Let $e_1, ..., e_D$ induce hyperplanes $H_1, ..., H_D$, respectively (i.e. H_i being the orthogonal complement of e_i). Without loss of generality, orient the space such that e_1 is the vertical

direction (i.e. so that we can use descriptions such as 'above' and 'below'). For each region in $\mathcal{A}$ that is bounded below, associate such a region to its lowest point. Then it follows that there are at most $\binom{n}{D}$ of these regions since they are the intersection of D hyperplanes.

We next count the regions unbounded below. Place H_1 below the lowest point corresponding the regions in $\mathcal{A}$ that were bounded below. Then we have that the regions unbounded below are $\{A \in \mathcal{A} : A \cap H_1 \neq \emptyset\}$. It thus remains now to count $\mathcal{A}_1 := \{A \cap H_1 : A \in \mathcal{A}, A \cap H_1 \neq \emptyset\}$.

We now orient the space so that e_2 corresponds to the vertical direction. Then we can repeat the same procedure and for each region in $\mathcal{A}_1$that is bounded below with the lowest point. There are at most $\binom{n}{D-1}$ since they are an intersection of $D-1$ hyperplanes in $\mathcal{H}$ along with H_1, and then placing e_2 sufficiently low, the remaining regions correspond to $\mathcal{A}_2 := \{A \cap H_1 \cap H_2 : A \in \mathcal{A}, A \cap H_1 \cap H_2 \neq \emptyset\}$.

Continuing this process, it follows that when we orient e_i to be the vertical direction, in order to count $\mathcal{A}_i := \{A \cap H_1 \cap \cdots \cap H_i : A \in \mathcal{A}, A \cap H_1 \cap \cdots \cap H_i \neq \emptyset\}$, the number of regions in $\mathcal{A}_i$ bounded below is at most $\binom{n}{D-i}$ and the remaining ones are correspond to $\mathcal{A}_{i+1}$.

It thus follows that $|\mathcal{A}| \leq \sum_{j=0}^{D} \binom{n}{j} \leq D \cdot n^D$, as desired. □

Proof of Theorem 1. We have

$$\begin{aligned}|f_k(x) - f(x)| &\leq \left| \frac{1}{|N_k(x)|} \sum_{i=1}^{n} (f(x_i) - f(x)) \cdot 1\left[x_i \in N_k(x)\right] \right| \\ &\quad + \left| \frac{1}{|N_k(x)|} \sum_{i=1}^{n} \xi_{x_i} \cdot 1\left[x_i \in N_k(x)\right] \right| \\ &\leq u_f(x, r_k(x)) + \left| \frac{1}{N_k(x)} \sum_{i=1}^{n} \xi_{x_i} \cdot 1\left[x_i \in N_k(x)\right] \right|.\end{aligned}$$

The first term can be viewed as the bias term and the second can be viewed as variance term.

By Lemma 2, we can bound the first term as follows with probability at least $1 - \delta/2$ uniformly in $x \in \mathcal{X}$: $u_f(x, r_k(x)) \leq u_f\left(x, \left(\frac{2k}{\gamma \cdot p_{X,0} \cdot v_D \cdot n}\right)^{1/D}\right)$. For the variance term, we have by Hoeffding's inequality that if $A_x := \left|\frac{1}{k}\sum_{i=1}^{n} \xi_{x_i} \cdot 1\left[x_i \in N_k(x)\right]\right|$ then $\mathbb{P}\left(A_x > \frac{\sqrt{2}\sigma \cdot t}{\sqrt{k}}\right) \leq \exp\left(-t^2\right)$.

Taking $t = \sqrt{D \log n + \log(2D/\delta)}$, then we have $\mathbb{P}\left(A_x > \frac{\sqrt{2}\sigma \cdot t}{\sqrt{k}}\right) \leq \delta/(2D \cdot n^D)$.

By Lemma 3 and union bound, it follows that $\mathbb{P}\left(\sup_{x \in \mathcal{X}} A_x > \frac{\sqrt{2}\sigma \cdot t}{\sqrt{k}}\right) \leq \delta/2$. Hence, we have with probability at least $1 - \delta$,

$$\begin{aligned}|f(x) - f_k(x)| &\leq u_f\left(x, \left(\frac{2k}{\gamma \cdot p_{X,0} \cdot v_d \cdot n}\right)^{1/D}\right) \\ &\quad + 2\sigma\sqrt{\frac{D \log n + \log(2/\delta)}{k}}.\end{aligned}$$

uniformly in $x \in \mathcal{X}$. □

It is easy to see that a simple modification to the proof of Theorem 1 will yield the following.

