Local Averaging Methods

Zhehao Li

May 29, 2022

Empirical Risk Minimization

In empirical risk minmization, our target is to approximate the Bayes predictor by minimizing the expected risk $\mathcal{R}(f) = \mathbb{E}_P[l(Y, f(X))]$. However, the joint distribution of real data $P(X, Y)$ remains unknown, so we have to minimize the empirical risk, which assign uniform weight $1/n$ to each (X_i, Y_i) pair,

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mathcal{R}}(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(X_i))$$

In spite of this, the empirical risk is still difficult to optimize as f could be any measurable function. Therefore, we constrain our choice in a hypothesis space $\mathcal{F}$ in order to make the optimization problem solvable.

- Binary Classification: $\mathcal{Y} = \{-1, 1\}$ with 0-1 loss $l(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$, then the expected risk is

$$\mathcal{R}(f) = \mathbb{E}[\mathbb{1}(Y \neq f(X))] = \mathbb{P}(Y \neq f(X))$$

- Regression: $\mathcal{Y} = \mathbb{R}$ with square loss $l(y, \hat{y}) = (y - \hat{y})^2$, and the expected risk

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

Bayes Risk and Predictor

Proposition (Bayes Predictor and Bayes Risk)

The conditional expected risk is minimized at a Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ satisfying for all $x \in \mathcal{X}$,

$$f^* \in \underset{\text{measurable } f}{\arg\min} \ \mathbb{E}\big[l(Y, f(X))\big]$$

The Bayes risk $\mathcal{R}^*$ is the risk of all Bayes predictors and is equal to

$$\mathcal{R}^* = \inf_{\text{measurable } f} \mathbb{E}\big[l(Y, f(X))\big]$$

Note that the Bayes predictor is not unique, but that all Bayes predictors lead to the same Bayes risk, and that the Bayes risk is usually nonzero (unless the dependence between X and Y is deterministic).

Binary Classification

Let $\mathcal{Y} = \{-1, 1\}$ and $l(y, \hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$, the Bayes predictor is equal to

$$f^*(X) \in \underset{y \in \{-1,1\}}{\arg\max}\ \mathbb{P}(Y = y \mid X) = \mathsf{sgn}(\eta(X) - 1/2)$$

where $\eta(X) = \mathbb{P}(Y = 1 \mid X)$, and multi-category case $f^*(X) \in \arg\max_{y \in \{1,\ldots,k\}}\ \mathbb{P}(Y = y \mid X)$. Moreover,

$$\mathcal{R}(f) - \mathcal{R}^* = \mathbb{E}\big[|2\eta(X) - 1| \cdot \mathbb{1}(f^*(X) \neq f(X))\big] \tag{1}$$

This is due to the fact that

$$\begin{aligned}\mathcal{R}(f) - \mathcal{R}^* &= \mathbb{E}\big[\mathbb{E}[\mathbb{1}(Y \neq f(X)) - \mathbb{1}(Y \neq f^*(X)) \mid X]\big] \\ &= \mathbb{E}\big[\mathbb{E}[\mathbb{1}(1 \neq f(X)) - \mathbb{1}(1 \neq f^*(X)) \mid X, Y = 1] \cdot \eta(X) + \mathbb{E}[\mathbb{1}(-1 \neq f(X)) - \mathbb{1}(-1 \neq f^*(X)) \mid X, Y = -1\end{aligned}$$

- $\eta(X) > 1/2$ and $f(X) = -1$, $\mathcal{R}(f) - \mathcal{R}^* = \eta(X) - (1 - \eta(X)) = 2\eta(X) - 1$
- $\eta(X) < 1/2$ and $f(X) = 1$, $\mathcal{R}(f) - \mathcal{R}^* = -\eta(X) + (1 - \eta(X)) = 1 - 2\eta(X)$

Regression

Let $\mathcal{Y} = \mathbb{R}$ and $l(y, \hat{y}) = (y - \hat{y})^2$, the Bayes predictor is

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

Moreover, with the square loss, we have

$$\begin{aligned}
\mathcal{R}(f) - \mathcal{R}^* &= \int_{\mathcal{X}} \Big\{ \mathbb{E}_Y[(Y - f(X))^2 \mid X = x] - \mathbb{E}_Y[(Y - f^*(X))^2 \mid X = x] \Big\} dP(x) \\
&= \int_{\mathcal{X}} \Big\{ \mathbb{E}_Y\big[2Y(f^*(X) - f(X)) + f(X)^2 - f^*(X)^2 \mid X = x\big] \Big\} dP(x) \\
&= \int_{\mathcal{X}} \Big\{ 2(f^*(x) - f(x)) \mathbb{E}_Y[Y \mid X = x] + f(x)^2 - f^*(x)^2 \Big\} dP(x) \qquad (2) \\
&= \int_{\mathcal{X}} \Big\{ 2f^*(x)^2 - 2f(x)f^*(x) + f(x)^2 - f^*(x)^2 \Big\} dP(x) \\
&= \int_{\mathcal{X}} \big(f(x) - f^*(x)\big)^2 dP(x) = ||f - f^*||^2_{L_2(\mathbb{P})}
\end{aligned}$$

To Sum Up

In a word, empirical risk minimization (ERM) is a method to approximate Bayes predictor f^*, knowing the training samples $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and the loss l, by minimizing the risk or excess risk $\mathcal{R}(f) - \mathcal{R}^*$.

Local Averaging Methods

Local averaging methods provide a different approach by minimizing the conditional expected risk $\mathbb{E}[l(Y, f(X)) \mid X]$ pointwisely, which leads to the Bayes predictor $f^*(X)$.

Proposition (Bayes Predictor and Bayes Risk)

The conditional expected risk is minimized at a Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ satisfying for all $x \in \mathcal{X}$,

$$f^*(X) \in \underset{f(X)\in\mathcal{Y}}{\arg\min}\ \mathbb{E}\big[l(Y, f(X)) \mid X\big]$$

The Bayes risk $\mathcal{R}^*$ is the risk of all Bayes predictors and is equal to

$$\mathcal{R}^* = \mathbb{E}_X\Big[\inf_{f(X)\in\mathcal{Y}} \mathbb{E}\big[l(Y, f(X)) \mid X\big]\Big]$$

Comparison of ERM and LAM

However, the conditional probability $\mathbb{P}(Y \mid X)$ is generally unknown. To overcome this obstacles, this time we approximate the $\mathbb{P}(Y \mid X)$ by some estimator $\hat{P}(Y \mid X)$, and the optimal predictor could be obtained by

$$\hat{f}(X) = \underset{f(X)\in\mathcal{Y}}{\arg\min}\ \hat{\mathbb{E}}\big[l(Y, f(X)) \mid X\big] = \underset{f(X)\in\mathcal{Y}}{\arg\min} \int_{\mathcal{Y}} l(y, f(X)) d\hat{P}(Y = y \mid X)$$

which are often called "plug-in" estimators. Recall what we done in ERM

$$\hat{f} = \underset{f\in\mathcal{F}}{\arg\min}\, \hat{\mathcal{R}}(f) = \underset{f\in\mathcal{F}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n} l(Y_i, f(X_i))$$

Running Examples I

In the usual cases, local averaging methods leads to the following prediction functions:

