Kernel Density Estimation

Liang Liang

Kernel Density Estimation

- $\{x_1, x_2, x_3, \ldots, x_N\}$ is a set of data samples, and $x_n \in \mathcal{R}^M$
- The PDF of the data can be approximated by the function:

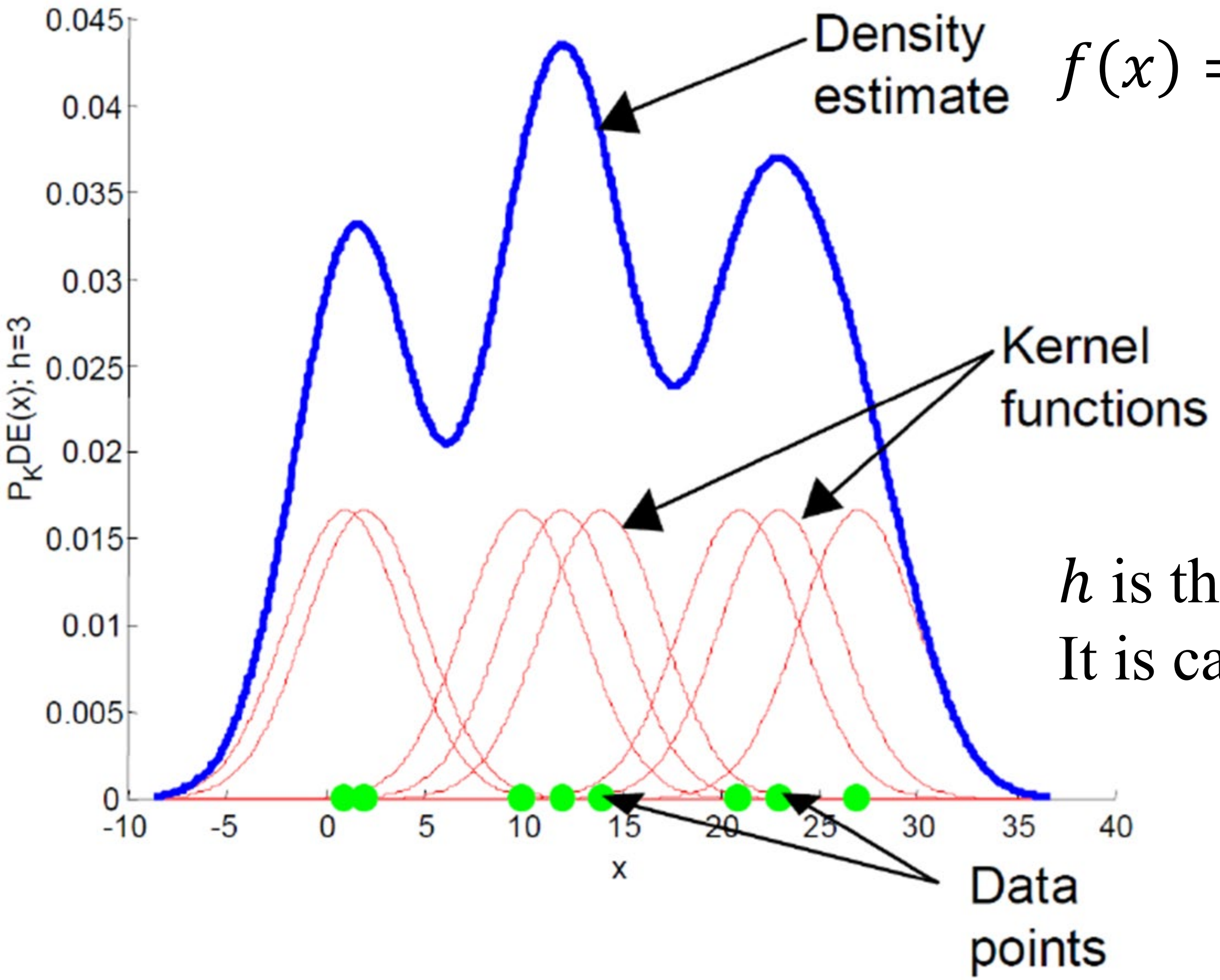
$$f(x) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(x, x_n)$$

$\mathcal{K}(x, x_n)$ is called *K*ernel function:

- $\mathcal{K}(x, x_n) \geq 0$
- $\int_{-\infty}^{\infty} \mathcal{K}(x, x_n) dx = 1$

