Local Averaging Methods

Regression: $X_1 \ldots X_n \in [0,1]^d$

$Y_1 \ldots Y_n \in \mathbb{R}$

$\rightsquigarrow \hat{f}(x) = \sum_{i=1}^{n} w_i(x) Y_i$] Local average of the output

where $\sum_{i=1}^{n} w_i(x) = 1$

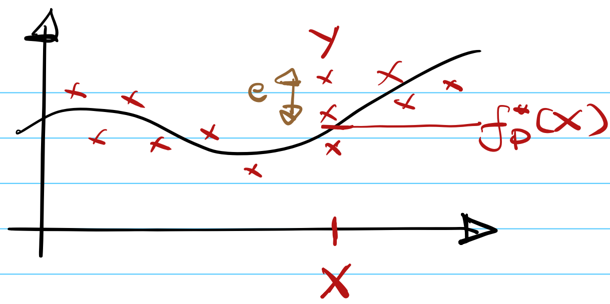
↳ if $\begin{cases} x \text{ close to } X_i : & w_i(x) \text{ large} \\ \text{not close} & w_i(x) \text{ small.} \end{cases}$

(1) The Regression problem

$(X, Y) \sim P$

$R_P(f) = \mathbb{E}[(f(X) - Y)^2]$

Bayes estimator: $f_P^*(x) = \mathbb{E}_P[Y \mid X = x]$.



Write $\quad Y = f_P^*(X) + e$

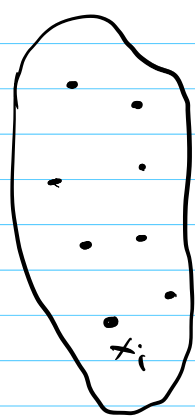
error term $= Y - f_P^*(x)$

By definition of f_P^*:

$$E_P[e \mid X] = E[Y \mid X] - f_P^*(X) = 0.$$

Example: $\quad X_1 \, \dots \, X_n$ street in a city

$Y_1 \, \dots \, Y_n$ CO2 concentration



$\leadsto$ $f_P^*(x) =$ Average CO2 concentration at street x.

$$Y = f_P^*(X) + e$$

Noise magnitude may depend on the street x.

Bayes Risk: $\quad \mathcal{R}_P^* = \mathcal{R}_P(f_P^*)$

$$= E[(f_P^*(x) - Y)^2]$$

$$= E[e^2].$$

What is the excess of risk of a function f ?

$$\mathcal{R}_P(f) = \mathbb{E}_P[(f(X)-Y)^2]$$

$$\mathbb{E}[(f(X)-Y)^2 \mid X] = \mathbb{E}[(f(X)-f_P^*(X)-e)^2 \mid X]$$

$$= \underbrace{\mathbb{E}[(f(X)-f_P^*(X))^2 \mid X]}_{=(f(X)-f_P^*(X))^2} + 2\,\mathbb{E}[(f(X)-f_P^*(X))e \mid X] + \mathbb{E}[e^2 \mid X]$$

$$2\,\mathbb{E}[(f(X)-f_P^*(X))e \mid X] = (f(X)-f_P^*(X))\underbrace{\mathbb{E}[e \mid X]}_{=0} = 0$$

$$= (f(X)-f_P^*(X))^2 + \mathbb{E}[e^2 \mid X].$$

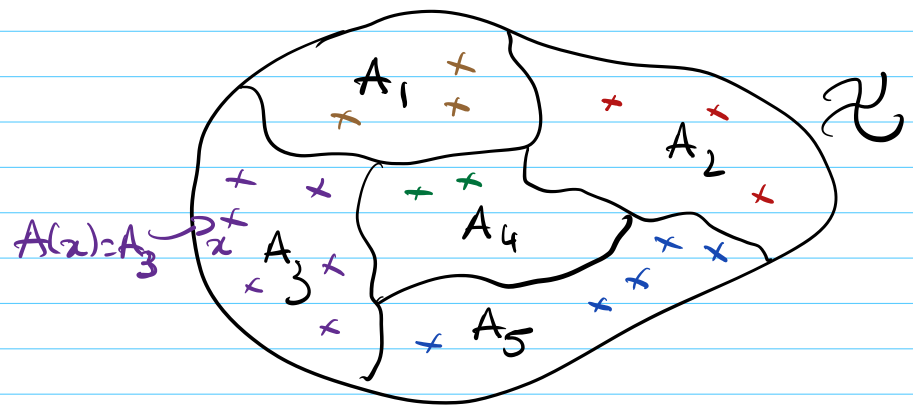
$$\Rightarrow \mathcal{R}_P(f) = \mathbb{E}[\mathbb{E}[(f(X)-f_P^*(X))^2 \mid X]]$$

