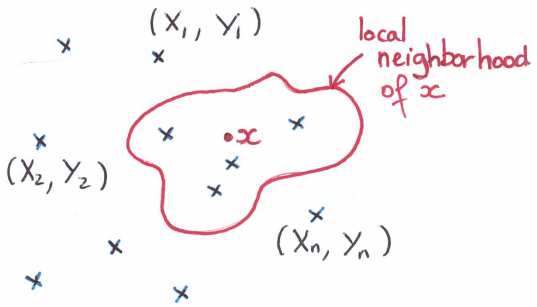
SL = K-NEAREST NEIGHBORS (K-NN)

The regression function $r(x) := \mathbb{E}(Y \mid X = x)$, introduced p. 8 in SL: FOUNDATIONS, provides an optimal predictor both in a regression context with square loss, and in the context of binary classification with 0-1 loss, see p. 8 and 10 of the same chapter.

A natural non-parametric approach to estimate $r(x)$ is based on local averaging: average the values of Y in the learning sample $\mathcal{L}_n := \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ associated with values of X belonging to a neighborhood of x.



↘ Different notions of neighborhood lead to different non-parametric estimators. However, local averaging estimators all take the form

$$f_n(x) = \sum_{i=1}^{n} W_{n,i}(x)\, Y_i ,$$

where $W_{n,i}$ = weight functions, and are usually taken such that $W_{n,i}(x) \geq 0$, and $\sum_{i=1}^{n} W_{n,i}(x) = 1$, $\forall x \in \mathcal{X}$.

Typically, $W_{n,i}(x) \geq W_{n,j}(x)$ if $d(x, X_i) \leq d(x, X_j)$ (some distance)

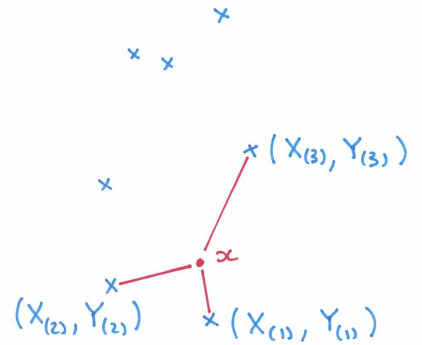
I_ K-NN ALGORITHM.

Fix an integer K in $\{1, \dots, n\}$, where n denotes the sample size. For a metric d on $\mathcal{X}$, and a point $x \in \mathcal{X}$, we define $\mathcal{N}(K, d, x)$ to be the random subset of $\{X_1, \dots, X_n\}$ composed of the K closest points to x. The K-NN estimate is defined to be the averaged value of the Y_is for which $X_i \in \mathcal{N}(K, d, x)$. The associated weights are $W_{n,i}(x) := \frac{1}{K} \mathbb{1}(X_i \in \mathcal{N}(K, d, x))$.

Alternatively, re-ordering the learning sample $\mathcal{L}_n$ such that $\| X_{(1)} - x \| \leq \dots \leq \| X_{(n)} - x \|$, $\mathcal{L}_n = \{(X_{(1)}, Y_{(1)}), \dots, (X_{(n)}, Y_{(n)})\}$, the K-NN estimate is

$$f_n(x) := \frac{1}{K} \sum_{i=1}^{K} Y_{(i)}$$

K-NN



Remarks

(i) Usually, the Euclidean distance is used

(ii) If $\| X_i - x \| = \| X_j - x \|$, we have a tie. Ties can be broken by indices: X_i is declared closer to X_j if $i < j$.

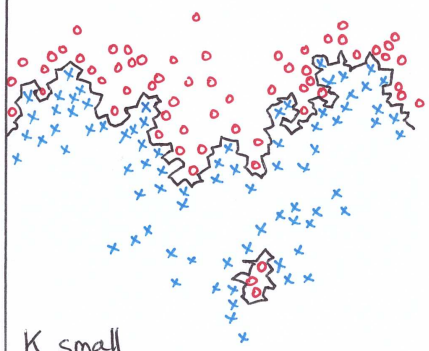
Alternatively, we may artificially introduce an additional Z independent of (X, Y), uniformly distributed on $[0, 1]$, and additional $Z_1, \dots, Z_n$, such that $(X_i, Z_i) \overset{d}{=} (X, Z)$. Ties can be broken depending on the value of Z. In this case, ties occur with probability 0. We assume in the remainder that this is the case.

