Introduction to Machine Learning
Lecture 3 - Total Derivative and Non-parametric estimations

1 Derivatives of Multivariate Functions

1.1 Scalar function

Consider the scalar function $f : \mathbb{R} \longrightarrow \mathbb{R}$:

$$f(x), \qquad x \in \mathbb{R}$$

The derivative of f is given by:

$$f' = \frac{\mathrm{d}f}{\mathrm{d}x}$$

That is, we can write:

$$\boxed{\mathrm{d}f = f' \cdot \mathrm{d}x}$$

1.2 Multivariate function - simple example

Consider the multivariate function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$:

$$f(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathbb{R}^2$$

The gradient of f is defined by the vector of partial derivatives:

$$\nabla f \triangleq \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$$

The total derivative of f is given by:

$$\boxed{\mathrm{d}f \triangleq \frac{\partial f}{\partial x_1}\mathrm{d}x_1 + \frac{\partial f}{\partial x_2}\mathrm{d}x_2}$$

Using inner product notation, we have:

$$\Rightarrow \boxed{\mathrm{d}f = \left\langle \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}, \begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{bmatrix} \right\rangle = \langle \nabla f, \mathrm{d}\boldsymbol{x} \rangle}, \qquad \mathrm{d}\boldsymbol{x} \triangleq \begin{bmatrix} \mathrm{d}x_1 \\ \mathrm{d}x_2 \end{bmatrix}$$

1.2.1 Example

Compute the gradient of the following function using both the direct and the total derivative methods:

$$f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x}, = \sum_{i=1}^{2} a_i x_i \qquad \boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^2$$

Method I - by definition	**Method II - total derivative**
$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1}\sum_{i=1}^{2} a_i x_i \\ \frac{\partial}{\partial x_2}\sum_{i=1}^{2} a_i x_i \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$ $\Rightarrow \boxed{\nabla f(\boldsymbol{x}) = \boldsymbol{a}}$	$\mathrm{d}f = \boldsymbol{a}^T\mathrm{d}\boldsymbol{x} = \langle \boldsymbol{a}, \mathrm{d}\boldsymbol{x} \rangle$ $\Rightarrow \boxed{\nabla f(\boldsymbol{x}) = \boldsymbol{a}}$

1.3 The general case

Consider the differential multivariate function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$,
If exist some $\boldsymbol{g} \in \mathbb{R}^n$ such that:

$$\mathrm{d}f = \langle \boldsymbol{g}, \mathrm{d}\boldsymbol{x} \rangle$$

then, $\boldsymbol{g}$ is the gradient of f, namely, $\boldsymbol{g} = \nabla f$.

1.3.1 Quadratic example:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$$

1. Show that without loss of generality, one can assume $\boldsymbol{A} = \boldsymbol{A}^T$ ($\boldsymbol{A}$ is symmetric matrix).
2. Find ∇f.

Solution:

1. We can write $\boldsymbol{A}$ as a sum of symmetric matrix and anti-symmetric matrix:

$$\boldsymbol{A} = \underbrace{\frac{\boldsymbol{A} + \boldsymbol{A}^T}{2}}_{\boldsymbol{S}} + \underbrace{\frac{\boldsymbol{A} - \boldsymbol{A}^T}{2}}_{\tilde{\boldsymbol{S}}} = \boldsymbol{S} + \tilde{\boldsymbol{S}}$$

now:

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{x}^T \left(\boldsymbol{S} + \tilde{\boldsymbol{S}} \right) \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{S} \boldsymbol{x} + \boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x}$$

Notice that $\left(\tilde{\boldsymbol{S}} = -\tilde{\boldsymbol{S}}^T \right)$:

$$\begin{aligned} \boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} &= \left(\boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} \right)^T, \qquad \boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} \in \mathbb{R} \\ \boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} &= -\boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} \end{aligned}$$

$$\Rightarrow \boldsymbol{x}^T \tilde{\boldsymbol{S}} \boldsymbol{x} = 0$$
$$\Rightarrow \boxed{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{S} \boldsymbol{x}}$$

Since this is true for any $\boldsymbol{x}$ we can consider without loss of generality only the symmetric part of $\boldsymbol{A}$.

