# Non-parametric methods
## contd…

Aarti Singh

Machine Learning 10-315
Oct 13, 2021

# Non-Parametric methods

- Typically don't make any distributional assumptions
- As we have more data, we should be able to learn more complex models
- Let number of parameters scale with number of training data

- Some nonparametric methods

  **Classification:** Decision trees, k-NN (k-Nearest Neighbor) classifier

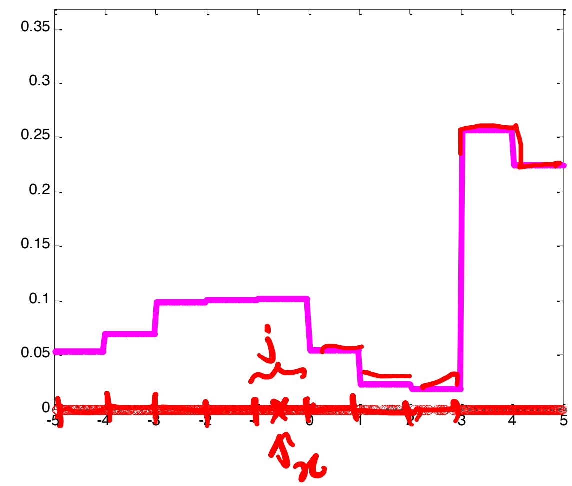
  **Density estimation:** k-NN, Histogram, Kernel density estimate

  **Regression:** Kernel regression

# Kernel density estimate

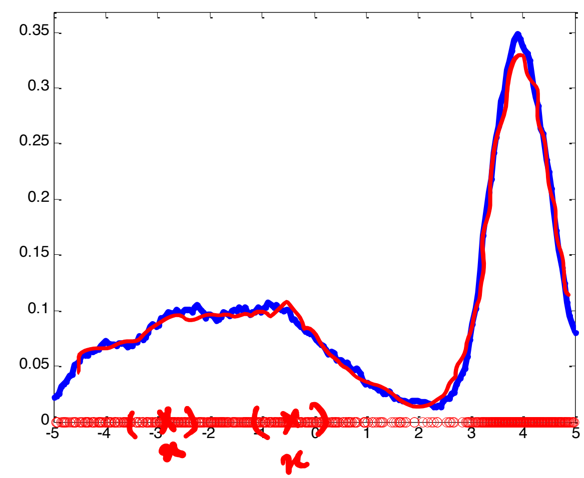
- Histogram – blocky estimate

$$\widehat{p}(x) = \frac{1}{\Delta}\frac{\sum_{j=1}^{n} \boxed{\mathbf{1}_{X_j \in \mathrm{Bin}_x}}}{n} = \frac{n_j}{\Delta n}$$




- Kernel density estimate aka "Parzen/moving window method"

$$\widehat{p}(x) = \frac{1}{\Delta}\underbrace{\frac{\sum_{j=1}^{n} \mathbf{1}_{||X_j - x|| \leq \Delta}}{n}}$$



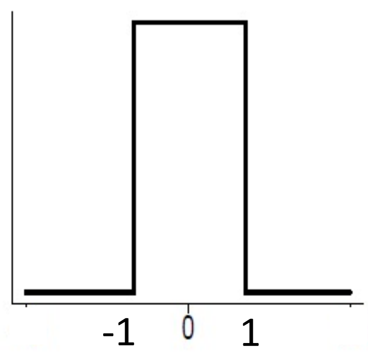
# Kernel density estimate

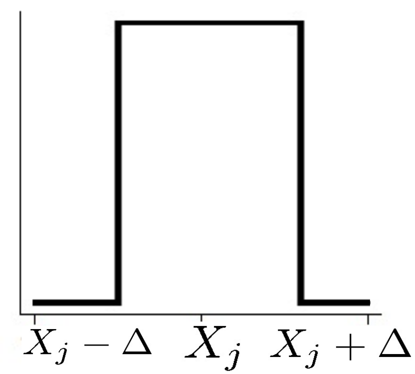
- $\widehat{p}(x) = \frac{1}{\Delta} \frac{\sum_{j=1}^{n} K\left(\frac{X_j - x}{\Delta}\right)}{n}$ more generally

$$K\left(\frac{X_j - x}{\Delta}\right) = 1_{\frac{\|X_j - x\|}{\Delta} \leq 1}$$

boxcar kernel :

