k-NN Regression Adapts to Local Intrinsic Dimension

Samory Kpotufe
Max Planck Institute for Intelligent Systems
samory@tuebingen.mpg.de

Abstract

Many nonparametric regressors were recently shown to converge at rates that depend only on the intrinsic dimension of data. These regressors thus escape the curse of dimension when high-dimensional data has low intrinsic dimension (e.g. a manifold). We show that k-NN regression is also adaptive to intrinsic dimension. In particular our rates are local to a query x and depend only on the way masses of balls centered at x vary with radius.

Furthermore, we show a simple way to choose $k = k(x)$ locally at any x so as to nearly achieve the minimax rate at x in terms of the unknown intrinsic dimension in the vicinity of x. We also establish that the minimax rate does not depend on a particular choice of metric space or distribution, but rather that this minimax rate holds for any metric space and doubling measure.

1 Introduction

We derive new rates of convergence in terms of dimension for the popular approach of Nearest Neighbor (k-NN) regression. Our motivation is that, for good performance, k-NN regression can require a number of samples exponential in the dimension of the input space $\mathcal{X}$. This is the so-called "curse of dimension". Formally stated, the curse of dimension is the fact that, for any nonparametric regressor there exists a distribution in $\mathbb{R}^D$ such that, given a training size n, the regressor converges at a rate no better than $n^{-1/O(D)}$ (see e.g. [1, 2]).

Fortunately it often occurs that high-dimensional data has low intrinsic dimension: typical examples are data lying near low-dimensional manifolds [3, 4, 5]. We would hope that in these cases nonparametric regressors can escape the curse of dimension, i.e. their performance should only depend on the intrinsic dimension of the data, appropriately formalized. In other words, if the data in $\mathbb{R}^D$ has intrinsic dimension $d \ll D$, we would hope for a better convergence rate of the form $n^{-1/O(d)}$ instead of $n^{-1/O(D)}$. This has recently been shown to indeed be the case for methods such as kernel regression [6], tree-based regression [7] and variants of these methods [8]. In the case of k-NN regression however, it is only known that 1-NN regression (where $k = 1$) converges at a rate that depends on intrinsic dimension [9]. Unfortunately 1-NN regression is not consistent. For consistency, it is well known that we need k to grow as a function of the sample size n [10] .

Our contributions are the following. We assume throughout that the target function f is Lipschitz. First we show that, for a wide range of values of k ensuring consistency, k-NN regression converges at a rate that only depends on the intrinsic dimension in a neighborhood of a query x. Our local notion of dimension in a neighborhood of a point x relies on the well-studied notion of *doubling measure* (see Section 2.3). In particular our dimension quantifies how the mass of balls vary with radius, and this captures standard examples of data with low intrinsic dimension. Our second, and perhaps most important contribution, is a simple procedure for choosing $k = k(x)$ so as to nearly achieve the minimax rate of $O\left(n^{-2/(2+d)}\right)$ in terms of the unknown dimension d in a neighborhood of x. Our final contribution is in showing that this minimax rate holds for any metric space and doubling measure. In other words the hardness of the regression problem is not tied to a particular choice of metric space $\mathcal{X}$ or doubling measure μ, but depends only on how the doubling measure μ expands on a metric space $\mathcal{X}$. Thus, for any marginal μ on $\mathcal{X}$ with expansion constant $\Theta\left(2^d\right)$, the minimax rate for the measure space $(\mathcal{X}, \mu)$ is $\Omega\left(n^{-2/(2+d)}\right)$.

1.1 Discussion

It is desirable to express regression rates in terms of a local notion of dimension rather than a global one because the complexity of data can vary considerably over regions of space. Consider for example a dataset made up of a collection of manifolds of various dimensions. The global complexity is necessarily of a worst case nature, i.e. is affected by the most complex regions of the space while we might happen to query x from a less complex region. Worse, it can be the case that the data is not complex locally anywhere, but globally the data is more complex. A simple example of this is a so-called *space filling curve* where a low-dimensional manifold curves enough that globally it seems to fill up space. We will see that the global complexity does not affect the behavior of k-NN regression, provided k/n is sufficiently small. The behavior of k-NN regression is rather controlled by the often smaller local dimension in a neighborhood $B(x, r)$ of x, where the neighborhood size r shrinks with k/n.

Given such a neighborhood $B(x, r)$ of x, how does one choose $k = k(x)$ optimally relative to the unknown local dimension in $B(x, r)$? This is nontrivial as standard methods of (global) parameter selection do not easily apply. For instance, it is unclear how to choose k by cross-validation over possible settings: we do not know reliable surrogates for the true errors at x of the various estimators $\{f_{n,k}(x)\}_{k\in[n]}$. Another possibility is to estimate the dimension of the data in the vicinity of x, and use this estimate to set k. However, for optimal rates, we have to estimate the dimension exactly and we know of no finite sample result that guarantees the exact estimate of intrinsic dimension. Our method consists of finding a value of k that balances quantities which control estimator variance and bias at x, namely $1/k$ and distances to x's k nearest neighbors. The method guarantees, uniformly over all $x \in \mathcal{X}$, a near optimal rate of $\widetilde{O}\left(n^{-2/(2+d)}\right)$ where $d = d(x)$ is exactly the unknown local dimension on a neighborhood $B(x, r)$ of x, where $r \to 0$ as $n \to \infty$.

2 Setup

We are given n i.i.d samples $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}_{i=1}^n$ from some unknown distribution where the input variable X belongs to a metric space $(\mathcal{X}, \rho)$, and the output Y is a real number. We assume that the class $\mathcal{B}$ of balls on $(\mathcal{X}, \rho)$ has finite VC dimension $\mathcal{V}_{\mathcal{B}}$. This is true for instance for any subset $\mathcal{X}$ of a Euclidean space, e.g. the low-dimensional spaces discussed in Section 2.3. The VC assumption is however irrelevant to the minimax result of Theorem 3.

We denote the marginal distribution on $\mathcal{X}$ by μ and the empirical distribution on $\mathbf{X}$ by μ_n.

2.1 Regression function and noise

The regression function $f(x) = \mathbb{E}\left[Y|X = x\right]$ is assumed to be λ-Lipschitz, i.e. there exists $\lambda > 0$ such that $\forall x, x' \in \mathcal{X}$, $|f(x) - f(x')| \leq \lambda\rho\left(x, x'\right)$.

We assume a simple but general noise model: the distributions of the noise at points $x \in \mathcal{X}$ have uniformly bounded tails and variance. In particular, Y is allowed to be unbounded. Formally:

$$\forall\delta > 0 \text{ there exists } t > 0 \text{ such that } \sup_{x\in\mathcal{X}} \mathbb{P}_{Y|X=x}\left(|Y - f(x)| > t\right) \leq \delta.$$

We denote by $t_Y(\delta)$ the infimum over all such t. Also, we assume that the variance of $(Y|X = x)$ is upper-bounded by a constant σ_Y^2 uniformly over all $x \in \mathcal{X}$.

