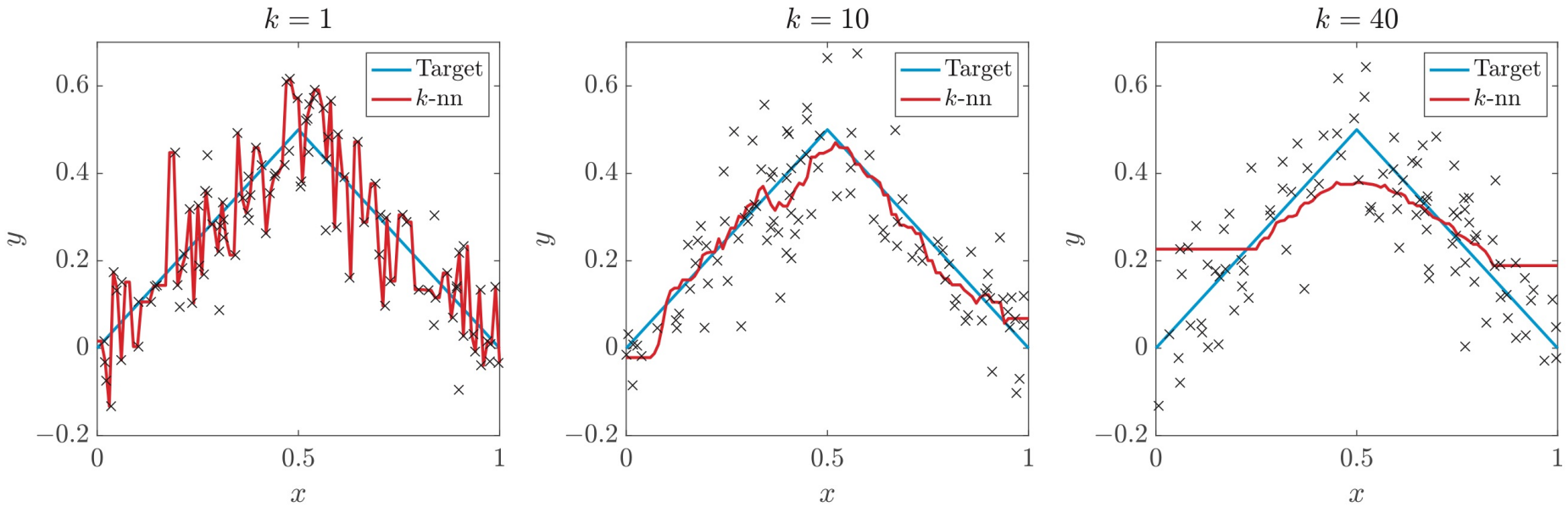
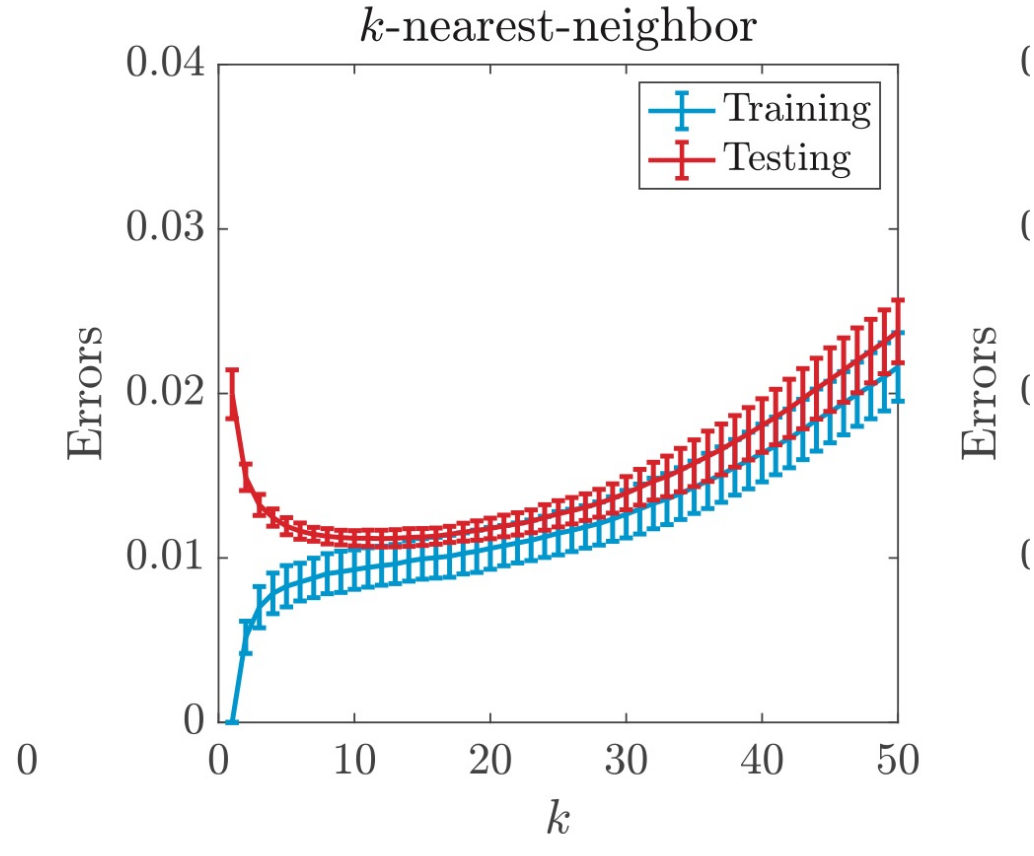