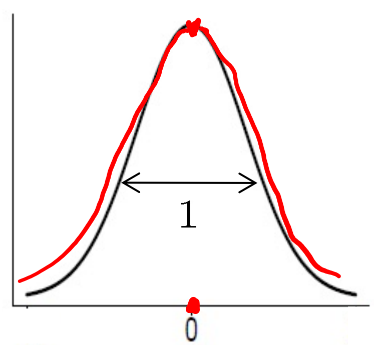
$$K(x) = \frac{1}{2} I(x),$$

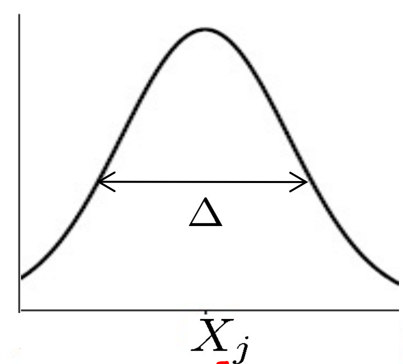



Gaussian kernel :

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

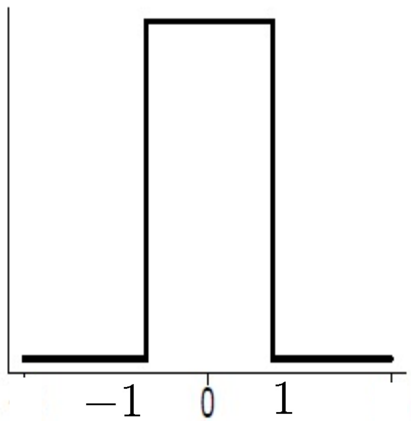


# Kernels

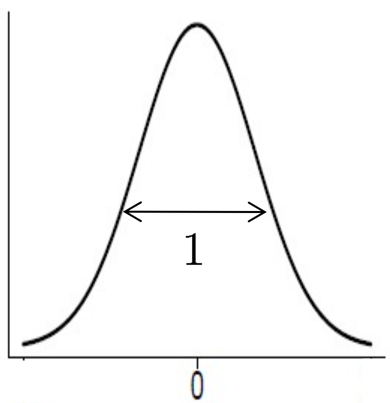
boxcar kernel :

$$K(x) = \frac{1}{2} I(x),$$




Gaussian kernel :

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$




Any kernel function that satisfies

$$K(x) \geq 0,$$

$$\int K(x)dx = 1$$

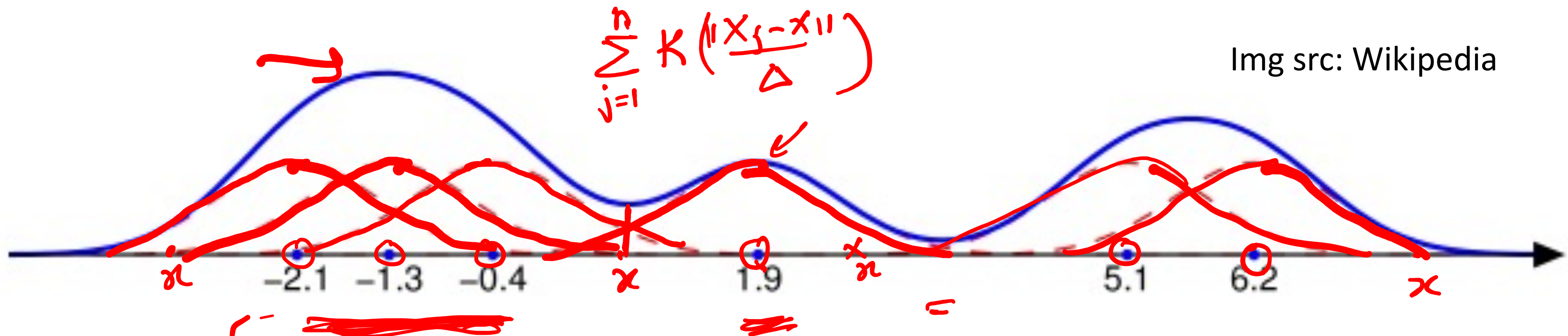
$p(x) \geq 0$

$\int p(x)\, dx = 1$

# Kernel density estimation

- Place small "bumps" at each data point, determined by the kernel function.
- The estimator consists of a (normalized) "sum of bumps".