To illustrate our noise assumptions, consider for instance the standard assumption of bounded noise, i.e. $|Y - f(x)|$ is uniformly bounded by some $M > 0$; then $\forall\delta > 0$, $t_Y(\delta) \leq M$, and can thus be replaced by M in all our results. Another standard assumption is that where the noise distribution has exponentially decreasing tail; in this case $\forall\delta > 0$, $t_Y(\delta) \leq O(\ln 1/\delta)$. As a last example, in the case of Gaussian (or sub-Gaussian) noise, it's not hard to see that $\forall\delta > 0$, $t_Y(\delta) \leq O(\sqrt{\ln 1/\delta})$.

2.2 Weighted k-NN regression estimate

We assume a kernel function $K : \mathbb{R}_+ \mapsto \mathbb{R}_+$, non-increasing, such that $K(1) > 0$, and $K(\rho) = 0$ for $\rho > 1$. For $x \in \mathcal{X}$, let $r_{k,n}(x)$ denote the distance to its k'th nearest neighbor in the sample $\mathbf{X}$. The regression estimate at x given the n-sample $(\mathbf{X}, \mathbf{Y})$ is then defined as

$$f_{n,k}(x) = \sum_i \frac{K\left(\rho(x, x_i)/r_{k,n}(x)\right)}{\sum_j K\left(\rho(x, x_j)/r_{k,n}(x)\right)} Y_i = \sum_i w_{i,k}(x) Y_i.$$

2.3 Notion of dimension

We start with the following definition of doubling measure which will lead to the notion of local dimension used in this work. We stay informal in developing the motivation and refer the reader to [**?**, 11, 12] for thorough overviews of the topic of metric space dimension and doubling measures.

Definition 1. *The marginal μ is a* **doubling measure** *if there exist $C_{db} > 0$ such that for any $x \in \mathcal{X}$ and $r \geq 0$, we have $\mu(B(x,r)) \leq C_{db}\mu(B(x, r/2))$. The quantity C_{db} is called an* **expansion constant** *of μ.*

An equivalent definition is that, μ is doubling if there exist C and d such that for any $x \in \mathcal{X}$, for any $r \geq 0$ and any $0 < \epsilon < 1$, we have $\mu(B(x,r)) \leq C\epsilon^{-d}\mu(B(x,\epsilon r))$. Here d acts as a dimension. It is not hard to show that d can be chosen as $\log_2 C_{\text{db}}$ and C as C_{db} (see e.g. [**?**]).

A simple example of a doubling measure is the Lebesgue volume in the Euclidean space $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$ and $r > 0$, $\text{vol}\left(B(x,r)\right) = \text{vol}\left(B(x,1)\right) r^d$. Thus $\text{vol}\left(B(x,r)\right) / \text{vol}\left(B(x,\epsilon r)\right) = \epsilon^{-d}$ for any $x \in \mathbb{R}^d$, $r > 0$ and $0 < \epsilon < 1$. Building upon the doubling behavior of volumes in $\mathbb{R}^d$, we can construct various examples of doubling *probability* measures. The following ingredients are sufficient. Let $\mathcal{X} \subset \mathbb{R}^D$ be a subset of a d-dimensional hyperplane, and let $\mathcal{X}$ satisfy for all balls $B(x,r)$ with $x \in \mathcal{X}$, $\text{vol}\left(B(x,r) \cap \mathcal{X}\right) = \Theta(r^d)$, the volume being with respect to the containing hyperplane. Now let μ be approximately uniform, that is μ satisfies for all such balls $B(x,r)$, $\mu(B(x,r) \cap \mathcal{X}) = \Theta(\text{vol}\left(B(x,r) \cap \mathcal{X}\right))$. We then have $\mu(B(x,r))/\mu(B(x,\epsilon r)) = \Theta(\epsilon^{-d})$.

Unfortunately a global notion of dimension such as the above definition of d is rather restrictive as it requires the same complexity globally and locally. However a data space can be complex globally and have small complexity locally. Consider for instance a d-dimensional submanifold $\mathcal{X}$ of $\mathbb{R}^D$, and let μ have an upper and lower bounded density on $\mathcal{X}$. The manifold might be globally complex but the restriction of μ to a ball $B(x,r)$, $x \in \mathcal{X}$, is doubling with local dimension d, provided r is sufficiently small and certain conditions on curvature hold. This is because, under such conditions (see e.g. the Bishop-Gromov theorem [13]), the volume (in $\mathcal{X}$) of $B(x,r) \cap \mathcal{X}$ is $\Theta(r^d)$.

The above example motivates the following definition of local dimension d.

Definition 2. *Fix $x \in \mathcal{X}$, and $r > 0$. Let $C \geq 1$ and $d \geq 1$. The marginal μ is* (C,d)**-homogeneous on** $B(x,r)$ *if we have $\mu(B(x,r')) \leq C\epsilon^{-d}\mu(B(x,\epsilon r'))$ for all $r' \leq r$ and $0 < \epsilon < 1$.*

The above definition covers cases other than manifolds. In particular, another space with small local dimension is a sparse data space $\mathcal{X} \subset \mathbb{R}^D$ where each vector x has at most d non-zero coordinates, i.e. $\mathcal{X}$ is a collection of finitely many hyperplanes of dimension at most d. More generally suppose the data distribution μ is a mixture $\sum_i \pi_i \mu_i$ of finitely many distributions μ_i with potentially different low-dimensional supports. Then if all μ_i supported on a ball B are (C_i, d)-homogeneous on B, i.e. all have local dimension d on B, then μ is also (C,d)-homogeneous on B for some C.

We want rates of convergence which hold uniformly over all regions where μ is doubling. We therefore also require (Definition 3) that C and d from Definition 2 are uniformly upper bounded. This will be the case in many situations including the above examples.

Definition 3. *The marginal μ is* (C_0, d_0)**-maximally-homogeneous** *for some $C_0 \geq 1$ and $d_0 \geq 1$, if the following holds for all $x \in \mathcal{X}$ and $r > 0$: suppose there exists $C \geq 1$ and $d \geq 1$ such that μ is (C,d)-homogeneous on $B(x,r)$, then μ is (C_0, d_0)-homogeneous on $B(x,r)$.*

We note that, rather than assuming as in Definition 3 that all local dimensions are at most d_0, we can express our results in terms of the subset of $\mathcal{X}$ where local dimensions are at most d_0. In this case d_0 would be allowed to grow with n. The less general assumption of Definition 3 allows for a clearer presentation which still captures the local behavior of k-NN regression.

3 Overview of results

3.1 Local rates for fixed k

The first result below establishes the rates of convergence for any $k \gtrsim \ln n$ in terms of the (unknown) complexity on $B(x,r)$ where r is any r satisfying $\mu(B(x,r)) > \Omega(k/n)$ (we need at least $\Omega(k)$ samples in the relevant neighborhoods of x).