$$= \mathbb{E}[(f(X)-f_P^*(X))^2] + \underbrace{\mathbb{E}[\mathbb{E}[e^2 \mid X]]}_{=\mathbb{E}[e^2] = \mathcal{R}_P^*}$$

$\Rightarrow$ Excess of Risk:

$$\boxed{\mathcal{R}_P(f) - \mathcal{R}_P^* = \mathbb{E}[(f(X)-f_P^*(X))^2] = \int (f(x)-f_P^*(x))^2 \, dP_X(x)}$$

② Partition estimator:



Partition $\mathcal{A} = (A_1 \ldots A_J)$ of a set $\mathcal{X}$

$$\begin{cases} A_j \cap A_{j'} = \emptyset & j \neq j' \\ \bigcup_{j=1}^{J} A_j = \mathcal{X} \end{cases}$$

$\mathcal{X} = [0,1]^d$

Def: $(X_1, Y_1) \ldots (X_n, Y_n)$

$x \in [0,1]^d$

$A(x) = A_j$ if $x \in A_j$.

Define

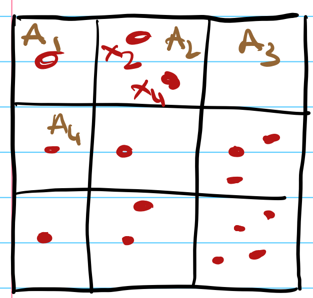
$$w_i(x) = \frac{\mathbb{1}\{X_i \in A(x)\}}{\sum_{i'=1}^{n} \mathbb{1}\{X_{i'} \in A(x)\}}$$

$$\Rightarrow \left[\hat{f}_{\mathcal{A}}(x) = \sum_{i=1}^{n} w_i(x) Y_i \right]$$

Partition estimator

Regressogram

⇝ Let us compute $\hat{f}_t(x)$:



$$\begin{cases} I_j = \text{indexes } i \text{ such that } X_i \in A_j \\ n_j = \text{size of } I_j . \end{cases}$$

$n_1 = 1 \quad n_2 = 2 \quad n_3 = 0$

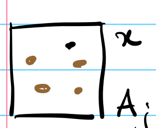
$I_2 = \{2, 4\}$

⇒ Let $x \in A_j$:

$$\hat{f}_t(x) = \sum_{i=1}^{n} w_i(x) Y_i$$

$$= \frac{\sum_{i=1}^{n} \mathbb{1}\{X_i \in A_j\} Y_i}{\underbrace{\sum_{i'=1}^{n} \mathbb{1}\{X_{i'} \in A_j\}}_{n_j}}$$

$$\boxed{\hat{f}_t(x) = \frac{1}{n_j} \sum_{i \in I_j} Y_i}$$



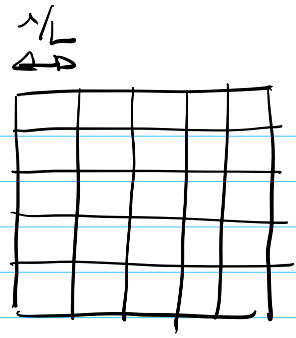
↳ Average of the outputs Y_i such that $X_i \in A_j$.

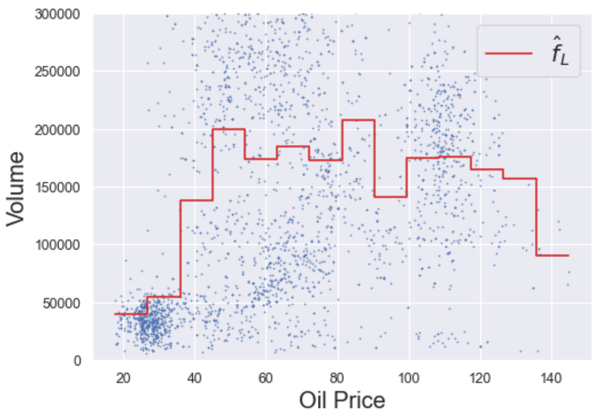
Example:

cube partition

L^d cubes

of side length $1/L$.