(iii) In practice, we allow K to grow with n. The value of K controls the smoothness of the K-NN regression estimate. We illustrate this in the context of binary classification. In this case, the K-NN estimate is compared with the value $1/2$ to produce the final classifier

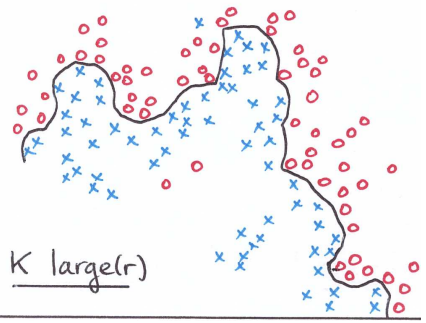
$$g_n(x) := \begin{cases} 1 & \text{if } f_n(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

A small value of K leads to a rough decision boundary, while increasing K leads to smoother boundaries

The usual bias-variance trade-off ↔



K small



K large(r)

Before presenting the properties of the K-NN estimator (consistency, convergence rates), we explain why local averaging techniques are expected to perform poorly in high-dimensional settings.

II. CURSE OF DIMENSIONALITY.

We explain through an example why non-parametric estimation of the regression function based on local averaging is particularly challenging if d is large, where d is the dimension of the input space (e.g. $\mathcal{X} = \mathbb{R}^d$). The main reason is that most points $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ are isolated: we cannot densely pack regions of $\mathcal{X}$ with "enough" observations to construct a reliable estimator of $\mathbb{E}(Y|X)$

Consider $X, X_1, \dots, X_n \in \mathbb{R}^d$ iid, uniformly distributed in $[0, 1]^d$. We propose to calculate a lower bound on the expected supremum-norm distance of X to its nearest neighbor in $X_1, \dots, X_n$:

$$d_\infty(d, n) := \mathbb{E}\left\{ \min_{1 \le \ell \le d} \| X - X_i \|_\infty \right\},$$

where $\|x\|_\infty := \max_{1 \le \ell \le d} |x^{(\ell)}|$; $x = (x^{(1)}, \dots, x^{(d)})^t \in \mathbb{R}^d$

We have

$$d_\infty(d, n) = \int_0^{+\infty} \mathbb{P}\left\{ \min_{1 \le \ell \le d} \| X - X_i \|_\infty > u \right\} du$$

$$= \int_0^{+\infty} \left(1 - \mathbb{P}\left\{ \min_{1 \le \ell \le d} \| X - X_i \|_\infty \le u \right\} \right) du .$$

Now,

$$\mathbb{P}\left\{\min_{1\le l\le d}\|X-X_i\|_\infty \le u\right\}$$

$$= \mathbb{P}\left(\bigcup_{1\le l\le d}\|X-X_i\|_\infty \le u\right)$$

(sub-additivity)

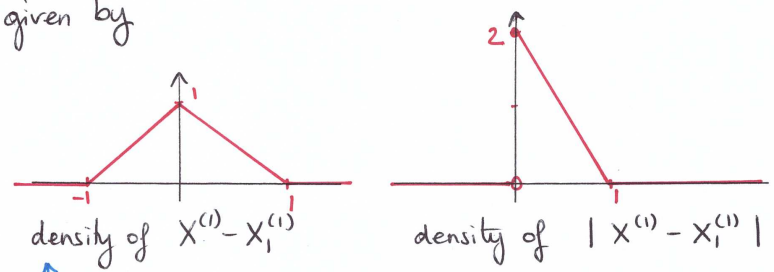
$$\le \sum_{l=1}^{d}\mathbb{P}\left(\|X-X_i\|_\infty \le u\right)$$