- Classification with 0-1 loss:

$$f^*(X) \in \underset{f(X)\in\mathcal{Y}}{\arg\min}\ \hat{\mathbb{E}}\big[\mathbb{1}(Y \neq f(X)) \mid X\big] = \underset{f(X)\in\mathcal{Y}}{\arg\min} \sum_{y=1}^{k} \mathbb{1}(y \neq f(X)) \cdot \hat{P}(Y = y \mid X)$$

  which is equivalent to

$$\hat{f}(X) \in \underset{y\in\{1,\ldots,k\}}{\arg\max}\ \hat{P}(Y = y \mid X)$$

Running Examples II

- Regression with square loss:

$$\min_{f(X)\in\mathcal{Y}} \int_{\mathcal{Y}} (y - f(X))^2 \, d\hat{P}(Y = y \mid X)$$

The first-order optimal condition yields

$$\int_{\mathcal{Y}} (2f(X) - 2y) \, d\hat{P}(Y = y \mid X) = 0$$

which implies

$$\hat{f}(X) = \int_{\mathcal{Y}} y \, d\hat{P}(Y = y \mid X) = \hat{\mathbb{E}}[Y \mid X]$$

In this way, we don't need to claim a hypothesis on the form of function f, but the tradeoff is we have to estimate the conditional distribution $P(Y \mid X)$ as well as the marginal distribution $P(X)$.

A Glance at Convergence Rate

As you shall see later, all of the methods we're going to introduce in this section can provably learn complex non-linear functions f with a convergence rate of the form

$$\mathcal{O}(n^{-2/(d+2)})$$

where d is the underlying dimension, leading to the curse of dimensionality.

Linear Estimators

In this section, we will consider "linear" estimators (which is linear in observations), where the conditional distribution is of the form

$$\hat{P}(Y = y \mid X) = \sum_{i=1}^{n} \hat{w}_i(X) \cdot \mathbb{1}(Y_i = y)$$

with its derviative

$$d\hat{P}(Y = y \mid X) = \sum_{i=1}^{n} \hat{w}_i(X) \cdot \delta_{Y_i}(y)dy$$

where δ_{Y_i} is the Dirac probability distribution at Y_i, and the weight function $\hat{w}_i : \mathcal{X} \mapsto \mathbb{R}$, $i = 1, \ldots, n$ depends on the input data only (for simplicity) and satisfy for all $i \in \{1, \ldots, n\}$ and $X \in \mathcal{X}$

$$\hat{w}_i(X) \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \hat{w}_i(X) = 1 \quad \text{almost surely in } X$$

These conditions ensure that for all $x \in \mathcal{X}$, $\hat{P}(Y \mid X)$ is a probability distribution.

Running Examples

For our running examples, we have:

- Binary Classification with category labels:

$$\hat{f}(X) \in \underset{y \in \{1,\ldots,k\}}{\arg\max}\ \hat{P}(Y = y \mid X) = \underset{y \in \{1,\ldots,k\}}{\arg\max} \sum_{i=1}^{n} \hat{w}_i(X) \cdot \mathbb{1}(Y_i = y)$$

  that is, each observation (X_i, Y_i) votes for its label with weight $\hat{w}_i(X)$.

- Regression on $\mathcal{Y} = \mathbb{R}$:

$$\hat{f}(X) = \int_{\mathcal{Y}} y \, d\hat{P}(Y = y \mid X) = \sum_{i=1}^{n} \hat{w}_i(X) \int_{\mathcal{Y}} y \cdot \delta_{Y_i}(y) dy = \sum_{i=1}^{n} \hat{w}_i(X) Y_i \tag{3}$$

  This is why the terminology "linear estimators" is sometimes used, since as function of the response vector in $\mathbb{R}^n$, the estimator is linear.

Weight Functions

Weight Functions. In most cases, for any i, the weight function $\hat{w}_i(X)$ is closed to 1 for training points X_i which are close to X (measure the similarity with X_i). We next show three classical ways of building them: (1) partition estimators, (2) Nearest-neighbors, and (3) Nadaraya-Waatson estimator (a.k.a. kernel regression).

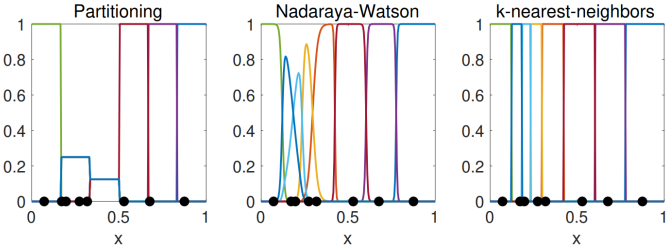


Figure: Weights of linear estimators in $d = 1$ estimation for three types of local averaging estimators. The $n = 8$ weight functions $\hat{w}_i(x)$ are plotted with the observations in black.

Partition Estimators I

If $\mathcal{X} = \bigcup_{j \in J} A_j$ is a partition (such that for all $j, j' \in J, A_j \cap A_{j'} = \emptyset$) of $\mathcal{X}$ with a countable index set J (which we assume finite for simplicity), then we can consider for any $X \in \mathcal{X}$ the corresponding element $A(X)$ of the partition (namely, $A(X)$ is the unique A_j such that $X \in A_j$), and define

$$\hat{w}_i(X) = \frac{\mathbb{1}_{X_i \in A(X)}}{\sum_{j=1}^{n} \mathbb{1}_{X_j \in A(X)}} \tag{4}$$

with the convention that if no training data points $\{X_i\}$ lies in $A(X)$, then $\hat{w}_i(X)$ is equal to $1/n$ for each $i \in \{1, \ldots, n\}$. Here we illustrate two standard applications of partition estimators:

- Fixed partitions:
  for example, when $\mathcal{X} = [0,1]^d$, then we choose the bandwidth h, with $|J| = h^{-d}$ (if $h = 1/5$ and $d = 2$, we have $|J| = 25$). Note here that the computation time for ech $X \in \mathcal{X}$ is not necessarily proportional to $|J|$, but to n (by considering the bins where the data lie). This estimator is some times called a "regressogram".

Partition Estimators II

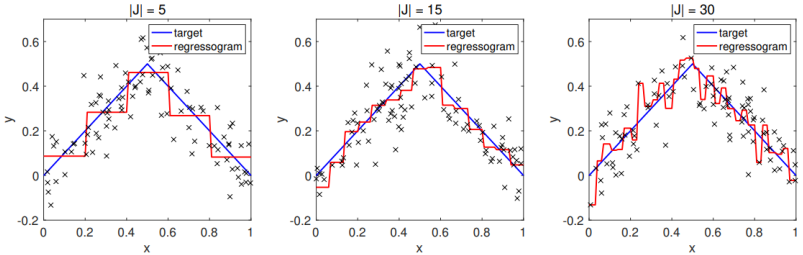


Figure: Regressograms in $d = 1$ dimension, with three different values of $|J|$. We can see both underfitting, or overfitting in this example. Note that the target function f^* is piecewise affine, and that on the affine parts, the estimator is far from linear, namely, the estimator cannot take advantage of extra-regularity.

Partition Estimators III

- Decision trees:
  for data in a hypercube, we can recursively partition it by selecting a variable to split leading to a maximum reduction in errors when defining the partitioning estimate. Note that now the partition depends on the labels (so the analysis below does not apply, unless the partitioning is learned on a different data than the one sued for the estimation).