Theorem 1. *Suppose μ is (C_0, d_0)-maximally-homogeneous, and $\mathcal{B}$ has finite VC dimension $\mathcal{V}_{\mathcal{B}}$. Let $0 < \delta < 1$. With probability at least $1 - 2\delta$ over the choice of $(\mathbf{X}, \mathbf{Y})$, the following holds simultaneously for all $x \in \mathcal{X}$ and k satisfying $n > k \geq \mathcal{V}_{\mathcal{B}} \ln 2n + \ln(8/\delta)$.*

Pick any $x \in \mathcal{X}$. Let $r > 0$ satisfy $\mu(B(x,r)) > 3C_0 k/n$. Suppose μ is (C,d)-homogeneous on $B(x,r)$, with $1 \leq C \leq C_0$ and $1 \leq d \leq d_0$. We have

$$|f_{n,k}(x) - f(x)|^2 \leq \frac{2K(0)}{K(1)} \cdot \frac{\mathcal{V}_{\mathcal{B}} \cdot t_Y^2(\delta/2n) \cdot \ln(2n/\delta) + \sigma_Y^2}{k} + 2\lambda^2 r^2 \left(\frac{3Ck}{n\mu(B(x,r))}\right)^{2/d}.$$

Note that the above rates hold uniformly over x, $k \gtrsim \ln n$, and any r where $\mu(B(x,r)) \geq \Omega(k/n)$. The rate also depends on $\mu(B(x,r))$ and suggests that the best scenario is that where x has a small neighborhood of large mass and small dimension d.

3.2 Minimax rates for a doubling measure

Our next result shows that the hardness of the regression problem is not tied to a particular choice of the metric $\mathcal{X}$ or the doubling measure μ. The result relies mainly on the fact that μ is doubling on $\mathcal{X}$. We however assume that μ has the same expansion constant everywhere and that this constant is tight. This does not however make the lower-bound less expressive, as it still tells us which rates to expect locally. Thus if μ is (C,d)-homogeneous near x, we cannot expect a better rate than $O\left(n^{-2/(2+d)}\right)$ (assuming a Lipschitz regression function f).

Theorem 2. *Let μ be a doubling measure on a metric space $(\mathcal{X}, \rho)$ of diameter 1, and suppose μ satisfies, for all $x \in \mathcal{X}$, for all $r > 0$ and $0 < \epsilon < 1$,*

$$C_1 \epsilon^{-d} \mu(B(x, \epsilon r)) \leq \mu(B(x,r)) \leq C_2 \epsilon^{-d} \mu(B(x, \epsilon r)),$$

where C_1, C_2 and d are positive constants independent of x, r, and ϵ. Let $\mathcal{Y}$ be a subset of $\mathbb{R}$ and let $\lambda > 0$. Define $\mathcal{D}_{\mu,\lambda}$ as the class of distributions on $\mathcal{X} \times \mathcal{Y}$ such that $X \sim \mu$ and the output $Y = f(X) + \mathcal{N}(0,1)$ where f is any λ-Lipschitz function from $\mathcal{X}$ to $\mathcal{Y}$. Fix a sample size $n > 0$ and let f_n denote any regressor on samples $(\mathbf{X}, \mathbf{Y})$ of size n, i.e. f_n maps any such sample to a function $f_{n|(\mathbf{X},\mathbf{Y})}(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$ in $L^2(\mu)$. There exists a constant C independent of n and λ such that

$$\inf_{\{f_n\}} \sup_{\mathcal{D}_{\mu,\lambda}} \frac{\mathbb{E}_{\mathbf{X},\mathbf{Y},x} \left|f_{n|(\mathbf{X},\mathbf{Y})}(x) - f(x)\right|^2}{\lambda^{2d/(2+d)} n^{-2/(2+d)}} \geq C.$$

3.3 Choosing k for near-optimal rates at x

Our last result shows a practical and simple way to choose k locally so as to nearly achieve the minimax rate at x, i.e. a rate that depends on the unknown local dimension in a neighborhood $B(x,r)$ of x, where again, r satisfies $\mu(B(x,r)) > \Omega(k/n)$ for good choices of k. It turns out that we just need $\mu(B(x,r)) > \Omega(n^{-1/3})$.

As we will see, the choice of k simply consists of monitoring the distances from x to its nearest neighbors. The intuition, similar to that of a method for tree-pruning in [7], is to look for a k that balances the variance (roughly $1/k$) and the square bias (roughly $r_{k,n}^2(x)$) of the estimate. The procedure is as follows:

Choosing k at x: Pick $\Delta \geq \max_i \rho(x, X_i)$, and pick $\theta_{n,\delta} \geq \ln n/\delta$.
Let k_1 be the highest integer in $[n]$ such that $\Delta^2 \cdot \theta_{n,\delta}/k_1 \geq r_{k_1,n}^2(x)$.
Define $k_2 = k_1 + 1$ and choose k as $\arg\min_{k_i, i\in[2]} \left(\theta_{n,\delta}/k_i + r_{k_i,n}^2(x)\right)$.

The parameter $\theta_{n,\delta}$ *guesses* how the noise in Y affects the risk. This will soon be clearer. Performance guarantees for the above procedure are given in the following theorem.

Theorem 3. *Suppose μ is (C_0, d_0)-maximally-homogeneous, and $\mathcal{B}$ has finite VC dimension $\mathcal{V}_\mathcal{B}$. Assume k is chosen for each $x \in \mathcal{X}$ using the above procedure, and let $f_{n,k}(x)$ be the corresponding estimate. Let $0 < \delta < 1$ and suppose $n^{4/(6+3d_0)} > (\mathcal{V}_\mathcal{B} \ln 2n + \ln(8/\delta))/\theta_{n,\delta}$. With probability at least $1 - 2\delta$ over the choice of $(\mathbf{X}, \mathbf{Y})$, the following holds simultaneously for all $x \in \mathcal{X}$.*

Pick any $x \in \mathcal{X}$. Let $0 < r < \Delta$ satisfy $\mu(B(x,r)) > 6C_0 n^{-1/3}$. Suppose μ is (C,d)-homogeneous on $B(x,r)$, with $1 \le C \le C_0$ and $1 \le d \le d_0$. We have

$$|f_{n,k}(x) - f(x)|^2 \le \left(\frac{2C_{n,\delta}}{\theta_{n,\delta}} + 2\lambda^2\right)\left(1 + 4\Delta^2\right)\left(\frac{3C\theta_{n,\delta}}{n\mu(B(x,r))}\right)^{2/(2+d)},$$

where $C_{n,\delta} = \left(V_\mathcal{B} \cdot t_Y^2(\delta/2n) \cdot \ln(2n/\delta) + \sigma_Y^2\right) K(0)/K(1)$.

Suppose we set $\theta_{n,\delta} = \ln^2 n/\delta$. Then, as per the discussion in Section 2.1, if the noise in Y is Gaussian, we have $t_Y^2(\delta/2n) = O(\ln n/\delta)$, and therefore the factor $C_{n,\delta}/\theta_{n,\delta} = O(1)$. Thus ideally we want to set $\theta_{n,\delta}$ to the order of $(t_Y^2(\delta/2n) \cdot \ln n/\delta)$.