$d = 1$

Is there a relation between price and volume of oil available on the market?

$\Rightarrow$ What is the excess of risk of the cube partition estimator?

We assume:

$\boxed{1}$ f_P^* is α-Lipschitz:

$$\| f_P^*(x) - f_P^*(x') \| \leq \alpha \|x - x'\|$$

$\boxed{2}$ f_P^* is bounded $\quad |f_P^*(x)| \leq \beta$.

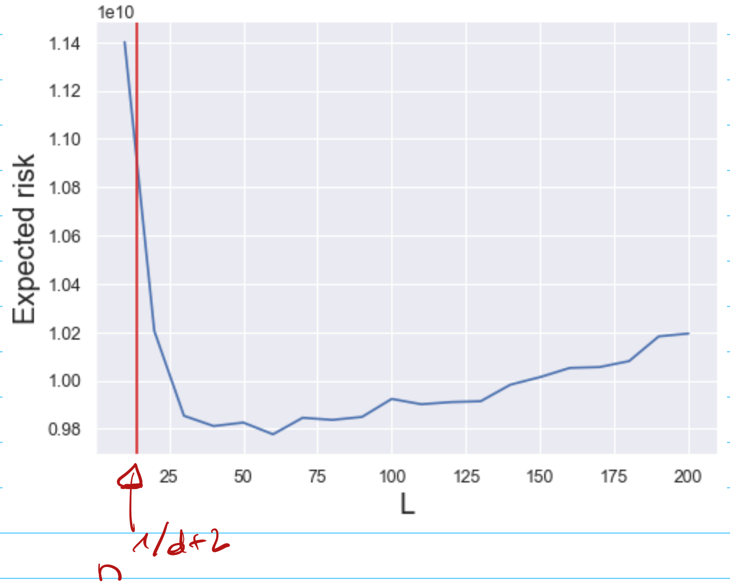
$\boxed{3}$ Bounded Noise: $|\varepsilon| \leq \sigma$.

THM : We have

$$\mathbb{E}\left[\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P^* \right] \lesssim \frac{\alpha^2}{L^2} + \frac{\sigma^2 L^d}{n}$$

$$L \simeq n^{1/d+2} \Longrightarrow \mathbb{E}\left[\mathcal{R}_P(\hat{f}_L) - \mathcal{R}_P^* \right] \lesssim n^{-\frac{2}{d+2}}$$

curse of dimensionality



$\Longrightarrow$ The theorem gives the order of magnitude of the optimal L.

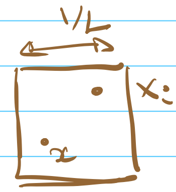
$\rightsquigarrow$ Cross-validation.

Proof Sketch

$\boxed{x \in A_j}$

$$\hat f_L(x) - f_P^*(x) = \frac{1}{n_j}\sum_{i\in I_j} \underbrace{Y_i}_{= f_P^*(X_i) + e_i} - f_P^*(x)$$

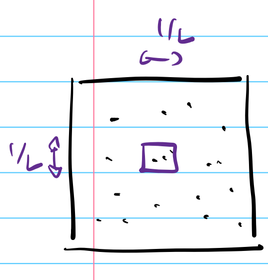
$$= \frac{1}{n_j}\sum_{i\in I_j} \underbrace{\left(f_P^*(X_i) - f_P^*(x)\right)}_{\substack{\le \alpha \|X_i - x\| \\ \le \alpha \frac{\sqrt{d}}{L}}} + \frac{1}{n_j}\sum_{i\in I_j} e_i$$



$$|\hat f_L(x) - f_P^*(x)| \le \alpha \frac{\sqrt{d}}{L} + \left| \frac{1}{n_j}\sum_{i\in I_j} e_i \right|$$

Conditionally on I_j, this is an average of n_j r.v.