Corollary 3 (k-NN Regression Upper and Lower Bounds)**.** *Let*

$$\begin{aligned}\hat{u}_f(x, r) &:= \sup_{x' \in B(x,r)} f(x') - f(x) \\ \check{u}_f(x, r) &:= \sup_{x' \in B(x,r)} f(x) - f(x') \\ \varepsilon_{var} &:= 2\sigma\sqrt{\frac{D \log n + \log(2/\delta)}{k}} \\ \varepsilon_k &:= \left(\frac{2k}{\gamma p_{X,0} v_D \cdot n}\right)^{1/D}.\end{aligned}$$

Suppose that Assumptions 1, 2, and 3 hold and that

$$k \geq 2^8 \cdot D \log^2(4/\delta) \cdot \log n.$$

Then probability at least $1-\delta$, the following holds uniformly in $x \in \mathcal{X}$.

$$\begin{aligned}f_k(x) &\leq f(x) + \hat{u}_f(x, \varepsilon_k) + \varepsilon_{var} \\ f_k(x) &\geq f(x) - \check{u}_f(x, \varepsilon_k) - \varepsilon_{var}.\end{aligned}$$

Proof of Theorem 2

We need the following guarantee on the volume of the intersection of a Euclidean ball and M; this is required to get a handle on the true mass of the ball under $\mathcal{P}$ in later arguments. The proof can be found in (Jiang 2017).

Lemma 4 (Ball Volume)**.** *If $0 < r < \min\{\tau/(4d), 1/\tau\}$, and $x \in M$ then*

$$1 - \tau^2 r^2 \leq \frac{vol_d(B(x, r) \cap M)}{v_d r^d} \leq 1 + 4d \cdot r/\tau,$$

where vol_d is the volume w.r.t. the uniform measure on M.

The next is the manifold analogue of Lemma 2.

Lemma 5. *Suppose that Assumptions 2, 3, and 4 hold. The following holds with probability at least $1 - \delta/2$. If*

$$2^8 \cdot D \log^2(4/\delta) \cdot \log n \leq k \leq \frac{1}{4}\left(\min\left\{\frac{\tau}{4d}, \frac{1}{\tau}\right\}\right)^d p_{X,0} \cdot v_d \cdot n.$$

then for all $x \in M$, $r_k(x) \leq \left(\frac{4k}{v_d \cdot n \cdot p_{X,0}}\right)^{1/d}$.

Proof. Let $r = \left(\frac{4k}{v_d \cdot n \cdot p_{X,0}}\right)^{1/d}$. We have

$$\begin{aligned}\mathcal{P}(B(x, r)) &\geq \inf_{x' \in B(x,r) \cap M} p_X(x') \cdot \mathrm{Vol}_d(B(x, r) \cap M) \\ &\geq p_{X,0} \cdot (1 - \tau^2 r^2) \cdot v_d r^d \geq \frac{1}{2} p_{X,0} v_d r^d \geq \frac{2k}{n}.\end{aligned}$$

By Lemma 7 of (Chaudhuri and Dasgupta 2010) and the condition on k, it follows that with probability $1 - \delta/2$, uniformly in $x \in \mathcal{X}$, $\mathcal{P}_n(B(x, r)) \geq \frac{k}{n}$. Hence, $r_k(x) < r$ and the result follows immediately. □

Theorem 2 now follows by replacing the usage of Lemma 2 with Lemma 5. We also note that an analogous result to Corollary 3 can also be established.

It is easy to see that a simple modification to the proof of Theorem 2 will yield the following.

Corollary 4 (k-NN Regression Upper and Lower Bounds)**.** *Let*

$$\hat{u}_f(x,r) := \sup_{x' \in B(x,r)} f(x') - f(x)$$
$$\check{u}_f(x,r) := \sup_{x' \in B(x,r)} f(x) - f(x')$$
$$\varepsilon_{var} := 2\sigma\sqrt{\frac{D\log n + \log(2/\delta)}{k}}$$
$$\varepsilon_k := \left(\frac{2k}{\gamma p_{X,0} v_D \cdot n}\right)^{1/d}.$$

Suppose that Assumptions 1, 2, and 3 hold and that

$$k \geq 2^8 \cdot D\log^2(4/\delta) \cdot \log n.$$

Then probability at least $1-\delta$*, the following holds uniformly in* $x \in \mathcal{X}$*.*

$$f_k(x) \leq f(x) + \hat{u}_f(x, \varepsilon_k) + \varepsilon_{var}$$
$$f_k(x) \geq f(x) - \check{u}_f(x, \varepsilon_k) - \varepsilon_{var}.$$

Proofs of Theorem 3 and 4

Proof of Theorem 3. We have that $E[\hat{\sigma}^2] = 2M^2 \geq 2\text{Var}(\xi^2) = 2\sigma^2$. Thus, when n is sufficiently large depending on ξ, f, and δ, we have by Bernstein-type concentration inequalities that with probability at least $1-\delta$, $2\sigma \leq \hat{\sigma} \leq \sqrt{5}M$.