Just as in Theorem 1, the rates of Theorem 3 hold uniformly for all $x \in \mathcal{X}$, and all $0 < r < \Delta$ where $\mu(B(x,r)) > \Omega(n^{-1/3})$. For any such r, let us call $B(x,r)$ an *admissible* neighborhood. It is clear that, as n grows to infinity, w.h.p. any neighborhood $B(x,r)$ of x, $0 < r < \sup_{x' \in \mathcal{X}} \rho(x, x')$, becomes admissible. Once a neighborhood $B(x,r)$ is admissible for some n, our procedure nearly attains the minimax rates in terms of the local dimension on $B(x,r)$, provided μ is doubling on $B(x,r)$. Again, the mass of an admissible neighborhood affects the rate, and the bound in Theorem 3 is best for an admissible neighborhood with large mass $\mu(B(x,r))$ and small dimension d.

4 Analysis

Define $\widetilde{f}_{n,k}(x) = \mathbb{E}_{\mathbf{Y}|\mathbf{X}} f_{n,k}(x) = \sum_i w_{i,k}(x) f(X_i)$. We will bound the error of the estimate at a point x in a standard way as

$$|f_{n,k}(x) - f(x)|^2 \le 2\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 + 2\left|\widetilde{f}_{n,k}(x) - f(x)\right|^2. \tag{1}$$

Theorem 1 is therefore obtained by combining bounds on the above two r.h.s terms (variance and bias). These terms are bounded separately in Lemma 2 and Lemma 3 below.

4.1 Local rates for fixed k: bias and variance at x

In this section we bound the bias and variance terms of equation (1) with high probability, uniformly over $x \in \mathcal{X}$. We will need the following lemma which follows easily from standard VC theory [14] results. The proof is given as supplement in the appendix.

Lemma 1. *Let $\mathcal{B}$ denote the class of balls on $\mathcal{X}$, with VC-dimension $\mathcal{V}_\mathcal{B}$. Let $0 < \delta < 1$, and define $\alpha_n = (\mathcal{V}_\mathcal{B} \ln 2n + \ln(8/\delta))/n$. The following holds with probability at least $1 - \delta$ for all balls in $\mathcal{B}$. Pick any $a \ge \alpha_n$. Then $\mu(B) \ge 3a \implies \mu_n(B) \ge a$ and $\mu_n(B) \ge 3a \implies \mu(B) \ge a$.*

We start with the bias which is simpler to handle: it is easy to show that the bias of the estimate at x depends on the radius $r_{k,n}(x)$. This radius can then be bounded, first in expectation using the doubling assumption on μ, then by calling on the above lemma to relate this expected bound to $r_{k,n}(x)$ with high probability.

Lemma 2 (Bias)**.** *Suppose μ is (C_0, d_0)-maximally-homogeneous. Let $0 < \delta < 1$. With probability at least $1 - \delta$ over the choice of $\mathbf{X}$, the following holds simultaneously for all $x \in \mathcal{X}$ and k satisfying $n > k \ge \mathcal{V}_\mathcal{B} \ln 2n + \ln(8/\delta)$.*

Pick any $x \in \mathcal{X}$. Let $r > 0$ satisfy $\mu(B(x,r)) > 3C_0 k/n$. Suppose μ is (C,d)-homogeneous on $B(x,r)$, with $1 \le C \le C_0$ and $1 \le d \le d_0$. We have:

$$\left|\widetilde{f}_{n,k}(x) - f(x)\right|^2 \le \lambda^2 r^2 \left(\frac{3Ck}{n\mu(B(x,r))}\right)^{2/d}.$$

Proof. First fix $\mathbf{X}$, $x \in \mathcal{X}$ and $k \in [n]$. We have:

$$\left|\widetilde{f}_{n,k}(x) - f(x)\right| = \left|\sum_i w_{i,k}(x)\left(f(X_i) - f(x)\right)\right| \leq \sum_i w_{i,k}(x)\left|f(X_i) - f(x)\right|$$
$$\leq \sum_i w_{i,k}(x)\lambda\rho\left(X_i, x\right) \leq \lambda r_{k,n}(x). \tag{2}$$

We therefore just need to bound $r_{k,n}(x)$. We proceed as follows.

Fix $x \in \mathcal{X}$ and k and pick any $r > 0$ such that $\mu(B(x,r)) > 3C_0k/n$. Suppose μ is (C,d)-homogeneous on $B(x,r)$, with $1 \leq C \leq C_0$ and $1 \leq d \leq d_0$. Define

$$\epsilon \doteq \left(\frac{3Ck}{n\mu(B(x,r))}\right)^{1/d},$$

so that $\epsilon < 1$ by the bound on $\mu(B(x,r))$; then by the local doubling assumption on $B(x,r)$, we have $\mu(B(x,\epsilon r)) \geq C^{-1}\epsilon^d\mu(B(x,r)) \geq 3k/n$. Let α_n as defined in Lemma 1, and assume $k/n \geq \alpha_n$ (this is exactly the assumption on k in the lemma statement). By Lemma 1, it follows that with probability at least $1-\delta$ uniform over x, r and k thus chosen, we have $\mu_n((B(x,\epsilon r)) \geq k/n$ implying that $r_{k,n}(x) \leq \epsilon r$. We then conclude with the lemma statement by using equation (2). □

Lemma 3 (Variance)**.** *Let* $0 < \delta < 1$. *With probability at least* $1 - 2\delta$ *over the choice of* $(\mathbf{X}, \mathbf{Y})$*, the following then holds simultaneously for all* $x \in \mathcal{X}$ *and* $k \in [n]$*:*

$$\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 \leq \frac{K(0)}{K(1)} \cdot \frac{\mathcal{V}_{\mathcal{B}} \cdot t_Y^2(\delta/2n) \cdot \ln(2n/\delta) + \sigma_Y^2}{k}.$$

Proof. First, condition on $\mathbf{X}$ fixed. For any $x \in \mathcal{X}$, $k \in [k]$, let $\mathbf{Y}_{x,k}$ denote the subset of $\mathbf{Y}$ corresponding to points from $\mathbf{X}$ falling in $B(x, r_{k,n}(x))$. For $\mathbf{X}$ fixed, the number of such subsets $\mathbf{Y}_{x,k}$ is therefore at most the number of ways we can intersect balls in $\mathcal{B}$ with the sample $\mathbf{X}$; this is in turn upper-bounded by $n^{\mathcal{V}_{\mathcal{B}}}$ as is well-known in VC theory.

Let $\psi(\mathbf{Y}_{x,k}) \doteq \left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|$. We'll proceed by showing that with high probability, for all $x \in \mathcal{X}$, $\psi(\mathbf{Y}_{x,k})$ is close to its expectation, then we bound this expectation.

