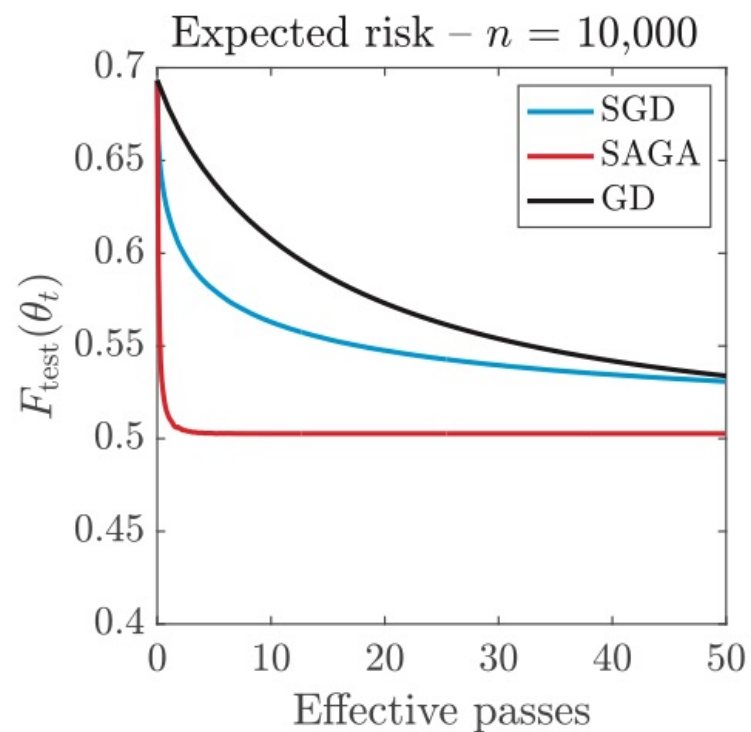
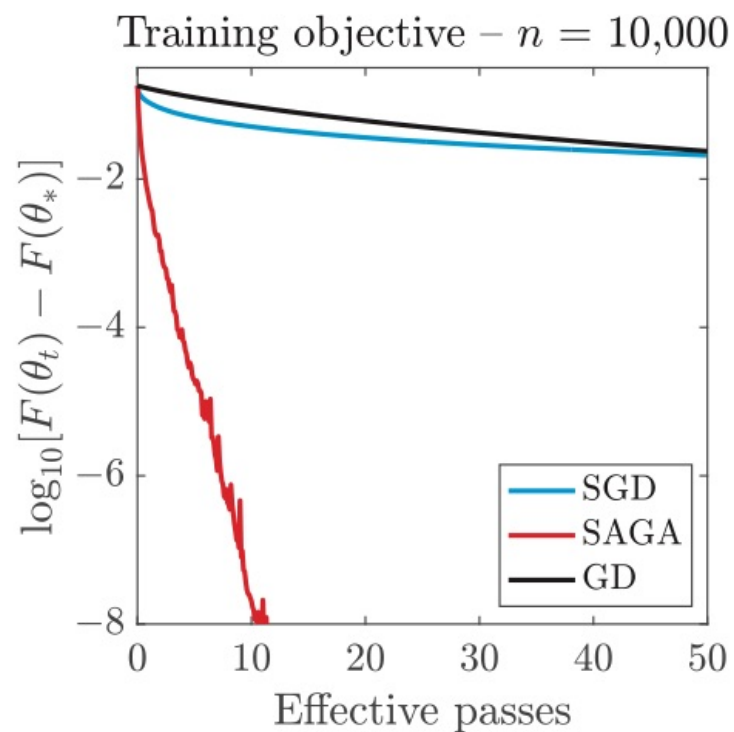
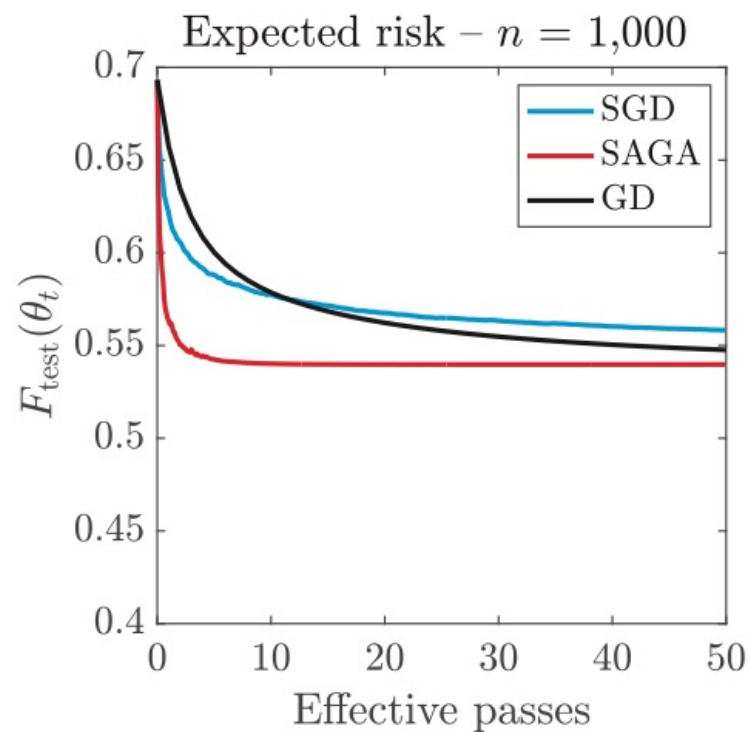
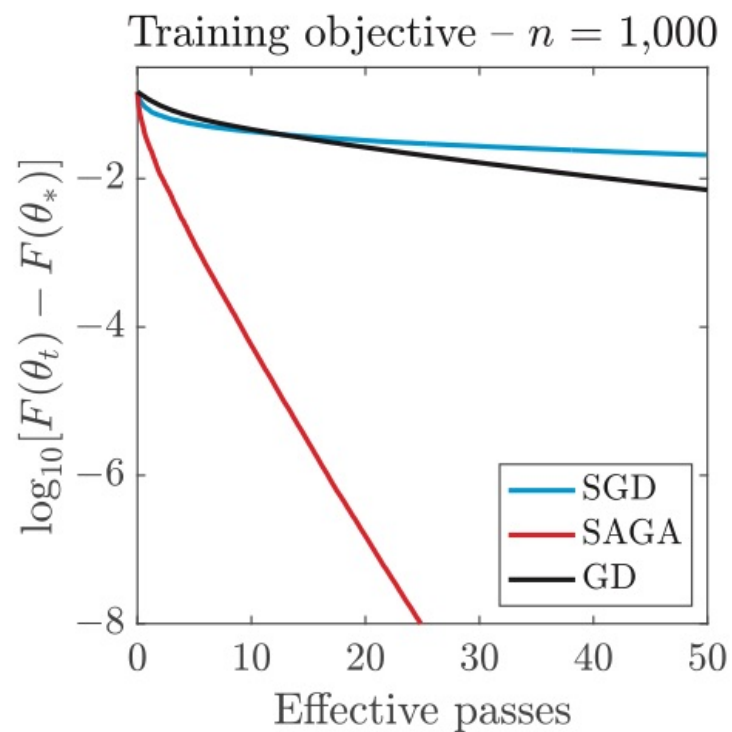
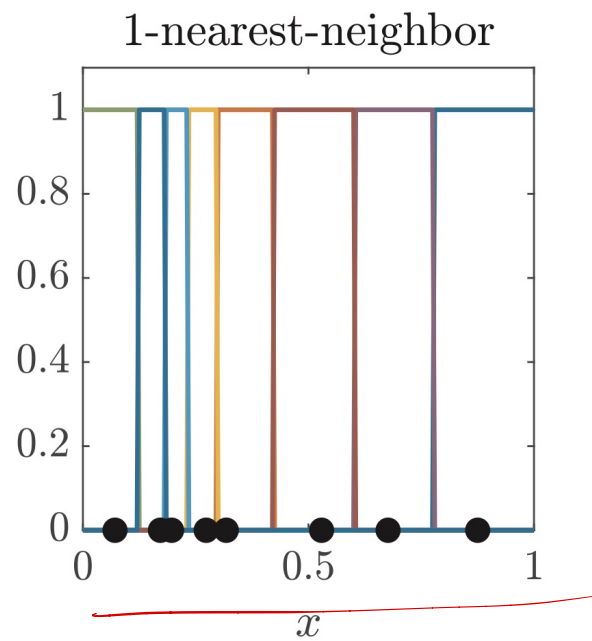
