Density Estimation

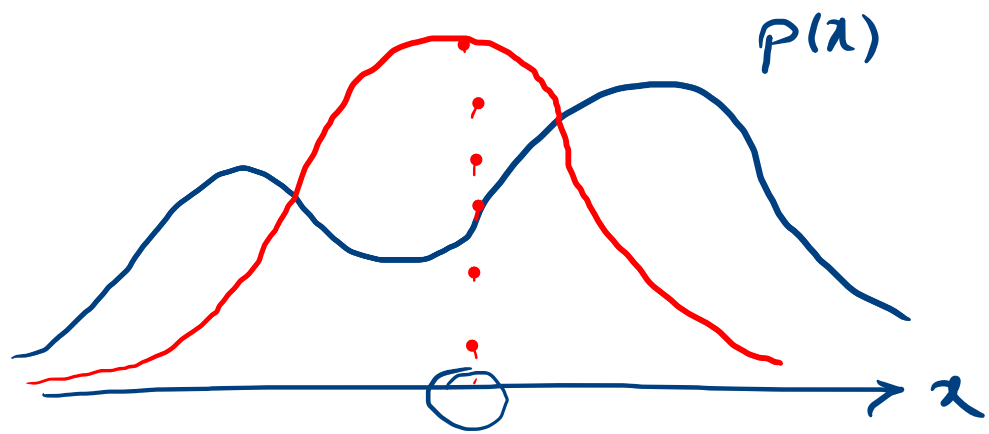
Parametric

$x \sim P(x|\theta)$

Estimate θ

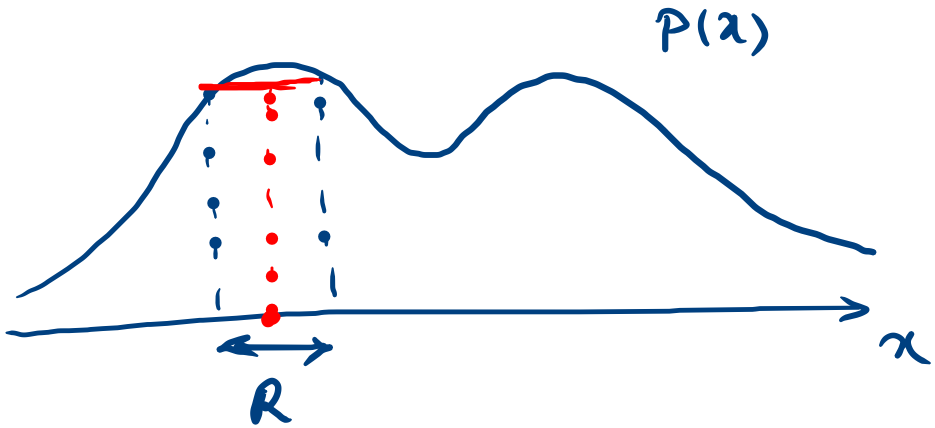
non-parametric

$P(x)$



$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$D = \{ x_1, \ldots, x_n \}$ $\qquad P(x) = ?$



$$P(x \in R) = \underbrace{\int_R p(x)\, dx}\simeq \frac{K}{n}$$

$$V\, P(x) = \frac{K}{n} \Rightarrow \boxed{P(x) = \frac{K}{nV}}$$

دلخواه

$$\boxed{P_n(x) = \frac{k_n}{n V_n}}$$

$n \longrightarrow \infty \overset{?}{\Longrightarrow} P_n(x) \overset{?}{\longrightarrow} P(x)$

① $\lim\limits_{n\to\infty} V_n \to 0$ $\quad V_n = \frac{1}{\sqrt{n}} \quad V_n = \frac{1}{\ln n}$

② $\lim\limits_{n\to\infty} k_n \to \infty$ $\quad k_n = \sqrt{n}$

③ $\lim\limits_{n\to\infty} \frac{k_n}{n} \to 0$ Or $\lim\limits_{n\to\infty} nV_n \to \infty$

✓

Conditions for convergence

- Let n be the **number of samples** used, R_n be the region used with n samples, $\mathbf{V_n}$ be the volume of R_n, k_n be the number of samples falling in R_n, and the estimate for p(x) be

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

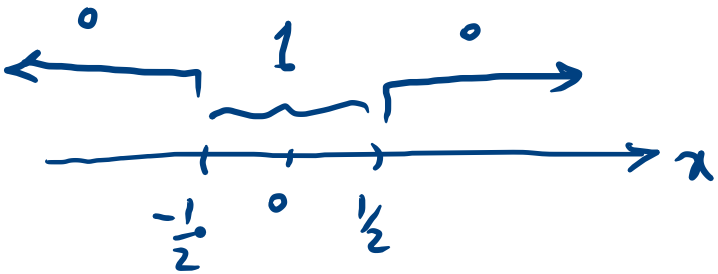
- If $p_n(x)$ is to **converge** to p(x), three **conditions** are required:

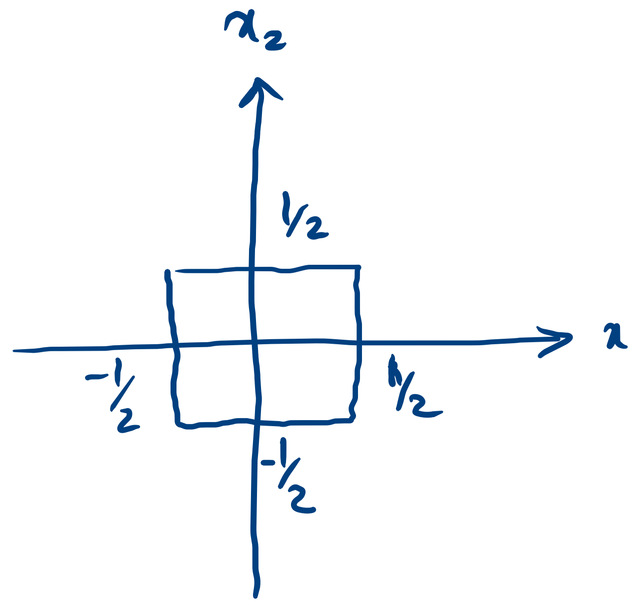
$$\lim_{n\to\infty} V_n = 0$$

$$\lim_{n\to\infty} k_n = \infty$$

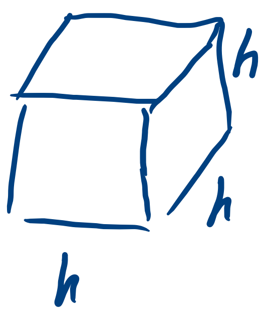
$$\lim_{n\to\infty} \frac{k_n}{n} = 0.$$