Let $\delta_0 \leq 1/2n$. We further condition on the event $\mathcal{Y}_{\delta_0}$ that for all n samples Y_i, $|Y_i - f(X_i)| \leq t_Y(\delta_0)$. By definition of $t_Y(\delta_0)$, the event $\mathcal{Y}_{\delta_0}$ happens with probability at least $1 - n\delta_0 \geq 1/2$. It follows that for any $x \in \mathcal{X}$

$$\mathbb{E}\,\psi(\mathbf{Y}_{x,k}) \geq \mathbb{P}\left(\mathcal{Y}_{\delta_0}\right) \cdot \mathop{\mathbb{E}}_{\mathcal{Y}_{\delta_0}} \psi(\mathbf{Y}_{x,k}) \geq \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{Y}_{\delta_0}} \psi(\mathbf{Y}_{x,k}),$$

where $\mathbb{E}_{\mathcal{Y}_{\delta_0}}[\cdot]$ denote conditional expectation under the event. Let $\epsilon > 0$, we in turn have

$$\mathbb{P}\left(\exists x, k,\ \psi(\mathbf{Y}_{x,k}) > 2\mathbb{E}\,\psi(\mathbf{Y}_{x,k}) + \epsilon\right) \leq \mathbb{P}\left(\exists x, k,\ \psi(\mathbf{Y}_{x,k}) > \mathop{\mathbb{E}}_{\mathcal{Y}_{\delta_0}} \psi(\mathbf{Y}_{x,k}) + \epsilon\right)$$
$$\leq \mathbb{P}_{\mathcal{Y}_{\delta_0}}\left(\exists x, k,\ \psi(\mathbf{Y}_{x,k}) > \mathop{\mathbb{E}}_{\mathcal{Y}_{\delta_0}} \psi(\mathbf{Y}_{x,k}) + \epsilon\right) + n\delta_0.$$

This last probability can be bounded by applying McDiarmid's inequality: changing any Y_i value changes $\psi(\mathbf{Y}_{x,k})$ by at most $w_{i,k} \cdot t_Y(\delta_0)$ when we condition on the event $\mathcal{Y}_{\delta_0}$. This, followed by a union-bound yields

$$\mathbb{P}_{\mathcal{Y}_{\delta_0}}\left(\exists x, k,\ \psi(\mathbf{Y}_{x,k}) > \mathop{\mathbb{E}}_{\mathcal{Y}_{\delta_0}} \psi(\mathbf{Y}_{x,k}) + \epsilon\right) \leq n^{\mathcal{V}_{\mathcal{B}}} \exp\left\{-2\epsilon^2/t_Y^2(\delta_0)\sum_i w_{i,k}^2\right\}.$$

Combining with the above we get

$$\mathbb{P}\left(\exists x \in \mathcal{X},\, \psi(\mathbf{Y}_{x,k}) > 2\mathbb{E}\,\psi(\mathbf{Y}_{x,k}) + \epsilon\right) \leq n^{\mathcal{V}_{\mathcal{B}}} \exp\left\{-2\epsilon^2/t_Y^2(\delta_0)\sum_i w_{i,k}^2\right\} + n\delta_0.$$

In other words, let $\delta_0 = \delta/2n$, with probability at least $1-\delta$, for all $x \in \mathcal{X}$ and $k \in [n]$

$$\begin{aligned}\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 &\leq 8\left(\mathop{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}}\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|\right)^2 + t_Y^2(\delta/2n)\left(\mathcal{V}_{\mathcal{B}}\ln(2n/\delta)\sum_i w_{i,k}^2\right)\\ &\leq 8\mathop{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}}\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 + t_Y^2(\delta/2n)\left(\mathcal{V}_{\mathcal{B}}\ln(2n/\delta)\sum_i w_{i,k}^2\right),\end{aligned}$$

where the second inequality is an application of Jensen's.

We bound the above expectation on the r.h.s. next. In what follows (second equality below) we use the fact that for i.i.d random variables z_i with zero mean, $\mathbb{E}\left|\sum_i z_i\right|^2 = \sum_i \mathbb{E}\left|z_i\right|^2$. We have

$$\begin{aligned}\mathop{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}}\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 &= \mathop{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}}\left|\sum_i w_{i,k}(x)\left(Y_i - f(X_i)\right)\right|^2\\ &= \sum_i w_{i,k}^2(x)\mathop{\mathbb{E}}_{\mathbf{Y}|\mathbf{X}}\left|Y_i - f(X_i)\right|^2 \leq \sum_i w_{i,k}^2(x)\sigma_Y^2.\end{aligned}$$

Combining with the previous bound we get that, with probability at least $1-\delta$, for all x and k,

$$\left|f_{n,k}(x) - \widetilde{f}_{n,k}(x)\right|^2 \leq \left(\mathcal{V}_{\mathcal{B}} \cdot t_Y^2(\delta/2n)\cdot\ln(2n/\delta) + \sigma_Y^2\right)\cdot\sum_i w_{i,k}^2(x). \tag{3}$$

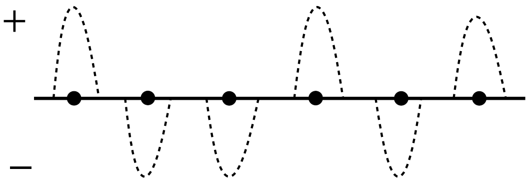
We can now bound $\sum_i w_{i,k}^2(x)$ as follows:

$$\begin{aligned}\sum_i w_{i,k}^2(x) &\leq \max_{i\in[n]} w_{i,k}(x) = \max_{i\in[n]}\frac{K\left(\rho(x,x_i)/r_{k,n}(x)\right)}{\sum_j K\left(\rho(x,x_j)/r_{k,n}(x)\right)} \leq \frac{K(0)}{\sum_j K\left(\rho(x,x_j)/r_{k,n}(x)\right)}\\ &\leq \frac{K(0)}{\sum_{x_j\in B(x,r_{k,n}(x))} K\left(\rho(x,x_j)/r_{k,n}(x)\right)} \leq \frac{K(0)}{K(1)k}.\end{aligned}$$

Plug this back into equation 3 and conclude. □

4.2 Minimax rates for a doubling measure

The minimax rates of theorem 2 (proved in the appendix) are obtained as is commonly done by constructing a regression problem that reduces to the problem of binary classification (see e.g. [1, 2, 10]). Intuitively the problem of classification is hard in those instances where labels (say $-1, +1$) vary wildly over the space $\mathcal{X}$, i.e. close points can have different labels. We make the regression problem similarly hard. We will consider a class of candidate regression functions such that each function f alternates between positive and negative in neighboring regions (f is depicted as the dashed line below).



The reduction relies on the simple observation that for a regressor f_n to approximate the right f from data it needs to at least identify the sign of f in the various regions of space. The more we can make each such f change between positive and negative, the harder the problem. We are however constrained in how much f changes since we also have to ensure that each f is Lipchitz continuous.

