LOCAL AVERAGING

distribution p on $\mathcal{X} \times \mathcal{Y}$

data (x_i, y_i) $i = 1, \dots, n$ i.i.d. from $p(x,y)$

loss function $\ell(y, z)$

Goal: find f such that $R(f) = \mathbb{E}\, \ell(y, f(x))$ for $f : \mathcal{X} \to \mathcal{Y}$

Expected risk

Bayes predictor: $f^*(x) = \underset{z \in \mathcal{Y}}{\operatorname{argmin}}\ \mathbb{E}_{p(y|x)}\, \ell(y, z)$

square loss: $\ell(y,z) = (y-z)^2$

$\mathcal{Y} \in \mathbb{R}$

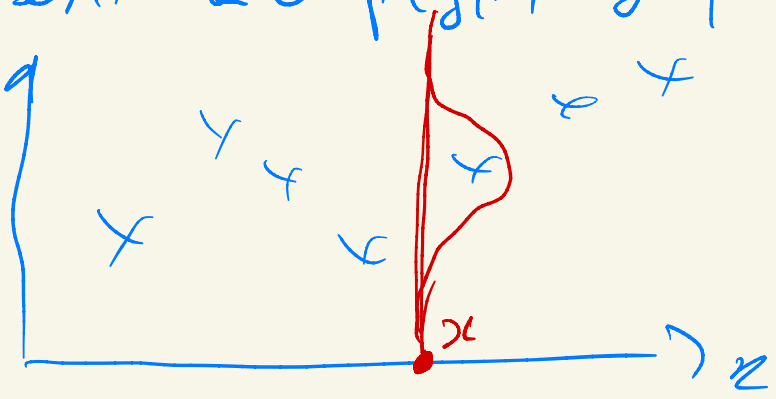
0-1 loss for classif: $f^*(x) = \underset{z \in \{1, \dots, k\}}{\operatorname{argmax}}\ p(y = k \mid x)$

$y \in \{1, \dots, k\}$

square loss: $f^*(x) = \mathbb{E}(y|x)$

$y \in \mathbb{R}$

Principle: estimate $p(y|x)$ by $\hat{p}(y|x)$ — plug-in estimators

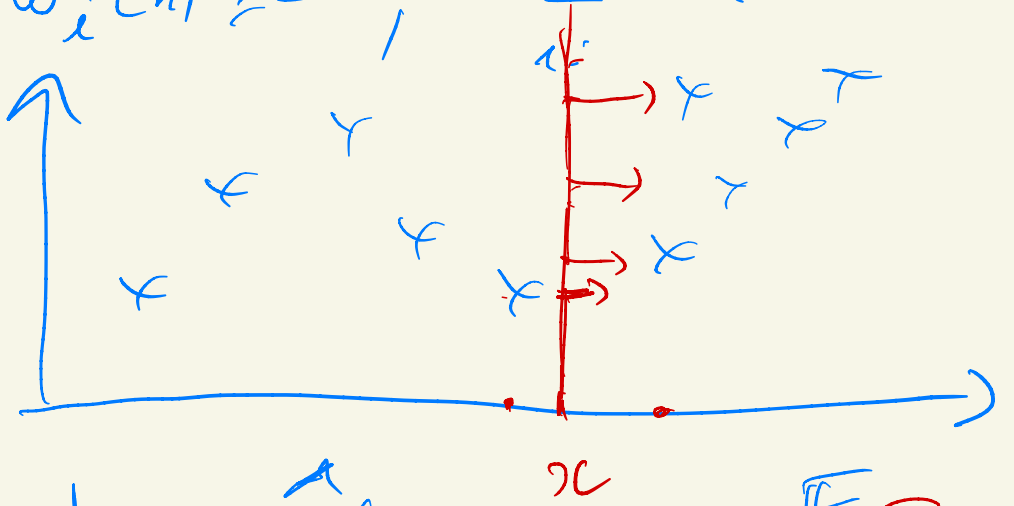


"Linear estimators": $\hat{p}(y|x) = \sum_{i=1}^{n} w_i(x)\, \delta_{y_i}(y)$ — δ_{y_i} Dirac at y_i

weight function $w_i : \mathcal{X} \to \mathbb{R}$

To get a prob. measure:

$\forall x \quad w_i(x) \geq 0$, $\sum_{i=1}^{n} w_i(x) = 1$



0-1 loss: $\hat{p}(y=3|x)$

$= \mathbb{E}_{\hat{p}(y|x)} 1_{y=3}$

$= \sum_{i=1}^{n} w_i(x)\, 1_{y_i=3}$

Majority vote

Estimators: $\hat{f}(x) = \arg\min_y \mathbb{E}_{\hat{p}(y|x)} \ell(y,3)$

$= \arg\min_y \sum_{i=1}^{n} w_i(x)\, \ell(y_i,3)$

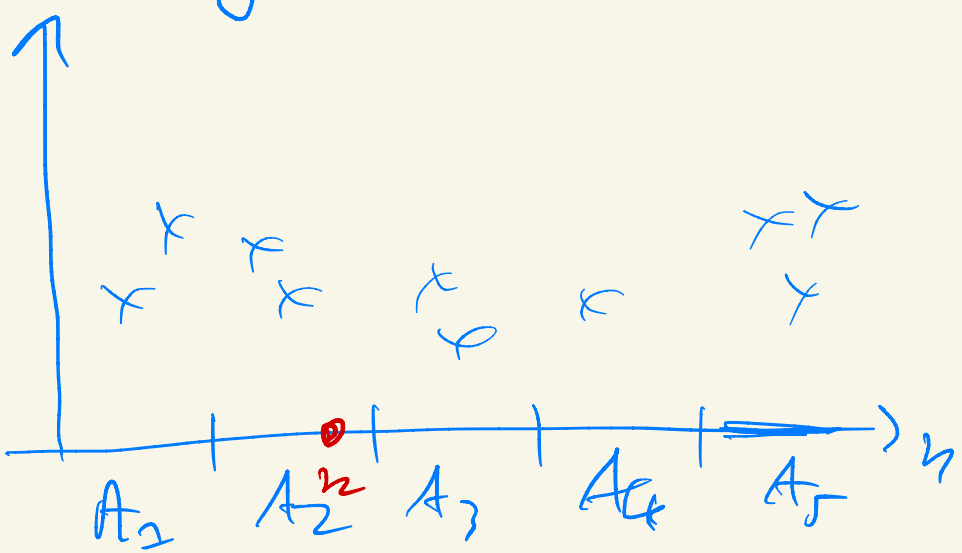
square loss. $\hat{f}(x) = \mathbb{E}_{\hat{p}(y|x)} y = \sum_{i=1}^{n} w_i(x) y_i$

weighted average of all labels

① Partitioning estimate

weight for cell A_j

v_j = number of observations in A_j

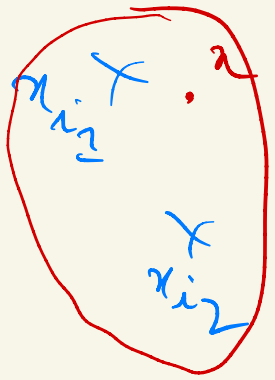


$w_i(x) = \frac{1}{v_j}$ if $x_i \in A_j$

② Nearest neighbor: need a metric d on $\mathcal{X}$

given x, order all $d(x, x_i)$: $d(x, x_{i_1}) \le \dots \le d(x, x_{i_k}) \le \dots \le d(x, x_{i_n})$