Equivalence with Linear-squares Regression I

Consider the case of regression where we use the linear estimator $\hat{f}(X) = \sum_{i=1}^{n} \hat{w}_i(X) Y_i$ with partition weights. This can be seen as a least-square estimator with feature vector

$$\varphi(X) = (\mathbb{1}_{X \in A_1}, \ldots, \mathbb{1}_{X \in A_J})^\top \in \mathbb{R}^J$$

in ERM approach. Indeed, from trainning data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we need to find the weight vector $\hat{\theta}$ through the normal equations

$$\sum_{i=1}^{n} \varphi(X_i)\varphi(X_i)^\top \theta = \sum_{i=1}^{n} Y_i \varphi(X_i)$$

It turns out that the matrix $\hat{\Sigma} = \sum_{i=1}^{n} \varphi(X_i)\varphi(X_i)^\top$ is diagonal with the jth component equals to n_{A_j}, the number of data points lying in cell A_j. This implies that for a non-empty cell A_j, θ_j is the average of all Y_i's for X_i lying in A_j, namely,

$$\theta_j = \frac{1}{n_{A_j}} \sum_{i=1}^{n} Y_i \cdot \mathbb{1}_{X_i \in A_j}$$

Equivalence with Linear-squares Regression II

Thus, for all $X \in A_j$, the prediction is exactly θ_j,

$$\hat{f}(X) = \sum_{i=1}^{n} \hat{w}_i(X) Y_i = \sum_{i=1}^{n} \frac{\mathbb{1}_{X_i \in A_j}}{n_{A_j}} Y_i$$

For empty cells, θ_j is not determined. Among the many OLS estimators, we select the one for which the variance of the vector θ is smallest, that is $\sum_{j \in J} (\theta_j - \frac{1}{|J|} \sum_{j' \in J} \theta'_j)^2$ is smallest. A short calculation shows that this exactly leads to

$$\theta_j = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

for these empty cells, which correspond to our chosen convention.

Nearest-Neighbors

Given an integer $k \geq 1$, and a distance d on $\mathcal{X}$, for any $X \in \mathcal{X}$, we can order the n samples so that

$$d(X_{i_1(X)}, X) \leq d(X_{i_2(X)}, X) \leq \cdots \leq d(X_{i_n(X)}, X)$$

where $\{i_1(X), \ldots, i_n(X)\} = \{1, \ldots, n\}$, and ties are broken randomly. We then define

$$\hat{w}_i(X) = \frac{1}{k}, \quad \text{if } i \in \{i_1(X), \ldots, i_k(X)\}$$

and $\hat{w}_i(X) = 0$ otherwise. Given a new input $X \in \mathcal{X}$, the nearest-neighbors predictor looks at the k nearest points X_i in the data set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ and predicts a majority vote among them for classification or simply the averaged response for regression. The number of nearest neighbors k is a hyperparameter which needs to be estimated (typically by cross-validation).

Algorithms for Nearest-Neighbors

Given a test point $X \in \mathcal{X}$, the naive algorithm looks at all training data points for computing the predicted response, thus the complexity is $O(nd)$ per test point in $\mathbb{R}^d$. When n is large, this is costly in time and memory. There exists indexing techniques for (potentially approximate) nearest-neighbor search, such as "k-dimensional-trees", with typically a logarithmic complexity in n (but with some additional compiling time).

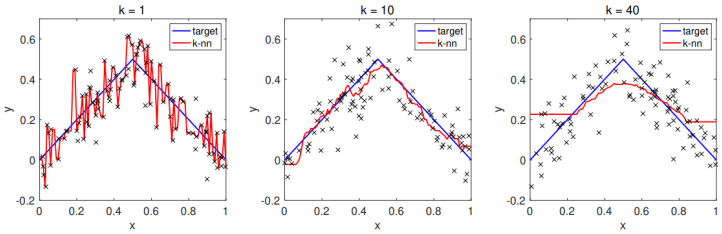


Figure: k-nearest neighbor regression in $d = 1$ dimension, with three values of k (the number of neighbors). We can see both underfitting (k too large), and overfitting (k too small).

Nadaraya-Watson Estimator (Kernel Regression)

Given a "kernel" function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}_+$, which is pointwise non-negative, we define

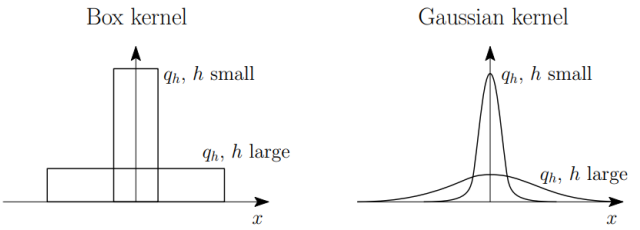
$$\hat{w}_i(X) = \frac{K(X_i, X)}{\sum_{j=1}^n K(X_j, X)}$$

with the convention that if $K(X_j, X) = 0$ for all $j \in \{1, \ldots, n\}$, then $\hat{w}_i(X)$ is equal to $1/n$ for each i. In most case where $\mathcal{X} \subset \mathbb{R}^d$, we take

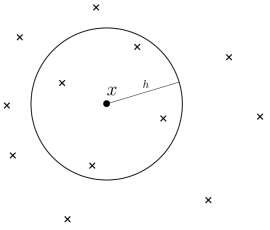
$$K(X, X') = \frac{q((X - X')/h)}{h^d}$$

for a certain function $q : \mathbb{R}^d \mapsto \mathbb{R}_+$ that has large values around 0, and $h > 0$ a bandwidth parameter to be selected. If we assume that q is integrable with integral equal to one, then $K(\cdot, X')$ is a probability density with mass around X', which gets more concentrated as h goes to zero.

Two Typical Examples



(a) Two typical kernel windows



(b) Box kernel in $d = 2$ dimensions

- Box kernel: $q(X) = \mathbb{1}_{||X||_2 \leq 1}$. See above for an illustration in $d = 2$ dimension
- Gaussian kernel: $q(X) = e^{-||X||^2/2}$, where we use the fact it is non-negative pointwise (as opposed to positive definitiness).

Algorithms

In terms of algorithms, with a naive algorithm, for every test point, all the input data have to be considered, that is, a complexity proportional to n. The same techniques used for efficient k-nearest-neighbor search (e.g. k-d-tress) can be applied here as well.

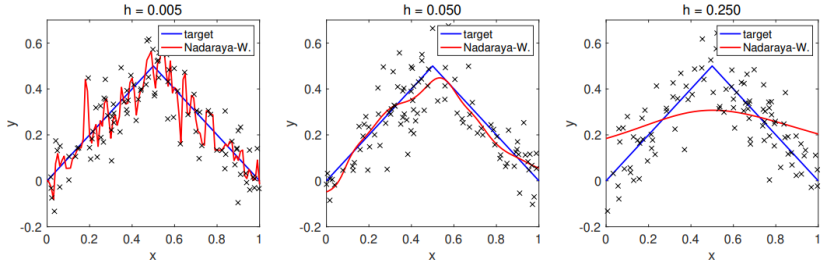


Figure: Nadaraya-Watson regression in $d = 1$ dimension, with three values of bandwidth h for the Gaussian kernel.