4.3 Choosing k for near-optimal rates at x

Proof of Theorem 3. Fix x and let r, d, C as defined in the theorem statement. Define

$$\kappa \doteq \theta_{n,\delta}^{d/(2+d)} \cdot \left(\frac{n\mu(B(x,r))}{3C}\right)^{2/(2+d)} \text{ and } \epsilon \doteq \left(\frac{3C\kappa}{n\mu(B(x,r))}\right)^{1/d}.$$

Note that, by our assumptions,

$$\mu(B(x,r)) > 6C\theta_{n,\delta}n^{-1/3} \geq 6C\theta_{n,\delta}n^{-d/(2+d)} = 6C\theta_{n,\delta}\frac{n^{2/(2+d)}}{n} \geq 6C\frac{\kappa}{n}. \tag{4}$$

The above equation (4) implies $\epsilon < 1$. Thus, by the homogeneity assumption on $B(x,r)$, $\mu(B(x,\epsilon r)) \geq C^{-1}\epsilon^d\mu(B(x,r)) \geq 3\kappa/n$. Now by the first inequality of (4) we also have

$$\frac{\kappa}{n} \geq \frac{\theta_{n,\delta}}{n}n^{4/(6+3d)} \geq \frac{\theta_{n,\delta}}{n}n^{4/(6+3d_0)} \geq \alpha_n,$$

where $\alpha_n = \left(\mathcal{V}_{\mathcal{B}}\ln 2n + \ln(8/\delta)\right)/n$ is as defined in Lemma 1. We can thus apply Lemma 1 to have that, with probability at least $1-\delta$, $\mu_n(B(x,\epsilon r)) \geq \kappa/n$. In other words, for any $k \leq \kappa$, $r_{k,n}(x) \leq \epsilon r$. It follows that if $k \leq \kappa$,

$$\frac{\Delta^2 \cdot \theta_{n,\delta}}{k} \geq \frac{\Delta^2 \cdot \theta_{n,\delta}}{\kappa} = \Delta^2\left(\frac{3C\kappa}{n\mu(B(x,r))}\right)^{2/d} \geq (\epsilon r)^2 \geq r_{k,n}^2(x).$$

Remember that the above inequality is exactly the condition on the choice of k_1 in the theorem statement. Therefore, suppose $k_1 \leq \kappa$, it must be that $k_2 > \kappa$ otherwise k_2 is the highest integer satisfying the condition, contradicting our choice of k_1. Thus we have (i) $\theta_{n,\delta}/k_2 < \theta_{n,\delta}/\kappa = \epsilon^2$. We also have (ii) $r_{k_2,n}(x) \leq 2^{1/d}\epsilon r$. To see this, notice that since $k_1 \leq \kappa < k_2 = k_1 + 1$ we have $k_2 \leq 2\kappa$; now by repeating the sort of argument above, we have $\mu(B(x, 2^{1/d}\epsilon r)) \geq 6\kappa/n$ which by Lemma 1 implies that $\mu_n(B(x, 2^{1/d}\epsilon r)) \geq 2\kappa/n \geq k_2/n$.

Now suppose instead that $k_1 > \kappa$, then by definition of k_1, we have (iii)

$$r_{k_1,n}(x)^2 \leq \frac{\Delta^2 \cdot \theta_{n,\delta}}{k_1} \leq \frac{\Delta^2 \cdot \theta_{n,\delta}}{\kappa} = (\Delta\epsilon)^2.$$

The following holds by (i), (ii), and (iii). Let k be chosen as in the theorem statement. Then, whether $k_1 > \kappa$ or not, it is true that

$$\left(\frac{\theta_{n,\delta}}{k} + r_{k,n}^2(x)\right) \leq \left(1 + 4\Delta^2\right)\epsilon^2 = \left(1 + 4\Delta^2\right)\left(\frac{3C\theta_{n,\delta}}{n\mu(B(x,r))}\right)^{2/(2+d)}.$$

Now combine Lemma 3 with equation (2) and we have that with probability at least $1 - 2\delta$ (accounting for all events discussed)

$$\begin{aligned}|f_{n,k}(x) - f(x)|^2 &\leq \frac{2C_{n,\delta}}{\theta_{n,\delta}}\frac{\theta_{n,\delta}}{k} + 2\lambda^2 r_{k,n}^2(x) \leq \left(\frac{2C_{n,\delta}}{\theta_{n,\delta}} + 2\lambda^2\right)\left(\frac{\theta_{n,\delta}}{k} + r_{k,n}^2(x)\right) \\ &\leq \left(\frac{2C_{n,\delta}}{\theta_{n,\delta}} + 2\lambda^2\right)\left(1 + 4\Delta^2\right)\left(\frac{3C\theta_{n,\delta}}{n\mu(B(x,r))}\right)^{2/(2+d)}.\end{aligned}$$

□

5 Final remark

The problem of choosing $k = k(x)$ optimally at x is similar to the problem of local bandwidth selection for kernel-based methods (see e.g. [15, 16]), and our method for choosing k might yield insights into bandwidth selection, since k-NN and kernel regression methods only differ in their notion of neighborhood of a query x.

Acknowledgments

I am grateful to David Balduzzi for many useful discussions.

References

[1] C. J. Stone. Optimal rates of convergence for non-parametric estimators. *Ann. Statist.*, 8:1348–1360, 1980.

[2] C. J. Stone. Optimal global rates of convergence for non-parametric estimators. *Ann. Statist.*, 10:1340–1353, 1982.

[3] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2000.

[4] J. Tenebaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[6] P. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Tech. Re. Dep. of Stats. UC Berkley*, 2006.

[7] S. Kpotufe. Escaping the curse of dimensionality with a tree-based regressor. *Conference On Learning Theory*, 2009.

[8] S. Kpotufe. Fast, smooth, and adaptive regression in metric spaces. *Neural Information Processing Systems*, 2009.

[9] S. Kulkarni and S. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41, 1995.

[10] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer, New York, NY, 2002.

[11] C. Cutler. A review of the theory and estimation of fractal dimension. *Nonlinear Time Series and Chaos, Vol. I: Dimension Estimation and Models*, 1993.

[12] K. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, 2005.

[13] M. do Carmo. *Riemannian Geometry*. Birkhauser, 1992.

[14] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their expectation. *Theory of probability and its applications*, 16:264–280, 1971.

[15] J. G. Staniswalis. Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, 84:284–288, 1989.

[16] R. Cao-Abad. Rate of convergence for the wild bootstrap in nonpara- metric regression. *Annals of Statistics*, 19:2226–2231, 1991.

Appendix

A Proof of Lemma 1

Lemma 4 (Relative VC bounds [14])**.** *Let $\mathcal{B}$ be a class of subsets of $\mathcal{X}$. Let $0 < \delta < 1$. Suppose a sample $\mathbf{X}$ of size n is drawn independently at random from a distribution μ over $\mathcal{X}$ with resulting empirical distribution μ_n. Define $\alpha_n = \left(\mathcal{V}_{\mathcal{B}} \ln 2n + \ln(8/\delta)\right)/n$.*

Then with probability at least $1 - \delta$ over the choice of $\mathbf{X}$, all $B \in \mathcal{B}$ satisfy

$$\mu(B) \leq \mu_n(B) + \sqrt{\mu_n(B)\alpha_n} + \alpha_n, \text{ and}$$
$$\mu_n(B) \leq \mu(B) + \sqrt{\mu(B)\alpha_n} + \alpha_n.$$

Lemma 1 is then obtained as the following corollary to Lemma 4 above.

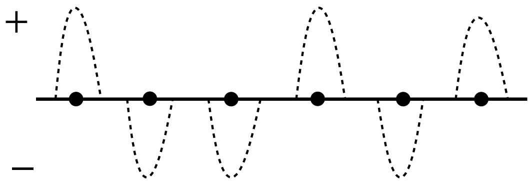
Corollary 1. *Let $\mathcal{B}$ denote the class of balls on $\mathcal{X}$. Let $0 < \delta < 1$, and as in Lemma 4 above, define $\alpha_n = \left(\mathcal{V}_{\mathcal{B}} \ln 2n + \ln(8/\delta)\right)/n$.*

The following holds with probability at least $1 - \delta$ for all balls in $\mathcal{B}$. Pick any $a \geq \alpha_n$.

