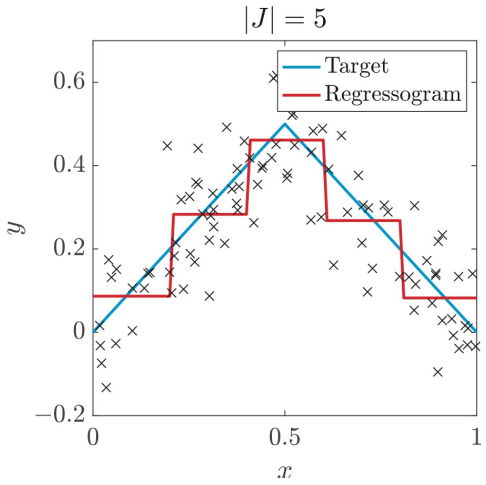
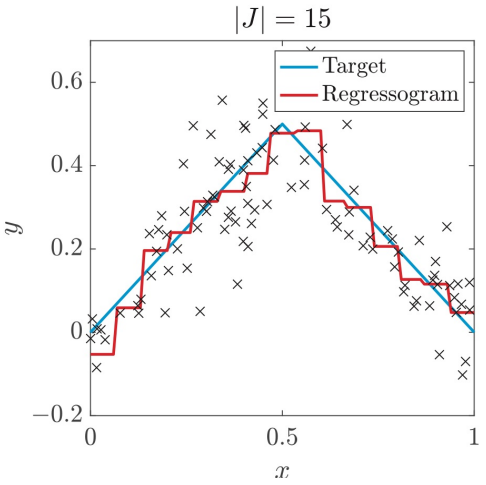
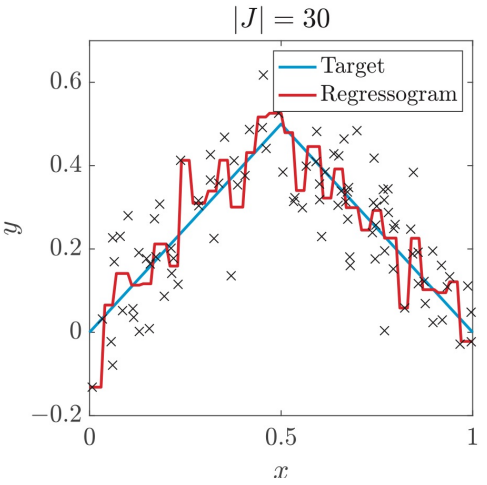
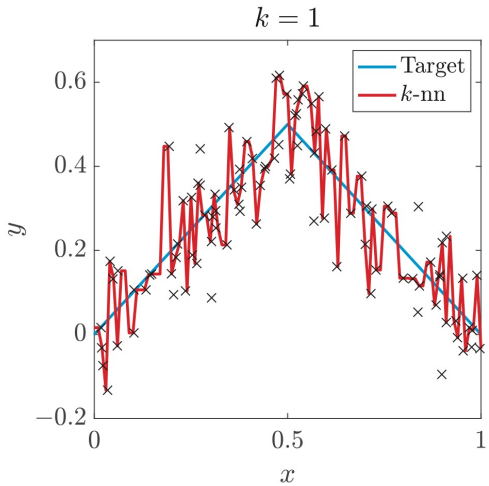
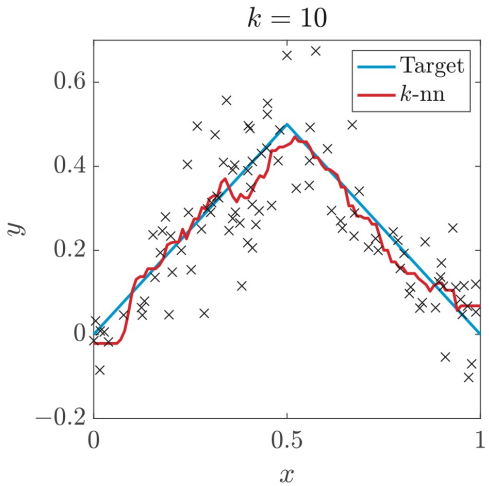
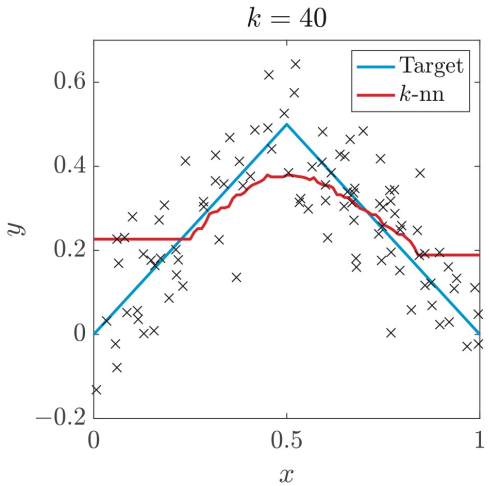