$$\phi(x) = \begin{cases} 1 & |x_i| \le 1/2 \\ 0 & \text{o.w.} \end{cases} \qquad i = 1, \dots, d \qquad x \in \mathbb{R}^d$$





$$\phi\left(\frac{x - x_0}{h}\right)$$



$$\int \phi\left(\frac{x - x_0}{h}\right) dx = h^d$$

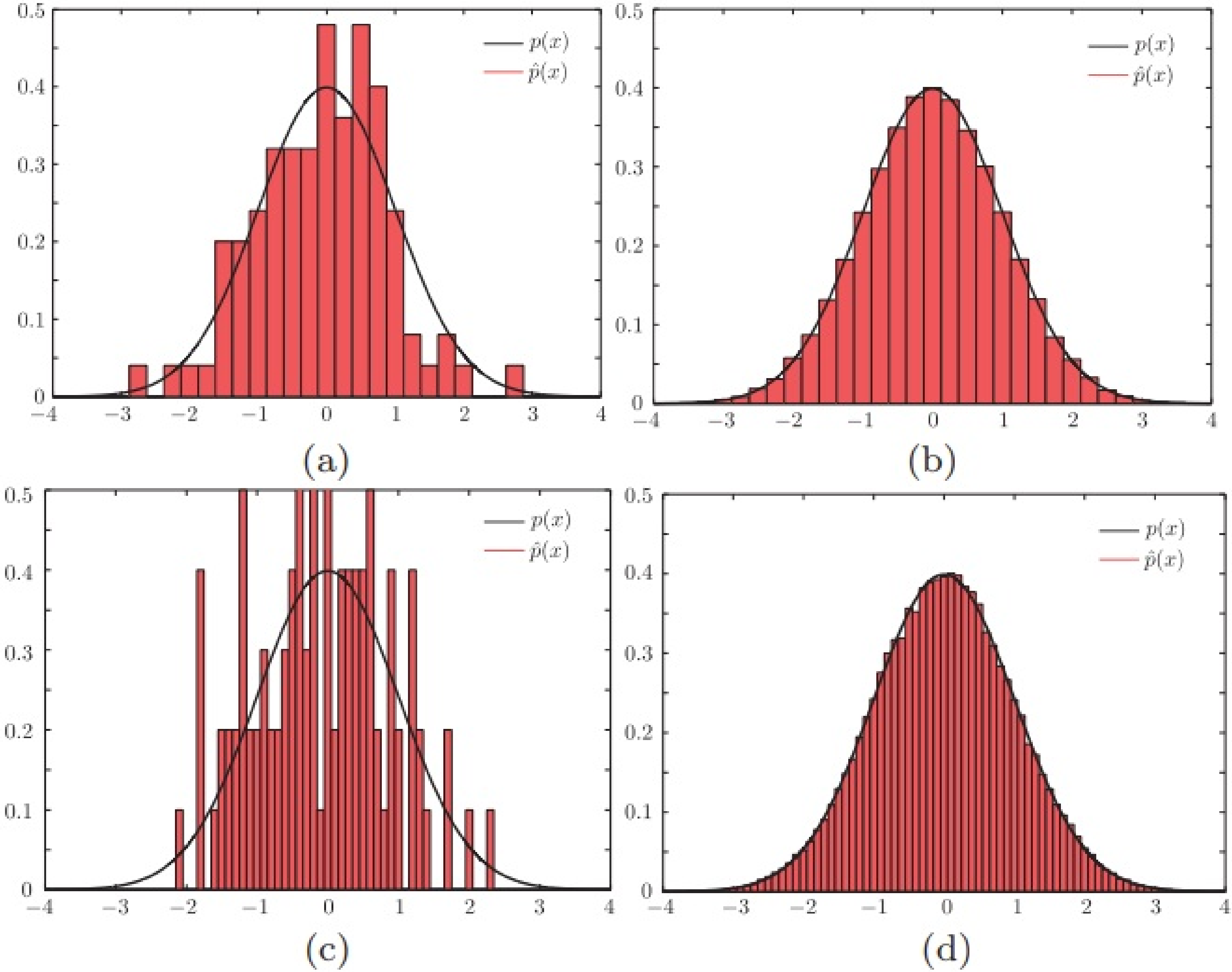
$$P_n(x) = \frac{k_n}{nV_n} = \frac{1}{nV_n} \underbrace{\sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h_n}\right)}_{k_n}$$

$$V_n = h_n^d$$

$$P(x)$$

$$\begin{cases} \int \phi(x)\,dx = 1 \\ \phi(x) \geq 0 \end{cases}$$

$$\boxed{\frac{1}{V_n} \phi\left(\frac{x - x_0}{h}\right)}$$



10^5

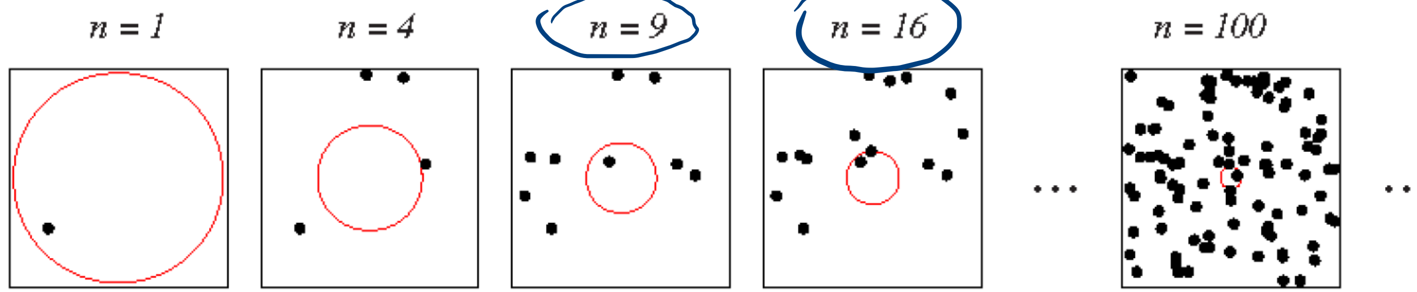
10^5

Two methods for estimating the density at a point **x** (at the center of each square)