2. Using the product rule

$$\begin{aligned} \mathrm{d}f &= \mathrm{d}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{A} \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{x}^T \boldsymbol{A}^T \mathrm{d}\boldsymbol{x} + \boldsymbol{x}^T \boldsymbol{A} \mathrm{d}\boldsymbol{x} \\ &= \boldsymbol{x}^T \left(\boldsymbol{A}^T + \boldsymbol{A} \right) \mathrm{d}\boldsymbol{x} \\ &= \left\langle \left(\boldsymbol{A} + \boldsymbol{A}^T \right) \boldsymbol{x}, \mathrm{d}\boldsymbol{x} \right\rangle \end{aligned}$$

$$\Rightarrow \boxed{\nabla f = \left(\boldsymbol{A} + \boldsymbol{A}^T \right) \boldsymbol{x}}$$

if we also consider $\boldsymbol{A} = \boldsymbol{A}^T$:

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$$

which is vary similar to the scalar case $\frac{d}{dx}\left(ax^2 \right) = 2ax$.

1.3.2 ℓ_2 - norm

Question 1 Find ∇f of:

$$f(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{b}\|_2^2$$

Solution:

$$f(\boldsymbol{x}) = \|\boldsymbol{x}-\boldsymbol{b}\|_2^2 = (\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{x}-\boldsymbol{b})$$

$$\begin{aligned}
\Rightarrow \mathrm{d}f &= \mathrm{d}\boldsymbol{x}^T(\boldsymbol{x}-\boldsymbol{b}) + (\boldsymbol{x}-\boldsymbol{b})^T \mathrm{d}\boldsymbol{x} \\
&= (\boldsymbol{x}-\boldsymbol{b})^T \mathrm{d}\boldsymbol{x} + (\boldsymbol{x}-\boldsymbol{b})^T \mathrm{d}\boldsymbol{x} \\
&= 2(\boldsymbol{x}-\boldsymbol{b})^T \mathrm{d}\boldsymbol{x} \\
&= \langle 2(\boldsymbol{x}-\boldsymbol{b}), \mathrm{d}\boldsymbol{x}\rangle
\end{aligned}$$

$$\Rightarrow \boxed{\nabla f(\boldsymbol{x}) = 2(\boldsymbol{x}-\boldsymbol{b})}$$

Question 2 Find ∇h of:

$$h(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2$$

Solution I:

$$h(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|_2^2 = (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})$$

$$\begin{aligned}
\Rightarrow \mathrm{d}h &= (\boldsymbol{A}\mathrm{d}\boldsymbol{x})^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + (\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T \boldsymbol{A}\mathrm{d}\boldsymbol{x} \\
&= 2(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T \boldsymbol{A}\mathrm{d}\boldsymbol{x} \\
&= \left\langle 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}), \mathrm{d}\boldsymbol{x}\right\rangle
\end{aligned}$$

$$\Rightarrow \boxed{\nabla h(\boldsymbol{x}) = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})}$$

Solution II - Using the chain rule:

$$h(\boldsymbol{x}) = f(\boldsymbol{A}\boldsymbol{x}) = f(\boldsymbol{u}), \qquad \boldsymbol{u} \triangleq \boldsymbol{A}\boldsymbol{x}$$

$$\begin{aligned}
\Rightarrow \mathrm{d}h &= \nabla^T f(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \\
&= \nabla^T f(\boldsymbol{u})\,\boldsymbol{A}\mathrm{d}\boldsymbol{x} \\
&= \left\langle \boldsymbol{A}^T \nabla f(\boldsymbol{u}), \mathrm{d}\boldsymbol{x}\right\rangle
\end{aligned}$$

$$\boxed{\nabla h(\boldsymbol{x}) = \boldsymbol{A}^T\nabla f(\boldsymbol{u}) = \boldsymbol{A}^T\nabla f(\boldsymbol{A}\boldsymbol{x}) = 2\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})}$$

1.3.3 Chain rule

Find ∇h of:

$$h\left(\boldsymbol{x}\right)=\varphi\left(f\left(\boldsymbol{x}\right)\right)$$

where $\varphi:\mathbb{R}\longrightarrow\mathbb{R}$ is a scalar function and $f,\nabla f$ are known.