$k = 1$
$k = 10$
$k = 40$
Target
k-nn
y
x
0.6
0.4
0.2
0
−0.2
0
0.5
1


k-nearest-neighbor
Training
Testing
Errors
k
0.04
0.03
0.02
0.01
0
0
10
20
30
40
50

SGD: objective $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$ Test error

(alternative $= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$ Training error)

Algorithm: $\theta_t = \theta_{t-1} - \gamma g_t \quad \in \mathbb{R}^d$

① Unbiased: $\mathbb{E}[g_t | \theta_{t-1}] = F'(\theta_{t-1})$

(ex: $g_t = \nabla_\theta \ell(y_t, f_\theta(x_{t+1}))\Big|_{\theta=\theta_{t-1}}$ $t \le n$)

② $\|g_t\|^2 \le B^2$ bounded.

Goal: bound $\mathbb{E}\left(F(\theta_t)\right) - F(\theta_*)$

minimizer of $F(\theta)$

$$\|\theta_t - \theta_*\|_2^2 = \|\theta_{t-1} - \theta_* - \gamma g_t\|_2^2 \qquad \text{because } \theta_t = \theta_{t-1} - \gamma g_t$$

$$= \|\theta_{t-1} - \theta_*\|_2^2 - 2\gamma g_t^\top (\theta_{t-1} - \theta_*) + \underbrace{\gamma^2 \|g_t\|_2^2}_{\le \gamma^2 B^2}$$

$$\mathbb{E}\left(\|\theta_t - \theta_*\|^2 \,|\, \theta_{t-1} \right) \le \|\theta_{t-1} - \theta_*\|_2^2 - 2\gamma \underbrace{F'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*)}_{\ge F(\theta_{t-1}) - F(\theta_*) \text{ by convexity}} + \gamma^2 B^2$$