- **Parzen window**

1/3 1/4

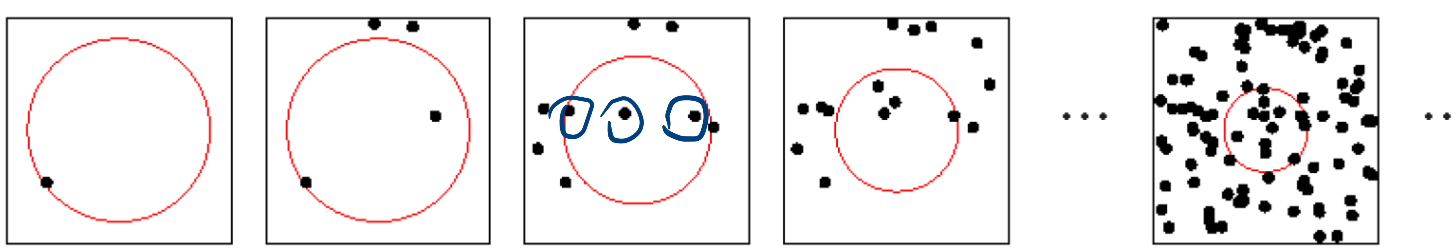
$V_n = 1/\sqrt{n}$



- **k-nearest neighbor**

$n \to 9$ $n \to 100$

$k_n = \sqrt{n}$

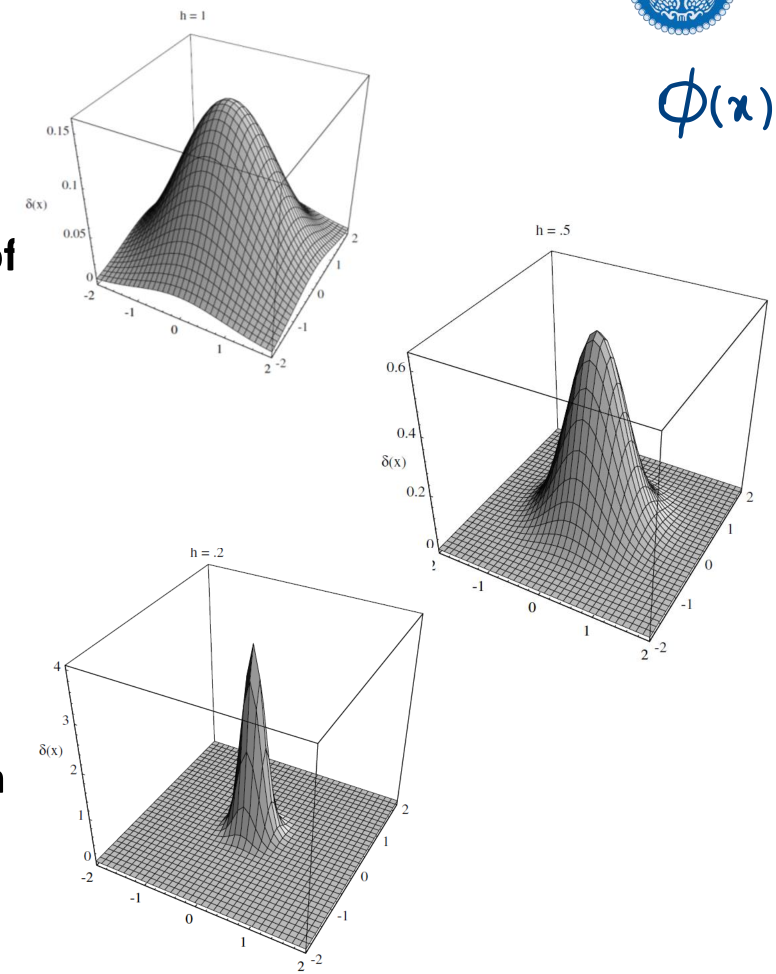


The role of h_n

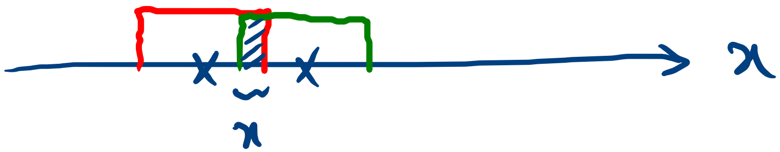
- If h_n is very **large**, $p_n(x)$ is the superposition of n **broad functions**, and is a smooth **“out-of focus”** estimate of p(x).
- If h_n is very **small**, $p_n(x)$ is the superposition of n **sharp pulses centered at the samples**, and is a “**noisy**” estimate of p(x).
- As h_n approaches zero, $d_n(x - x_i)$ approaches a **Dirac delta** function centered at x_i, and $p_n(x)$ is a superposition of delta functions.