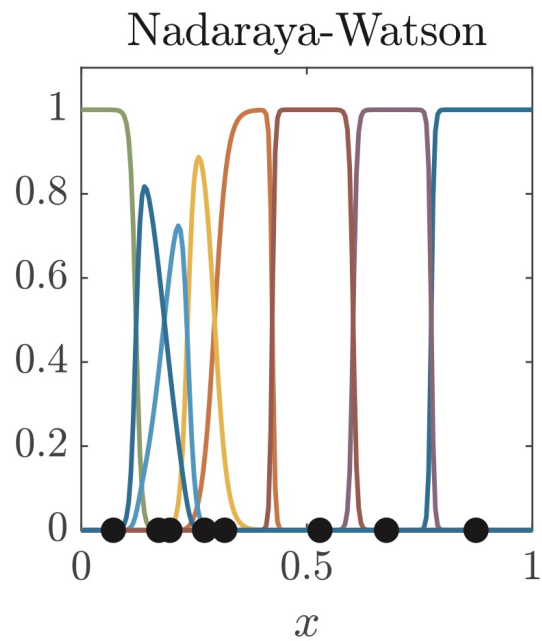
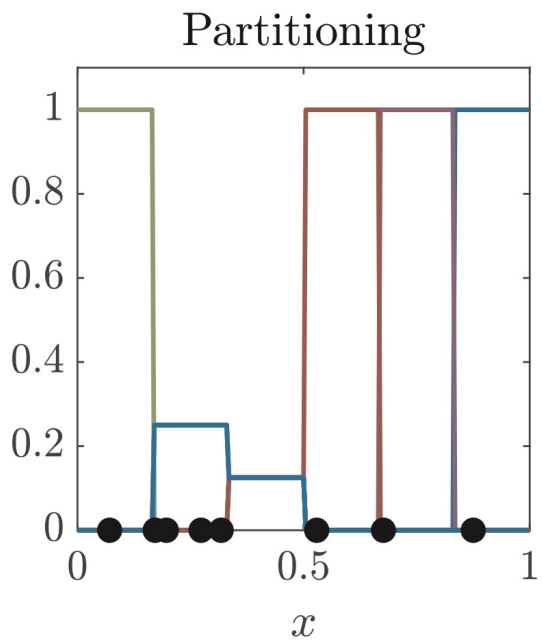
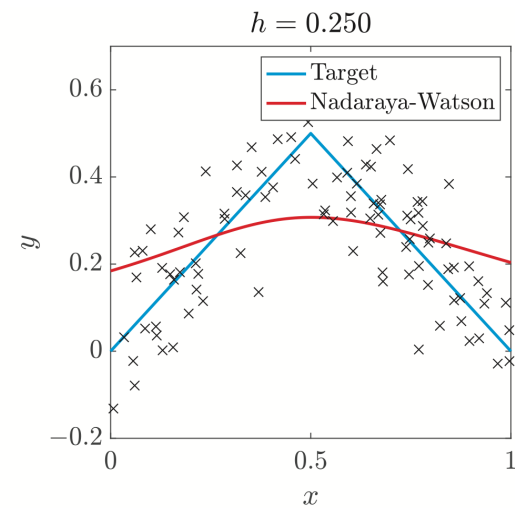
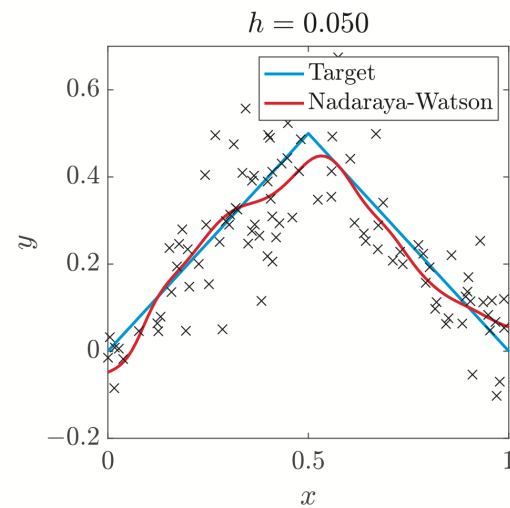
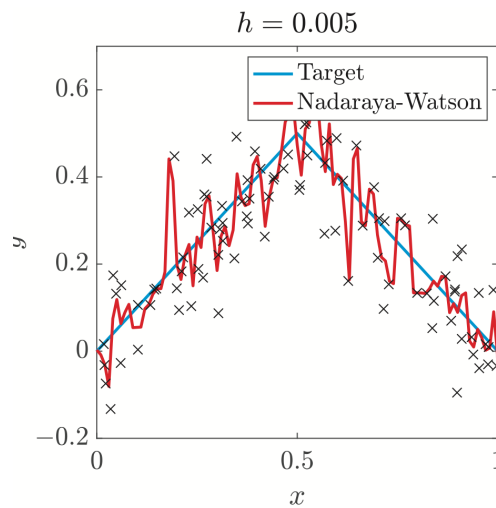