(towering law of expectata)

$$\le \|\theta_{t-1} - \theta_*\|^2 - 2\gamma \left[F(\theta_{t-1}) - F(\theta_*) \right] + \gamma^2 B^2$$

$$\mathbb{E}\left[\|\theta_t - \theta_*\|_2^2 \right] \le \mathbb{E}\left[\|\theta_{t-1} - \theta_*\|^2 \right] - 2\gamma \left(\mathbb{E} F(\theta_{t-1}) - F(\theta_*) \right) + \gamma^2 B^2$$

$$\mathbb{E}\left(F(\theta_{t-1}) \right) - F(\theta_*) \le \frac{\mathbb{E}\left(\|\theta_{t-1} - \theta_*\|^2 \right) - \mathbb{E}\left(\|\theta_t - \theta_*\|_2^2 \right)}{2\gamma} + \frac{\gamma^2 B^2}{2}$$

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[F(\theta_{t-1}) \right] - F(\theta_*) \le \frac{1}{2\gamma T} \left(\mathbb{E}\left(\|\theta_0 - \theta_*\|_2^2 \right] - \underbrace{\mathbb{E}\left(\|\theta_T - \theta_*\|_2^2 \right]}_{\ge 0} + \frac{\gamma B^2}{2}$$

(Jensen for finite sum)

$$\mathbb{E}\left[F\left(\frac{1}{T} \sum_{t=1}^{T} \theta_{t-1} \right) \right] - F(\theta_*) \le \frac{1}{2\gamma T} \|\theta_0 - \theta_*\|_2^2 + \frac{\gamma B^2}{2}$$

$\le R/\sqrt{T}$ $\qquad \gamma \propto \frac{1}{\sqrt{T}}$

Smoothness vs Lipschitz-continuity of $F: \mathbb{R} \to \mathbb{R}$

$|F'(\theta) - F'(\eta)| \leq L|\theta - \eta|$

$\Downarrow$

$|F''(\theta)| \leq L$

γ constant is allowed

$|F(\theta) - F(\eta)| \leq B|\theta - \eta| \quad \forall \theta, \eta$

$\Downarrow$

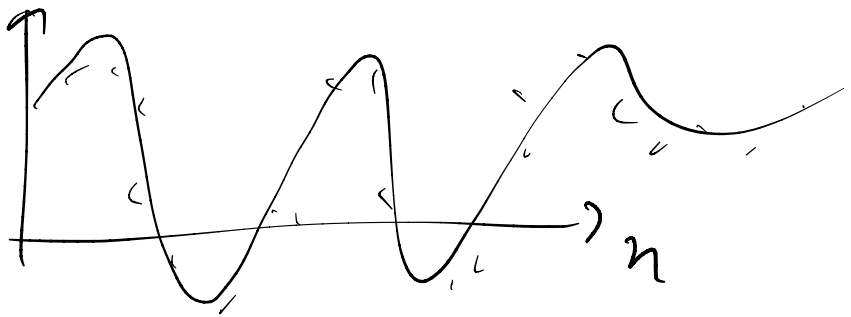
$|F'(\theta)| \leq B \quad \forall \theta$

γ has to decay with t.

Linear model $f_\theta = \theta^T \varphi(x) \quad \varphi: \mathcal{X} \to \mathbb{R}^d$.

Estimation error (Rademacher)

Optimization (Convex optimization)