Generic Assumptions

We consider for simplicity the **regression** case. For classification, calibration techniques such as those used in convexification can be used (with then a square root calibration function on top of the least-squares excess risk), but better rate can be obtained directly (please refer the textbook).

Except for the requirement of square-integrable for target function $f(x)$, we make extra generic assumptions here:

1. **Bounded noise**: there exists $\sigma \geq 0$ such that $|Y - \mathbb{E}[Y \mid X]|^2 \leq \sigma^2$ almost surely.
2. **Regular target function**: the target function $f^*(x) = \mathbb{E}[Y \mid X = x]$ is B-Lipschitz-continuous with respect to a distance d.

The Difference of Target Functions I

Recall the target function $f^*(x) = \mathbb{E}[Y \mid X = x]$ and the predictor $\hat{f}(x) = \sum_{i=1}^{n} \hat{w}_i(x) \cdot Y_i$ in Eq.(3) at a test point $x \in \mathcal{X}$. Using that the summation of weights $w_i(x)$ is one, we have:

$$
\begin{aligned}
\hat{f}(X) - f^*(X) &= \sum_{i=1}^{n} \hat{w}_i(x) \cdot Y_i - \mathbb{E}[Y \mid X = x] \\
&= \sum_{i=1}^{n} \hat{w}_i(x) \cdot (Y_i - \mathbb{E}[Y_i \mid X_i]) + \sum_{i=1}^{n} \hat{w}_i(x) \cdot (\mathbb{E}[Y_i \mid X_i] - \mathbb{E}[Y \mid X = x]) \\
&= \underbrace{\sum_{i=1}^{n} \hat{w}_i(x) \cdot \big(Y_i - \mathbb{E}[Y_i \mid X_i]\big)}_{\text{part I}} + \underbrace{\sum_{i=1}^{n} \hat{w}_i(x) \cdot (f^*(X_i) - f^*(x))}_{\text{part II}}
\end{aligned}
$$

The Difference of Target Functions II

Conditioning on $X_1, \ldots, X_n$ and because we have assumed the weight functions do not depend on the labels $\{Y_i\}$, the first term (part I) has zero expectation (with respect to sample S_n)

$$\begin{aligned}\mathbb{E}_{S_n}\Big[\text{part I} \mid X_1, \ldots, X_n\Big] &= \hat{w}_i(x) \cdot \mathbb{E}\big[Y_i - \mathbb{E}[Y \mid X_i] \mid X_i\big] \\ &= \hat{w}_i(x) \cdot \big(\mathbb{E}[Y_i \mid X_i] - \mathbb{E}[Y_i \mid X_i]\big) = 0\end{aligned}$$

while the second term (part II) is deterministic and the variance is therefore zero

$$\operatorname*{Var}_{S_n}\Big(\text{part II} \mid X_1, \ldots, X_n\Big) = 0$$

Empirical Excess Risk I

We thus have, using the independencies of samples $(X_i, Y_i).i = 1, \ldots, n$ and weights $w_i(x)$ sum to one:

$$
\begin{aligned}
&\mathbb{E}_{S_n}\left[\left(\hat{f}(x) - f^*(x)\right)^2 \mid X_1, \ldots, X_n\right] \\
&= \left(\mathbb{E}_{S_n}\left[\hat{f}(x) - f^*(x) \mid X_1, \ldots, X_n\right]\right)^2 + \operatorname*{Var}_{S_n}\left(\hat{f}(x) - f^*(x) \mid X_1, \ldots, X_n\right) \\
&= \left(\mathbb{E}_{S_n}\left[\text{part II} \mid X_1, \ldots, X_n\right]\right)^2 + \operatorname*{Var}_{S_n}\left(\text{part I} \mid X_1, \ldots, X_n\right) \\
&= \underbrace{\left(\sum_{i=1}^{n} \hat{w}_i(x) \cdot \left(f^*(X_i) - f^*(x)\right)\right)^2}_{\text{bias}} + \underbrace{\sum_{i=1}^{n} \hat{w}_i(x)^2 \cdot \mathbb{E}_{S_n}\left[\left(Y_i - \mathbb{E}[Y_i \mid X_i]\right)^2 \mid X_i\right]}_{\text{variance}}
\end{aligned}
$$

Empirical Excess Risk II

with a "bias" term which is zero if f^* is constant over $\mathcal{X}$, and a "variance" term which is zero, when Y is a deterministic function of X. We can further bound these as

$$\begin{aligned}
\mathbb{E}_{S_n}\left[\left(\hat{f}(x) - f^*(x)\right)^2 \mid X_1, \ldots, X_n\right] &\leq \left(\sum_{i=1}^n \hat{w}_i(x) \cdot \left|f^*(X_i) - f^*(x)\right|\right)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(X)^2 \quad \text{(Assumption 1)} \\
&\leq \left(\sum_{i=1}^n \hat{w}_i(x) \cdot B \cdot d(X_i, x)\right)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2 \quad \text{(Assumption 2)} \\
&\leq B^2 \sum_{i=1}^n \hat{w}_i(x) \cdot d(X_i, x)^2 + \sigma^2 \sum_{i=1}^n \hat{w}_i(x)^2 \quad \text{(Jensen's inequality)}
\end{aligned} \tag{5}$$

Expected Excess Risk

According to Eq.(2), the expected excess risk for regression case is:

$$\begin{aligned}\mathcal{R}(\hat{f}) - \mathcal{R}^* &= \mathbb{E}\left[l(Y,\hat{f}(X)) - l(Y,f^*(X))\right] = \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\left(\hat{f}(X) - f^*(X)\right)^2 \mid X = x\right] dP(x) \\ &\leq \underbrace{B^2 \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^n \hat{w}_i(x) d(X_i,x)^2\right] dP(x)}_{\text{bias term}} + \underbrace{\sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^n \hat{w}_i(x)^2\right] dP(x)}_{\text{variance term}} \end{aligned} \tag{6}$$

- A bias term $B^2 \int_{\mathcal{X}} \mathbb{E}\left[\sum_{i=1}^n \hat{w}_i(x) d(X_i,x)^2\right] dP(x)$, which depends on the regularity of the target function.
- A variance term $\sigma^2 \sum_{i=1}^n \int_{\mathcal{X}} \mathbb{E}\left[\hat{w}_i(x)^2\right] dP(x)$, that depends on the noise on top of the optimal predictions. Since the weights sum to one, we can write

$$\sum_{i=1}^n \mathbb{E}[\hat{w}_i(x)^2] = \sum_{i=1}^n \mathbb{E}[(\hat{w}_i(x) - 1/n)^2] + 2/n - 1/n^2$$

Hence, up to vanishing constant, the variance term measures the deviation ot the uniform weights.

To Wrap Up

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* \leq \underbrace{B^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x)}_{\text{bias term}} + \underbrace{\sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x)^2 \right] dP(x)}_{\text{variance term}}$$

Both variance and bias have to go to zero when n grows, and this corresponds to two simple quantities on the weights.

- For the variance, the worst case scenario is that $\hat{w}_i(x)^2 \approx \hat{w}_i(x)$, that is, weights are putting all the mass in to a single label (different for different testing sample), thus leading to overfitting.
- For the bias, the worst case scenario is that weights are uniform (the distance $d(X_i, X)$ are different, may lead to underfitting).