$$= n\,\mathbb{P}\left(\|X-X_1\|_\infty \le u\right)$$

$$= n\,\mathbb{P}\left(\bigcap_{l=1}^{d}|X^{(l)}-X_1^{(l)}| \le u\right)$$

$$= n\left\{\mathbb{P}\left(|X^{(1)}-X_1^{(1)}| \le u\right)\right\}^d.$$

Making use of $X^{(1)}, X_1^{(1)} \sim \mathcal{U}[0,1]$, independent, and the fact that the distribution of $X^{(1)}-X_1^{(1)}$ and $|X^{(1)}-X_1^{(1)}|$ is given by



Indeed, $f_{X^{(1)}-X_1^{(1)}}(x) = f_{X^{(1)}} * f_{-X_1^{(1)}}(x)$, with $X^{(1)} \sim \mathcal{U}[0,1]$, $-X_1^{(1)} \sim \mathcal{U}[-1,0]$.

($x \in [-1,1]$)

$$= \int f_{X^{(1)}}(u)\, f_{-X_1^{(1)}}(x-u)\,du$$

$$= \int_0^1 \mathbb{1}\left(x-u \in [-1,0]\right)du$$

$$= \int_0^1 \mathbb{1}\left(u \in [x, x+1]\right)du$$

- If $x \in [-1,0]$, then $[x,x+1] \cap [0,1] = [0, x+1]$, and $f_{X^{(1)}-X_1^{(1)}}(x) = \int_0^{x+1} du = x+1$
- If $x \in [0,1]$, then $[x,x+1] \cap [0,1] = [x, 1]$, and $f_{X^{(1)}-X_1^{(1)}}(x) = \int_x^1 du = 1-x$.

It follows that $\mathbb{P}\left\{|X^{(1)}-X_1^{(1)}| \le u\right\} = \int_0^u 2(1-x)\,dx$

$$= 2u - u^2 \le 2u$$

and

$$\mathbb{P}\left\{\min_{1\le l\le d}\|X-X_i\|_\infty \le u\right\} \le n(2u)^d.$$

$$\Rightarrow d_\infty(d,n) \ge \int_0^{+\infty}\left(1-n(2u)^d\right)\mathbb{1}(\dots \ge 0)\,du$$

($\dots \ge 0 \Leftrightarrow u \le \frac{1}{2n^{1/d}}$)

$$= \int_0^{\frac{1}{2n^{1/d}}}\left(1-n(2u)^d\right)du$$

$$= \left[u - n2^d\frac{u^{d+1}}{d+1}\right]_0^{1/2n^{1/d}}$$

$$\boxed{d_\infty(d,n) \ge \frac{d}{2(d+1)}n^{-1/d}}$$

↖ As d gets larger, $n^{-1/d}$ converges slower to zero.

The following values are presented in Györfi & al:

	$n=100$	$n=1000$	$n=100,000$
$d_\infty(1,n)$	≥ 0.0025	≥ 0.00025	≥ 0.0000025
$d_\infty(10,n)$	≥ 0.28	≥ 0.22	≥ 0.14
$d_\infty(20,n)$	≥ 0.37	≥ 0.34	≥ 0.26 ↖ SLOW !

⇒ For a new x, the available data in $\mathcal{L}_n$ are not close to x; and so it is hard to estimate $\mathbb{E}(Y|X=x)$ without imposing further assumptions on the shape or properties of the regression function $r(x)$. Alternatively, consider reducing the dimension of the input space as a pre-processing step ($\rightarrow$ PCA, ...).

Note that a similar lower bound can be derived if we replace the supremum norm by the Euclidean norm. Specifically, consider

$$d_2(d,n) := \mathbb{E}\left\{ \min_{1\leq i\leq n} \| X - X_i \|_2 \right\}.$$

Similar calculations show that we are lead to bound

$$\mathbb{P}\left(\| X - X_1 \|_2 \leq u \right) \leq \text{Volume of the } d\text{-dimensional sphere, with radius } u$$

$$= \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} u^d$$

Components of $(X - X_1)$ are independent, whose density ≤ 1.