LOCAL AVERAGING METHODS

Training data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$

Expected risk $R(f) = \mathbb{E}[\ell(y, f(x))]$

Testing error

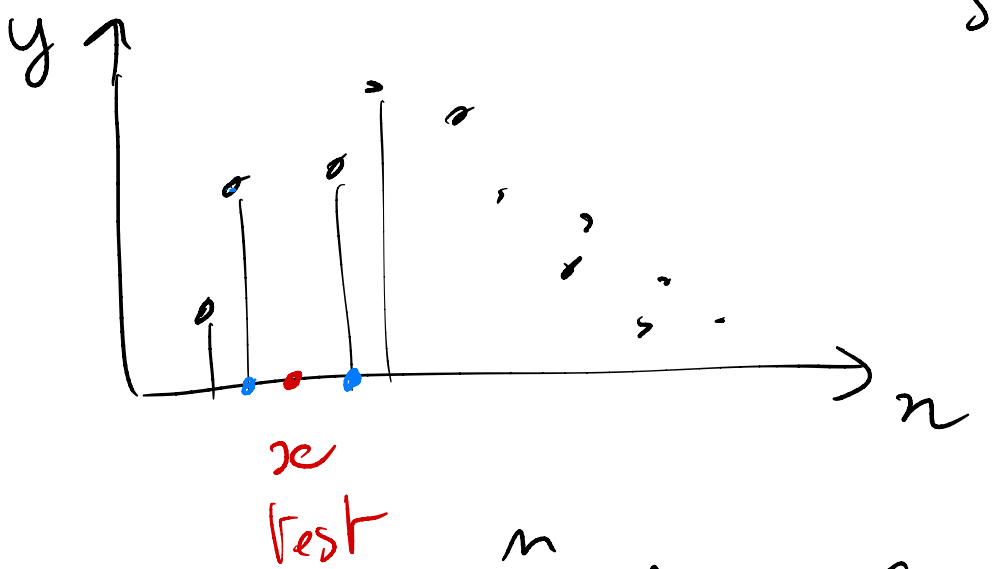
Optimal predictor (Bayes) $f^*(x) = \arg\min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z) \mid x]$

Key examples:

- regression with square loss
  $f^*(x) = \mathbb{E}[y \mid x]$
- classification with 0-1 loss
  $f^*(x) = \arg\max_{k} \mathbb{P}(y = k \mid x)$
- convex surrogates for binary classification
  $\ell(y, f(x)) = \log(1 + e^{-y f(x)})$ logistic regression
  $y \in \{-1, 1\}$

$$f_*(x) = \arg\min_{z \in \mathcal{Y}} \mathbb{E}\left[\ell(y,z) \mid x\right] = \int \ell(y,z)\, dp(y|x)$$

Local averaging : estimate directly $p(y|x)$



by $\hat{p}(y|x)$

$(\mathbb{E}_{\delta_{y_i}} y = y_i)$

① Regression

$\hat{f}(x) = \hat{\mathbb{E}}(y|x)$ /

$= \sum_{i=1}^{n} \hat{w}_i(x)\, y_i$ Majority vote

② classification

$\hat{f}(x) = \arg\max_k \hat{p}(y=k|x)$

$\sum_{i=1}^{n} \hat{w}_i(x)\, 1_{y_i=k}$

$$\hat{p}(y|x) = \sum_{i=1}^{n} \hat{w}_i(x)\, \delta_{y_i}$$

Dirac at y_i

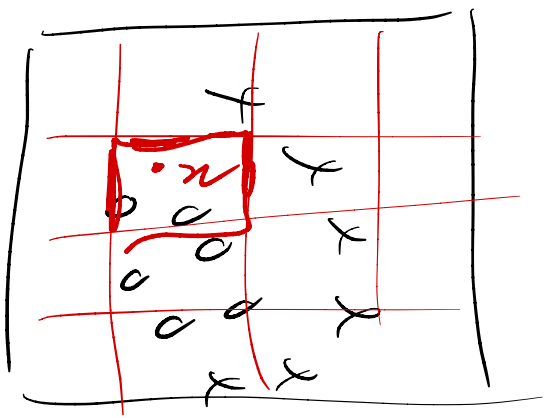
weight function $\forall x$ $\hat{w}_i(x) \geq c$