$|J| = 5$
Target
Regressogram
y
x

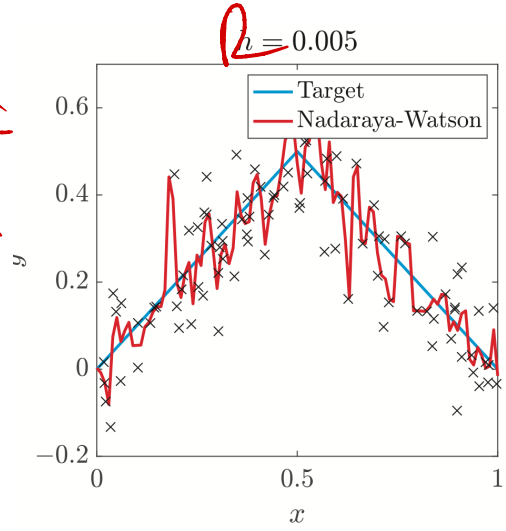

$|J| = 15$
Target
Regressogram
y
x

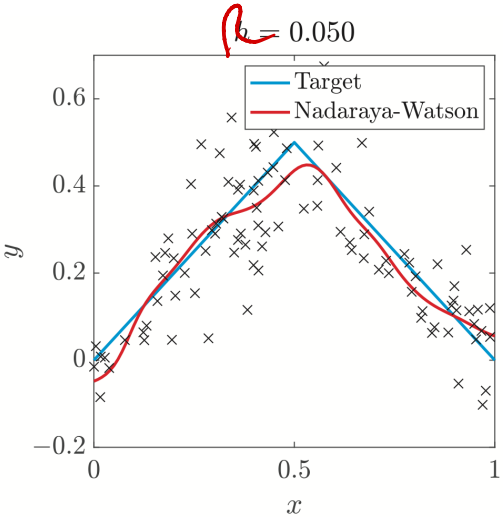

$|J| = 30$
Target
Regressogram
y
x

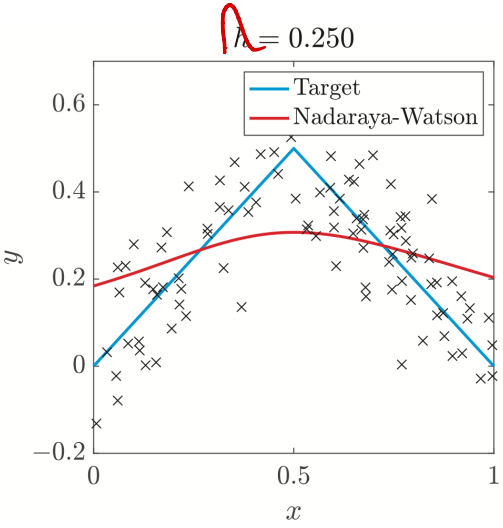

$k = 1$
Target
k-nn
y
x


$k = 10$
Target
k-nn
y
x


$k = 40$
Target
k-nn
y
x


$h = 0.005$
Target
Nadaraya-Watson
y
x


$h = 0.050$
Target
Nadaraya-Watson
y
x


$h = 0.250$
Target
Nadaraya-Watson
y
x

optimization: objective function $F: \mathbb{R}^d \to \mathbb{R}$ → empirical risk $F(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$ ($\in \mathbb{R}$)

→ expected risk $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$ ↑ test distribution

key assumption: convexity of F ⟸ convexity of ℓ and linearity of f_θ

① **smoothness** (bounded Hessians)

$\forall \theta \in \mathbb{R}^d$, $F''(\theta)$ has eigenvalues in $[0, L] \subset \mathbb{R}_+$