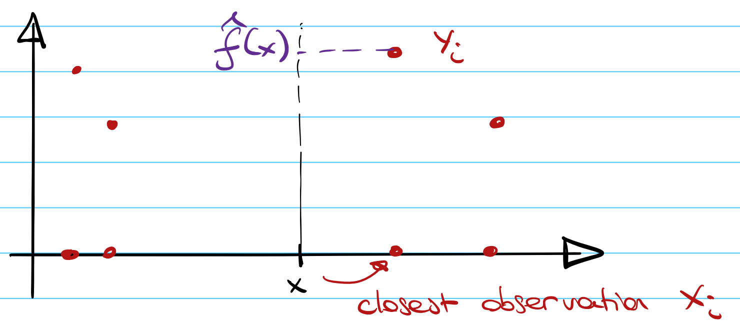
$$\mathbb{E}\left[\left(\frac{1}{n_j}\sum_{i\in I_j} e_i\right)^2 \Big| I_j\right] = \frac{1}{n_j^2}\sum_{i\in I_j} \mathbb{E}\left[e_i^2 | I_j\right] \le \frac{\sigma^2}{n_j}$$



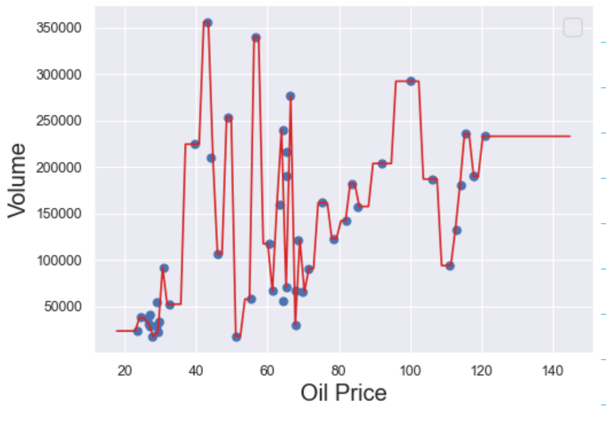
$$n_j \approx n \times L^{-d}$$

$$\Rightarrow \mathbb{E}\left[|\hat f_L(x) - f_P^*(x)|^2\right] \lesssim \frac{1}{L^2} + \frac{\sigma^2}{n L^{-d}}. \quad \blacksquare$$

③ Nearest-Neighbor Methods:

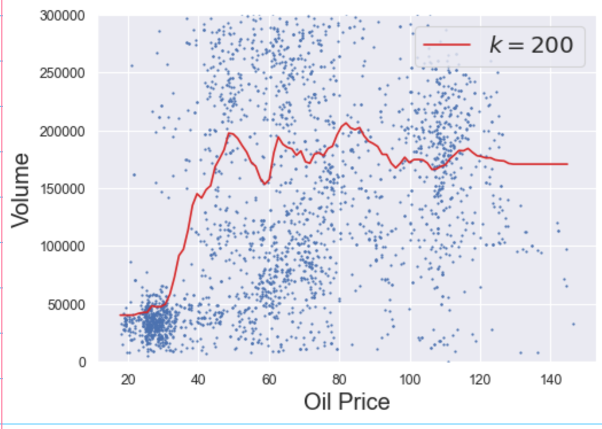


Prediction at x = output Y_i of the nearest neighbor X_i.



1 Nearest Neighbor Estimator

Reasonable idea ... but overfitting.



k - Nearest - Neighbor

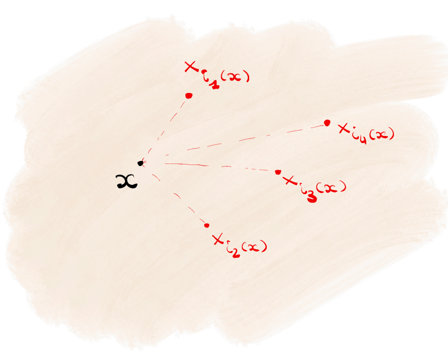
Prediction at x = average of the k outputs Y_i corresponding to the k NN of x.