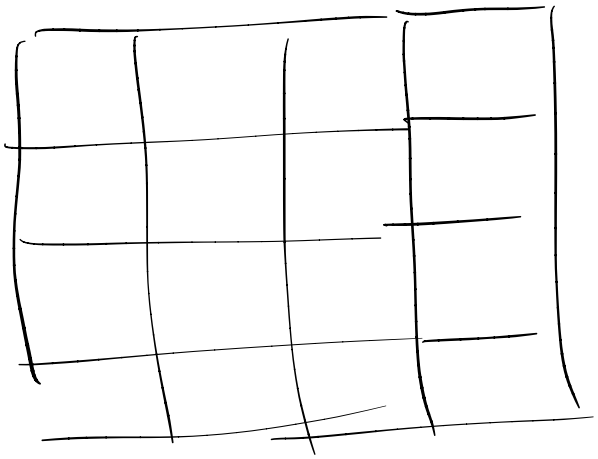
$$\sum_{i=1}^{n} \hat{w}_i(x) = 1$$

Estimator :

$$\hat{f}(x) = \arg\min_{z \in \mathcal{Y}} \int \ell(y,z)\, d\hat{p}(y|x)$$

Example 1 : partition estimator



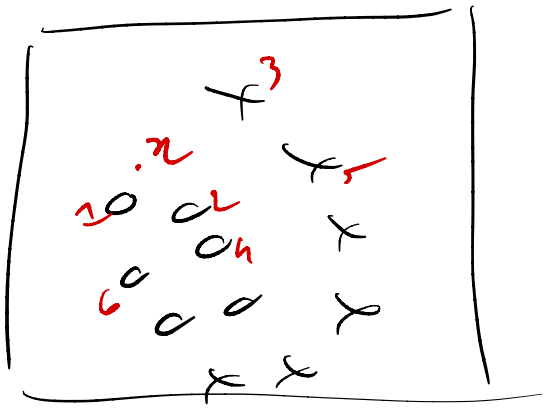


$\mathcal{X} = \mathbb{R}^2$

$\mathcal{Y} = \{o, x\}$.

NB : simplified decision tree

Example 2 : k-nearest neighbor



$\mathcal{X} = \mathbb{R}^2$

$\mathcal{Y} = \{o, x\}$

$i_1(x)$ nearest-neighbor

$i_k(x)$ k-nearest neighbor

$\Delta(x, x_{i_1(x)}) \leq \Delta(x, x_{i_2(x)}) \leq \dots$

↳ distance over $\mathcal{X}$.

- Running-time complexity

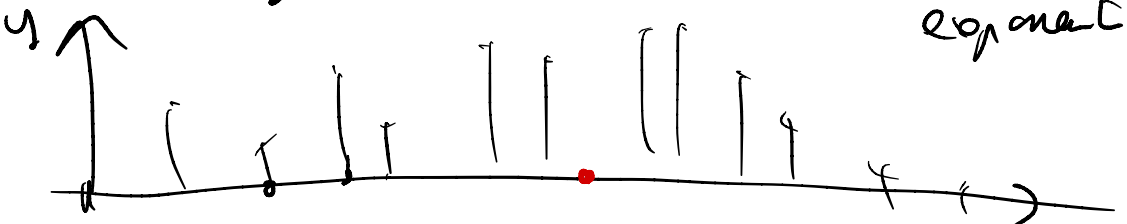
n points $x_i \in \mathbb{R}^d$

$O(nd)$ per test point (time)

$O(nd)$ memory $\Rightarrow$ store the data

"indexing" technique : $\log(n)$ complexity

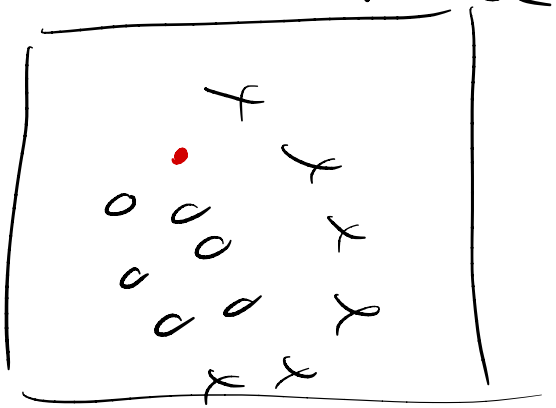
exponential in d



$$w_i(x) = \begin{cases} \frac{1}{k} & \text{if } i \in \{i_1(x), i_2(x), \dots, i_k(x)\}, \\ 0 & \text{otherwise} \end{cases}$$