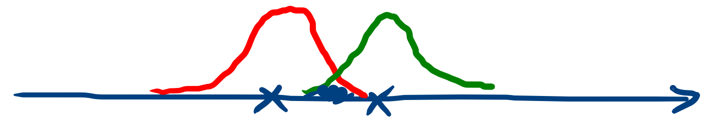
$\phi(x) = N(0,1)$

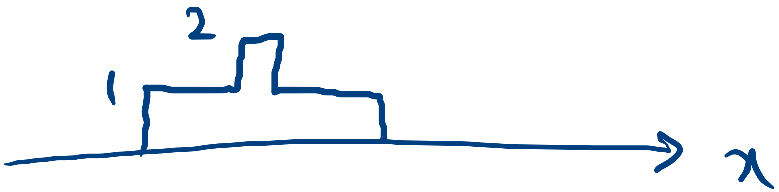
$\phi\left(\frac{x - x_0}{h}\right)$

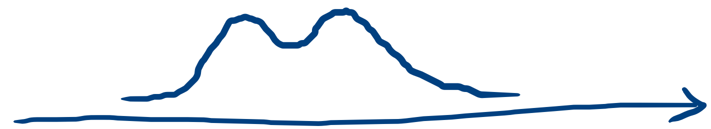


f(u) as a **function** of h_n





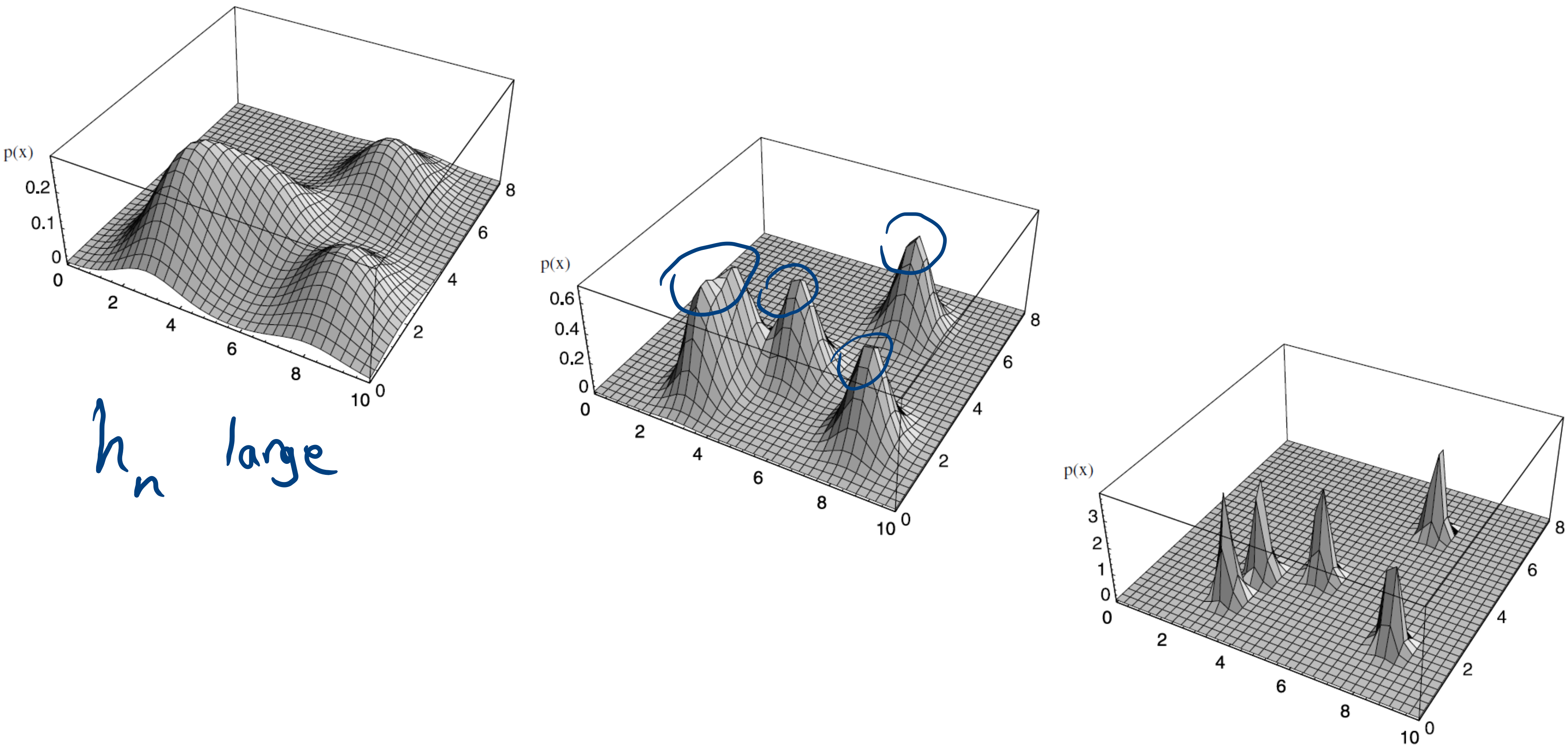




$$P_n(x) = \frac{1}{n V_n} \sum \Phi\left(\frac{x - x_i}{h}\right)$$

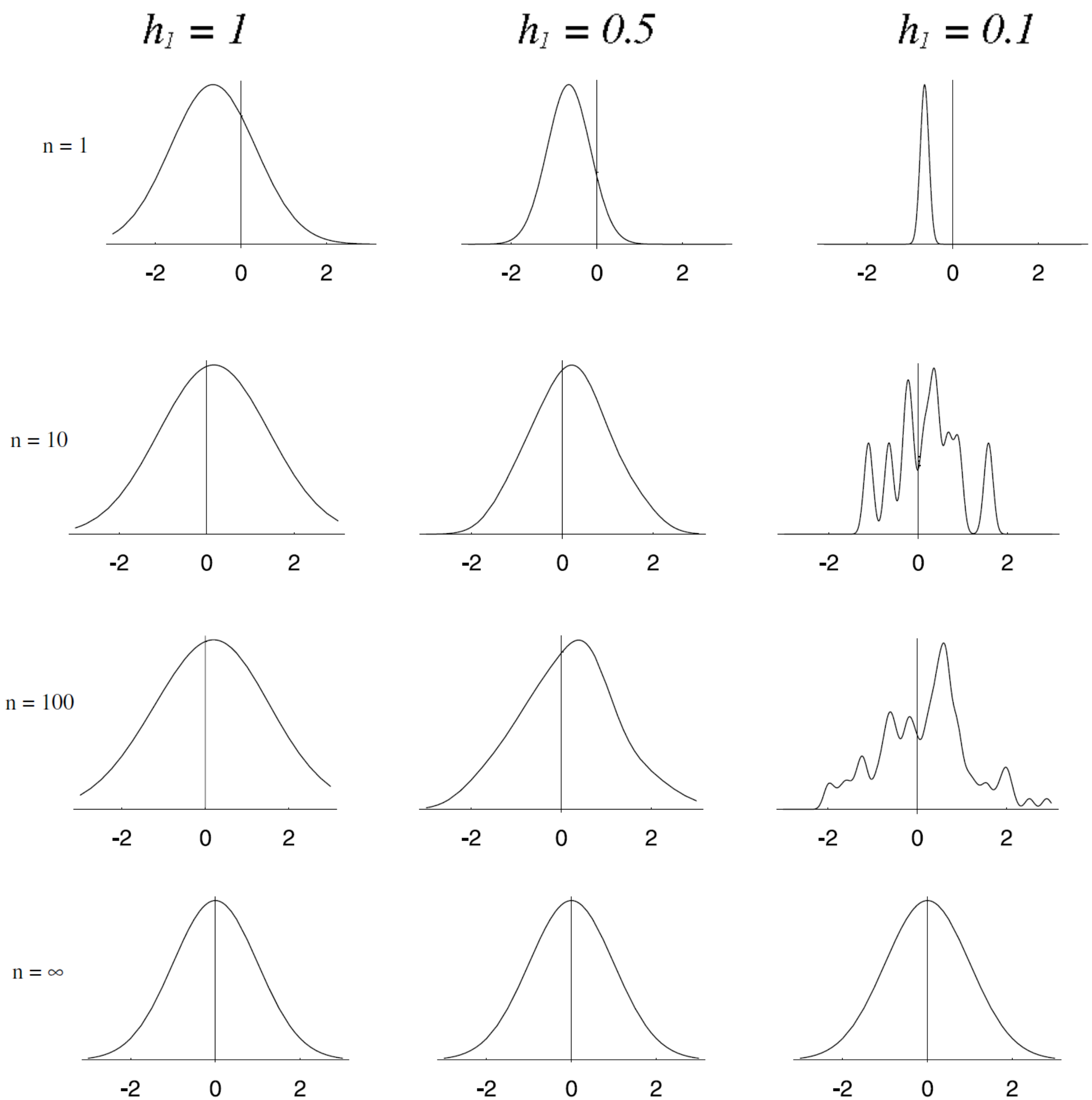
$p_n(x)$ as a function of h_n

Parzen window density estimates based on the same set of five samples using the window functions in the previous figure