Assign weight $\frac{1}{k}$ to the k-nearest



Running time complexity

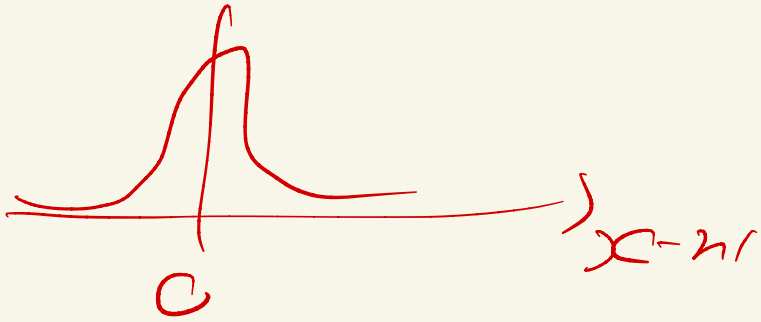
$O(nd)$ for each test point

kd-trees

③ kernel regression / Nadaraya-Watson estimators

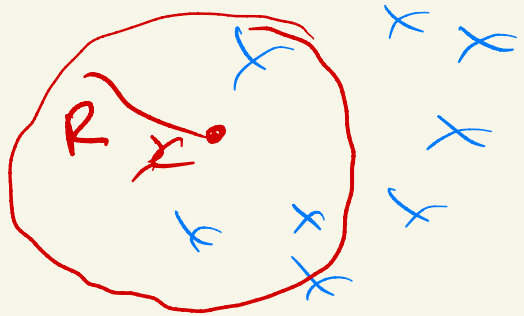
Two types of kernels. compact: $k(x, x') \geq 0$

usually: $k(x, x') = q(x - x')$



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

$\Rightarrow$



ex: $k(x, x') = 1$ if $|x - x'| \leq R$

$= 0$ otherwise

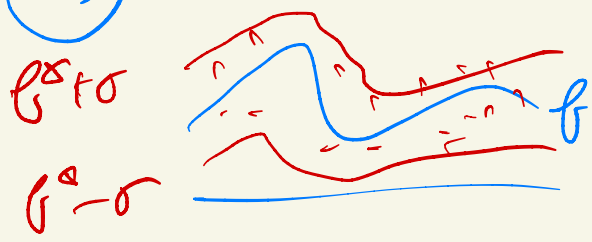
Analysis of local averaging

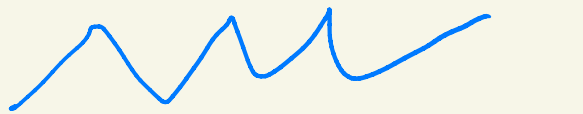
Linear reg: $f^*(x) = \theta^T x$ / $\frac{\sigma^2 d}{n}$

(H1) $Y = f(X) + $ square loss, with $f^*(x) = \mathbb{E}(Y|X)$

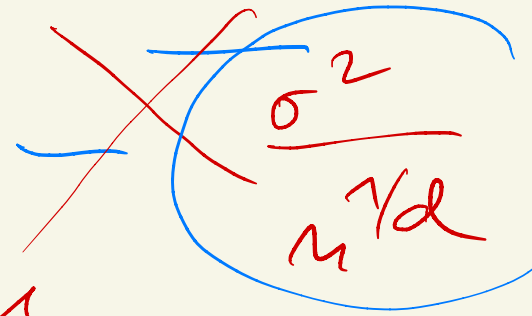
(H2) f^* is B-lipschitz continuous, $|f^*(x) - f^*(x')| \le B\, d(x, x')$

(H3) Bounded noise: $|Y - f^*(x)| \le \sigma$ almost surely





$\Rightarrow$



$\frac{1}{n^{1/d}} = a.1$

$\hat{f}(x) = \sum_{i=1}^{n} w_i(x)\, y_i$ — w_i only depends on $x_1, \dots, x_n$

Goal: $R(\hat{f}) - R(f^*)$ is small $\Rightarrow$ (rest dist) $\Rightarrow$ small

with $R(f) = \mathbb{E}(Y - f(X))^2$

$$R(\hat{f}) - R(f^*) = \mathbb{E}\left[(\hat{f}(x) - f^*(x))^2\right] = \int (\hat{f}(x) - f^*(x))^2 p(x)\, dx$$

$\|\hat{f} - f^*\|_{L^2}$

$n^{1/c}$ $\quad$ $\frac{1}{n^{2/c}}$

Curse of dimensionality

Goal: $\mathbb{E}\, R(\hat{f}) - R(f^*)$

$$\hat{f}(x) - f^*(x) = \sum_{i=1}^{n} w_i(x) y_i - f^*(x)$$

x fixed

$$= \sum_{i=1}^{n} w_i(x) \left[y_i - \mathbb{E}(y_i | x_i) \right] + \sum_{i=1}^{n} w_i(x) \left[\mathbb{E}(y_i | x_i) - f^*(x) \right]$$

($f^*(x_i)$ above the first bracket)

(because $\sum_i w_i(x) = 1$)

$$\mathbb{E}\left(\left(\hat{f}(x) - f^*(x) \right)^2 | x_1, \dots, x_n \right) = \left(\sum_i w_i(x) \left| \left(f^*(x_i) - f^*(x) \right) \right| \right)^2 + \sum_{i=1}^{n} w_i^2(x) \, \mathbb{E}\left(\left(y_i - \mathbb{E}(y_i | x_i) \right)^2 | x_i \right)$$