The function may have some parameters

Kernel Density Estimation (KDE) using Gaussian Kernels



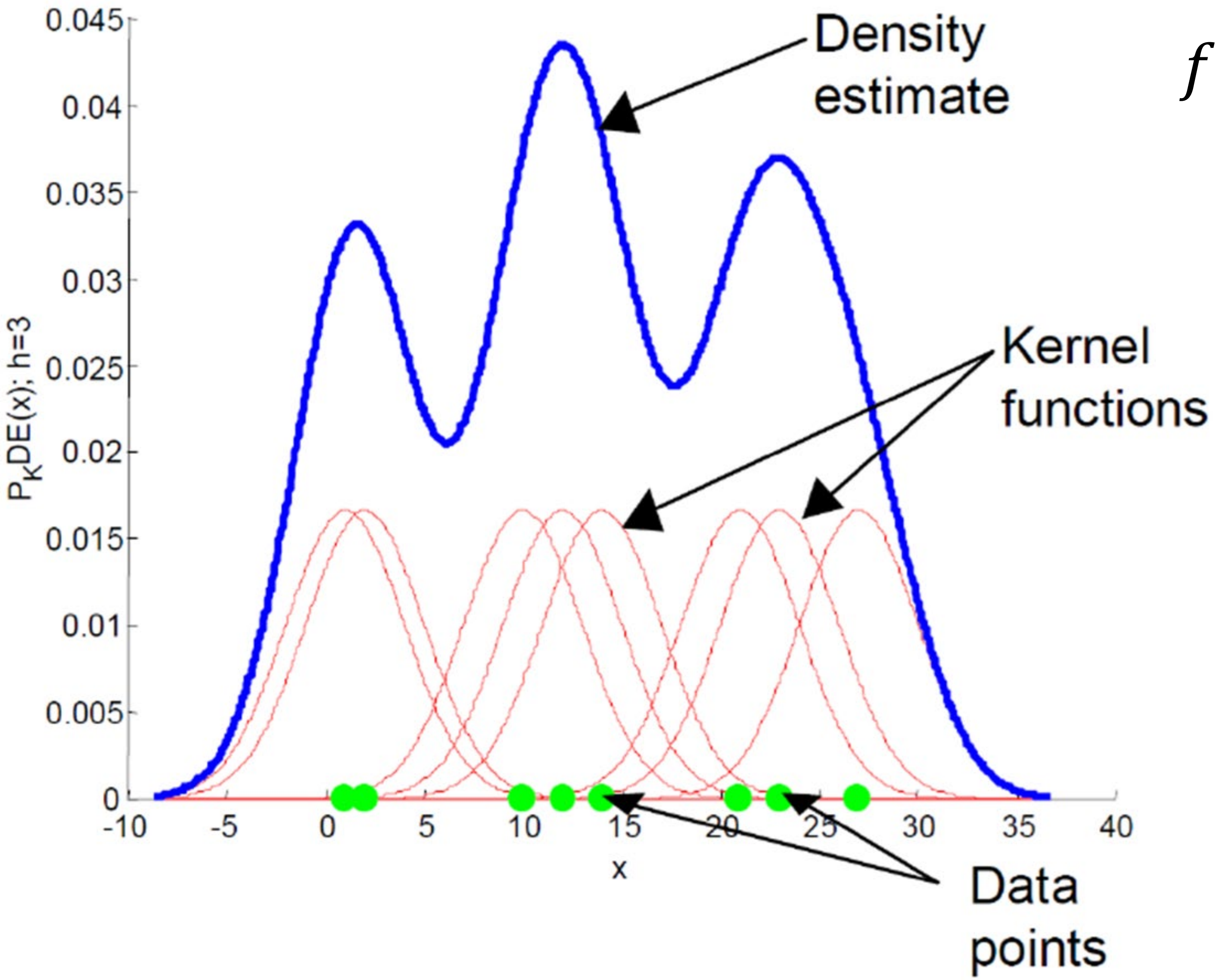
$$f(x) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\sqrt{2\pi}h}e^{-\frac{(x-x_n)^2}{2h^2}}$$

$$\mathcal{K}(x, x_n) = \frac{1}{\sqrt{2\pi}h}e^{-\frac{(x-x_n)^2}{2h^2}}$$

h is the same for every Gaussian kernel
It is called "bandwidth"

https://docs.scipy.org/doc/scipy/reference/reference/generated/scipy.stats.gaussian_kde.html#scipy.stats.gaussian_kde