Gamma function $\Gamma(x) := \int_0^\infty u^{x-1} e^{-u}\, du$

Thus $\mathbb{P}\left(\min_{1\leq i\leq n} \| X - X_i \|_2 \leq u \right) \leq n \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} u^d$,

so that

$$d_2(d,n) \geq \int_0^{+\infty} \left(1 - n \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} u^d \right) \mathbb{1}(\ldots \geq 0)\, du,$$

$$d_2(d,n) \geq \int_0^{C n^{-1/d}} \left(1 - \left(\frac{u}{C}\right)^d \right) du \;;$$

with $C := \left(\frac{\Gamma\left(\frac{d}{2}+1\right)}{\pi^{d/2}} \right)^{1/d}$

$$= \frac{d}{d+1} \left(\frac{\Gamma\left(\frac{d}{2}+1\right)^{1/d}}{\pi^{1/2}} \right) n^{-1/d}$$

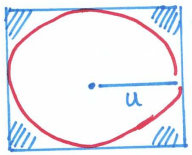
Same rate of convergence as for the supremum norm.

Remark:

Volume of a ball in $\mathbb{R}^d$ of radius u is $\frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)} u^d$

Volume of a cube in $\mathbb{R}^d$ of side $2u$ is $(2u)^d$

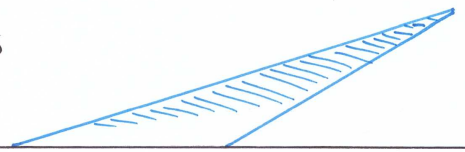
Ratio of the volume of the ball with volume of the cube is $\frac{\pi^{d/2}}{2^d\, \Gamma\left(\frac{d}{2}+1\right)} \longrightarrow 0$ as $d \to +\infty$ (use Stirling to approximate Γ)



← Consequence: in high dimension, most of the volume of the cube is concentrated in its corners.

In addition, the distance from the center to the edge is $u\sqrt{d}$ (direct application of Pythagora's theorem).

⇒ Corners of the cube become more & more pointy as d increases



III. PROPERTIES OF THE K-NN ESTIMATOR

Section III.1 states some properties of the K-NN rule, when K is held fixed. In Sections III.2 and III.3, we give consistency & convergence rates results when K is allowed to grow with n, $K = K_n \to \infty$, and $K_n/n \to 0$, as $n \to \infty$. Recall that the K-NN rule is given by

$$f_n(x) = \frac{1}{K} \sum_{i=1}^{K} Y_{(i)} .$$

↖ estimate of $r(x) = \mathbb{E}(Y \mid X = x)$.

In a binary classification context, the final classification rule is

$$g_n(x) = \begin{cases} 1 & \text{if } f_n(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

↖ not symmetric, unless K is odd.

→ The performance of f_n is evaluated under a square loss function: $\mathcal{E}(f_n) = R(f_n) - R^*$

$\left[R(f_n) = \mathbb{E}\{(Y - f_n(X))^2 \mid \mathcal{L}_n\} \; ; \; R^* = R(r) \right]$

→ The performance of g_n is evaluated under the 0-1 loss: $\mathcal{E}_{0-1}(g_n) = R_{0-1}(g_n) - R^*_{0-1}$

$\left[R_{0-1}(g_n) = \mathbb{P}(Y \neq g_n(X) \mid \mathcal{L}_n) \; ; \; R^*_{0-1} = \text{Bayes risk} \right]$

Recall that $\mathcal{E}_{01}(g_n) \leq 2\sqrt{\mathcal{E}(f_n)}$, see p.12 in SL: FOUNDATIONS.

Most results presented in this section can be found in [DGL] and [GKKW].

III.1. K fixed

The results in this section are presented in the context of binary classification. We assume that $Y \in \{0, 1\}$.