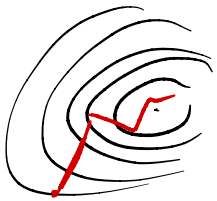
$\gamma = \frac{1}{L}$ for GD: $\theta_t = \theta_{t-1} - \gamma F'(\theta_{t-1})$

result: $F(\theta_t) - F(\eta_*) \leq \left(1 - \frac{\mu}{L}\right)^t [F(\theta_0) - F(\eta_*)]$ (exponential) cvce (linear)

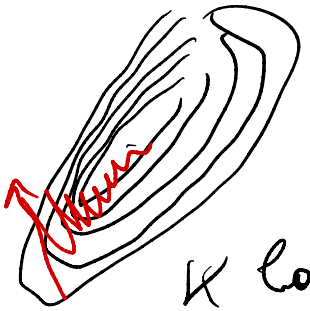
$\leq \frac{L}{2t} \|\theta_0 - \eta_*\|_2^2$

η_* ↑ minimizer of F

$\kappa = \frac{L}{\mu}$ condition number ≥ 1



κ small



κ large

② "**non smooth**" problems (bounded gradients)

$\gamma \sim \frac{1}{\sqrt{t}}$ or $\frac{1}{\sqrt{T}}$ ↳ horizon

$F\left(\frac{1}{T}\sum_{t=0}^{T-1} \theta_t\right) - F(\eta_*) \leq \frac{D}{\sqrt{T}}$ / if strongly cx $\frac{1}{\mu t}$

$\bar{\theta}_T$ averaged iterate

uniform deviations $\frac{D}{\sqrt{n}}$

SGD : $\theta_t = \theta_{t-1} - \gamma_t g_t$ with $\mathbb{E}[g_t(\theta_{t-1}] = F'(\theta_{t-1})$

$\|g_t\|_2^2 \le B^2$

Example 1 : $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i))$

$g_t = \frac{\partial}{\partial \theta}\left[\ell(y_{i(t)}, f_\theta(x_{i(t)})\right]_{\theta=\theta_{t-1}}$

↳ uniform choice at random in $\{1, \dots, n\}$.

$\mathbb{E}[F(\bar{\theta}_t) - F(\theta_*)] \le$

↳ empirical risk minimizer

Example 2 : $F(\theta) = \mathbb{E}[\ell(y, f_\theta(x))]$. $g_t = \frac{\partial}{\partial \theta}\left[\ell(y_t, f_\theta(x_t))\right]_{\theta = \theta_{t-1}}$

only for $t \le n$

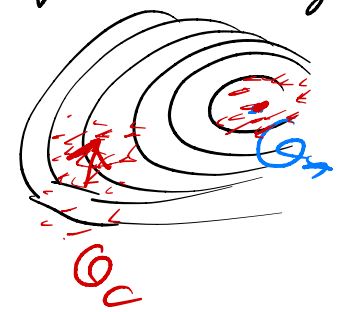
$\mathbb{E}[g_t | \theta_{t-1}] = F'(\theta_{t-1})$

$\mathbb{E}(F(\bar{\theta}_t)) - F(\theta_*) \le$

↳ test error minimizer

$\theta_{t+1} = \theta_t - \gamma_t g_t$ with $\mathbb{E}[g_t \mid \theta_t] = F'(\theta_t)$, $\|g_t\|^2 \le B^2$

θ_* minimizer of F



Research method



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

$$\mathbb{E}\left(\|\theta_t - \theta_*\|_2^2 \mid \theta_{t-1}\right) \le \|\theta_{t-1} - \theta_*\|_2^2 - 2\gamma_t F'(\theta_{t-1})^\top (\theta_{t-1} - \theta_*) + \gamma_t^2 B^2$$

$\ge F(\theta_{t-1}) - F(\theta_*)$ by convexity

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

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

(taking $\gamma_t = \gamma$)

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

(Jensen)

$$\mathbb{E}\left(F(\bar{\theta}_T)\right) - F(\theta_*) \le$$

A few extensions: ① $\min_{\theta} F(\theta) + \frac{\mu}{2}\|\theta\|_2^2 = G(\theta)$

SGD: $\theta_t = \theta_t - \gamma_t[g_t + \mu\theta_{t-1}]$

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

If $\gamma_t = \frac{1}{\mu t}$ then $\mathbb{E}[G(\bar{\theta}_t)] - \min G \leq \log t \frac{1}{\mu t}$

② What about finite sums?