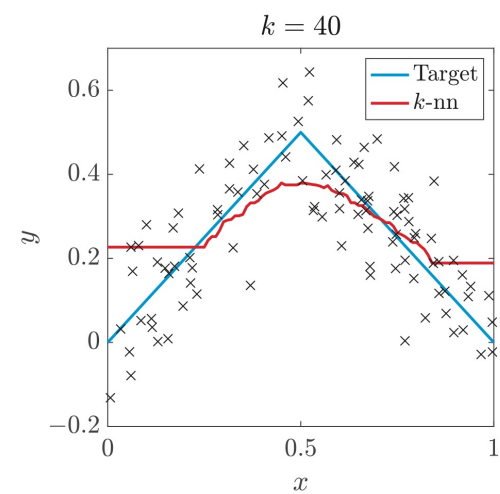
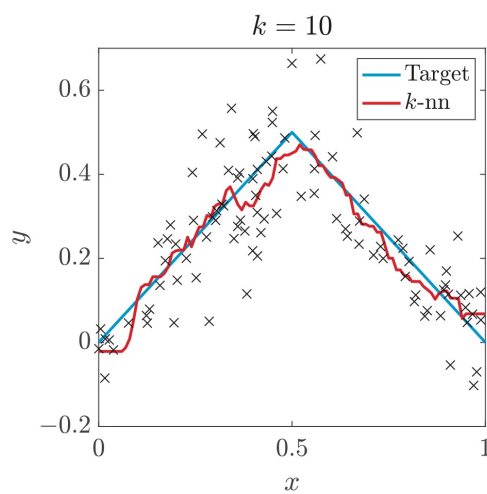
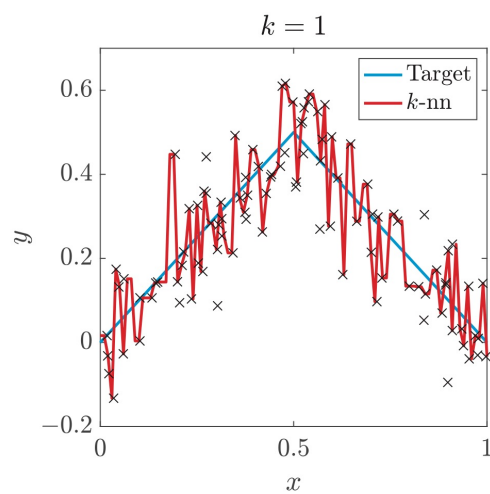
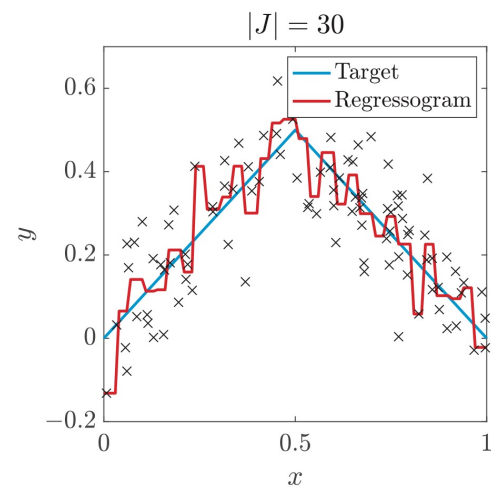