h_n large

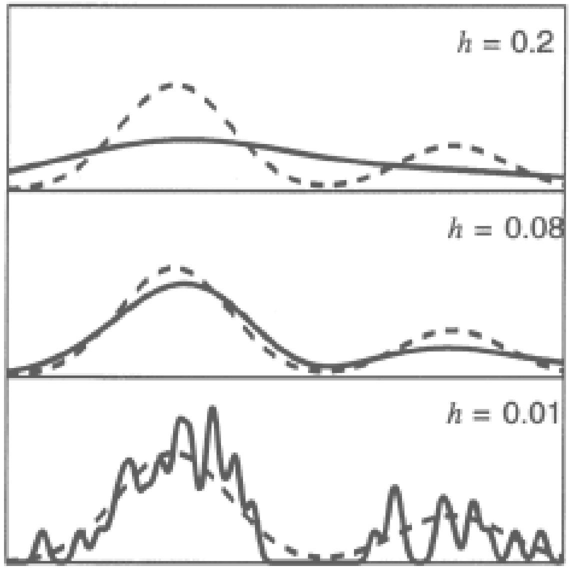
both p(**x**) and f(u) are Gaussian



$p_n(x)$ is **an average of normal** densities centered at the samples

$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}.$$

$$h_n = h_1/\sqrt{n},$$

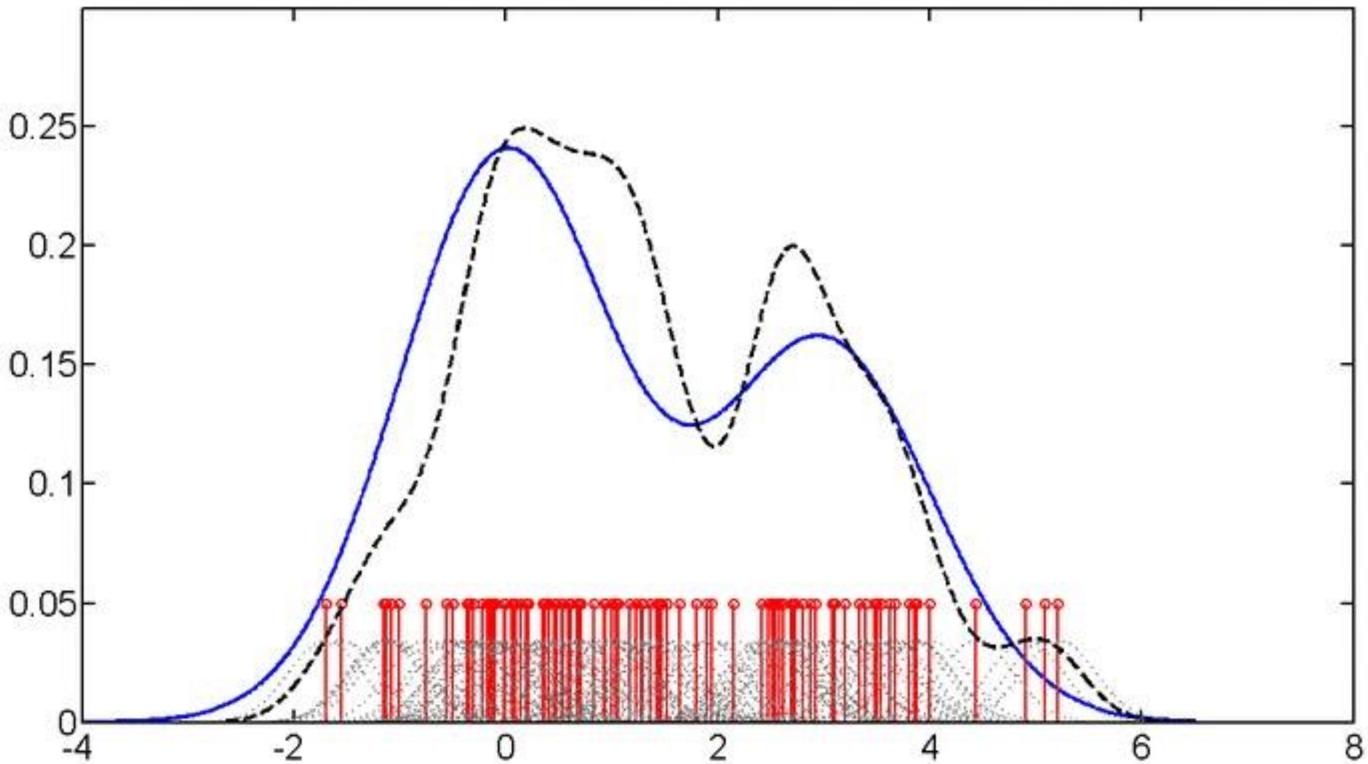


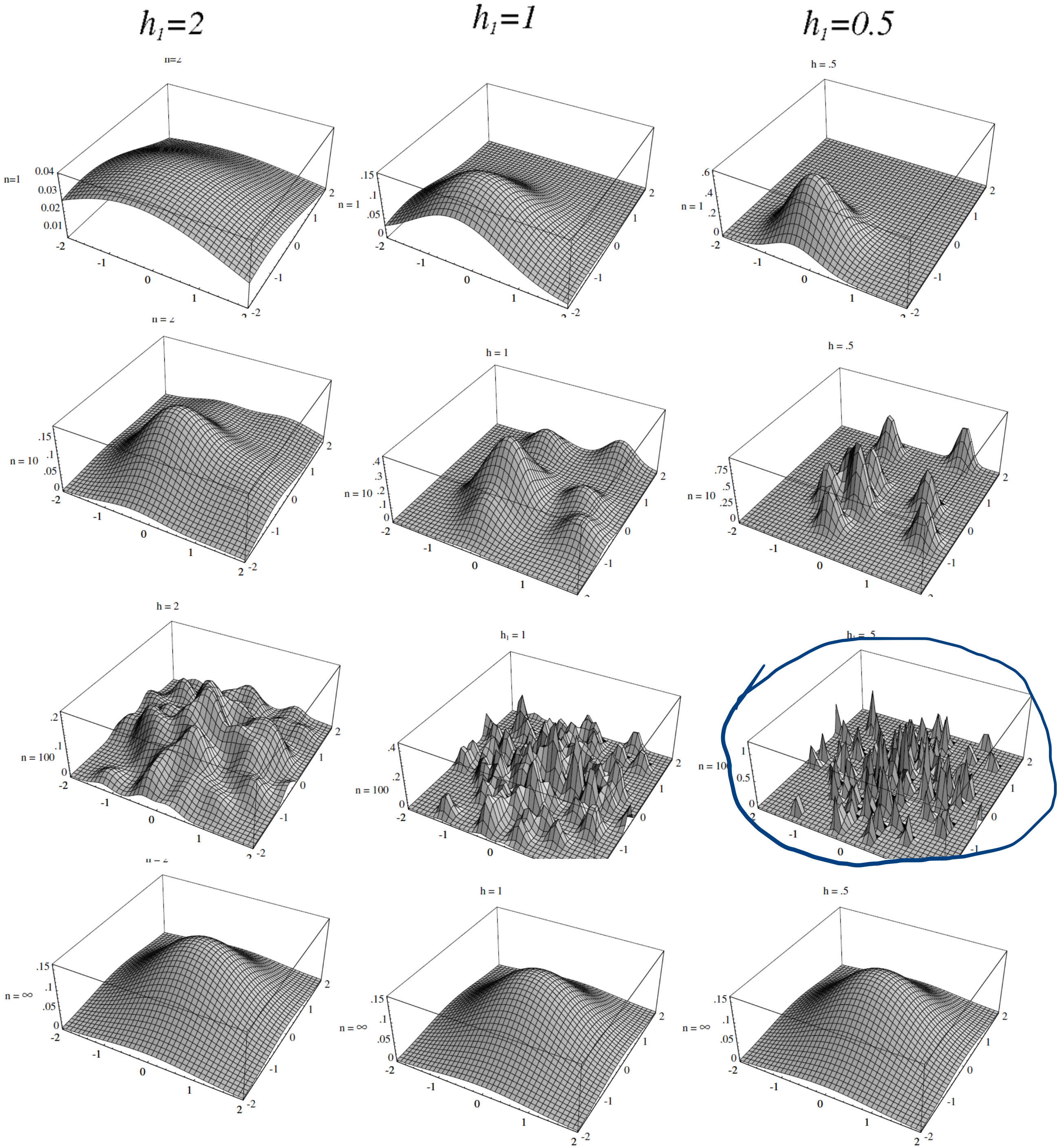
Parzen window

Blue curve: true density is mixture of two Gaussians centered around 0 and 3

In each frame, 100 samples are generated from the distribution, shown in **red**