Generic Analysis for Fixed Partition

Proposition (Convergence rate for partition estimates)

Assume bounded noise (A1) and a Lipschitz-continuous target function (A2), and a partition $\mathcal{X} = \bigcup_{j \in J} A_j$; then for the partitioning estimate $\hat{f}$, we have

$$\int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\left(\hat{f}(x) - f^*(x) \right)^2 \right] dP(x) \leq \left(8\sigma^2 + \frac{B^2}{2} \mathit{diam}(\mathcal{X})^2 \right) \frac{|J|}{n} + B^2 \max_{j \in J} \mathit{diam}(A_j)^2 \tag{7}$$

where the diameter of set $\mathcal{X}$ is defined as $\mathit{diam}(\mathcal{X}) = \sup\{d(x, x') \mid x, x' \in \mathcal{X}\}$.

Before we look at the proof, there is a tradeoff between bias and variance, and we need to balance the terms (up to constants) $\max_{j \in J} \mathrm{diam}(A_j)^2$ and $|J|/n$.

A Simple Example of Bound

Consider the case of unit-cube $[0,1]^d$, with $|J| = h^{-d}$ cubes of length h, we have $|J|/n \approx 1/(nh^d)$ and $\max_{j \in J} \text{diam}(A_j)^2 \approx h^2$, which are equal when $h \approx n^{-1/(2+d)}$, leads to a rate proportional to $n^{-2/(2+d)}$.

- While optimal, this is a very slow rate, and a typical example of the curse of dimensionality. For this rate to be small, n has to be exponentially large in dimension. This is unavoidable with so little regularity (only bounded first-order derivatives).
- In Kernel Methods, we show how to leverage smoothness to get significantly improved bounds. In Sparse Methods, we will leverage dependence on a small number of variables.

Proof of Convergence Rate for Partition Estimates I

We consider an element A_j of the partition with at least one observation in it (a non-empty cell). Then for the test point $x \in A_j$ and i among the indices of training points lying in A_j, we have

$$\hat{w}_i(x) = 1/n_{A_j}$$

where $n_{A_j} \in \{1, \ldots, n\}$ is the number of data points lying in A_j.

- **Variance**. From Eq.(6), the variance term is bounded from above by σ^2 times

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

Proof of Convergence Rate for Partition Estimates II

If A_j contains no input observations, then all weights are equal to $1/n$ and this sum is equal to $n \times (1/n^2) = 1/n$ for all $X \in A_j$. Thus we get

$$\begin{aligned}\int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^{n} \hat{w}_i(x)^2\right] dP(x) &= \int_{\mathcal{X}} \sum_{j \in J} \mathbb{1}(x \in A_j) \cdot \mathbb{E}_{S_n}\left[\frac{1}{n_{A_j}} \mathbb{1}(n_{A_j} > 0) + \frac{1}{n} \mathbb{1}(n_{A_j} = 0)\right] dP(x) \\ &= \sum_{j \in J} \mathbb{P}(A_j) \cdot \mathbb{E}_{S_n}\left[\frac{1}{n_{A_j}} \mathbb{1}(n_{A_j} > 0) + \frac{1}{n} \mathbb{1}(n_{A_j} = 0)\right]\end{aligned} \tag{8}$$

Intuitively, by the law of large numbers, n_{A_j}/n tends to $\mathbb{P}(A_j)$, so the variance term is expected to be of the order $\sigma^2 \sum_{j \in J} \mathbb{P}(A_j) \frac{1}{n\mathbb{P}(A_j)} = \sigma^2 \frac{|J|}{n}$, which is to be expected as this is essentially equivalent to the least-squares regression with features $(\mathbb{1}(X \in A_j))_{j \in J}$.

Proof of Convergence Rate for Partition Estimates III

More formally, we have $\mathbb{P}(n_{A_j} = 0) = (1 - \mathbb{P}(A_j))^n$. Using Bernstein's inequality for the random variable $\mathbb{1}(X \in A_j)$, which have mean and variance upper-bounded by $\mathbb{P}(A_j)$ (namely, $\mu, \sigma^2 \leq \mathbb{P}(A_j)$), we have

$$\begin{aligned}
\mathbb{P}\left(\frac{n_{A_j}}{n} \leq \frac{\mathbb{P}(A_j)}{2}\right) &= \mathbb{P}\left(\frac{n_{A_j}}{n} - \mathbb{P}(A_j) \leq -\frac{\mathbb{P}(A_j)}{2}\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \in A_j) - \mathbb{E}[\mathbb{1}(X \in A_j)]\right| \geq \frac{\mathbb{P}(A_j)}{2}\right) \\
&\leq \exp\left(-\frac{n\mathbb{P}(A_j)^2/4}{2\mathbb{P}(A_j) + 2(\mathbb{P}(A_j)/2)/3)}\right) \\
&\leq \exp\left(-\frac{n\mathbb{P}(A_j)}{10}\right) \leq \frac{5}{n\mathbb{P}(A_j)}
\end{aligned}$$

Proof of Convergence Rate for Partition Estimates IV

where the last inequality holds by knowing the fact that $e^{-x} \leq 1/(2x)$ when $x > 0$. Such result leads to the bound

$$\begin{aligned}
&\sum_{j \in J} \mathbb{P}(A_j) \cdot \mathbb{E}_{S_n}\left[\frac{1}{n_{A_j}} \mathbb{1}(n_{A_j} > 0) + \frac{1}{n} \mathbb{1}(n_{A_j} = 0)\right] \\
&= \sum_{j \in J} \mathbb{P}(A_j) \cdot \mathbb{E}_{S_n}\left[\frac{1}{n_{A_j}} \mathbb{1}(n_{A_j} > 0) \cdot \left(\mathbb{1}\left(n_{A_j} \leq \frac{n\mathbb{P}(A_j)}{2}\right) + \left(n_{A_j} > \frac{n\mathbb{P}(A_j)}{2}\right)\right) + \frac{1}{n} \mathbb{1}(n_{A_j} = 0)\right] \\
&\leq \sum_{j \in J} \mathbb{P}(A_j) \cdot \left[\mathbb{P}\left(\frac{n_{A_j}}{n} \leq \frac{\mathbb{P}(A_j)}{2}\right) + \frac{2}{n\mathbb{P}(A_j)} + \frac{1}{n}\mathbb{P}(n_{A_j} = 0)\right] \\
&\leq \sum_{j \in J} \mathbb{P}(A_j) \cdot \left[\frac{5}{n\mathbb{P}(A_j)} + \frac{2}{n\mathbb{P}(A_j)} + \frac{1}{n\mathbb{P}(A_j)}\right] \leq \frac{8|J|}{n}
\end{aligned}$$