$\leq B \, d(x_i, x)$; $\leq \sigma^2$

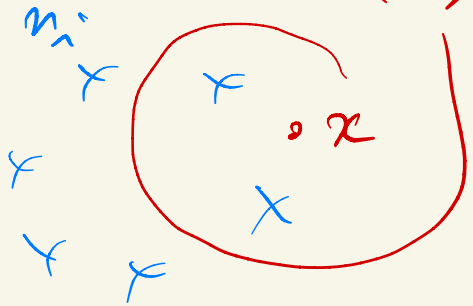
$$\mathbb{E}Z^2 = (\mathbb{E}Z)^2 + \mathbb{E}(Z - \mathbb{E}Z)^2$$

$$\leq \left(\sum_i w_i(x) B \, d(x_i, x) \right)^2 + \sigma^2 \sum_i w_i^2(x)$$

$$\leq \underbrace{B^2 \sum_i w_i(x) \, d(x, x_i)^2}_{\text{bias}} + \underbrace{\sigma^2 \sum_{i=1}^{n} w_i^2(x)}_{\text{variance}}$$

$MW = \frac{1}{n}$

$\sum_i w_i(x) = 1$



$$\sigma^2 \left[\sum_i \left(w_i(x) - \frac{1}{n} \right)^2 - n \times \frac{1}{n^2} + \frac{2}{n} \right]$$

measure of non uniformity

Task 1 : compute variance for k-nn $\Rightarrow$ $\boxed{\sigma^2/k}$

Task 2 : $\int \sum_{i=1}^{n} w_i(x)\, d(x, x_i)^2 p(x)\, dx$ for knn $\Big\{$ $w_i(x) = \frac{1}{k}$ for $i = 1 \to k$; 0 otherwise

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

$\frac{k}{n}$ if $d = 1$

Trade-off : $B^2 diam(\mathcal{X})^2 \left(\frac{k}{n}\right)^{2/d} + \frac{\sigma^2}{k}$

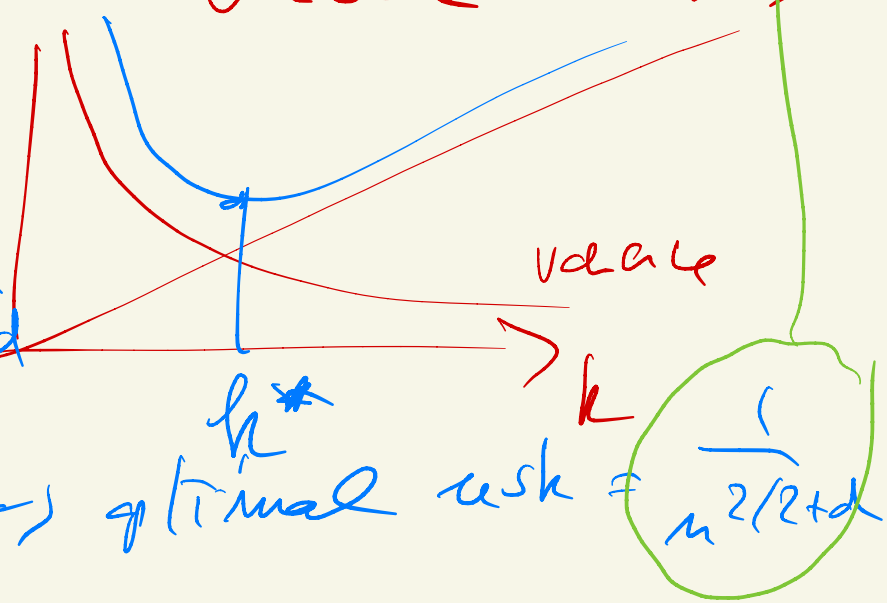
bias — variance — optimal

"optimal" choice

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

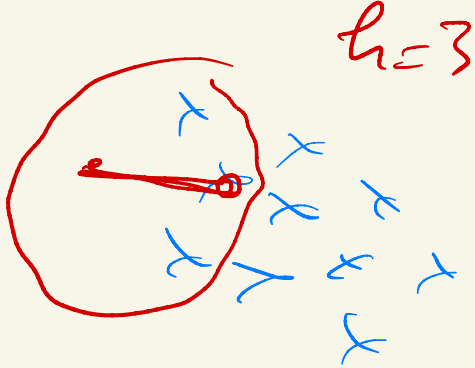
$k^{\frac{2}{d}+1} = n^{2/d} \Leftrightarrow k = n^{\frac{2/d}{2/d+1}} = n^{2/2+d}$