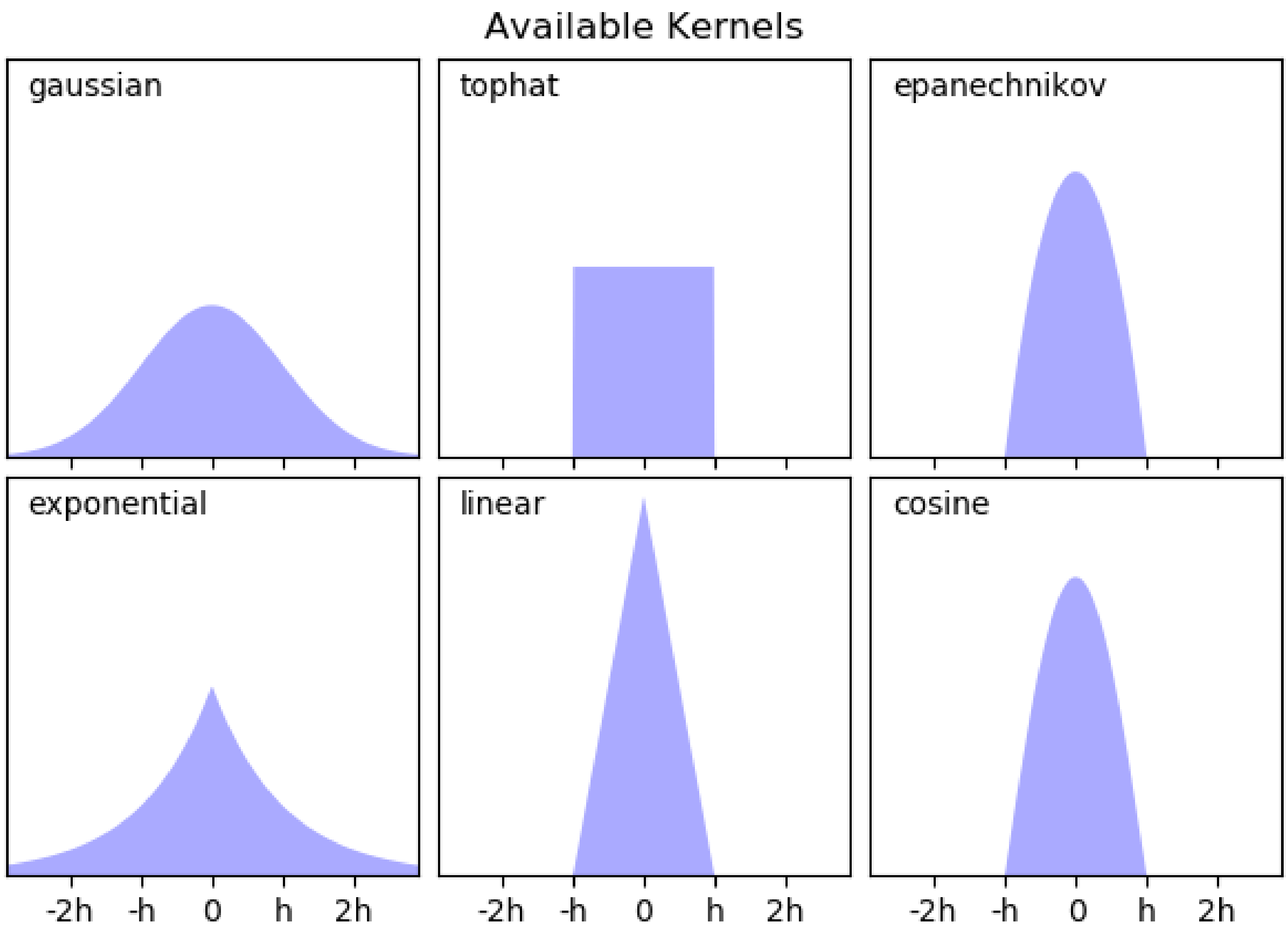
Kernel Density Estimation (KDE) using General Kernel Functions



$$f(x) = \frac{1}{N} \sum_{n=1}^{N} \mathcal{K}(x - x_n)$$

$$\mathcal{K}(x - x_n)$$

https://scikit-learn.org/stable/modules/density.html



Kernel function:

- $u = x - x_n$
- $\mathcal{K}(u) \geq 0$
- $\int_{-\infty}^{\infty} \mathcal{K}(u) du = 1$

Apply KDE on the Iris Data Set (demo KDE.ipynb)

- Iris dataset contains 3 classes, and each class has 50 samples
- A class refers to a type of iris plant.
- A data sample has 4 features/attributes
- We will build a KDE model for each of the 3 classes

  - $f(x) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(x, x_n)$ using the data points in class-0
  - $f(x) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(x, x_n)$ using the data points in class-1
  - $f(x) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(x, x_n)$ using the data points in class-2