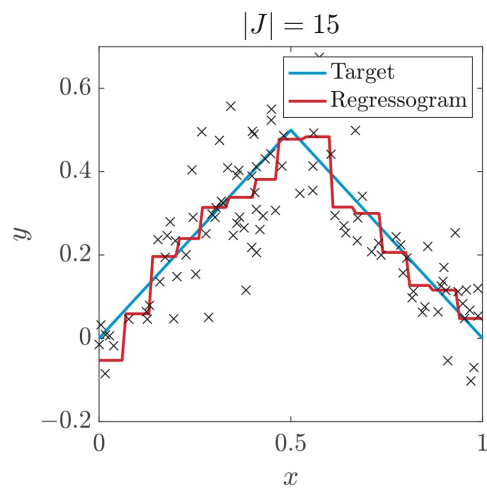
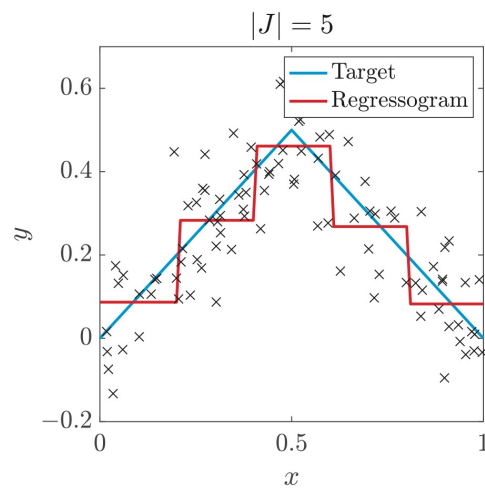
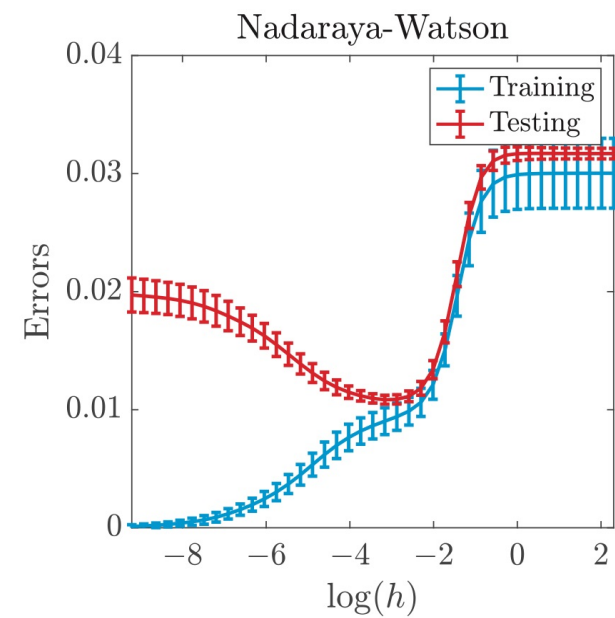
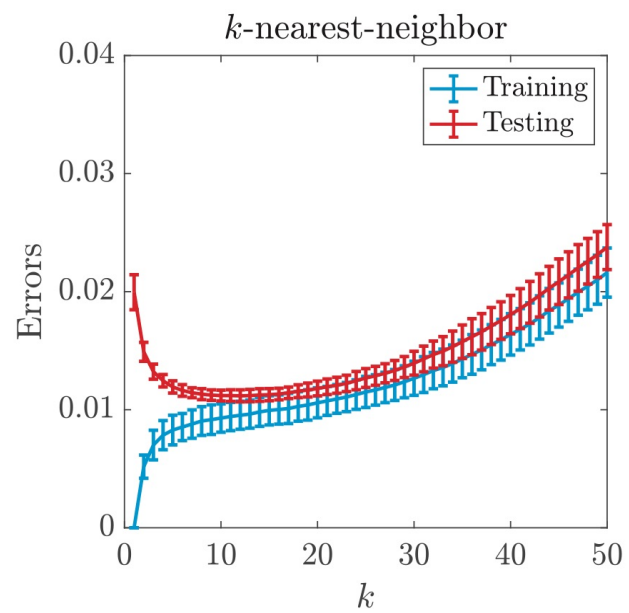
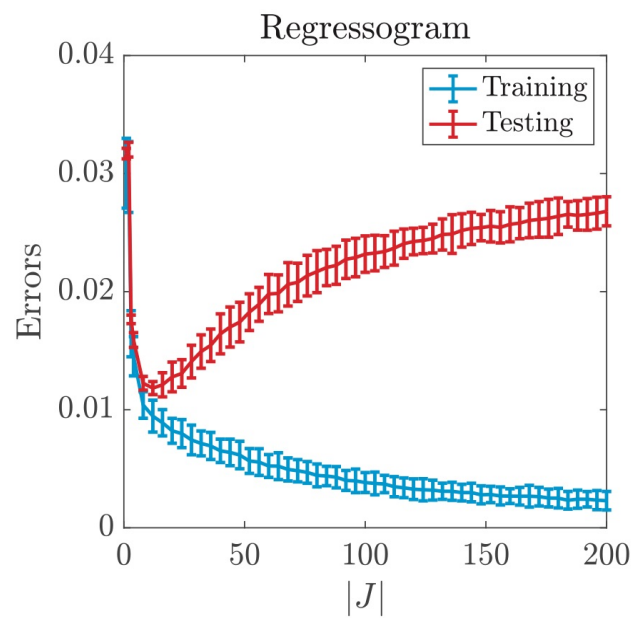