Img src: Wikipedia



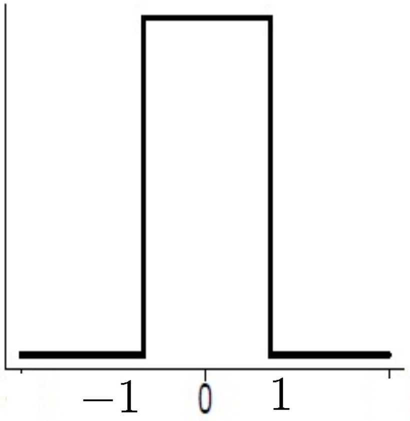

Gaussian bumps (red) around six data points and their sum (blue)

- Note that where the points are denser the density estimate will have higher values.

# Choice of Kernels

boxcar kernel :

$$K(x) = \frac{1}{2} I(x),$$

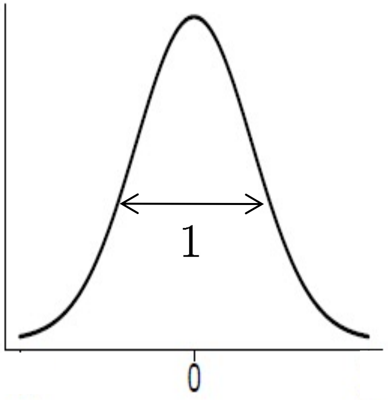



Finite support

– only need local points to compute estimate

Gaussian kernel :