$F(\theta) = \frac{1}{n}\sum \ell(y_i, f_\theta(x_i))$

variance reduction leads to exponential rates

for strongly convex problems

③ Adam / Adagrad: $\theta_t = \theta_{t-1} - \gamma_t$ [M] g_t

Diagonal / full matrix.

preconditioning



Summary of linear models : $f_\theta(x) = \theta^T \varphi(x)$

(1) Estimation error $\leq$ uniform deviation

order $\frac{RG}{\sqrt{n}}$

R : explicitly computable

band on $\|\varphi\|_2$

G : $\|\theta_*\|_2$

(2) GD an empirical risk optimization error is $\frac{RG}{\sqrt{t}}$

(3) SGD on testing risk $\Rightarrow$ excess risk of $\frac{RG}{\sqrt{n}}$ (single pass)

dimension-free bounds

$\Rightarrow$ beyond linear models $\Rightarrow$ boosting, kernel methods, sparse methods, neural nets

Local averaging: loss function $\ell(y,z)$ — square loss if $y=\mathbb{R}$
0-1 loss if y finite
(convex surrogates if $y=\{-1,1\}$)

goal: $\min_{f: X \to \mathbb{R}} \mathbb{E}[\ell(y, f(x))]_{=R(f)}$ eq'

test distribution

solution: (Bayes predictor, target function) $f^*(x') = \arg\min_{z \in Y} \mathbb{E}[\ell(y,z) \mid x = x']$

$= \int \ell(y,z)\, dp(y \mid x=x')$

$= \mathbb{E}[y \mid x = x']$ for square loss

$= \arg\max_z \mathbb{P}(y=z \mid x=x')$ for 0-1 loss

$\arg\min_f R(f)$

Local averaging = formula to estimate

$\hat{p} = p(y \mid x=x')$ based on data (x_i, y_i), $i=1,\dots,n$ IID

plug-in approach = predict using $\hat{p}$

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

Training data: $(x_1, y_1), \dots, (x_n, y_n) \in X \times Y$

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

δ_z dirac measure at z

$\delta_z(A) = 1 \cdot 1_{z \in A} = 1_{z \in A}$

$\hat{w}$ depends only on inputs $x_1, \dots, x_n$

$$\operatorname*{argmin}_{z \in Y} \int \ell(y,z)\, d\hat{p}(y|x=x') \;?$$

Regression $Y \in \mathbb{R}$

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

"linear" estimator

Classification

$$\hat{f}(x) = \operatorname*{argmax}_{z \in Y} \hat{P}(y=z|x)$$

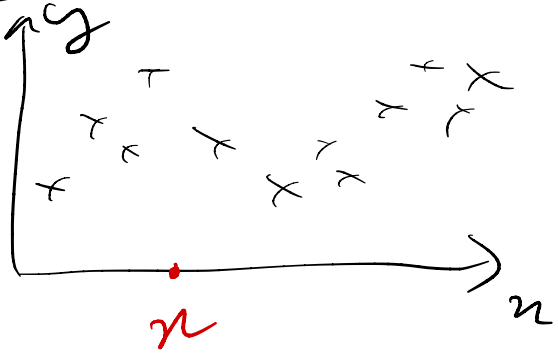
$$\hat{P}(y=z|x) = \int 1_{y=z}\, d\hat{p}(y|x)$$

$$= \operatorname*{argmax}_{z \in Y} \sum_{i=1}^{n} \hat{w}_i(x)\, 1_{z=y_i}$$

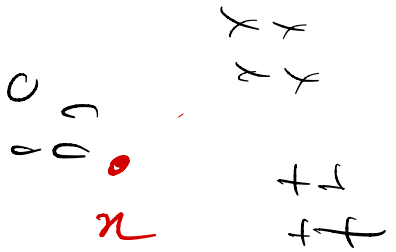
weighted majority vote

$|C^n| < \infty \to |C^n| < \infty$

Regressia : $\mathcal{X} = \mathbb{R}$



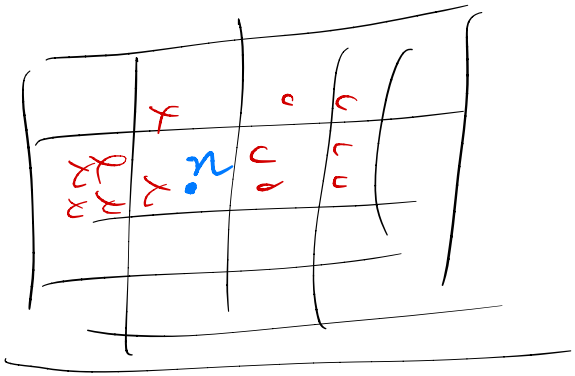
Classificata : $\mathcal{X} = \mathbb{R}^2$