$w_\lambda(2)$







Optimization : objective $F: \mathbb{R}^d \rightarrow \mathbb{R}$ $\rightarrow$ empirical risk $F(\theta) = \frac{1}{n} \sum_i \ell(y_i, f_{\theta}(x_i))$

key assumption : convexity of F
 $(\Leftrightarrow)$ linearity of f_{θ}

expected risk $F(\theta) = \mathbb{E}(\ell(y, f_{\theta}(x)))$

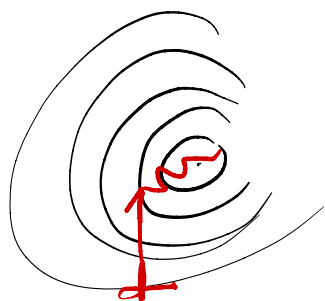
① smoothness (bounded Hessians) with GD : $\theta_t = \theta_{t-1} - \gamma_t F'(\theta_{t-1})$

$$\gamma_t = \frac{1}{L}$$

$$F(\theta_t) - F(\eta_*) \leq (1 - \frac{\kappa}{L})^t (F(\theta_0) - F(\eta_*))$$

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

strong
convexity
parameter



$\kappa = \frac{L}{\mu}$ small

condition number



κ large

② "non smoothness" (bounded gradients)

$$\gamma_t \sim \frac{1}{\sqrt{t}}$$

$$F(\theta_t) - F(\eta_*) \leq$$

$$\frac{B}{\sqrt{t}}$$

$$(\kappa \approx n)$$

linked to
uniform deviation

$$\frac{1}{\sqrt{n}}$$

SGD: $\Theta_t = \Theta_{t-1} - \gamma g_t$ with $E[g_t | \Theta_{t-1}] = F'(\Theta_{t-1})$

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

$g_t = \frac{\partial \ell(g_t, \theta)}{\partial \theta} \Big|_{\theta = \Theta_{t-1}}$

random, uniform in $\{1, \dots, n\}$
 $\hookrightarrow$ independent

$\|g_t\|^2 \leq B^2$

Assuming data fixed

$E[g_t | \Theta_{t-1}] = \frac{1}{n} \sum_{i=1}^n \frac{\partial \ell(g_i, \theta)}{\partial \theta} \Big|_{\theta = \Theta_{t-1}} = F'(\Theta_{t-1})$

$E[F(\Theta_H) - F(\Theta_0)]$

$\hookrightarrow$ Empirical risk minimizer

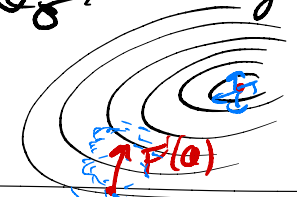
Example 2: $F(\theta) = E[\ell(g, \theta)]$

$g_t = \frac{\partial \ell(g_t, \theta)}{\partial \theta} (g_t, \theta)$ where $(n_1, g_1), \dots, (n_n, g_n)$ are iid observations

$E[g_t | \Theta_{t-1}] = E\left[\frac{\partial \ell(g_t, \theta)}{\partial \theta} \Big|_{\theta = \Theta_{t-1}} \Big| \Theta_{t-1}\right] = \frac{\partial F(\Theta_{t-1})}{\partial \theta} = F'(\Theta_{t-1})$

Θ_{t-1} is a function of $n_1, g_1, \dots, n_{t-1}, g_{t-1}$: independent of (n_t, g_t)
 only for $t \leq n$

$E[F(\Theta_H) - F(\Theta_0)] \rightarrow$ minimizer of the Expected risk.

$$Q_t = Q_{t-1} - \gamma g_t \text{ with } \begin{cases} E[g_t | Q_{t-1}] = F'(Q_{t-1}) & | Q_0 \text{ min of } F \\ \|g_t\|^2 \leq B^2 \end{cases}$$


trick: $\|Q_t - Q_0\|_2^2 = \|Q_{t-1} - Q_0\|_2^2 - 2\gamma \langle g_t, Q_{t-1} - Q_0 \rangle + \gamma^2 \|g_t\|^2$

$\leq B^2$

$$E(\|Q_t - Q_0\|^2 | Q_{t-1}) \leq \|Q_{t-1} - Q_0\|^2 - 2\gamma \langle F'(Q_{t-1}), Q_{t-1} - Q_0 \rangle + \gamma^2 B^2$$

$$E(\|Q_t - Q_0\|^2 | Q_{t-1}) \leq \|Q_{t-1} - Q_0\|^2 - 2\gamma (F(Q_{t-1}) - F(Q_0)) + \gamma^2 B^2$$

$$\underline{E}(\|Q_t - Q_0\|^2) \leq \underline{E}(\|Q_{t-1} - Q_0\|^2) - 2\gamma (\underline{E}F(Q_{t-1}) - F(Q_0)) + \gamma^2 B^2$$

$$\Rightarrow \underline{E}F(Q_{t-1}) - F(Q_0) \leq \underline{E}(\|Q_{t-1} - Q_0\|^2) - \underline{E}(\|Q_t - Q_0\|^2) + \frac{\gamma B^2}{2}$$

$$\sum_{t=1}^{T-1} \left(\underline{E} \left(\frac{1}{T} \sum_{t=1}^T F(Q_{t-1}) \right) - F(Q_0) \right) \leq \frac{\|Q_0 - Q_0\|^2}{2\gamma T} + \frac{\gamma B^2}{2}$$