Def: $(X_1, Y_1) \quad \dots \quad (X_n, Y_n) \qquad k \geq 1.$

$x \in [0,1]^d$

⇝ We order the inputs :

$$\| x - X_{i_1(x)} \| \leq \| x - X_{i_2(x)} \| \leq \dots \leq \| x - X_{i_n(x)} \|$$



⇝ $I_k(x) = \{ i_1(x), \dots i_k(x) \}$ indexes of the k-Nearest-Neighbors

$$W_i(x) = \begin{cases} 1/k & \text{if } i \in I_k(x) \\ 0 & \text{otherwise.} \end{cases}$$

$$\boxed{\begin{aligned}\Rightarrow\quad \hat{f}_k^{NN}(x) &= \sum_{i=1}^{n} w_i(x) Y_i \\ &= \frac{1}{k} \sum_{i \in I_k(x)} Y_i\end{aligned}}$$

THM: Assume:

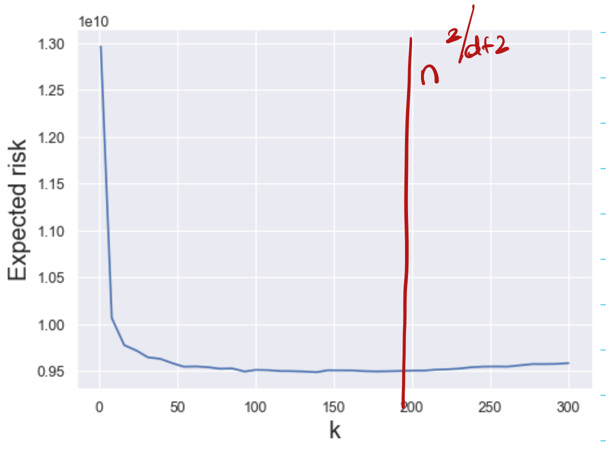
1. f_P^* is α-Lipschitz
2. Noise e is bounded: $|e| \leq \sigma$
3. X is uniform on $[0,1]^d$.

Then $$\mathbb{E}\left[R_P(\hat{f}_k^{NN}) - R_P^*\right] \lesssim \alpha^2 \left(\frac{k}{n}\right)^{2/d} + \frac{\sigma^2}{k}.$$

For $k \approx n^{2/d+2}$: we obtain

$$\mathbb{E}\left[R_P(\hat{f}_k^{NN}) - R_P^*\right] \lesssim n^{-2/d+2}$$

Curse of dimensionality



Select k with cross-validation in practice.

Why does this work ?

$\rightarrow \begin{cases} \text{The } k\text{NN of } x \text{ is close to } x. \\ f_P^* \text{ is Lipschitz} \end{cases}$

$$\Longrightarrow \| f_P^*(x) - f_P^*(X_{i_k(x)}) \| \leq \alpha \| x - X_{i_k(x)} \|$$

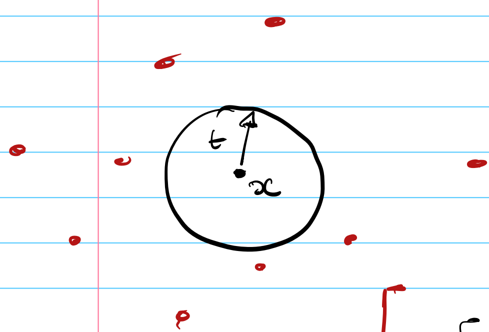
What is the distance between x and its kNN ?

$k=1$

$$\mathbb{P}(\|x - X_{i_1(x)}\| \geq t) = \mathbb{P}(\forall i, \|x - X_i\| \geq t)$$

$$= (1 - \mathbb{P}(B(x,t)))^n$$

$$\leq \exp(-n \underbrace{\mathbb{P}(B(x,t))}_{\simeq t^d})$$



$$\Downarrow$$

$$\boxed{\mathbb{E}\left[\|x - X_{i_1(x)}\|^2\right] \lesssim n^{-2/d}}$$

General fact: $\mathbb{E}[Z^2] = \mathbb{E}\left[\int_0^{+\infty} \mathbb{1}\{t \leq Z^2\} dt\right]$

$$= \int_0^{+\infty} \mathbb{P}(Z^2 \geq t) dt$$

$$= 2 \int_0^{+\infty} u \mathbb{P}(Z \geq u) du$$

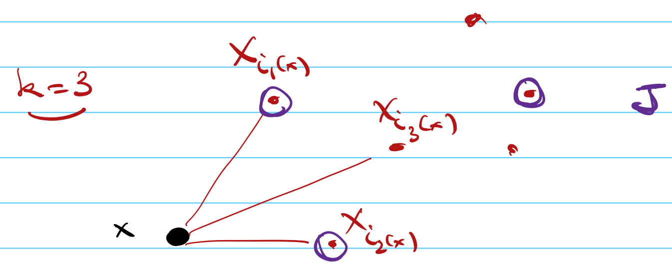
Proof Sketch :