Let $\tilde{r} := 2(2\epsilon/\check{C})^{1/\beta}$ and let us use the notation introduced in Corollary 3. It suffices to show that (1) $\widehat{L}_f(\lambda) \subseteq L_f(\lambda) \oplus \tilde{r}$ and (2) $L_f(\lambda) \subseteq \widehat{L}_f(\lambda) \oplus \tilde{r}$. We begin with (1). We have

$$\begin{aligned}
\sup_{x \in \mathcal{X} \setminus (L_f(\lambda) \oplus \tilde{r})} f_k(x) &\leq \sup_{x \in \mathcal{X} \setminus (L_f(\lambda) \oplus \tilde{r})} (f(x) + \hat{u}_f(x, \epsilon_k)) + \varepsilon_{\text{var}} \\
&\leq \sup_{x \in \mathcal{X} \setminus (L_f(\lambda) \oplus \tilde{r})} \sup_{x' \in B(x, \epsilon_k)} f(x') + \varepsilon_{\text{var}} \\
&= \sup_{x \in \mathcal{X} \setminus (L_f(\lambda) \oplus (\tilde{r} - \epsilon_k))} f(x) + \varepsilon_{\text{var}} \\
&\leq \lambda - \check{C}(\tilde{r} - \epsilon_k)^\beta + \varepsilon_{\text{var}} \leq \lambda - \epsilon,
\end{aligned}$$

where the first inequality holds by Corollary 3, the second-to-last inequality holds by β-regularity and that $\tilde{r} < r_M$, and the last inequality holds by the conditions on k (which in particular imply $\epsilon \geq 2\varepsilon_{\text{var}}$ and $\epsilon_k < (2\epsilon/\check{C})^{1/\beta}$). Thus, if $x \notin L_f(\lambda) \oplus \tilde{r}$, then $f_k(x) < \lambda - \epsilon$. Therefore, $\widehat{L}_f(\lambda) \subseteq L_f(\lambda) \oplus \tilde{r}$, which establishes (1).

We now show (2). Let $\bar{r} = \epsilon_k$. Since $\bar{r} < \tilde{r}$, it suffices to show that $L_f(\lambda) \subseteq \widehat{L}_f(\lambda) \oplus \bar{r}$. For any $x \in L_f(\lambda)$, we have

$$\mathcal{P}(B(x, \bar{r})) \geq \frac{2k}{n} \geq \frac{16\log(4/\delta) D \log n}{n},$$

where the last inequality holds by the conditions on k. Hence, by Lemma 7 of (Chaudhuri and Dasgupta 2010), we have $\mathcal{P}_n(B(x, \bar{r})) > 0$. Thus, for any $x \in L_f(\lambda)$, there exists a sample point in $B(x, \bar{r})$. Furthermore, we have

$$\begin{aligned}
\inf_{x' \in B(x,\bar{r})} f_k(x') &\geq \inf_{x' \in B(x,\bar{r})} f(x) - \check{u}_f(x, \epsilon_k) - \varepsilon_{\text{var}} \\
&\geq \inf_{x' \in B(x,\bar{r})} \inf_{x'' \in B(x', \epsilon_k)} f(x'') - \varepsilon_{\text{var}} \\
&= \inf_{x' \in B(x, \bar{r} + \epsilon_k)} f(x') - \varepsilon_{\text{var}} \\
&\geq \lambda - \hat{C}(\bar{r} + \epsilon_k)^\beta - \varepsilon_{\text{var}} \geq \lambda - \epsilon.
\end{aligned}$$

where the first inequality holds by Corollary 3, the second last inequality holds by β-regularity, and the final inequality holds by the conditions on k.