See also post by ORABONA for bound on $F(Q_T)$

herigan

Jasak

$$E F \left(\frac{1}{T} \sum_{t=1}^T Q_{t-1} \right) - F(Q_0) \leq \frac{\|Q_0 - Q_0\|^2}{2\gamma T} + \frac{\gamma B^2}{2} \propto \frac{1}{\sqrt{T}}$$

(in practice $\gamma = \frac{1}{\sqrt{T}}$)

$\propto \gamma \propto \frac{1}{\sqrt{T}}$

if adding regularizer $\min_{\Theta} F(\Theta) + \frac{\mu}{2} \|\Theta\|_2^2 = G(\Theta)$

$$\Theta_t = \Theta_{t-1} - \gamma_t (g_t + \mu \Theta_{t-1})$$

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

if $\gamma_t = \frac{1}{\mu t}$, then

$$\mathbb{E}(G(\overline{\Theta}_t)) - \min G \leq \frac{(\gamma_t + \frac{\beta^2}{\mu t})}{\mu t}$$

↑
average
error

(compared to $\frac{1}{\sqrt{t}}$)

only an improvement if μ large enough

if optimizing over finite sums

$\delta(G)$ can be improved by variable reduction

$$\begin{aligned} \text{ex: } \mu &= 10^{-6} \\ \frac{1}{\mu t} &< \frac{1}{\sqrt{t}} \\ t &> 10^{12} \end{aligned}$$

LOCAL AVERAGING: loss function $\ell(y, z) = \text{square } (y-z)^2$

Goal: $\min \mathbb{E} \ell(y, f(x))$ *expected risk*

$f: \mathcal{X} \rightarrow \mathcal{Y}$

0-1 $\mathbb{1}_{y \neq z}$
cross entropy

test distribution

Bayes predictor (optimal) $\left| f^*(x) = \arg \min_{z \in \mathcal{Y}} \mathbb{E}(\ell(y, z) | x = x') \right|$

$\int \ell(y, z) dP(y | x = x')$
($= \mathbb{E}(y | x = x')$ for square loss
 $= \arg \max_z P(y = z | x = x')$ for 0-1 loss)

in local averaging techniques:

formula to estimate $p(y | x = x')$
based on input data close to x

as $\hat{p}(y | x = x')$

for regression $\hat{f}(x) = \int y d\hat{p}(y | x)$
for classification $\hat{f}(x) = \arg \max_z \hat{P}(y = z | x)$

"plug-in" estimator

Definition : training data $(x_1, y_1), \dots, (x_n, y_n)$

$\delta_z(y)$ ^{code} measure on y

$\delta_A(y) = 1$ if $y \in A$
0 if $y \notin A$

x

o

o

binary
classification
in $X = \mathbb{R}^2$

predictions

= regression:

classification

$\hat{f}(x) = \arg \max_z$

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

majority vote

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

(Direct point masses)

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

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

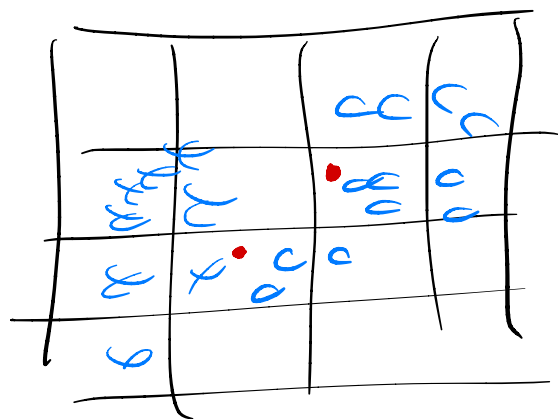


regression
in $X = \mathbb{R}^2$

$$\begin{aligned} \hat{f}(x) &= \int y d\hat{p}(y|x) = \sum_{i=1}^n \hat{w}_i(x) \int y \delta_{y_i}(y) \\ &= \sum_{i=1}^n \hat{w}_i(x) y_i \end{aligned}$$