For the nearest neighbor rule ($K = 1$), we have that

$$\mathbb{E}\{R_{0-1}(g_n)\} \xrightarrow{n \to \infty} 2\,\mathbb{E}\{r(X)(1 - r(X))\} =: R_{1\text{-NN}}$$

[Thm 5.1 in DGL] (Cover & Hart '67)

Compare this expression with Bayes risk

$$R^* = \inf_f \mathbb{P}(Y \neq f(X)) = \mathbb{E}\{\min(r(X), 1 - r(X))\}$$

This expression can be deduced from calculations in the proof of the theorem p.10 in SL: FOUNDATIONS.

Indeed,

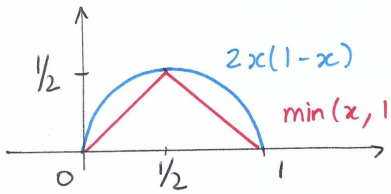
$$\begin{aligned} R^* &= \mathbb{P}(Y \neq f^*(X)) \qquad f^* = \text{Bayes classifier} \\ &= \mathbb{E}\{1 - \mathbb{1}(f^*(X) = 1) r(X) - \mathbb{1}(f^*(X) = 0)(1 - r(X))\} \qquad \text{(p.11)} \\ &= \mathbb{E}\{1 - \mathbb{1}(r(X) > 1/2) r(X) - \mathbb{1}(r(X) \leq 1/2)(1 - r(X))\} . \end{aligned}$$

→ If $r(X) > 1 - r(X)$, then $r(X) > 1/2$, and $\{\dots\} = 1 - r(X) = \min(r(X), 1 - r(X))$

→ If $r(X) \leq 1 - r(X)$, then $r(X) \leq 1/2$, and $\{\dots\} = 1 - (1 - r(X)) = r(X) = \min(r(X), 1 - r(X))$

⇒ $R^* = \mathbb{E}\{\min(r(X), 1 - r(X))\}$, as required. ∎

$\forall x \in [0,1], \quad 2x(1-x) \geq \min(x, 1-x)$



Put $\psi(X) := \min(r(X), 1-r(X))$. We have:

$$R^* = \mathbb{E}\{\min(r(X), 1-r(X))\}$$

$\leq 2\,\mathbb{E}\{r(X)(1-r(X))\}$ ← $2x(1-x) \geq \min(x, 1-x)$

$= 2\,\mathbb{E}\{\psi(X)(1-\psi(X))\}$ ← since if $\psi(X) = r(X)$, $1-\psi(X) = 1-r(X)$; while if $\psi(X) = 1-r(X)$, $1-\psi(X) = r(X)$

$\leq 2\,\underbrace{\mathbb{E}\{\psi(X)\}}_{R^*}\,\underbrace{\mathbb{E}\{1-\psi(X)\}}_{1-R^*}$ ← (*)

$= 2R^*(1-R^*)$

$\leq 2R^*$

(*) If f is monotone ↑, g is monotone ↓

Then $\mathbb{E}\{f(X)g(X)\} \leq \mathbb{E}f(X)\,\mathbb{E}g(X)$

Summurizing,

$$R^* \leq R_{1\text{-NN}} \leq 2R^*(1-R^*) \leq 2R^*$$

↖ The expected misclassification probability converges to a number between R^* and $2R^*$. The nearest neighbor classifier is asymptotically at most twice as bad as Baye's classifier.

If K is odd and fixed, then [Thm 5.2 in DGL]

$$\mathbb{E}\{R_{0\text{-}1}(g_n)\} \xrightarrow{n\to\infty} R_{K\text{-NN}},$$

where

$$R_{K\text{-NN}} = \mathbb{E}\left\{r(X)\,\mathbb{P}\left(Z(X) < \frac{K}{2} \mid X\right)\right\} + \mathbb{E}\left\{(1-r(X))\,\mathbb{P}\left(Z(X) > \frac{K}{2} \mid X\right)\right\},$$

$Z(X) \sim Bi(K, r(X))$.