metrica di distanza $d(x, x')$

independent of data

① Partitia estimator



$(c,1)^2$

celle $A_1, \dots, A_m$ s.t. $A_1 \sqcup \dots \sqcup A_m = \mathcal{X}$

partitia di $\mathcal{X}$

x belongs to a unique A_j $j \in (1, \dots, m)$

$w_i(x) = c$ if $x_i \notin A_j$

$= \frac{1}{h_j}$ if $i = j$ with

$h_j = \#$ of elements in cell A_j

see section in book

$\in$ least-squares $\in$ regression

② k-nearest neighbors

order distances $\Delta(x_i, x)$

$$\Delta(x_{i(1)}, x) \leq \Delta(x_{i(2)}, x) \leq \cdots \leq \Delta(x_{i(n)}, x)$$

$\hat{w}_i(x) = \frac{1}{k}$ if i among the k closest

$\hat{w}_i(x) = 0$ otherwise

hyperparameter: $k \in \{1, \ldots, n\}$

complexity: $O(nd)$ for every test point

but can be reduced by "indexing" techniques



(3) Nadaraya - Watson / "kernel" regression



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

$$\hat{w}_i(x) = \frac{h(x, x_i)}{\sum_{j=1}^{n} h(x, x_j)} \Rightarrow \bar{w} \in \text{simplex}$$

$$h(x, x') = \mathbb{1}_{\|x - x'\| \le R} = e^{-\|x - x'\|^2 / R^2}$$

(any norm)



"positive" ≠ "positive"

pointwise: $h(x, x') \in \mathbb{R}_+$

positive definite

hyperparameter: R

Analysis for least-squares: $\hat{f}(x) = \sum_{i=1}^n \hat{w}_i(x) y_i$ → depends only on $x_1, \dots, x_n$

Assumptions:

1. $\hat{w}(x)$ depends only on $x_1, \dots, x_n$
2. $\hat{w}(x) \in$ simplex for all $x \in X$
3. $f_*(x) = \mathbb{E}[y|x]$ is B-Lipschitz continuous w.r.t. distance Δ. ("weak") $|f_*(x) - f_*(x')| \le B\,\Delta(x, x')$
4. $|y - f_*(x)| \le \sigma$ almost surely ("weak")

ε s.t. $\mathbb{E}(\varepsilon|x) = 0$

Let $x \in X$ be a test point (fixed)

$\hat{w}$ sums to 1

$$\hat{f}(x) - f_*(x) = \sum_{i=1}^n \hat{w}_i(x) y_i - f_*(x) = \sum_{i=1}^n \hat{w}_i(x)\,[y_i - f_*(x)] = \sum_{i=1}^n \hat{w}_i(x)\,[y_i - f_*(x_i)] + \sum_{i=1}^n \hat{w}_i(x)\,[f_*(x_i) - f_*(x)]$$

$$\mathbb{E}\left[\left(\hat{f}(x) - f_*(x)\right)^2 \,\middle|\, x_1, \dots, x_n\right] = \left(\sum_{i=1}^n \hat{w}_i(x)\,(f_*(x_i) - f_*(x_i))\right)^2 + \mathrm{var}\left(\sum_{i=1}^n \hat{w}_i(x)\,[y_i - f_*(x_i)]\right) \quad | \; x_1 \to x_n$$

only training outputs

$$= \sum_{i=1}^n \hat{w}_i(x)^2 \mathrm{var}(y_i|x_i)$$

because of (1)

$$\le \left(\sum_{i=1}^n \hat{w}_i(x)\,|f_*(x_i) - f_*(x_i)|\right)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2$$

($\hat{w}_i(x) \ge 0$)

because of (4)