Dashed black curve: averaging the Gaussians yields the density estimate



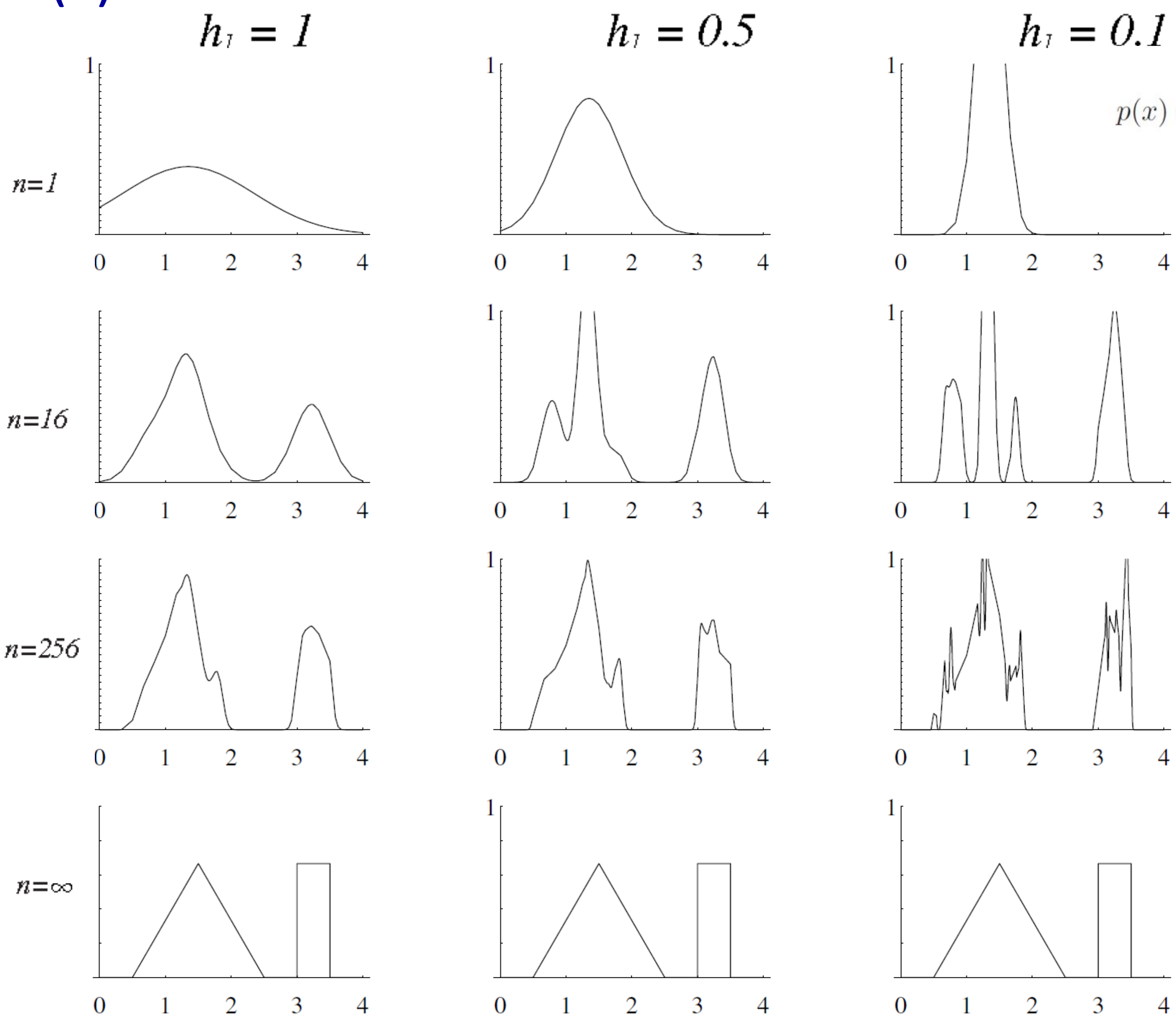


Parzen-window estimates of a **bivariate** normal

$$\varphi(\mathbf{u}) = N(\mathbf{0}, \mathbf{I})$$

$$h_n = h_1/\sqrt{n}.$$

$p(\mathbf{x})$ consists of a uniform and triangular density and f(x) is Gaussian



$$p(x) = \begin{cases} 1 & -2.5 < x < -2 \\ 1/4 & 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

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

$$h_n = h_1/\sqrt{n},$$