Proof of Convergence Rate for Partition Estimates V

- **Bias**. We have, for $x \in A_j$ and a non-empty cell A_j,

$$\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \leq \mathsf{diam}(A_j)^2$$

and $\sum_{i=1}^{n} \hat{w}_i d(X_i, x)^2 = \frac{1}{n} \sum_{i=1}^{n} d(X_i, x)^2 \leq \mathsf{diam}(\mathcal{X})^2$ from empty-cells. Thus, separating the cases $n_{A_j} = 0$ and $n_{A_j} > 0$:

$$\begin{aligned}
\int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x) &= \int_{\mathcal{X}} \sum_{j \in J} \mathbb{1}(x \in A_j) \cdot \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x) \\
&\leq \int_{\mathcal{X}} \sum_{j \in J} \mathbb{1}(x \in A_j) \cdot \mathbb{E}_{S_n} \Big[\mathsf{diam}(A_j)^2 \cdot \mathbb{1}(n_{A_j} > 0) + \mathsf{diam}(\mathcal{X})^2 \cdot \mathbb{1}(n_{A_j} = 0) \Big] dP(x) \\
&= \sum_{j=1}^{J} \mathbb{P}(A_j) \cdot \Big[\mathsf{diam}(A_j)^2 \cdot \mathbb{P}(n_{A_j} > 0) + \mathsf{diam}(\mathcal{X})^2 \cdot \mathbb{P}(n_{A_j} = 0) \Big]
\end{aligned}$$

Proof of Convergence Rate for Partition Estimates VI

Notice that $(1-x)^n \leq 1/(2nx)$ for $x \in (0,1]$, we have

$$\begin{aligned}
\int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i,x)^2\right] dP(x) &= \sum_{j=1}^{J} \mathbb{P}(A_j) \cdot \left[\mathsf{diam}(A_j)^2 \cdot \mathbb{P}(n_{A_j} > 0) + \mathsf{diam}(\mathcal{X})^2 \cdot \mathbb{P}(n_{A_j} = 0)\right] \\
&\leq \sum_{j=1}^{J} \mathbb{P}(A_j) \cdot \left[\mathsf{diam}(A_j)^2 + \mathsf{diam}(\mathcal{X})^2 \cdot \left(1 - \mathbb{P}(A_j)\right)^n\right] \\
&\leq \sum_{j=1}^{J} \mathbb{P}(A_j) \cdot \left[\mathsf{diam}(A_j)^2 + \mathsf{diam}(\mathcal{X})^2 \cdot \frac{1}{2n\mathbb{P}(A_j)}\right] \qquad (1) \\
&= \sum_{j \in J} \mathbb{P}(A_j) \cdot \mathsf{diam}(A_j)^2 + \frac{|J|}{2n} \cdot \mathsf{diam}(\mathcal{X})^2 \\
&\leq \max_{j \in J} \mathsf{diam}(A_j)^2 + \frac{|J|}{2n} \cdot \mathsf{diam}(\mathcal{X})^2
\end{aligned}$$

Proof of Convergence Rate for Partition Estimates VII

Combining the upper-bounds of variance and bias term together leads to the desire results.

$$\int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\left(\hat{f}(x) - f^*(x)\right)^2\right] dP(x) \leq \left(8\sigma^2 + \frac{B^2}{2}\mathsf{diam}(\mathcal{X})^2\right)\frac{|J|}{n} + B^2 \max_{j\in J}\mathsf{diam}(A_j)^2$$

K-nearest Neighbors

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* \leq \underbrace{B^2 \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2\right] dP(x)}_{\text{bias term}} + \underbrace{\sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[\sum_{i=1}^{n} \hat{w}_i(x)^2\right] dP(x)}_{\text{variance term}}$$

$$\hat{w}_i(x) = \frac{1}{k}, \quad \text{if } i \in \{i_1(x), \ldots, i_k(x)\}$$

In this case, we immediately have $\sum_{i=1}^{n} \hat{w}_i(x)^2 = 1/k$, so

- the variance term will go down as soon as k tends to infinity;
- for the bias term, the needed term $\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2$ is equal to the average squared distances between test point X and its k-nearest neighbors with training samples $\{X_1, \ldots, X_n\}$, and this is less than the expected distance to the k-nearest neighbor

Next we will introduce two lemmas give an estimate for the l_∞-distance, and thus for all distances by equivalence of norms on $\mathbb{R}^d$.

Distance to Nearest Neighbor I

Lemma (Distance to nearest neighbor)

Consider a probability distribution with compact support in $\mathcal{X} \subset \mathbb{R}^d$. Consider $n+1$ points $X_1, \ldots, X_n, X_{n+1}$ sampled i.i.d. from $\mathcal{X}$. Then the ***expected*** *squared l_∞-distance between X_{n+1} and its first-nearest-neighbor is less than*

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

for $d \geq 2$, and less than $\frac{2}{n} diam(\mathcal{X})^2$ for $d = 1$.

By symmetry, we aim at computing the value $\frac{1}{n+1}\sum_{i=1}^{n+1}\mathbb{E}[||X_i - X_{(i)}||_\infty^2]$, where $X_{(i)}$ is a nearest neighbor of X_i among the other n points. Denote $R_i = ||X_i - X_{(i)}||_\infty$, then the sets
$B_i = \{X \in \mathbb{R}^d \mid ||X - X_i||_\infty < R_i/2\}$ are disjoint.

Distance to Nearest Neighbor II

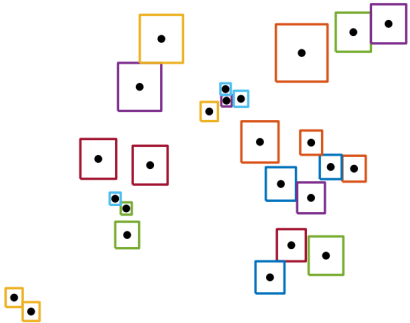


Figure: Distance to nearest neighbors

Distance to Nearest Neighbor III

Moreover, their union has diameter less than $\text{diam}(\mathcal{X}) + \text{diam}(\mathcal{X}) = 2\text{diam}(\mathcal{X})$. Thus by comparing volumes, we have: $\sum_{i=1}^{n+1} R_i^d \leq (2\text{diam}(\mathcal{X}))^d$. Therefore, by Jensen's inequality, for $d \geq 2$,

$$\left(\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2\right)^{d/2} \leq \frac{1}{n+1}\sum_{i=1}^{n+1}(R_i)^d \leq \frac{2^d \text{diam}(\mathcal{X})^d}{n+1}$$

leading to the desired result. For $d = 1$, we simply have

$$\left(\frac{1}{n+1}\sum_{i=1}^{n+1} R_i^2\right) \leq \text{diam}(\mathcal{X})\left(\frac{1}{n+1}\sum_{i=1}^{n} R_i\right) \leq \frac{2}{n+1}\text{diam}(\mathcal{X})^2$$

Distance to k-nearest Neighbors

Lemma (Distance to k-nearest neighbor)

Let $k \geq 1$. Consider a probability distribution with compact support in $\mathcal{X} \subset \mathbb{R}^d$. Consider $n+1$ points $X_1, \ldots, X_n, X_{n+1}$ sampled i.i.d. from $\mathcal{X}$. Then the expected squared l_∞-distance between X_{n+1} and its k-nearest neighbor is less than

$$8 diam(\mathcal{X})^2 \left(\frac{2k}{n}\right)^{2/d}$$

for $d \geq 2$, and less than $diam(\mathcal{X})^2 \frac{2k}{n}$ for $d = 1$.