Solution:

let $u\triangleq f\left(\boldsymbol{x}\right)$

$$\mathrm{d}h=\varphi'\left(u\right)\mathrm{d}u=\varphi'\left(f\left(\boldsymbol{x}\right)\right)\left\langle\nabla f\left(\boldsymbol{x}\right),\mathrm{d}\boldsymbol{x}\right\rangle=\left\langle\varphi'\left(f\right)\nabla f,\mathrm{d}\boldsymbol{x}\right\rangle$$

$$\Rightarrow\nabla h\left(\boldsymbol{x}\right)=\varphi'\left(f\left(\boldsymbol{x}\right)\right)\nabla f\left(\boldsymbol{x}\right)$$

1.4 Matrix derivative example

Find $\nabla_A f$ of:

$$f\left(\boldsymbol{A}\right)=\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}$$

Solution:

$$\begin{aligned}\mathrm{d}f&=\boldsymbol{x}^T\mathrm{d}\boldsymbol{A}\boldsymbol{y}\\&=\mathrm{Tr}\left\{\boldsymbol{x}^T\mathrm{d}\boldsymbol{A}\boldsymbol{y}\right\}\\&=\mathrm{Tr}\left\{\boldsymbol{y}\boldsymbol{x}^T\mathrm{d}\boldsymbol{A}\right\}\\&=\left\langle\boldsymbol{x}\boldsymbol{y}^T,\mathrm{d}\boldsymbol{A}\right\rangle\end{aligned}$$

$$\Rightarrow\boxed{\nabla_A f=\boldsymbol{x}\boldsymbol{y}^T}$$

2 Non-parametric Estimation

Consider a random variable X, with some unknown probability function of p_X.
In Lecture 2, we assumed some model for p_X (X is Gaussian, X is uniform, etc') and we only estimated the model's parameters.
In this section, we will estimate p_X without any model assumptions.

2.1 Cumulative Distribution Function (CDF) estimation

Reminder The CDF of the random variable X is given by:

$$F_X(x) \triangleq \Pr\{X \leq x\}$$

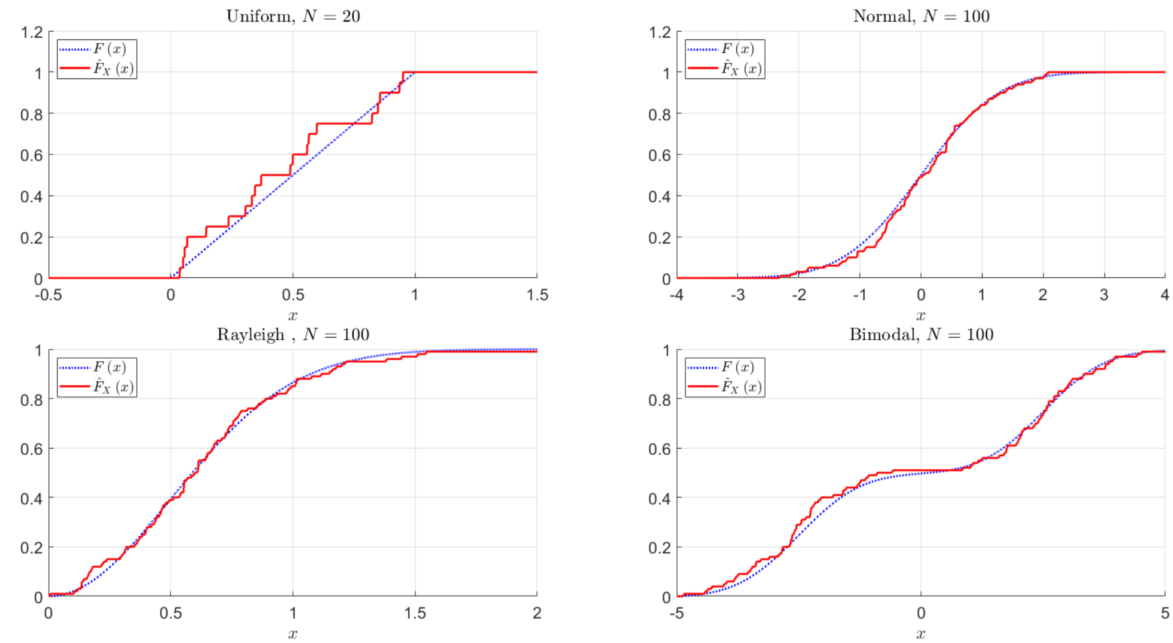
and the probability density p_X (assuming X is continuous) is given by:

$$p_X(x) = F'(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$$

CDF non-parametric estimation Consider $\{x_i\}_{i=1}^N$, N i.i.d realizations of X and the following estimation for F_X:

$$\hat{F}_X(x) = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{I}\{x_i \leq x\}$$

Example We generate N points $\{x_i\}_{i=1}^N$ from different distribution and plot the CDF estimation.