Tensioning (2)

$$\le \sum_i \hat{w}_i(x)\,|f_*(x_i) - f_*(x_i)|^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2$$

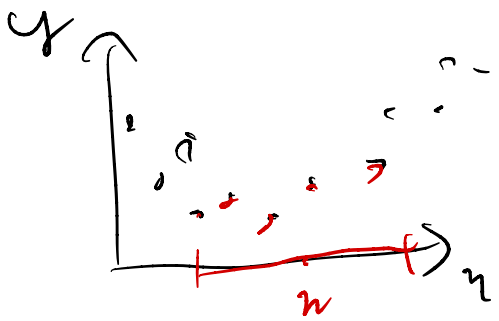
$$\le B^2 \sum_i \hat{w}_i(x)\,\Delta(x, x_i)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2$$

Lemma: $\mathbb{E}(Z - a)^2 = (\mathbb{E}Z - a)^2 + \mathrm{var}(Z)$

We have: $\mathbb{E}\left((f_{\mathcal{S}}(x) - \hat{f}(x))^2 \mid x_1, \dots, x_n\right) \le B^2 \sum_{i=1}^n \hat{w}_i(x) \Delta(x, x_i)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2$

$\forall x \in \mathcal{X}$

(squared) Bias — Variance (error when $f^* = \text{constant}$)



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

Excess risk. target fct

$$\mathbb{E}\left[R(\hat{f}) - R(f_{\mathcal{S}})\right] = \mathbb{E}\left[\int |\hat{f}(x) - f_{\mathcal{S}}(x)|^2 dp(x)\right]$$

training — expected risk — training — testing

$$= \mathbb{E}_{\text{training}} \mathbb{E}_{\text{testing}} |\hat{f}(x) - f_{\mathcal{S}}(x)|^2$$

$$\text{Excess risk} \le B^2 \int \underbrace{\mathbb{E}\left[\sum_{i=1}^n \hat{w}_i(x) \Delta(x, x_i)^2\right]}_{\text{training}} dp(x) + \sigma^2 \sum_{i=1}^n \int \mathbb{E}\left(\hat{w}_i(x)^2\right) \text{training } dp(x)$$

average distance between x and k nearest neighbors

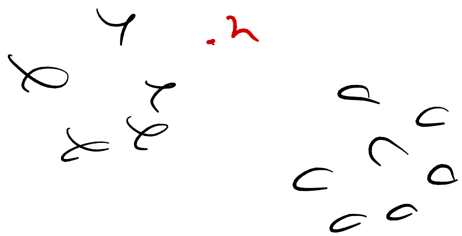
$+ \sigma^2/k$

k-NN =

Lemma ① The average distance to the nearest neighbors

(see book)

$$\leq \frac{\mathrm{diam}(\mathcal{X})^2}{n^{2/d}} \quad \text{if } d \geq 2$$



② The average distance to the k-th NN

$$\leq \mathrm{diam}(\mathcal{X})^2 \left(\frac{k}{n}\right)^{2/d}$$

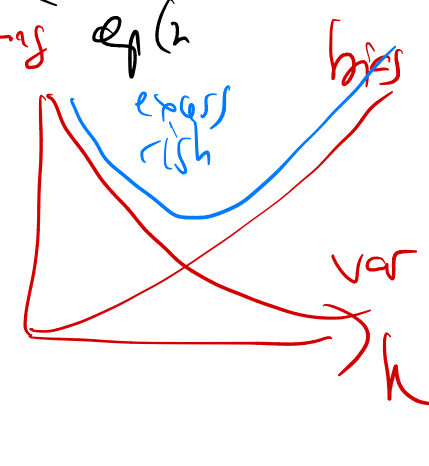
Excess risk

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

trains   trains

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

bias   variance



Excess risk $\leq R^2 \left(\frac{k}{n}\right)^{2/d} + \sigma^2 k$

"optimal" balance will make the two terms equal

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

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

$$= n^{2/2+d}$$

$\Rightarrow$ optimized rate $\frac{1}{n^{2/2+d}}$

Excess risk

+ : goes to zero when $n \to +\infty$ (weak assumptions)

− : very slow (curse of dimensionality)

Unavailable

① assume more smoothness: s derivatives, kernels $\frac{1}{n^{s/d}}$

② latent variables: NN