Convergence Rate for k-nearest Neighbors

Proposition (Convergence rate for k-nearest neighbors)

Assume bounded noise (A1) and a B-Lipchitz-continuous target function $(A2)$. Then for the k-nearest neighbor estimate $\hat{f}$ with the l_∞-norm, we have, for $d \geq 2$

$$\int_{\mathcal{X}} \mathbb{E}_{S_n} \left[(\hat{f}(x) - f^*(x))^2 \right] dP(x) \leq \frac{\sigma^2}{k} + 8B^2 \mathit{diam}(\mathcal{X})^2 \left(\frac{2k}{n} \right)^{2/d} \tag{9}$$

Kernel Regression

In this section, we assume that $\mathcal{X} = \mathbb{R}^d$, and for simplicity, we assume that marginal distribution $P(x)$ has a density $p(x)$ with respect to the Lebesgue measure. We also assume that

$$K(X, X') = q_h(X - X') = \frac{q((X - X')/h)}{h^d}$$

for a probability desntiy $q : \mathbb{R}^d \mapsto \mathbb{R}_+$. The function q_h is also a density, which is the density of hZ when random variable Z has density $q(Z)$ (it is thus gets more concentrated around 0 as h tends to zero). With these notations, the weights can be written as

$$\hat{w}_i(x) = \frac{K(X_i, x)}{\sum_{j=1}^n K(X_j, x)} = \frac{q_h(x - X_i)}{\sum_{j=1}^n q_h(x - X_j)}$$

Smoothing by Convolution

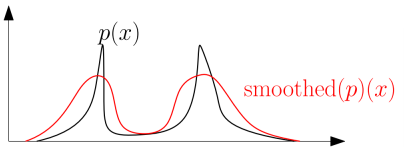
When performing kernel smoothing, quantities like $\frac{1}{n}\sum_{i=1}^{n} q_h(x - X_i)g(X_i)$ naturally appear. When the number n of observations goes to infinity, by the law of large numbers, it tends to almost surely to the mean

$$\int_{\mathbb{R}^d} q_h(x - z)g(z)p(z)dz$$

which is exactly the convolution between the function q_h and the function $x \mapsto p(x)g(x)$.

- The function q_h is a probability density that is putting all most its weights at range of values which are of order h, e.g., for kernels like the Gaussian kernel or the box kernel.
- Thus convolution will smooth the function pg by averaging values which are at range h.

Thus, when h goes to zero, it converges to the function pg itself. See an example below for $g = 1$.



Variance Term I

Recall that

$$\mathcal{R}(\hat{f}) - \mathcal{R}^* \leq \underbrace{B^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x)}_{\text{bias term}} + \underbrace{\sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x)^2 \right] dP(x)}_{\text{variance term}}$$

We have, for a fixed $X \in \mathcal{X}$:

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

Using the law of large numbers and the smoothing reasoning above, this summation $n \sum_{i=1}^{n} \hat{w}_i(x)^2$ is converging almost surely to

$$\frac{\int_{\mathbb{R}^d} q_h(x - z)^2 p(z) dz}{\left(\int_{\mathbb{R}^d} q_h(x - z) p(z) dz \right)^2} = \frac{q_h^2 * p(x)}{(q_h * p(x))^2}$$

Variance Term II

When h tends to zero, then the denominator above $(q_h * p(x))^2$ tends to $p(x)^2$ because the bandwidth of the smoothing goes to zero ($q_h \to \delta(x)$). The numerator above corresponds to the smoothing of p by the density $x \mapsto \frac{q_h(x)^2}{\int_{\mathbb{R}^d} q_h(u)^2 du}$ (where $\int_{\mathbb{R}^d} q_h(u)^2 du$ is just a normalize constant), and is thus asymptotically equivalent to

$$q_h^2 * p(x) \to p(x) \int_{\mathbb{R}^d} q_h(u)^2 du = p(x) h^{-d} \int_{\mathbb{R}^d} q(u)^2 du$$

where $q_h(X - X') = \frac{q((X-X')/h)}{h^d}$. Overall, when n tends to infinity, and h tends to zero, we get:

$$\sum_{i=1}^n \hat{w}_i(x)^2 \sim \frac{1}{nh^d} \frac{1}{p(x)} \int_{\mathbb{R}^d} q(u)^2 du$$

and thus

$$\int_{\mathcal{X}} \left[\sum_{i=1}^n \hat{w}_i(x)^2 \right] p(x) dx \sim \frac{1}{nh^d} \text{vol}(\text{supp}(P)) \int_{\mathcal{R}^d} q(u)^2 du$$

Bias Term

With the same intuitive reasoning, we get, when n tends to infinity:

$$\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \to \frac{\int_{\mathcal{R}^d} q_h(x-z)||x-z||_2^2 p(z) dz}{\int_{\mathbb{R}^d} q_h(x-z) p(z) dz}$$

The denominator has the same shape for the variance term and tends to $p(X)$ when h tends to zero. With the change of variable $u = \frac{1}{h}(x-z)$, the numerator is equal to

$$\int_{\mathbb{R}^d} q_h(x-z)||x-z||_2^2 p(z) dz = h^2 \int_{\mathbb{R}^d} q(u)||u||_2^2 p(x-hu) du$$

which is equivalent to $h^2 p(x) \int_{\mathbb{R}^d} q(u)||u||_2^2 du$ when h tends to zero. Overall, when n tends to infinity, and h tends to zero, we get:

$$\int_{\mathcal{X}} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] p(x) dx \sim h^2 \int_{\mathbb{R}^d} q(u)||u||_2^2 du$$

Overall Results

Combining the results from variance and bias terms, we get an *asymptotic bound* of $\mathcal{R}(\hat{f}) - \mathcal{R}^*$ proportional to (up to constants depending on q):

$$\frac{\sigma^2}{nh^d} + B^2 h^2$$

leading to the same upper-bound as for partitioning estimates, by setting $h \approx n^{-1/(d+2)}$.

We can make the informal reasoning above more formal using **concentration inequalities**, leading to non-asymptotic bounds of the same nature (simply more complicated), that make explicit the joint dependence on n and h.

Convergence rate for Nadaraya-Watson Estimation

Proposition (Convergence rate for Nadaraya-Watson estimation)

Assume bounded noise (A1) and a Lipschitz-continuous target function (A2), and a function $q : \mathbb{R}^d \mapsto \mathbb{R}$ *such that* $\int_{\mathbb{R}^d} q(z)dz = 1$, *and* $||q||_\infty = \sup_{z\in\mathbb{R}^d} q(z)$ *is finite. We also assume that* $p(x) \in [p_{\min}, p_{\max}]$ *for all* $x \in \mathcal{X}$. *Then for the Nadaraya-Watson estimate* $\hat{f}$, *we have:*

$$\mathcal{R}(\hat{f}) - \mathbb{R}^* = \int_{\mathcal{X}} \mathbb{E}_{S_n}\left[(\hat{f}(X) - f^*(X)) \mid X = x\right] dP(x) \leq \frac{4||q||_\infty}{p_{\min}} \cdot \frac{2\sigma^2 + B\,\mathrm{diam}(\mathcal{X})^2}{nh^d} + 2h^2 \cdot \frac{p_{\max}}{p_{\min}} \int_{\mathcal{R}^d} q(u)||u||_2^2 du \tag{10}$$

Universal Consistency

In the previous discussion, we have required the following conditions on the weights:

- $\int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^n \hat{w}_i(x) d(X_i, x)^2\right] dP(x) \to 0$ when n tends to infinity, to ensure that the bias goes to zero.
- $\int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^n \hat{w}_i(x)^2\right] dP(x) \to 0$ when n tends to infinity, to ensure that the variance goes to zero.