Exercise

1. Compute the bias of $\hat{F}_X(x_0)$ (for some x_0).
2. Compute the $\text{MSE}(F_X(x_0))$ (for some x_0).
3. What is the MSE for $N \to \infty$?

Solution:

1. $$\begin{aligned}\mathbb{E}\left[\hat{F}_X(x_0)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\{x_i \leq x_0\}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{I}\{x_i \leq x_0\}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\Pr\{x_i \leq x_0\} = \Pr\{x_1 \leq x_0\} \\ &= F_X(x_0)\end{aligned}$$

$$\Rightarrow b\left(\hat{F}_X(x_0)\right) = \mathbb{E}\left[\hat{F}_X(x_0)\right] - F_X(x_0) = 0$$

   Hence, $\hat{F}_X$ is unbiased.

2. In Lecture 1 we proved that:
$$\text{MSE}\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = b^2\left(\hat{\theta}\right) + V\left(\hat{\theta}\right)$$

$$\begin{aligned}\Rightarrow \text{MSE}\left(\hat{F}_X(x_0)\right) &= \underbrace{b^2\left(\hat{F}_X(x_0)\right)}_{=0} + \text{Var}\left(\hat{F}_X(x_0)\right) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\underbrace{\boldsymbol{I}\{x_i \leq x_0\}}_{\triangleq Y_i}\right) \\ &= \frac{1}{N^2}\text{Var}\left(\sum_{i=1}^{N}Y_i\right), \qquad \{Y_i\}_i \text{ are i.i.d} \\ &= \frac{1}{N^2}N\text{Var}(Y_1) = \frac{\mathbb{E}\left[Y_1^2\right] - \mathbb{E}^2[Y_1]}{N} \\ &= \frac{F_X(x_0) - F_X^2(x_0)}{N}\end{aligned}$$

3. Using the previous result, we have:
$$\text{MSE}\left(\hat{F}_X(x_0)\right) \xrightarrow[N\to\infty]{} 0$$

2.2 Histogram

Let $\{x_i\}_{i=1}^{N}$ be N i.i.d realizations of $X \in \mathcal{X}$.
We can split the domain $\mathcal{X}$ into K **disjoint** intervals $\{R_k\}_{k=1}^{K}$ (see figure below) such that:

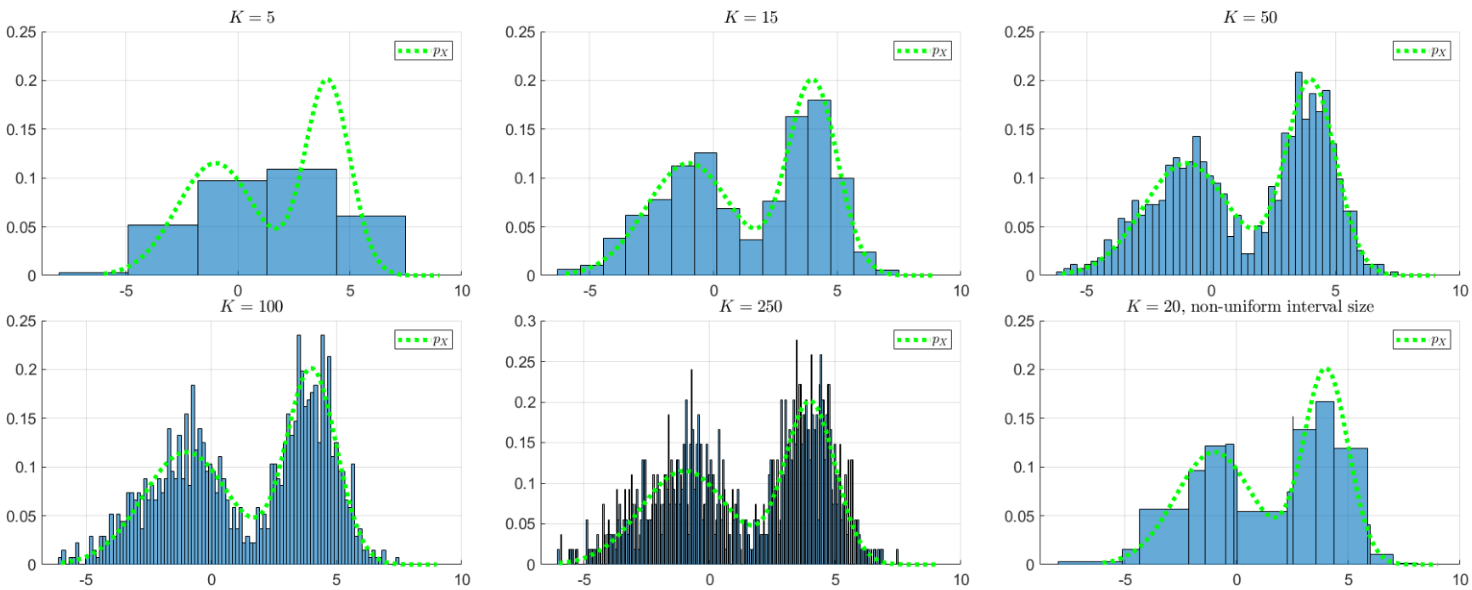
$$\mathcal{X} = \bigsqcup_{k=1}^{K} R_k, \qquad \text{(disjoint union)}$$