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$




Infinite support

- need all points to compute estimate
-But quite popular since smoother

# Choice of kernel bandwidth

$\Delta \equiv k$

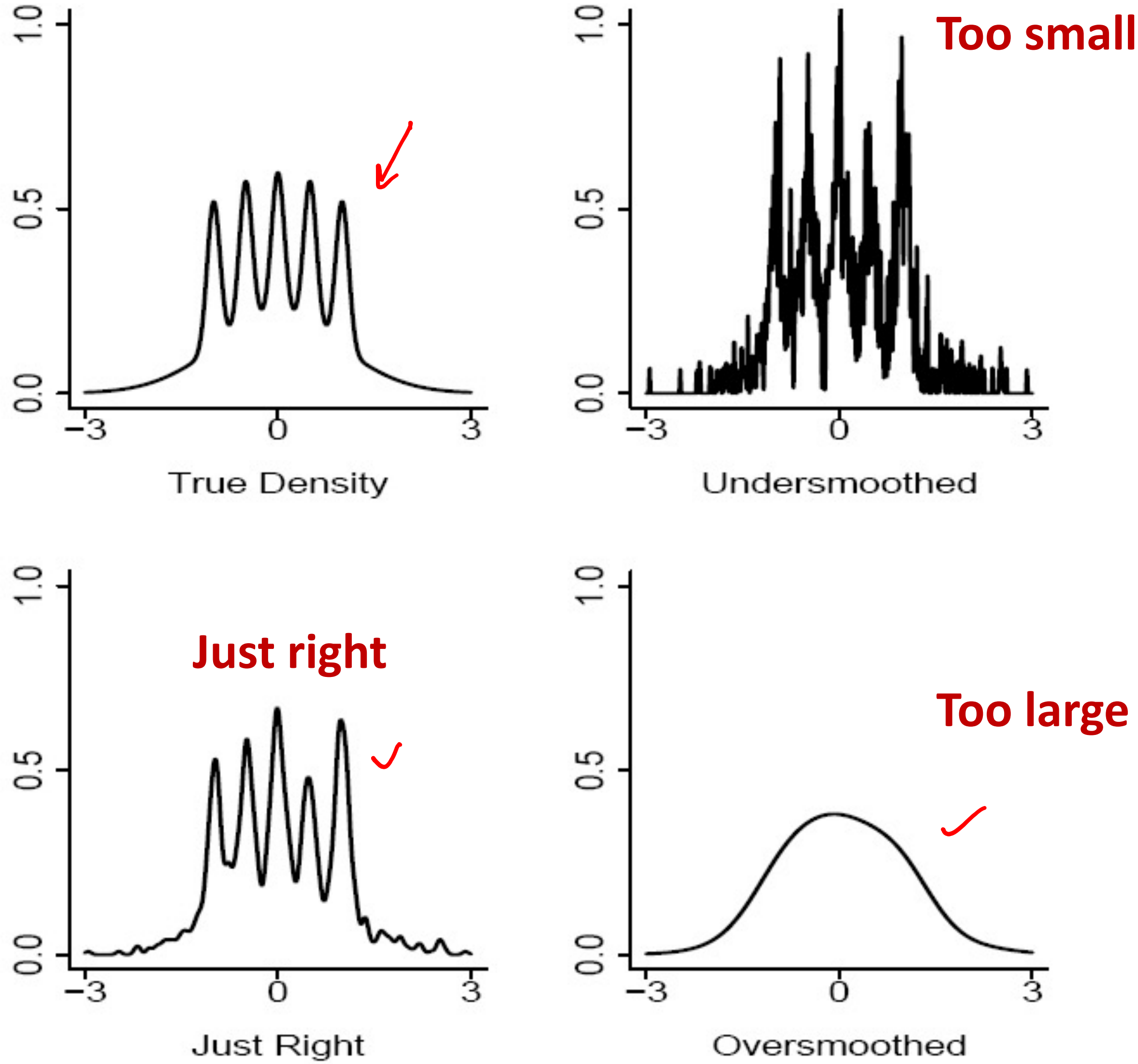
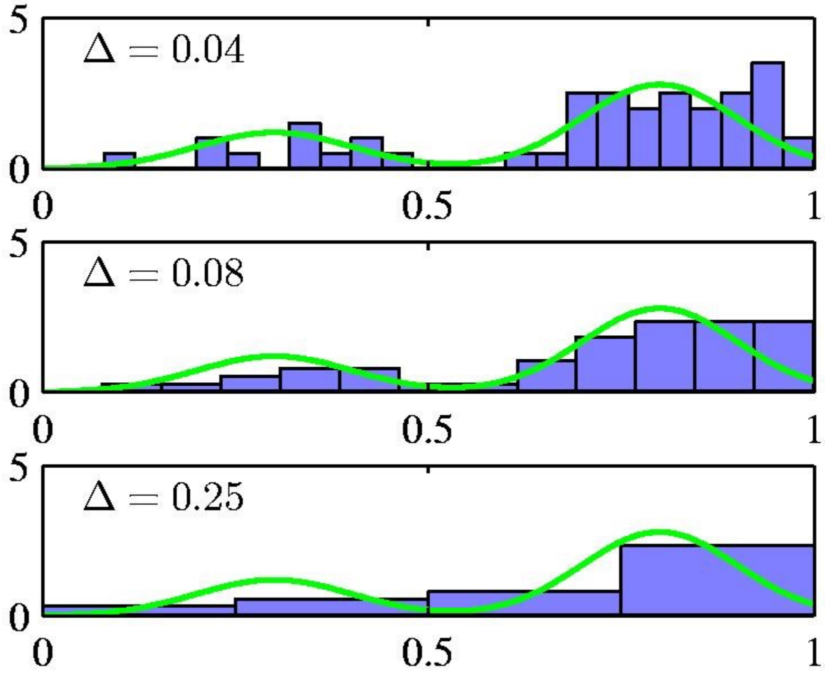
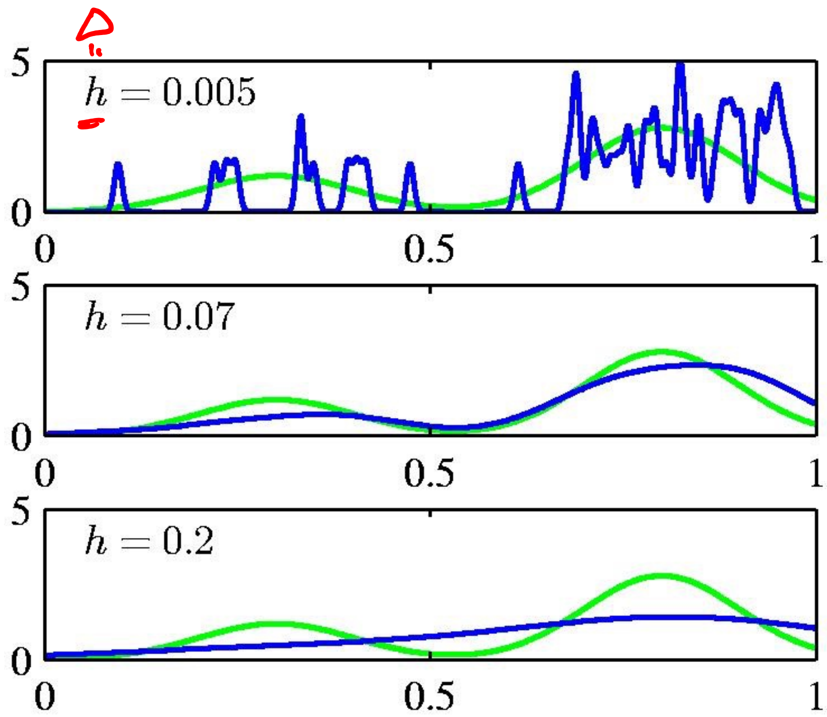