These were enough to show consistency when the target function is Lipschitz-continuous in $\mathbb{R}^d$ and these also led to a precise rate of convergence (which turned out to be optimal).

In order to show universal consistency for any **square-intergrable** functions, we need an extra (technical) assumption, which was first outlined in Stone's theorem, namely that there exists $c > 0$ such that for any non-negative integrable function $h : \mathcal{X} \mapsto \mathbb{R}$,

$$\int_{\mathcal{X}} \sum_{i=1}^n \mathbb{E}_{S_n} \left[\hat{w}_i(x) h(X_i)\right] dP(x) \leq c \cdot \int_{\mathcal{X}} h(x) dP(x) \tag{11}$$

Again, we only take the expectation with respect to the training data, while we use the integral notation to take the expectation with respect to the training distribution P.

Universal Consistency Analysis I

In the following contents, h will be the **squared deviation (distance) between two functions**. Then for any $\varepsilon > 0$, and for any $f^* \in L_2(P(x))$, we can find a function g which is $B(\varepsilon)$-Lipschitz-continuous and such that $||f^* - g||_{L_2(P(x))} \leq \varepsilon$, that is because the set of Lipschitz-continuous functions is dense in $L_2(P(x))$. Then for a given $X \in \mathcal{X}$, we have

$$
\begin{aligned}
&\mathbb{E}_{S_n}\left(\sum_{i=1}^{n}\hat{w}_i(x)[f^*(X_i) - f^*(X)]\right)^2 \\
&\leq \mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\Big[\big|f^*(X_i) - g(X_i)\big| + \big|g(X_i) - g(x)\big| + \big|g(x) - f^*(X)\big|\Big]\right)^2 \\
&\leq 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\big|f^*(X_i) - g(X_i)\big|\right)^2 + 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\big|g(X_i) - g(x)\big|\right)^2 + 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\big|g(x) - f^*(x)\big|\right)^2 \quad (1) \\
&\leq 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\big|f^*(X_i) - g(X_i)\big|\right)^2 + 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)B(\varepsilon)d(X_i, x)\right)^2 + 3\mathbb{E}\big|g(x) - f^*(x)\big|^2 \quad (2)
\end{aligned}
$$

Universal Consistency Analysis II

$$\leq 3\mathbb{E}\left(\sum_{i=1}^{n}\hat{w}_i(x)\left|f^*(X_i)-g(X_i)\right|\right)^2+3B(\varepsilon)^2\cdot\mathbb{E}\left[\sum_{i=1}^{n}\hat{w}_i(x)d(X_i,x)^2\right]+3\mathbb{E}\left|g(x)-f^*(x)\right|^2 \quad (3)$$

$$\leq 3c\cdot\mathbb{E}\left|f^*(x)-g(x)\right|^2+3B(\varepsilon)^2\cdot\mathbb{E}\left[\sum_{i=1}^{n}\hat{w}_i(x)d(X_i,x)^2\right]+3\mathbb{E}\left|g(x)-f^*(x)\right|^2$$

The inequality (1) is obtained using that $(a+b+c)^2\leq 3a^2+3b^2+3c^2$, inequality (2) is due to the fact that weights summing to one and function g is Lipschitz-continuous, while inequality (3) applys the Jensen's inequality to the second term, and the last inequality is a direct result of Eq.(11)

Expected Excess Risk

We can now integrate with respect to $X = x$ and utilize the assumption that $||f^* - g||_{L_2(P(x))} \leq \varepsilon$ to get

$$\int_{\mathcal{X}} \mathbb{E}_{S_n} \left(\sum_{i=1}^{n} \hat{w}_i(x)[f^*(X_i) - f^*(x)] \right)^2 dP(x) \leq 3(c+1)\varepsilon^2 + 3B(\varepsilon)^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x)$$

We can then substitue the upper-bound of bias term into the first inequality of Eq.(5), which is

$$\begin{aligned} \int_{\mathcal{X}} \mathbb{E}_{S_n} (\hat{f}(x) - f^*(x))^2 dP(x) &\leq \int_{\mathcal{X}} \mathbb{E}_{S_n} \left(\sum_{i=1}^{n} \hat{w}_i(x) \cdot \left| f^*(X_i) - f^*(X) \right| \right)^2 dP(x) + \sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x)^2 \right] dP(x) \\ &\leq 3B(\varepsilon)^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x) + \sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x)^2 \right] dP(x) \\ &\quad + 3(c+1)\varepsilon^2 \end{aligned} \tag{12}$$

Proving Universal Consistency I

In order to prove universal consistency, we fix a certain ε, from which we obtain some $B(\varepsilon)$. For such a $B(\varepsilon)$, we know how to obtain an overall term

$$3B(\varepsilon)^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x) d(X_i, x)^2 \right] dP(x) + \sigma^2 \int_{\mathcal{X}} \mathbb{E}_{S_n} \left[\sum_{i=1}^{n} \hat{w}_i(x)^2 \right] dP(x)$$

for a well chosen hyperparameter and number of observations n as previous section. Thus, if the extra condition in Eq.(11) is satisfied, these three methods are universally consistent.

Case Study

We can now look at the three cases by reviewing the techinical asumption

$$\int_{\mathcal{X}} \sum_{i=1}^{n} \mathbb{E}_{S_n} \left[\hat{w}_i(x) h(X_i) \right] dP(x) \leq c \cdot \int_{\mathcal{X}} h(x) dP(x)$$

- Partitioning: Let $c = 2$ in Eq.(11) and we get universal consistency. This is because

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}_{S_n} [\hat{w}_i(x) f(X_i)] &= \sum_{j \in J} \sum_{i=1}^{n} \mathbb{E}_{S_n} \left[\hat{w}_i(x) \mathbb{1}_{x \in A_j} f(X_i) \right] \\
&= \sum_{j \in J} \mathbb{E}_{S_n} \left[\mathbb{1}_{x \in A_j} \left(\mathbb{1}_{n_{A_j} > 0} \cdot \frac{1}{n_{A_j}} \sum_{i \in A_j} f(X_i) + \mathbb{1}_{n_{A_j} = 0} \cdot \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right) \right] \\
&\leq \sum_{j \in J} \mathbb{E}_{S_n} \left[\mathbb{1}_{x \in A_j} \left(\mathbb{E}[f(Z) \mid Z \in A_j] + \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right) \right] \quad (X_i, i = 1, \ldots, n \text{ i.i.d}) \\
&= 2 \sum_{j \in J} \mathbb{1}_{x \in A_j} \mathbb{E}_{S_n} [f(X)] = 2 \mathbb{E}_{S_n} [f(X)]
\end{aligned}$$

Case Study Continue

- Kernel regression: it can be shown using the same type of techiniques outlined for oonsistency for Lipschitz-continuous functions
- k-nearest neighbor: the condition in Eq.(11) is not easy to show, and is often referred to as Stone's lemma.

Acknowledgment

Thank You!

Here I would like to thank Professor Mao for his generous help and valuable advice during my prepartion.