Then, for any $x \in R_k$ we estimate the PDF by:

$$\hat{p}_X(x) = \frac{1}{|R_k|}\frac{1}{N}\underbrace{\sum_{i=1}^{N} \boldsymbol{I}\{x_i \in R_k\}}_{(*)}, \qquad x \in R_k$$

where $|R_k|$ is the length (or volume) of R_k.
In words, $(*)$ is the number of realizations inside the interval R_k.

Example:
Let $\{x_i\}_{i=1}^{N}$ be $N = 1,000$ i.i.d realizations from an unknown p_X.
We plot the histogram for different values of K:



Notes:

- Remember the bias variance tradeoff.

$$\text{MSE}\left(\hat{\theta}\right) = b^2\left(\hat{\theta}\right) + V\left(\hat{\theta}\right)$$

  For small values of K the variance is small but the bias is large,
  whereas for large values of K the bias is small but the variance is large.

- A reasonable choice is $K = \sqrt{N}$.

Exercise:
Show that $\hat{p}_X$ is a valid density function, namely show that:

1. $\hat{p}_X(x) \geq 0, \forall x$
2. $\int_{\mathcal{X}} \hat{p}_X(x)\,\mathrm{d}x = 1$

Solution:

1. $\hat{p}_X(x) \geq 0$ is immediate from the definition

$$\hat{p}_X(x) = \frac{1}{|R_k|}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\{x_i \in R_k\} \geq 0$$

2.

$$\begin{aligned}
\int_{\mathcal{X}} \hat{p}_X(x)\,\mathrm{d}x &= \sum_{k=1}^{K}\int_{R_k}\hat{p}_X(x)\,\mathrm{d}x \\
&= \sum_{k=1}^{K}\int_{R_k}\frac{1}{|R_k|}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\{x_i \in R_k\}\,\mathrm{d}x \\
&= \sum_{k=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\{x_i \in R_k\}\underbrace{\int_{R_k}\frac{1}{|R_k|}\mathrm{d}x}_{=1} \\
&= \frac{1}{N}\underbrace{\sum_{k=1}^{K}\sum_{i=1}^{N}\boldsymbol{I}\{x_i \in R_k\}}_{=N} \\
&= 1
\end{aligned}$$

2.3 Kernel Density Estimation

When there are not enough data points to properly estimate p_X, one should consider using KDE.

For a given kernel h, the Kernel Density Estimation (KDE) is given by:

$$\hat{p}_h(x) = \frac{1}{N}\sum_{i=1}^{N} h(x - x_i)$$

where:

1. $h \geq 0$
2. $\int_{\mathcal{X}} h(x)\,\mathrm{d}x = 1$

Common choices for h are:

1. Rectangular window:

$$h(x) = \frac{1}{\alpha}\begin{cases} 1 & |x| \leq \frac{\alpha}{2} \\ 0 & \text{else} \end{cases}, \qquad \alpha > 0$$

2. Gaussian window:

$$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \qquad \sigma^2 > 0$$

- Note: We can write $\hat{p}_h(x)$ as a convolution with h (δ is the Dirac delta function):

$$\hat{p}_h(x) = \frac{1}{N}\sum_{i=1}^{N} h(x - x_i) = \left(\frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)\right) * h$$

Example:

Let $\{x_i\}_{i=1}^{N}$ be $N = 100$ i.i.d realizations from the unknown p_X.