$$|\hat f_k^{NN}(x) - f_P^*(x)|^2 = \Big| \frac{1}{k} \sum_{i \in I_k(x)} \overset{f_P^*(X_i) + e_i}{Y_i} - f_P^*(x) \Big|^2$$

$$\lesssim \Big(\frac{1}{k} \sum_{i \in I_k(x)} \underbrace{|f_P^*(X_i) - f_P^*(x)|}_{\leq \alpha \|x - X_i\|} \Big)^2 + \Big| \frac{1}{k} \underbrace{\sum_{i \in I_k(x)} e_i}_{\text{average of } k \text{ error terms}} \Big|^2$$

JENSEN

$$\leq \frac{\alpha^2}{k} \sum_{i \in I_k(x)} \|x - X_i\|^2 \qquad \Rightarrow \quad \approx \frac{\sigma^2}{k}.$$



If J is any set of k index:

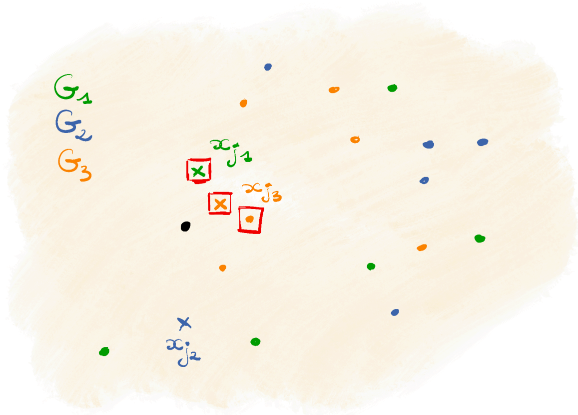
$$\frac{1}{k} \sum_{i \in I_k} \|x - X_i\|^2 \leq \frac{1}{k} \sum_{i \in J} \|x - X_i\|^2$$

→ We split the observations in k groups :

$$\underbrace{X_1 \, --- \, X_{\frac{n}{k}}}_{G_1} \quad \underbrace{X_{\frac{n}{k}+1} \, --- \, X_{2\frac{n}{k}}}_{G_2} \quad ---- \quad \underbrace{X_{\frac{(k-1)}{k}n+1} \, --- \, X_n}_{G_k}$$

j_1 = index of closest to x in G_1 . . . . . j_k = index of closest to x in G_k

→ $J = \{j_1 \, --- \, j_k\}$



$$\mathbb{E}\left[\frac{1}{k}\sum_{i\in I_k}\|x-X_i\|^2\right] \leq \underbrace{\mathbb{E}\left[\frac{1}{k}\sum_{m=1}^{k}\|x-X_{j_m}\|^2\right]}$$

$$= \frac{1}{k}\sum_{m=1}^{k}\mathbb{E}\big[\|x-\underbrace{X_{j_m}}_{\text{closest to } x \text{ in the group } G_m.}\|^2\big]$$

$$\Longrightarrow \mathbb{E}[\|\ \ \|^2] \leq \frac{1}{(n/k)^{2/d}}$$

$$\lesssim \left(\frac{k}{n}\right)^{2/d}.$$

Conclusion:

$$= \int (\hat{f}_P(x) - f_P^*(x))^2 \, dP_X(x)$$

$$\boxed{\mathbb{E}\left[R_P(\hat{f}_k^{NN}) - R_P^*\right] \lesssim \alpha^2\left(\frac{k}{n}\right)^{2/d} + \frac{\sigma^2}{k}.}$$

Conclusion :

* Two local averaging methods :
  Partition estimators + kNN
* Easy to compute and interpret.
* Depends on one parameter that needs to be tuned.
* Suffers from the curse of dimensionality.
  ↳ Rate of cv $n^{-2/d+2}$

Other local averaging method:
Nadaraya - Watson : see Lecture Notes.