(Several representations for $R_{K\text{-NN}}$ exist and can be found in [DGL])

But there is more [Thm 5.4 in DGL] ← $\forall \mathbb{P}$

$$R^* \leq \ldots \leq R_{(2K+1)\text{-NN}} \leq R_{(2K-1)\text{-NN}} \leq \ldots \leq R_{3\text{-NN}} \leq R_{1\text{-NN}} \leq 2R^*$$

↳ The decreasing nature of the sequence $\{R_{K\text{-NN}};\ K \text{ odd}\}$ suggests that we should allow K to grow with n, to ensure better performance of the K-NN estimator.

The case K odd is easier to handle since voting ties are avoided (when K is odd, there are always more 0s than 1s, or more 1s than 0s, strictly). When K is even, we define the K-NN rule as

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{2K} Y_{(i)} > K \\ 0 & \text{if } \text{—}''\text{—} < K \\ Y_{(1)} & \text{if } \text{—}''\text{—} = K \end{cases}.$$

↖ "symmetric" rule

↑ Break ties using the closest neighbor.

For such a rule, we have [Thm 5.5 in DGL]

$$R_{(2K-1)\text{-NN}} = R_{2K\text{-NN}}$$

(even values of K do not decrease the asymptotic risk)

Several other results relating $R_{K\text{-NN}}$ and R^* exist:

[Thm 5.6 in DGL] $\forall K$ odd $\forall \mathbb{P}$ $\quad R_{K\text{-NN}} \leq R^* + \frac{1}{\sqrt{Ke}}$

[Thm 5.7 in DGL] $\forall K$ odd $\forall \mathbb{P}$ $\quad R_{K\text{-NN}} \leq R^* + \sqrt{\frac{2R_{1\text{-NN}}}{K}}$

III.2. Weak & Strong Consistency.

The first universal consistency result, due to Stone, appeared in 1977. Before this date, consistency results were imposing conditions on the distribution of (X, Y), not verifiable in practice.

Theorem [Weak Universal Consistency] [Thm 6.1 in GKKW]

If $K = K_n \to \infty$ and $K_n / n \to 0$ as $n \to \infty$, then the K_n-NN rule is Weakly Universally Consistent for any distribution such that $\mathbb{E} Y^2 < \infty$ & where (distance) ties occur with probability 0; that is

$$\lim_{n\to\infty} \mathbb{E}\{\mathcal{E}(f_n)\} = 0$$

↑ $\mathcal{E}(f_n)$ is computed under a square loss.

Note that if $Y \in \{0, 1\}$, consistency for the binary classification problem follows from

$$\mathbb{E}\{\mathcal{E}_{0,1}(g_n)\} \leq 2\,\mathbb{E}\sqrt{\mathcal{E}(f_n)} \leq 2\sqrt{\mathbb{E}\,\mathcal{E}(f_n)} \to 0$$

(Jensen)

Theorem [Strong Consistency] [Thm 23.7 in GKKW]

Assume that $\mathbb{P}\{|Y| \leq L\} = 1$ for $L < \infty$, and that $\|X - x\|$ is absolutely continuous $\forall x$ (no distance ties).

If $K_n \to \infty$ and $K_n / n \to 0$ as $n \to \infty$, then the K_n-NN rule is strongly consistent ($\mathcal{E}(f_n) \to 0$ a.s.)

↑ Assumptions of the theorem require boundedness of Y ⇒ restrictions on $\mathbb{P}_{X,Y}$ apply ⇒ not universally strongly consistent.
However, in the case $Y \in \{0, 1\}$; i.e. in the binary classification problem, we have universal strong consistency as the condition $\mathbb{P}(|Y| \leq L) = 1$ is immediately verified (we need to adapt definitions). In fact, it is proved in [Thm 11.1 in DGL] that $\exists c > 0$ $\forall \varepsilon > 0$ $\exists n_0$ s.t. $\forall n \geq n_0$,

$$\mathbb{P}\left(R_{a_1}(g_n) - R^* > \varepsilon\right) \leq e^{-nc\varepsilon^2}$$

↑ + Borel-Cantelli ⇒ $R_{a_1}(g_n) \to R^*$ a.s.