Image Source:
Larry's book – All of Nonparametric Statistics

**Bart-Simpson Density**

# Histograms vs. Kernel density estimation




$\Delta = \text{h}$ acts as a smoother.

# Nonparametric density estimation

- Histogram $\qquad \widehat{p}(x) = \frac{n_i}{n\Delta}\mathbf{1}_{x \in \mathrm{Bin}_i}$
- Kernel density est $\qquad \widehat{p}(x) = \frac{n_x}{n\Delta}$

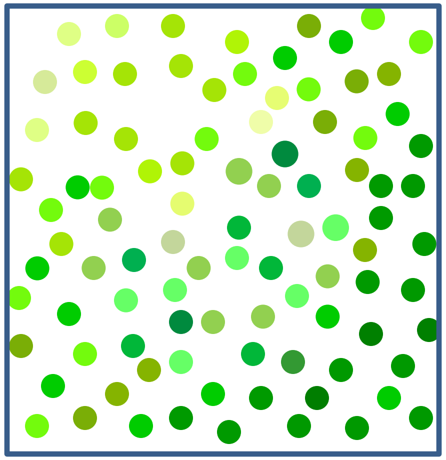
Fix Δ, estimate number of points within Δ of x ($n_i$ or $n_x$) from data

Fix $n_x = k$, estimate Δ from data (volume of ball around x that contains k training pts)

- k-NN density est $\qquad \widehat{p}(x) = \frac{k}{n\Delta_{k,x}}$

# Local Kernel Regression

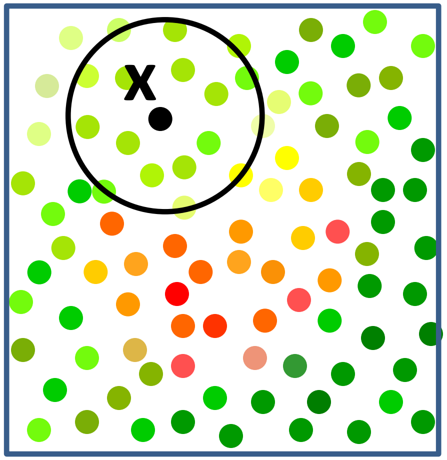
- What is the temperature in the room?



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

**Average**

at location x?