Classification using kernel-based density estimation (Bayesian decision rule)

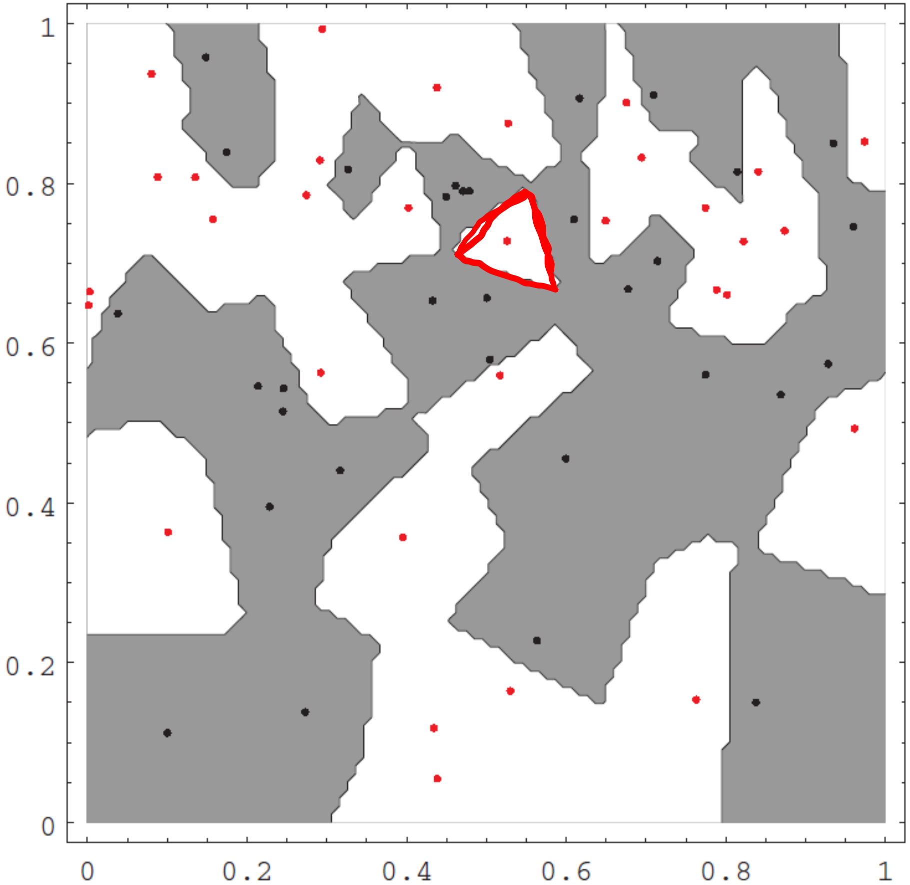
- Estimate density for **each class**.
- **Classify** a test point by computing the **posterior probabilities** and **picking the max**.
- The **decision regions** depend on the choice of the **kernel function** and $\boldsymbol{h_n}$.
- The training error can be made **arbitrarily low** by making the window width **sufficiently small**.
- However, the **goal is to classify novel** patterns so the window width **cannot be made** too small.