linear est. ^{y_i} mult. ⁱⁿ labels

Example 1: partition est. matrix



partition of X as $X = A_1 \cup \dots \cup A_m$
 $m \geq 0$

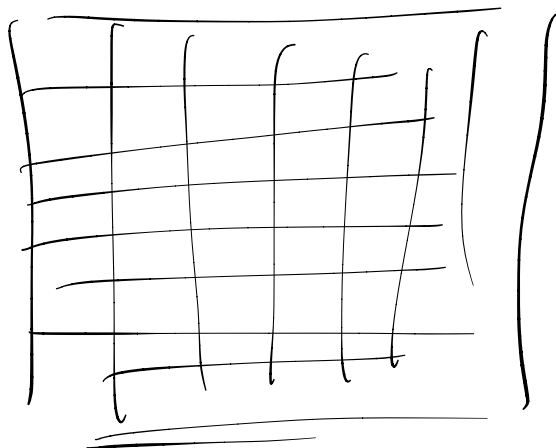
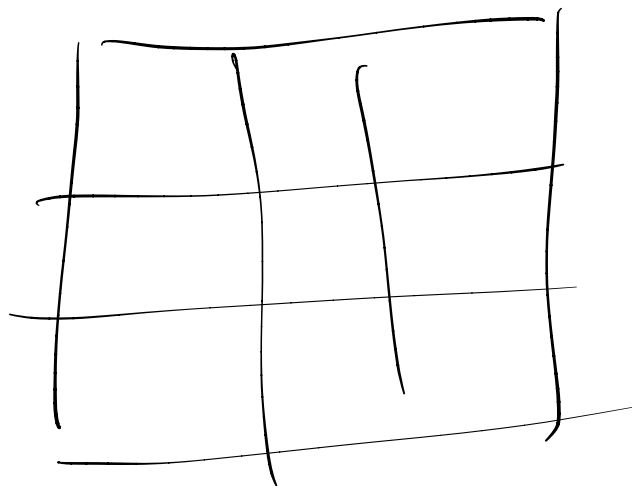
if $x \in A_0$

$w_i(x) = 0$ if $i \neq j$

= combine otherwise = 1

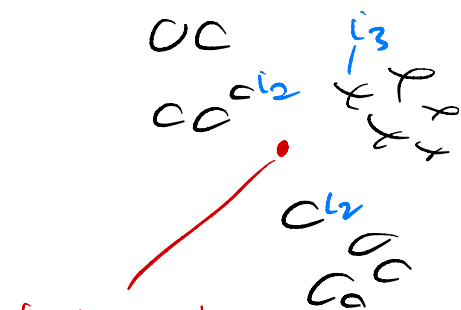
where h is # of
 observations in A_j

hyperparameter



Example 2: h -nearest neighbors. h -nn
need distance Δ or χ .

order $\Delta(x_{i_1(n)}, \underline{x}) \leq \Delta(x_{i_2(n)}, x) \leq \dots \leq \Delta(x_{i_h(n)}, x)$



test point
 x

$h=3$ $\hat{f}(x) = 0$

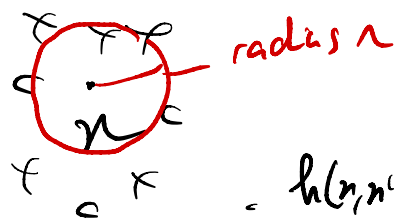
$$\hat{w}_i(x) = \frac{1}{h} \text{ if } i \text{ among the } h \text{ closest ones.}$$

$$= 0 \text{ otherwise}$$

Example 3: Nadaraya-Watson estimator
(kernel regression)

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

$$[h \geq 0]$$



$h(x, x_i) = 1$ if $\|x - x_i\|_2 \leq r$ Box kernel

$h(x, x_i) = \exp\left(-\frac{\|x - x_i\|_2^2}{\sigma^2}\right)$ gaussian kernel

! positive = positive

$(1, 3, 3)$
 $(1, 5, 5)$

local averaging: $\hat{f}(x) = \sum_{i=1}^n \hat{w}_i(x) y_i \Rightarrow$ analysis for regression with square loss

- Assumptions
- ① $\hat{w}_i(x)$ depends only on inputs $x_1, \dots, x_n$
 - ② $\hat{w}_i(x) \geq 0$ and sum to 1 $\forall x$
 - ③ $f_0(x) = \mathbb{E}[y|x]$ B-Lipschitz-continuous wrt distance Δ .
 - ④ $|y - f_0(x)| \leq r$ almost surely (could be replaced by $\mathbb{E}(\epsilon^2|x) \leq \sigma^2$)
- $\mathbb{E}(\epsilon|x) = 0$

Let $x \in X$ a test point

$$\hat{f}(x) - f_0(x) = \sum_{i=1}^n \hat{w}_i(x) y_i - f_0(x)$$

$$= \sum_{i=1}^n \hat{w}_i(x) (y_i - f_0(x_i)) + \sum_{i=1}^n \hat{w}_i(x) (f_0(x_i) - f_0(x))$$

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

$$\mathbb{E}((\hat{f}(x) - f_0(x))^2 | x_1, \dots, x_n) = \left(\sum_{i=1}^n \hat{w}_i(x) (f_0(x_i) - f_0(x)) \right)^2 + \text{var} \left(\sum_{i=1}^n \hat{w}_i(x) (y_i - \mathbb{E}(y_i|x_i)) \mid x_1, \dots, x_n \right)$$

↑
training output

$$= \frac{\left(\sum_{i=1}^n \hat{w}_i(x) (f_0(x_i) - f_0(x)) \right)^2}{\sum_{i=1}^n \hat{w}_i(x)^2} + \sum_{i=1}^n \hat{w}_i(x)^2 \text{var}(y_i | x_i) \leq \sigma^2$$

$$\mathbb{E}((y_i - f_0(x_i))^2 | x_i)$$

$$\leq \left(\sum_{i=1}^n \hat{w}_i(x) |f_0(x_i) - f_0(x)| \right)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2$$

Jensen's
 $w_i \geq 0$
sum to one

$$\leq \sum_{i=1}^n \hat{w}_i(x) |f_0(x_i) - f_0(x)|^2 + \frac{\left(\sum_{i=1}^n \hat{w}_i(x) |f_0(x_i) - f_0(x)| \right)^2}{\sum_{i=1}^n \hat{w}_i(x)}$$

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

$$\leq B^2 \underbrace{\sum_{i=1}^n \hat{w}_i(x) \Delta(x, x_i)^2}_{\text{bias term favors "locality"}} + \underbrace{\sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2}_{\text{variance term}}$$