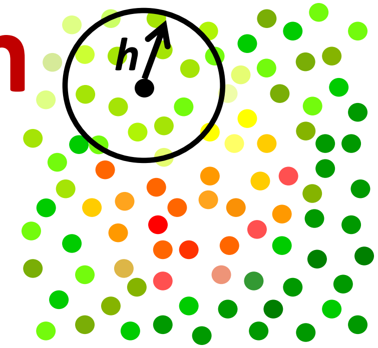



$$\widehat{T}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}_{||X_i - x|| \le h}}{\sum_{i=1}^{n} \mathbf{1}_{||X_i - x|| \le h}}$$

$n_x$

**"Local" Average**

# Local Kernel Regression



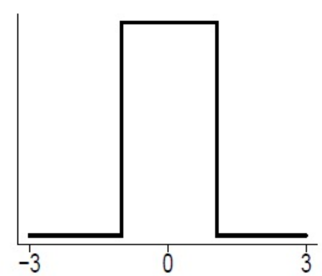

- Nonparametric estimator
- Nadaraya-Watson Kernel Estimator

$$\widehat{f}_n(X) = \sum_{i=1}^{n} w_i Y_i \quad \text{Where} \quad w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)}$$

- Weight each training point based on distance to test point
- Boxcar kernel yields local average

boxcar kernel :