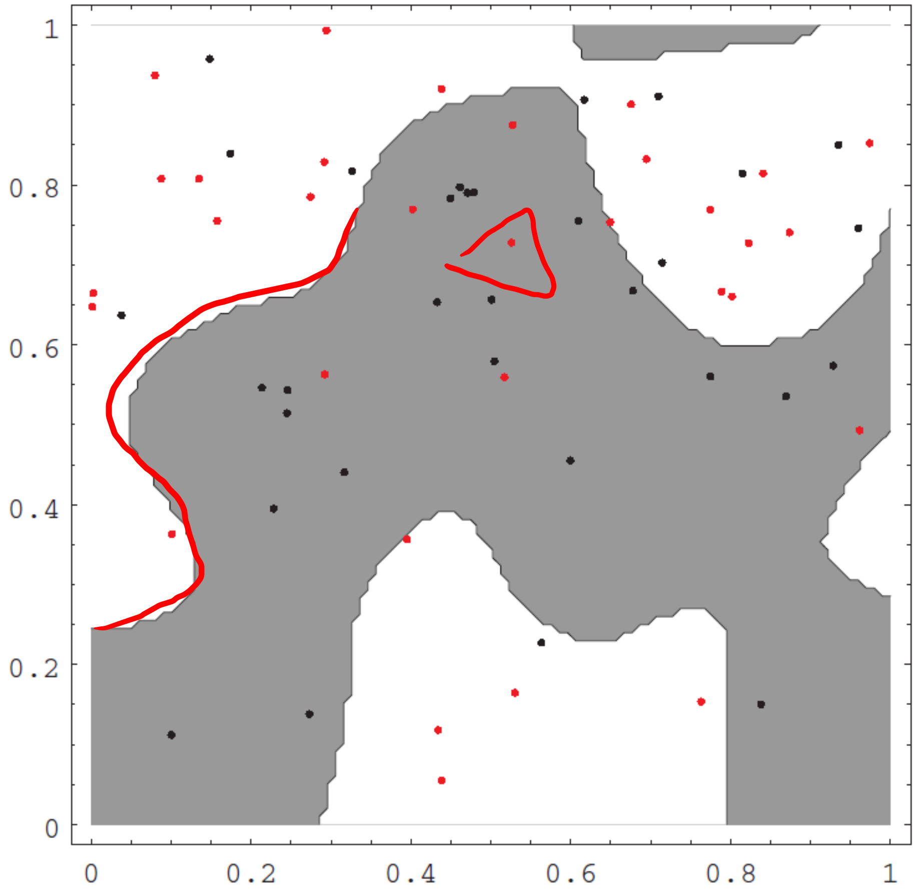
$$\frac{p(y=1 \mid x)}{p(y=2 \mid x)} = \frac{p(x \mid y=1)\, p(y=1)}{p(x \mid y=2)\, p(y=2)}$$

dimensional Parzen-window dichotomizer

- small h_n
- Boundaries are more complicated

- large h_n
- No single window width is ideal overall





Drawbacks of kernel-based methods

- Require a large number of samples.
- Require all the samples to **be stored**.
- Evaluation of the density could be **very slow** if the number of data points is large.
- Possible **solution**:
  - use **fewer kernels** and **adapt the positions** and widths in response to the data (e.g., mixtures of Gaussians!)

$$\lim_{n\to\infty} P_n(x) \longrightarrow P(x) \qquad x \sim P(x)$$

$$\mathrm{Var}(X)$$

$$\begin{cases} E[P_n(x)] = P(x) \\ \mathrm{Var}(P_n(x)) = 0 \end{cases}$$

R.V.

$$E[P_n(x)] = E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E\left[\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right]}_{g(x_i)}$$

$$= E\left[\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right] = \int \underbrace{\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)}_{\delta(x-x_i)} P(x_i)\,dx_i = \int \delta(x-x_i)\,P(x_i)\,dx_i$$

$$V_n \to 0$$

$$= P(x)$$

$$var(X_1 + X_2) = var(X_1) + var(X_2) \qquad var(X) = E[X^2] - E^2[X]$$

$$var(P_n(x)) = var\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} var\left(\frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)\right)$$

$$= \frac{1}{n} var\left(\frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)\right) = \frac{1}{n} E\left[\frac{1}{V_n^2}\phi^2\left(\frac{x - x_i}{h_n}\right)\right] - \frac{1}{n} E^2\underbrace{\left[\frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)\right]}_{\bar{P}_n(x)}$$

$$= \frac{1}{n}\int \frac{1}{V_n^2}\underbrace{\phi^2\left(\frac{x - x_i}{h_n}\right)}_{\phi_{max}\,\phi\left(\frac{x - x_i}{h_n}\right)} P(x_i)\, dx_i - \underbrace{\frac{1}{n}\bar{P}_n^2(x)}_{\geq 0}$$

$$\leq \frac{1}{n}\frac{\phi_{max}}{V_n}\int \overbrace{\frac{1}{V_n}\phi\left(\frac{x - x_i}{h_n}\right)}^{\delta(x - x_i)} p(x_i)\, dx_i$$

$$\lim_{n\to\infty} \operatorname{var}(p_n(x)) \leq \frac{\phi_{max}\, p(x)}{n V_n} \to 0$$

$$\lim_{n\to\infty} n V_n \to \infty$$

$$\boxed{\phi_{max} < \infty}$$