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

favors uniform weights

bias term
favors "locality"

variance
term

Excess test risk: $\mathbb{E}[\mathcal{R}(\hat{g}) - \mathcal{R}_*]$

$\uparrow$ training data
 $\uparrow$ (expected) risk
 $\uparrow$ test distribution

$$= \mathbb{E} \left(\int |\hat{g}(x) - g_*(x)|^2 d\mathbb{P}(x) \right)$$

training data

$$\approx \left(\mathbb{E}_{\text{training}} \mathbb{E}_{\text{testing}} |\hat{g}(x) - g_*(x)|^2 \right)$$

depends on training data.

$$= \int \mathbb{E} |\hat{g}(x) - g_*(x)|^2 d\mathbb{P}(x)$$

training data

$$\mathbb{E} |\hat{g}(x) - g_*(x)|^2 \leq B^2 \mathbb{E} \left(\sum_{i=1}^n \hat{w}_i(x) \Delta(x, x_i)^2 \right) + \sigma^2 \mathbb{E} \left(\sum_{i=1}^n \hat{w}_i(x)^2 \right)$$

Expected Excess Test Risk $\leq B^2 \left(\int \mathbb{E} \left(\sum_{i=1}^n \hat{w}_i(x) \Delta(x, x_i)^2 \right) d\mathbb{P}(x) \right) + \sigma^2 \left(\int \mathbb{E} \left(\sum_{i=1}^n \hat{w}_i(x)^2 \right) d\mathbb{P}(x) \right)$

B^2 : bias
 Expected distance to h -nearest neighbors

variance
 for h -nn $= \left\lfloor \frac{\sigma^2}{h} \right\rfloor$

Lemma: distance to nearest neighbor: $h=1$; $d \geq 2$
 let p be a distribution on $X \subset \mathbb{R}^d$ with X bounded.
 $n+1$ points iid from $p: x_1, \dots, x_{n+1}$.

$$\mathbb{E} [\ell_\infty\text{-distance}^2 (x_{n+1}, \text{nearest neighbor of } x_{n+1} \text{ within } \{x_1, \dots, x_n\})]$$

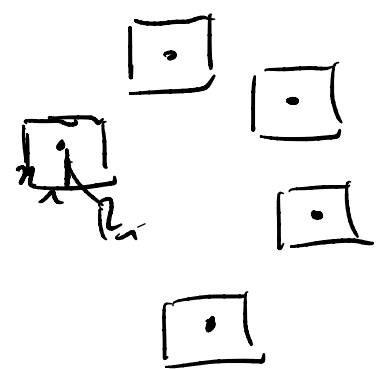
$$e = \mathbb{E} \left(\min_{i \in \{1, \dots, n\}} \|x_{n+1} - x_i\|_\infty^2 \right) \leq \frac{4}{n^{2/d}} \text{diam}(X)^2$$

Proof: let $x_{(i)}$ be the nearest neighbor of x_i among $x_j, j \neq i, j \in \{1, \dots, n+1\}$.

$$e = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{E} \left(\|x_i - x_{(i)}\|_\infty^2 \right) \quad \text{by exchangeability}$$

$= e$

$$R_i = \|x_i - x_{(i)}\|_\infty > 0$$



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

$$B_i \cap B_j = \emptyset \text{ because } \|x_i - x_j\|_\infty \geq \max\{R_i, R_j\} \geq \frac{R_i + R_j}{2}$$

$$\Rightarrow \text{vol} \left(\bigcup_{i=1}^{n+1} B_i \right) = \sum_{i=1}^{n+1} \text{vol}(B_i) = \sum_{i=1}^{n+1} R_i^d$$

Jensen's

$$\text{goal: } \frac{1}{n+1} \sum_{i=1}^{n+1} R_i^2 \leq \left(\frac{1}{n+1} \sum_{i=1}^{n+1} R_i^{2 \cdot d/2} \right)^{2/d} \leq \frac{1}{(n+1)^{2/d}} (2 \text{diam}(X))^2 \quad \text{if } \frac{d}{2} \geq 1$$

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

Lemma: Expected distance to h -nn

$$\leq 8 \operatorname{diam}(X)^2 \left(\frac{2k}{n}\right)^{2/d} \quad k-d \geq 2$$

Proof: see Lemma 6.2 in book.

Excess risk for h -nn

$$\leq \underbrace{\frac{\sigma^2}{k}}_{\text{variance}} + B^2 8 \operatorname{diam}(X)^2 \left(\frac{2k}{n}\right)^{2/d}.$$

opt.-mat trade off: $\frac{1}{k} \sim \left(\frac{h}{n}\right)^{2/d}$ ^{tries} $(\Rightarrow) h^{1+\frac{d}{2}} = n$
 $h = n^{2/d+2}$

with opt.-mat value $\propto \frac{1}{n^{2/d+2}}$

very slow
but
"optimal"

Curse of dimensionality

Two extra assumptions

① f_{θ} with more derivatives
 f_{θ} Lipschitz $\Rightarrow |f'(\alpha)|$ bounded
 $\Rightarrow \|f_{\theta}^{(s)}\|$ bounded $\Rightarrow$

$$\frac{1}{n^{2s/d}}$$

kernel
methods

② f_{θ} depends on a ^{unknown} low-dimensional
projection of x

kernel
nets

$$\frac{1}{n^{2s/r}}$$

where
 $r = \dim$