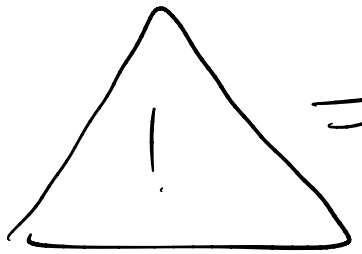
Example 3 : Nadaraya-Watson estimator

kernel regression



$\mathcal{X} = \mathbb{R}^2$

$\mathcal{Y} = \{o, x\}$



$\neq$ kernel methods

$$w_i(x) = \frac{k(x, x_i)}{\sum_{j=1}^{n} k(x, x_j)}$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$

$$k(x, x') = \exp\left(-\frac{\|x - x'\|_2^2}{h^2}\right)$$

10^4 2s

10^4 4s

Estimator = $\hat{f}(x) = \sum_{i=1}^{n} \hat{w}_i(x)\, y_i$ (Regression)

Criterion: excess expected risk

$$R(\hat{f}) - R(f_*) = \mathbb{E}\left|\hat{f}(x) - f_*(x)\right|^2$$

Assumptions: ① bounded noise:

$$|y - \underbrace{f_*(x)}_{\mathbb{E}(y|x)}| \le \sigma \quad \text{almost surely}$$

② Regularity = f_* is Lipschitz-continuous with constant B wrt distance Δ

$$\hat{\beta}(x) - \beta_\sigma(x) = \sum_{i=1}^{n} \hat{w}_i(x) y_i - \left(\sum_{i=1}^{n} \hat{w}_i(x)\right) \beta_\sigma(x) \qquad \text{for } x \in \mathcal{X} \text{ a vertex}$$

$$= \sum_{i=1}^{n} \hat{w}_i(x)(y_i - \beta_\sigma(x_i)) + \sum_{i=1}^{n} \hat{w}_i(x) \left[\beta_\sigma(x_i) - \beta_\sigma(x)\right]$$

weights sum to 1

$$\mathbb{E}\left[(\hat{\beta}(x) - \beta_\sigma(x))^2 \mid x_1, \dots, x_n\right] = \left(\mathbb{E}(\hat{\beta}(x) - \beta_\sigma(x) \mid x_1, \dots, x_n)\right)^2 + \mathrm{var}(\hat{\beta}(x) \mid x_1, \dots, x_n)$$

$$= \left(\sum_i \hat{w}_i(x)\left[\beta_\sigma(x_i) - \beta_\sigma(x)\right)\right)^2 + \mathrm{var}\left(\sum_{i=1}^{n} \hat{w}_i(x)(y_i - \beta_\sigma(x_i) \mid x_1, \dots, x_n\right)$$

$$= \left(\sum_i \hat{w}_i(x)\left[\beta_\sigma(x_i) - \beta_\sigma(x)\right)\right)^2 + \sum_{i=1}^{n} \hat{w}_i(x)^2 \,\mathrm{var}(y_i - \beta_\sigma(x_i) \mid x_i)$$

"bias" — variance

$$\leq \left(\sum_i \hat{w}_i(x) \left|\beta_\sigma(x_i) - \beta_\sigma(x)\right|\right)^2 + \sigma^2 \sum_{i=1}^{n} \hat{w}_i(x)^2$$