$\propto n^{2/2+d} \Rightarrow$ optimal risk $\neq \frac{1}{n^{2/2+d}}$



goal: $\int p(x) \sum_{i=1}^{k} w_i(x)\, d(x, x_i)^2$

$\leq d(x, k\text{-th nearest neighbors})$



$n+1$ points: $\underbrace{x_1, \dots, x_n}_{\text{training data}}, \underbrace{x_{n+1}}_{\text{test point}}$

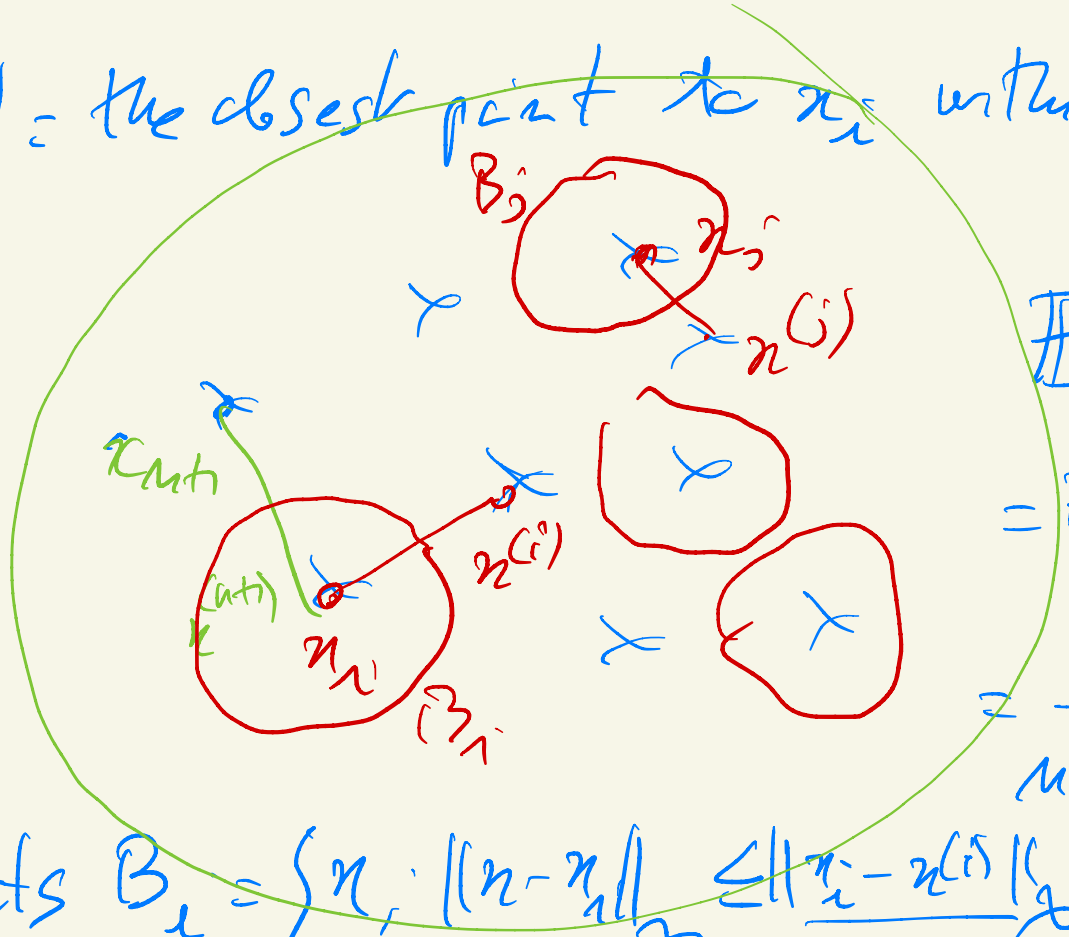
Lemma: Let $x_1, \dots, x_n, x_{n+1}$ $n+1$ iid points from $p(x)$

then $E\, d(x_{n+1}, \text{1st nearest neighbor in } \{x_1, \dots, x_n\})^2 \leq \frac{1}{n^{2/d}} \cdot \text{const}$ for $d \geq 2$