Theorem [Strong Universal Consistency] [Thm 23.8 in GKKW]

Assume that $\|X - x\|$ is AC for all $x \in X$.

If $K_n / \log n \to \infty$ and $K_n / n \to 0$ as $n \to \infty$, then the K_n-NN rule is strongly universally consistent ($\mathcal{E}(f_n) \to 0$ a.s.)

↑ Recall: $\forall P_{X,Y}$ s.t. $\mathbb{E} Y^2 < +\infty$

III.3. Convergence rates.

Without further restrictions on the distribution of (X, Y), we cannot construct f_n from the data with $\mathbb{E}\{\mathcal{E}(f_n)\} = \mathbb{E}_{\mathcal{D}_n}\left\{\int (f_n(x) - r(x))^2 \mathbb{P}_X(dx)\right\}$ tending to 0 with a guaranteed rate of convergence. Specifically, let $\{a_n\}$ be a sequence of positive numbers converging to 0. Then, for any sequence of regression estimates $\{f_n\}$, there exists a distribution $\mathbb{P}_{X,Y}$ of (X, Y) such that

- $X \sim \mathcal{U}([0,1])$ "X has a nice distribution"
- $Y = \mathbb{E}(Y|X) = r(X) \in \{-1, 1\}$ classification problem "noiseless case" "ideal"

and

$$\limsup_{n\to\infty} \frac{\mathbb{E}\{\mathcal{E}(f_n)\}}{a_n} \geq 1 \qquad \text{[Thm 3.1 in GKKW]}$$

"$\mathbb{E}\{\mathcal{E}(f_n)\}$ converges to zero slower than a_n"

↑ Irrespectively of your algorithm output, there exists a distribution of (X, Y) for which convergence is arbitrarily slow.

Remark: Since the classification task is easier than the regression task, similar bounds can be deduced in the context of regression. Let $\{a_n\}$ be a sequence of positive numbers converging to 0, and such that $1/64 \geq a_1 \geq a_2 \geq \dots$. Then, for every sequence of regression estimates $\{f_n\}$, there exists a distribution $\mathbb{P}_{X,Y}$ of (X, Y) such that $X \sim \mathcal{U}[0,1]$, $Y = \mathbb{E}(Y|X) = r(X)$, and $\mathbb{E}\{\mathcal{E}(f_n)\} \geq a_n \ \forall n$ [see p. 36 in GKKW]. Such results were discussed on p.17 in SL: FOUNDATIONS, and are known as "no-free lunch thms"

Convergence rates are usually derived under smoothness assumptions of the regression function $r(x)$, such as Lipschitz continuity: $|r(x) - r(y)| \leq C\|x - y\|$, for some $C > 0$. ($x, y \in \mathbb{R}^d$)

For such regression functions, we have the following result:

Theorem [Thm 6.2 in GKKW]

Assume that
- X is bounded
- $\sigma^2(x) := \mathrm{Var}(Y|X=x) \leq \sigma^2$, $x \in \mathbb{R}^d$
- $|r(x) - r(y)| \leq C\|x - y\|$
- $d \geq 3$

For the k_n-NN estimate, with $k_n \sim n^{\frac{2}{d+2}}$, there exists a constant $c_1 > 0$ such that

$$\mathbb{E}\{\mathcal{E}(f_n)\} \leq c_1 n^{-\frac{2}{d+2}}$$

↑ Slower rate of convergence, as the dimension of the input space increases.

It turns out that the k_n-NN rule achieves the best possible convergence rate for this class of probability distributions: one can show that for lipschitz continuous regression functions, the minimax rate (see p. 16 in SL: FOUNDATIONS) is precisely $n^{-\frac{2}{d+2}}$. However, the k_n-NN rule is not optimal for smoother r, as it is unable to capture information regarding the derivative of a differentiable function; see discussion p. 96 in [GKKW], and [Thm 3.2 in GKKW].