$\leq B\,\Delta(x_i, x)$

$$\leq \sum_i \hat{w}_i(x) B^2 \Delta(x_i, x)^2 + \sigma^2 \sum_{i=1}^{n} \hat{w}_i(x)^2$$

by Jensen's inequality

$$\mathbb{E}\left(\left(\hat{f}_S(x) - \hat{f}(x)\right)^2\right) \le B^2 \sum_{i=1}^{n} \mathbb{E}\left[\hat{w}_i(x)\,\Delta(x, x_i)^2\right] + \sigma^2 \sum_{i=1}^{n} \mathbb{E}\left(\hat{w}_i(x)^2\right)$$

expected test error at $x \in \mathcal{X}$

$$\int \mathbb{E}\left[\left(\hat{f}_S(x) - \hat{f}(x)\right)^2\right] dq_1(x) \le B^2 \sum_{i=1}^{n} \int \mathbb{E}\left[\hat{w}_i(x)\,\Delta(x, x_i)^2\right] dq_1(x) + \sigma^2 \sum_{i=1}^{n} \int \mathbb{E}\left[\hat{w}_i(x)^2\right] dq_1(x)$$

$\mathbb{E}$: training data; $dq_1(x)$: test point

excess risk

minimized for $w_i(x)$ very localized.

$$\sum_{i=1}^{n} \hat{w}_i(x)^2$$

$$= \sum_{i} \left(\hat{w}_i(x) - \frac{1}{n}\right)^2 + \frac{1}{n}$$

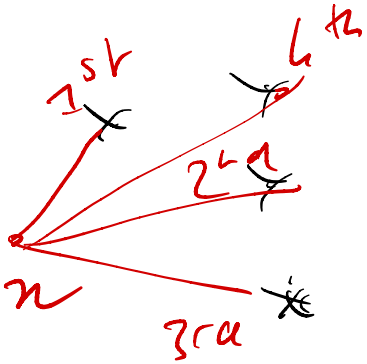
minimized for uniform weights

maximized for all weights on single x_i

k-nn $= \sum_{i=1}^{n} \hat{w}_i(x)^2 = k \cdot \frac{1}{k^2} = \frac{1}{k}$

$\Rightarrow$ variance term $= \sigma^2/k$

goal: bound $\mathbb{E} \int \sum_{i=1}^{n} \hat{w}_i(x) \, \Delta(x, x_i)^2 \, dp(x)$



$$\mathbb{E} \int \frac{1}{h} \sum_{j=1}^{h} \Delta^2(x, x_{i_j(x)}) \, dp(x)$$

$x_{i_j(x)}$: j-th nearest neighbor

Lemma: $\mathcal{X} \subset \mathbb{R}^d$, $\Delta = \|\cdot\|_2$ distance, $\mathcal{X}$ bounded

$x_1, \dots, x_{n+1}$ sampled iid from a distribution

Expected square distance between x_{n+1} and its nearest neighbor ($h=1$) $\leq \frac{4}{n^{2/d}} \operatorname{diam}(\mathcal{X})^2$

(among $x_1, \dots, x_n$)

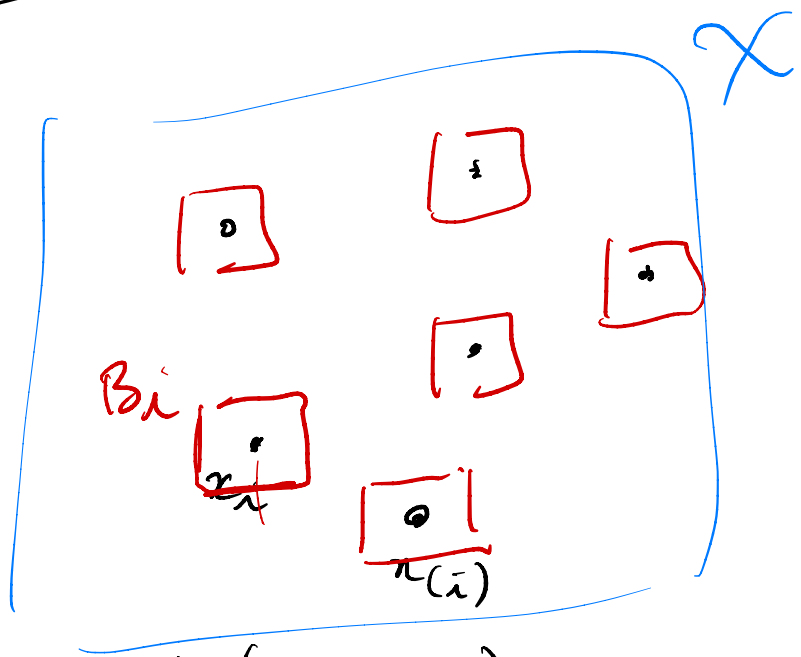
$$\operatorname{diam}(\mathcal{X}) = \sup_{x, x' \in \mathcal{X}} \Delta(x, x')$$