$$K(x) = \frac{1}{2} I(x),$$



# Choice of kernel bandwidth h

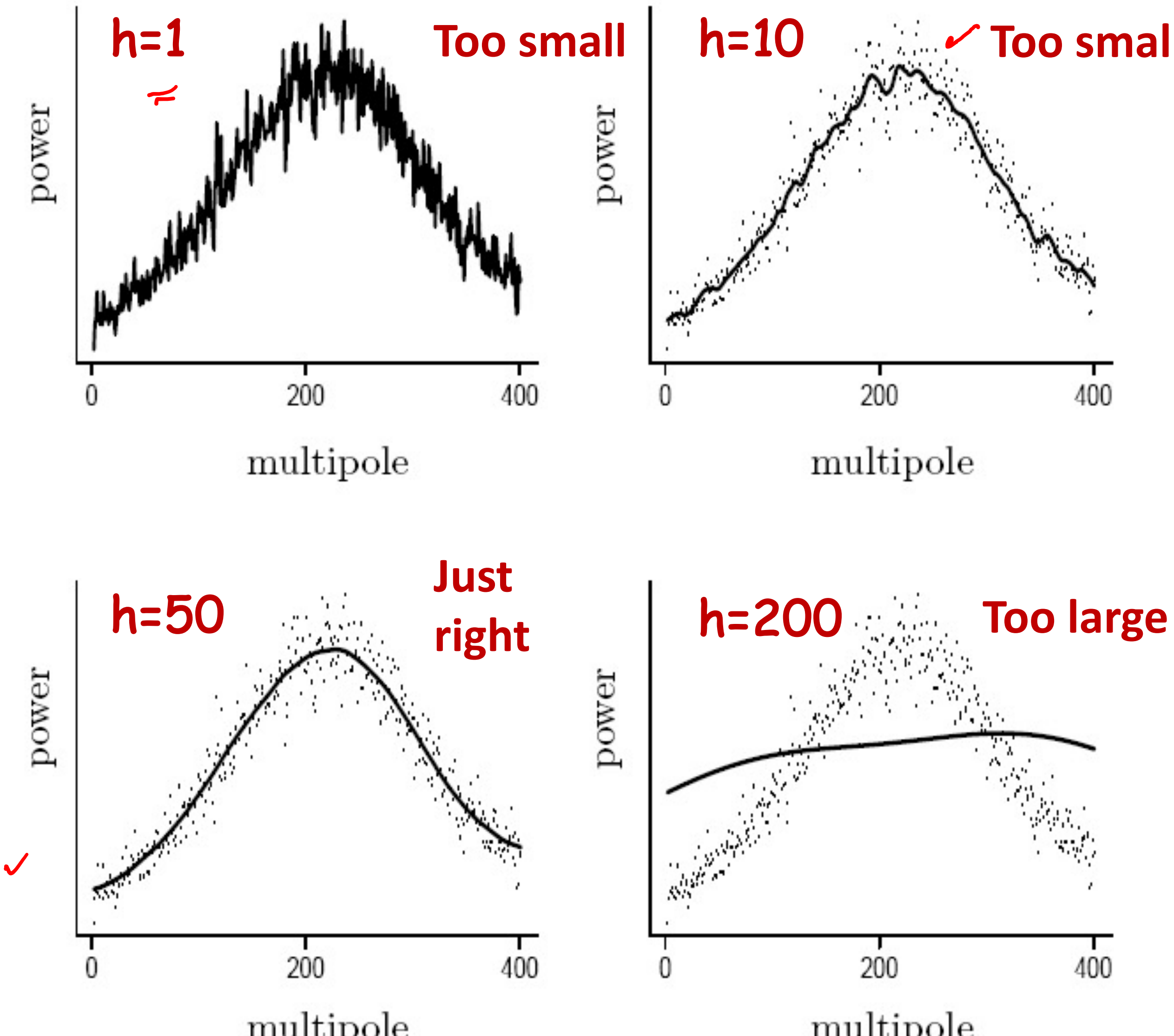



Image Source: Larry's book – All of Nonparametric Statistics

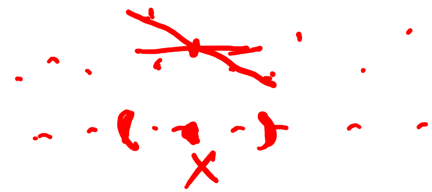
# Kernel Regression as Weighted Least Squares

$X_i\beta$

$$\min_f \sum_{i=1}^{n} \boxed{w_i}(f(X_i) - Y_i)^2$$

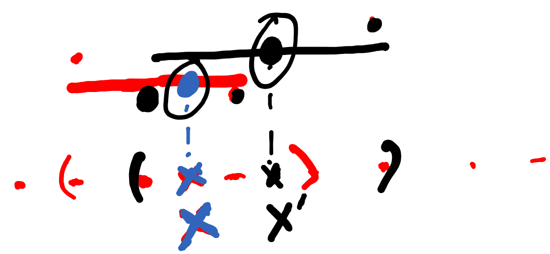
Weighted Least Squares

$$\boxed{w_i(X)} = \frac{K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}$$



Kernel regression corresponds to locally constant estimator obtained from (locally) weighted least squares

i.e. set $f(X_i) = \beta$ (a constant)



# Kernel Regression as Weighted Least Squares

$\beta X_i + \beta_0$

set $f(X_i) = \beta$ (a constant)

$$\min_{\beta} \sum_{i=1}^{n} w_i(\underbrace{\beta}_{\text{constant}} - Y_i)^2$$

$\sum_i w_i(x) = 1$

$$w_i(X) = \frac{K\left(\frac{X - X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)}$$

$$\frac{\partial J(\beta)}{\partial \beta} = 2\sum_{i=1}^{n} w_i(\beta - Y_i) = 0$$

$\Rightarrow \cancel{\sum_{i=1}^{n} w_i}^{\,1} \beta = \boxed{\sum_{i=1}^{n} w_i Y_i}$

Notice that $\displaystyle\sum_{i=1}^{n} w_i = 1$

$$\Rightarrow \widehat{f}_n(X) = \widehat{\beta} = \sum_{i=1}^{n} w_i Y_i$$

# Local Linear/Polynomial Regression

$$\min_f \sum_{i=1}^{n} w_i(f(X_i) - Y_i)^2 \qquad w_i(X) = \frac{K\left(\frac{X-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X-X_i}{h}\right)}$$

Weighted Least Squares

Local Polynomial regression corresponds to locally polynomial estimator obtained from (locally) weighted least squares

i.e. set $f(X_i) = \beta_0 + \beta_1(X_i - X) + \frac{\beta_2}{2!}(X_i - X)^2 + \cdots + \frac{\beta_p}{p!}(X_i - X)^p$

(local polynomial of degree p around X)

# Summary

- Non-parametric approaches

**Four things make a nonparametric/memory/instance based/lazy learner:**

1. *A distance metric, dist(x,X$_i$)*
   **Euclidean (and many more)**
2. *How many nearby neighbors/radius to look at?*
   **k, Δ/h**
3. *A weighting function (optional)*
   **W based on kernel K**
4. *How to fit with the local points?*
   **Average, Majority vote, Weighted average, Poly fit**

# Summary

- Parametric vs Nonparametric approaches

    - Nonparametric models place very mild assumptions on the data distribution and provide good models for complex data

      Parametric models rely on very strong (simplistic) distributional assumptions

    - Nonparametric models (not histograms) requires storing and computing with the entire data set.

      Parametric models, once fitted, are much more efficient in terms of storage and computation.