- $\mu(B) \geq 3a \implies \mu_n(B) + \sqrt{\mu_n(B)\alpha_n} \geq \mu(B) - \alpha_n \geq a + \sqrt{a\alpha_n} \implies \mu_n(B) \geq a.$
- $\mu_n(B) \geq 3a \implies \mu(B) + \sqrt{\mu(B)\alpha_n} \geq \mu_n(B) - \alpha_n \geq a + \sqrt{a\alpha_n} \implies \mu(B) \geq a.$

B Proof of Theorem 2

The minimax rates shown here are obtained as is commonly done by constructing a regression problem that reduces to the problem of binary classification (see e.g. [1, 2, 10]). Intuitively the problem of classification is hard in those instances where labels (say $-1, +1$) vary wildly over the space $\mathcal{X}$, i.e. close points can have different labels. We make the regression problem similarly hard. We will consider a class of candidate regression functions such that each function f alternates between positive and negative in neighboring regions (f is depicted as the dashed line below).



The reduction relies on the simple observation that for a regressor f_n to approximate the right f from data it needs to at least identify the sign of f in the various regions of space. The more we can make each such f change between positive and negative, the harder the problem. We are however constrained in how much f changes since we also have to ensure that each f is Lipchitz continuous. Thus if f is to be positive in some region and negative in another, these regions cannot be too close. We therefore have to break up the space $\mathcal{X}$ into a set of regions whose centers are far enough to ensure smoothness of the candidates f, yet in a way that the set is large so that each f could be made to alternate a lot. We will therefore pick the centers of these regions as an r-net of $\mathcal{X}$ for some appropriate choice of r; by definition the centers would be r far apart and together they would form an r-cover over space. As it turns out, any r-net is large under some tightness conditions on the expansion constants of μ.

We start with the following lemma which upper and lower bounds the size of an r-net under some tightness conditions on the doubling behavior of μ. Results of this type appear in different forms in the literature. In particular similar upper-bounds on r-net size are discussed in [12, **?**]. Here, we are mainly interested in the lower-bound on r-net size but show both upper and lower bounds for completion. Both upper and lower bounds relie on union-bounds over sets of balls centered on a net.

Lemma 5. *Let μ be a measure on $\mathcal{X}$ such that for all $x \in \mathcal{X}$, for all $r > 0$ and $0 < \epsilon < 1$,*

$$C_1\epsilon^{-d}\mu(B(x,\epsilon r)) \leq \mu(B(x,r)) \leq C_2\epsilon^{-d}\mu(B(x,\epsilon r)),$$

where C_1, C_2 and d are positive constants independent of x, r, and ϵ. Then, there exist C_1', C_2' such that, for all $x \in \mathcal{X}$, for all $r > 0$ and $0 < \epsilon < 1$, an (ϵr)-net of $B(x,r)$ has size at least $C_1'\epsilon^{-d}$ and at most $C_2'\epsilon^{-d}$.

Proof. Fix $B(x,r)$, and consider an (ϵr)-net Z of $B(x,r)$. First we handle the upper-bound on $|Z|$.

Since, by a triangle inequality, we have for any $z \in Z$, $B(x,r) \subset B(z,2r)$, it follows by the assumption on μ that,

$$\mu\left(B\left(z, \frac{\epsilon}{2}r\right)\right) \geq C_2^{-1}4^{-d}\epsilon^d\mu\left(B(z,2r)\right) \geq C_2^{-1}4^{-d}\epsilon^d\mu\left(B(x,r)\right).$$

Now, since for any such z, $B\left(z, \frac{\epsilon}{2}r\right) \subset B(x,2r)$, and the balls $B\left(z, \frac{\epsilon}{2}r\right)$ for $z \in Z$ are disjoint (centers are ϵr apart) we have

$$|Z|\, C_2^{-1}4^{-d}\epsilon^d\mu\left(B(x,r)\right) \leq \mu\left(\bigcup_{z\in Z} B\left(z, \frac{\epsilon}{2}r\right)\right) \leq \mu(B(x,2r)) \leq C_2 2^d\mu(B(x,r)),$$

implying that $|Z| \leq C_2^2 8^d\epsilon^{-d}$.

The lower-bound on $|Z|$ is handled similarly as follows. We have for all $z \in Z$

$$\mu(B(z,\epsilon r)) \leq C_1^{-1}\epsilon^d\mu(B(z,r)) \leq C_1^{-1}\epsilon^d\mu(B(x,2r)).$$

Thus, applying a union-bound we obtain

$$|Z|\, C_1^{-1}\epsilon^d\mu(B(x,2r)) \geq \mu\left(\bigcup_{z\in Z} B\left(z, \epsilon r\right)\right) \geq \mu(B(x,r)) \geq C_1 2^{-d}\mu(B(x,2r)),$$

implying that $|Z| \geq C_1^2 2^{-d}\epsilon^d$. □

The proof of the minimax theorem follows.

Proof of Theorem 2. In what follows, for any function $g : \mathcal{X} \mapsto \mathbb{R}$, define $\|g\|^2 = \mathbb{E}_x |g(x)|^2$, i.e. the squared norm of g in $L^2(\mu)$.

Let $r_n = (\lambda^2 n)^{-1/(2+d)}$, and let Z be an r_n-net of $\mathcal{X}$. Define $\tau = \min\left\{\frac{1}{3}C_1^{1/d}, \frac{1}{4}\right\}$. For every $z \in Z$ we define the following function

$$g_z(x) \doteq \frac{\lambda}{5}\left(\tau r_n - \rho\left(x,z\right)\right)_+ \text{ i.e. } g_z(x) = 0 \text{ whenever } \rho\left(x,z\right) \geq \tau r_n.$$

It is easy to verify that for all $x, x' \in \mathcal{X}$

$$|g_z(x) - g_z(x')| \leq \frac{\lambda}{5}\left|\rho\left(x,z\right) - \rho\left(x',z\right)\right| \leq \frac{\lambda}{5}\rho\left(x,x'\right),$$

that is, g_z is $\frac{\lambda}{5}$-Lipschitz.