(cf $d \geq 2$)

$h > 1$: $8 \operatorname{diam}(\mathcal{X})^2 \left(\frac{2h}{n}\right)^{2/d}$

$\{x, \|x - a\|_2 \leq r\}$ => diameter: $2r$

Prob: $x_{(i)}$ the nearest neighbor of x_i among the $n+n$ points.



Goal: $\mathbb{E}\,\Delta(x_{n+i}, x_{(n+i)})^2$

$= \mathbb{E}\,\Delta(x_i, x_{(i)})^2$

$= \frac{1}{n+n} \sum_{i=1}^{n+n} \mathbb{E}\left[\Delta(x_i, x_{(i)})^2\right] = \frac{1}{n+n} \sum_i R_i^2$ $\forall i$

$R_i = \Delta(x_i, x_{(i)}) > 0$

sets $B_i = \{x \in \mathbb{R}^d,\ \|x - x_i\|_\infty < \frac{R_i}{2}\}$

Note that $\Delta(x_i, x_j) \geq \max(R_i, R_j) \Rightarrow$ the balls B_i are disjoint

$\text{volume}\left(\bigcup B_i\right) \leq \text{volume}\left(\{x+q,\ x \in X,\ \|s\|_\infty \leq \text{diam}(X)\}\right)$

$\leq \left(2\,\text{diam}(X)\right)^d$

|| because disjoint

$= \sum_i \text{vol}(B_i) = \sum_i R_i^d$

$\Rightarrow \boxed{\sum_i R_i^d \leq \left(2\,\text{diam}(X)\right)^d}$

GoR : bound on $\frac{1}{n+n} \sum R_i^2$

we know $\sum_i R_i^d \leq (2 \operatorname{diam}(\mathcal{X}))^d$

$$\frac{1}{n+n} \sum_{i=1}^{n} \underbrace{(R_i^2)^{\frac{d}{2}}}_{R_i^d} \geq \left(\frac{1}{n+n} \sum_i R_i^2 \right)^{d/2}$$

$\frac{d}{2} \geq 1$

Jensen's applied to $x \mapsto x^{d/2}$

$$\Rightarrow \left(\frac{1}{n+n} \sum_i R_i^2 \right) \leq \left(\frac{1}{n+n} \left([2 \operatorname{diam}(\mathcal{X})]^d \right) \right)^{2/d}$$

$$\leq \frac{1}{n^{2/d}} 4 \operatorname{diam}(\mathcal{X})^2$$

Summary

Excess risk of k-nn $\leq \underbrace{\frac{\sigma^2}{k}}_{\text{variance}} + \underbrace{B^2 \, \delta \, \text{diam}(\mathcal{X})^2 \left(\frac{2k}{n}\right)^{2/d}}_{\text{bias}}$

decreasing in k — increasing in k

$$\frac{\boxed{A}}{k} + \boxed{B}\left(\frac{k}{n}\right)^{2/d}$$

optimal balance $= \frac{1}{k} \sim \left(\frac{k}{n}\right)^{2/d} \iff k^{d/2} \sim \frac{n}{k}$

$$\iff k \sim n^{\frac{1}{\frac{d}{2}+1}}$$

$$k \sim n^{\frac{2}{2+d}}$$

increasing in k

optimal value $\sim \frac{1}{n^{2/(2+d)}}$

curse of dimensionality