Proof: in $\mathbb{R}^d$ with $d(x, x') = \|x - x'\|_\infty$

Notation: $x^{(i)}$ the nearest-neighbor of x_i in $x_1, \dots, x_{n+1}$ $\setminus x_i$

$x^{(i)}$ = the closest point to x_i within $x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{n+1}$



$$\mathbb{E}\,\|x_{n+1} - x^{(n+1)}\|_\infty^2 = \mathbb{E}\,\|x_k - x^{(k)}\|_2^2 \quad \forall k$$

$$= \frac{1}{n+1} \sum_{k=1}^{n+1} \mathbb{E}\,\|x_k - x^{(k)}\|_\infty^2$$

Sets $B_i = \left\{ x : \|x - x_i\|_\infty \le \|x_i - x^{(i)}\|_\infty \right\}$

The balls are disjoint

$$\sum_i \text{volume}(B_i) \le \text{volume}(Z)$$

where

$$Z = \{ a : \|a - y\|_2 \le \text{diam}(\mathcal{X}) \}$$

$$\frac{1}{n+1} \sum_i \left[\frac{\|x_i - x^{(i)}\|_\infty}{2} \right]^d \le \left(2\,\text{diam}(\mathcal{X}) \right)^d \quad \text{diam}(Z) \le 2\,\text{diam}(\mathcal{X})$$

Then, with Jensen's inequality

$$\sum_{i=1}^{m+1} \frac{\|x_i - x^{(i)}\|^d}{2^d} \leq \operatorname{diam}(\mathcal{X})^d$$

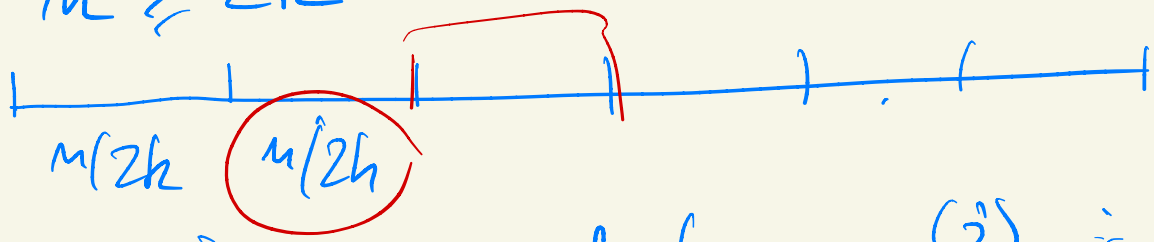
Jensen: $\frac{1}{m+1} \sum_{i=1}^{m+1} \left(\|x_i - x^{(i)}\|^2\right)^{d/2} \geq \left(\frac{1}{m+1} \sum_{i=1}^{m+1} \|x_i - x^{(i)}\|^2\right)^{d/2}$

$$\frac{1}{m+1} \sum_{i=1}^{m+1} \|x_i - x^{(i)}\|^2 \leq \left(\frac{4^d \operatorname{diam}(\mathcal{X})^d}{m+1}\right)^{2/d}$$

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

Lemma: the average squared distance to the k-NN is less than $8\left(\frac{k}{n}\right)^{2/d} \operatorname{diam}(\mathcal{X})^2$ if $d \geq 2$

Proof: $n \geq 2k$



$\Rightarrow 2k$ pieces, notation: $x^{(j)}$ is the nearest neighbor of x_{n+1} within block j

$$\|x_{n+1} - k\text{-nn}\|^2 \leq \frac{1}{k} \sum_{i=1}^{2k} \|x_{n+1} - x^{(j)}\|^2$$

$$\leq \frac{1}{k} \sum_{i} \frac{8 \operatorname{diam}(\mathcal{X})^2}{(n/2k)^{2/d}} \leq \left(\frac{2k}{n}\right)^{2/d} 8 \operatorname{diam}(\mathcal{X})^2$$

Neural networks = Empirical risk minimization

- → kernel methods
- → sparse
- → neural networks