Now consider the random vector $\varsigma = \{\varsigma_z\}_{z\in Z}$ where the coordinates $\varsigma_z \in \{-1, 1\}$ are independent Bernoulli r.v.s that are 1 with probability $1/2$. For every possible value of ς define

$$f_\varsigma(x) \doteq \sum_{z\in Z}\varsigma_z g_z(x).$$

Next we verify that f_ς is λ-Lipschitz. First pick x and x' from the same ball $B(z, \tau r_n)$. It is clear that

$$|f_\varsigma(x) - f_\varsigma(x')| = |g_z(x) - g_z(x')| \leq \lambda\rho\left(x,x'\right).$$

Now suppose x or x' or both are outside all balls $\{B(z, \tau r_n)\}_{z\in Z}$, then again it is easy to see that $|f_\varsigma(x) - f_\varsigma(x')| \leq \lambda \rho\,(x, x')$. We now check the final case where $x \in B(z, \tau r_n)$ and $x' \in B(z', \tau r_n)$, $z \neq z'$. To this end, first notice that the ring $B(z, r_n/2) \setminus B(z, \tau r_n)$ is non-empty since $\mu(B(z, r_n/2)) \geq C_1(2\tau)^{-d}\mu(B(z, \tau r_n)) > \mu(B(z, \tau r_n))$. Pick x'' in this ring, and notice that x'' is outside both balls $B(z, \tau r_n)$ and $B(z', \tau r_n)$. Thus, $g_z(x'') = g_{z'}(x'') = 0$, and we can write

$$\begin{aligned}
|f_\varsigma(x) - f_\varsigma(x')| &= |g_z(x) - g_{z'}(x')| \leq |g_z(x) - g_z(x'')| + |g_{z'}(x'') - g_{z'}(x')| \\
&\leq \frac{\lambda}{5}\left(\rho\,(x, x'') + \rho\,(x'', x')\right) \leq \frac{\lambda}{5}\left(2\rho\,(x, x'') + \rho\,(x, x')\right) \\
&\leq \frac{\lambda}{5}\left(2r_n + \rho\,(x, x')\right) \leq \frac{\lambda}{5}\left(4\rho\,(x, x') + \rho\,(x, x')\right) = \lambda\rho\,(x, x')\,.
\end{aligned}$$

At this point we can define $\mathcal{D}$ as the class of distributions on $\mathcal{X} \times \mathcal{Y}$, where $X \sim \mu$ and $Y = f_\varsigma(X) + \mathcal{N}(0, 1)$, for some f_ς as constructed above. Clearly $\mathcal{D} \subset \mathcal{D}_{\mu,\lambda}$ and we just have to show that

$$\inf_{\{f_n\}} \sup_{\mathcal{D}} \frac{\mathbb{E}\,\|f_n - f_\varsigma\|^2}{\lambda^2 r_n^2} \geq O(1).$$

Fix a regressor f_n, that is f_n maps any sample $(\mathbf{X}, \mathbf{Y})$ to a function in $L^2(\mu)$, which, for simplicity of notation, we also denote by f_n. For $(\mathbf{X}, \mathbf{Y})$ fixed, we denote by $f_{n,Z}$ the projection (in the Hilbert space $L^2(\mu)$) of f_n onto the orthonormal system $\{g_z/\|g_z\|\}_{z\in Z}$. In other words, $f_{n,Z} = \sum_{z\in Z} \frac{<f_n, g_z>}{\|g_z\|^2} g_z = \sum_{z\in Z} w_{n,z} g_z$. Thus, for f_ς fixed, we have

$$\mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \|f_n - f_\varsigma\|^2 \geq \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \|f_{n,Z} - f_\varsigma\|^2 = \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \sum_{z\in Z} (w_{n,Z} - \varsigma_z)^2 \|g_z\|^2\,.$$

To bound $\|g_z\|$, notice that g_z is at least $\lambda\tau r_n/10$ on the ball $B(z, \tau r_n/2)$. This ball in turn has mass at least $C_2^{-1}(\tau r_n/2)^d$, so $\|g_z\| \geq C_3/\sqrt{n}$, for C_3 appropriately chosen. We therefore have

$$\begin{aligned}
\sup_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \|f_n - f_\varsigma\|^2 &\geq \frac{C_3^2}{n} \sup_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \sum_{z\in Z} (w_{n,z} - \varsigma_z)^2 \geq \frac{C_3^2}{n} \sup_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \sum_{z\in Z} \mathbb{1}\{w_{n,z} \cdot \varsigma_z < 0\} \\
&\geq \frac{C_3^2}{n} \mathop{\mathbb{E}}_{\varsigma} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \sum_{z\in Z} \mathbb{1}\{w_{n,z} \cdot \varsigma_z < 0\} \geq \frac{C_3^2}{n} \mathop{\mathbb{E}}_{\mathbf{X}} \sum_{z\in Z} \mathop{\mathbb{E}}_{Y,\varsigma|\mathbf{X}} \mathbb{1}\{w_{n,z} \cdot \varsigma_z < 0\}\,.
\end{aligned}$$

For $z \in Z$ fixed, $\mathbb{E}_{Y,\varsigma|\mathbf{X}}\, \mathbb{1}\{w_{n,Z} \cdot \varsigma_z < 0\}$ is the probability of error of a classifier (which outputs sign$(w_{n,z})$) for the following prediction task. Let $x_{(1)}, x_{(2)}, \ldots x_{(m)}$ denote the values of X falling in $B(z, \tau r_n)$ where g_z is non zero. Then

$$(Y_{(1)}, \ldots Y_{(m)}) = \varsigma_z(g_z(x_{(1)}), \ldots, g_z(x_{(m)})) + \mathcal{N}(0, I_m)$$

is a random vector sampled from the equal-weight mixture of two spherical Gaussians in $\mathbb{R}^m$ centered at $u \doteq (g_z(x_{(1)}), \ldots, g_z(x_{(m)}))$ and $-u$. The prediction task is that of identifying the right mixture component from the single sample $(Y_{(1)}, \ldots Y_{(m)})$. The smallest possible error for this task is that of the Bayes classifier and is well known to be $\Phi(-\|u\|) \geq \Phi\left(-\sqrt{\sum_{i=1}^n g_z^2(X_i)}\right)$. Since $\Phi(-\sqrt{\cdot})$ is convex, we can apply Jensen's inequality as follows.

$$\begin{aligned}
\sup_{\mathcal{D}} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \|f_n - f_\varsigma\|^2 &\geq \frac{C_3^2}{n} \sum_{z\in Z} \mathop{\mathbb{E}}_{\mathbf{X}} \Phi\left(-\sqrt{\sum_{i=1}^n g_z^2(X_i)}\right) \geq \frac{C_3^2}{n} \sum_{z\in Z} \Phi\left(-\sqrt{\sum_{i=1}^n \mathop{\mathbb{E}}_{\mathbf{X}} g_z^2(X_i)}\right) \\
&= \frac{C_3^2}{n} \sum_{z\in Z} \Phi\left(-\sqrt{n\,\|g_z\|^2}\right).
\end{aligned}$$

The norm $\|g_z\|$ is at most $C_3'/\sqrt{n}$ since $g_z \leq \lambda\tau r_n$ everywhere, and non zero only on the ball $B(z, \tau r_n)$ which has mass at most $C_1^{-1}(\tau r_n)^d$. Finally, remember that by Lemma 5, $|Z| \geq C_1' r_n^{-d}$. Using these two facts we have

$$\sup_{\mathcal{D}} \frac{1}{\lambda^2 r_n^2} \mathop{\mathbb{E}}_{\mathbf{X},\mathbf{Y}} \|f_n - f_\varsigma\|^2 \geq \frac{C_3^2}{\lambda^2 r_n^2 n} \sum_{z\in Z} \Phi(-C_3') \geq C_3^2 C_1' \tau^d \Phi(-C_3'),$$

which concludes the proof since f_n is arbitrarily chosen. □