Thus, for any $x \in L_f(\lambda)$, not only does there exists a sample point in $B(x, \bar{r})$, but any such sample point will have f_k value at least $\lambda - \epsilon$ and thus is in $\widehat{L}_f(\lambda)$. Therefore, $L_f(\lambda) \subseteq \widehat{L}_f(\lambda) \oplus \bar{r}$, as desired. □

Proof of Theorem 4. The proof is the same as that of Theorem 3 but with the full-dimensional k-NN regression bounds replaced by the manifold versions, and is omitted here. □

Proof of Theorem 5

Proof of Theorem 5. Define the following.

$$\varepsilon_{\text{var}} := 2\sigma\sqrt{\frac{D\log n + \log(2/\delta)}{k}}, \varepsilon_k := \left(\frac{2k}{\gamma \cdot p_{X,0} v_D \cdot n}\right)^{1/D}$$
$$\tilde{r}^2 := \max\{16\varepsilon_{\text{var}}/\hat{C}, (2\epsilon_k/c)^2\},$$

where $c^2 = \check{C}/8\hat{C}$. The goal is now to show $|x - x_0| \leq \tilde{r}$. The proof now mirrors that of Theorem 1 of Dasgupta and Kpotufe (2014). It suffices to show that

$$\sup_{x \in \mathcal{X} \setminus B(x_0, \tilde{r})} f_k(x) < \inf_{x \in B(x_0, r_n)} f_k(x),$$

where $r_n = d(x_0, X)$. We have by Corollary 3:

$$\begin{aligned}
\sup_{x \in \mathcal{X} \setminus B(x_0,\tilde{r})} f_k(x) &\leq \sup_{x \in \mathcal{X} \setminus B(x_0,\tilde{r})} f(x) + \hat{u}_f(x, \varepsilon_k) + \varepsilon_{\text{var}} \\
&\leq \sup_{x \in \mathcal{X} \setminus B(x_0,\tilde{r})} f(x) + \hat{u}_f(x, \tilde{r}/2) + \varepsilon_{\text{var}} \\
&\leq \sup_{x \in \mathcal{X} \setminus B(x_0,\tilde{r}/2)} f(x) + \varepsilon_{\text{var}} \\
&\leq f(x_0) - \check{C}(\tilde{r}/2)^2 + \varepsilon_{\text{var}}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\inf_{x \in B(x_0, r_n)} f_k(x) &\geq \inf_{x \in B(x_0, r_n)} f(x) - \check{u}_f(x, \varepsilon_k) - \varepsilon_{\text{var}} \\
&\geq \inf_{x \in B(x_0, c\tilde{r}/2)} f(x) - \check{u}_f(x, c\tilde{r}/2) - \varepsilon_{\text{var}} \\
&\geq \inf_{x \in B(x_0, c\tilde{r})} f(x) - \varepsilon_{\text{var}} \\
&\geq f(x_0) - \hat{C}(c\tilde{r})^2 - \varepsilon_{\text{var}}.
\end{aligned}$$

The result now follows from our choice of $\tilde{r}$. □

Conclusion: We provided finite-sample sup-norm bounds for k-NN regression under standard nonparametric assumptions for both the full-dimensional and manifold setting. We then applied our results to level-set and global maxima estimation.

References

Balakrishnan, S.; Narayanan, S.; Rinaldo, A.; Singh, A.; and Wasserman, L. 2013. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, 2679–2687.

Biau, G.; Cérou, F.; and Guyader, A. 2010. Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Transactions on Information Theory* 56(4):2034–2040.

Chaudhuri, K., and Dasgupta, S. 2010. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, 343–351.

Cheng, P. E. 1984. Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis* 15(1):63–72.

Dasgupta, S., and Kpotufe, S. 2014. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, 2555–2563.

Devroye, L.; Gyorfi, L.; Krzyzak, A.; and Lugosi, G. 1994. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics* 1371–1385.

Devroye, L. 1978. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory* 24(2):142–151.

Genovese, C.; Perone-Pacifico, M.; Verdinelli, I.; and Wasserman, L. 2012. Minimax manifold estimation. *Journal of machine learning research* 13(May):1263–1291.

Jiang, H. 2017. Density level set estimation on manifolds with dbscan. *International Conference on Machine Learning (ICML)*.

Kpotufe, S. 2011. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, 729–737.

Kudraszow, N. L., and Vieu, P. 2013. Uniform consistency of knn regressors for functional variables. *Statistics & Probability Letters* 83(8):1863–1870.

Kulkarni, S. R., and Posner, S. E. 1995. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory* 41(4):1028–1039.

Lian, H. 2011. Convergence of functional k-nearest neighbor regression estimate with functional responses. *Electronic Journal of Statistics* 5:31–40.

Mack, Y.-p., and Silverman, B. W. 1982. Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields* 61(3):405–415.

Romano, J. P. 1988. On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics* 629–647.

Singh, A.; Scott, C.; and Nowak, R. 2009. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics* 37(5B):2760–2782.

Stone, C. J. 1977. Consistent nonparametric regression. *The annals of statistics* 595–620.

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

Tsybakov, A. B. 1997. On nonparametric estimation of density level sets. *The Annals of Statistics* 25(3):948–969.

Willett, R. M., and Nowak, R. D. 2007. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing* 16(12):2965–2979.