PARAMETER ESTIMATION-
PROBABILITY DISTRIBUTION ESTIMATION-
BAYESIAN INFERENCE

ESTIMATION OF UNKNOWN PROBABILITY DENSITY FUNCTIONS

❖ Maximum Likelihood

- Let $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N$ known and independent
- Let $p(\underline{x})$ known within an unknown vector parameter $\underline{\theta}$: $p(\underline{x}) \equiv p(\underline{x};\underline{\theta})$
- $X = \{\underline{x}_1, \underline{x}_2, \ldots \underline{x}_N\}$
- $p(X;\underline{\theta}) \equiv p(\underline{x}_1, \underline{x}_2, \ldots \underline{x}_N;\underline{\theta})$

$$= \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})$$

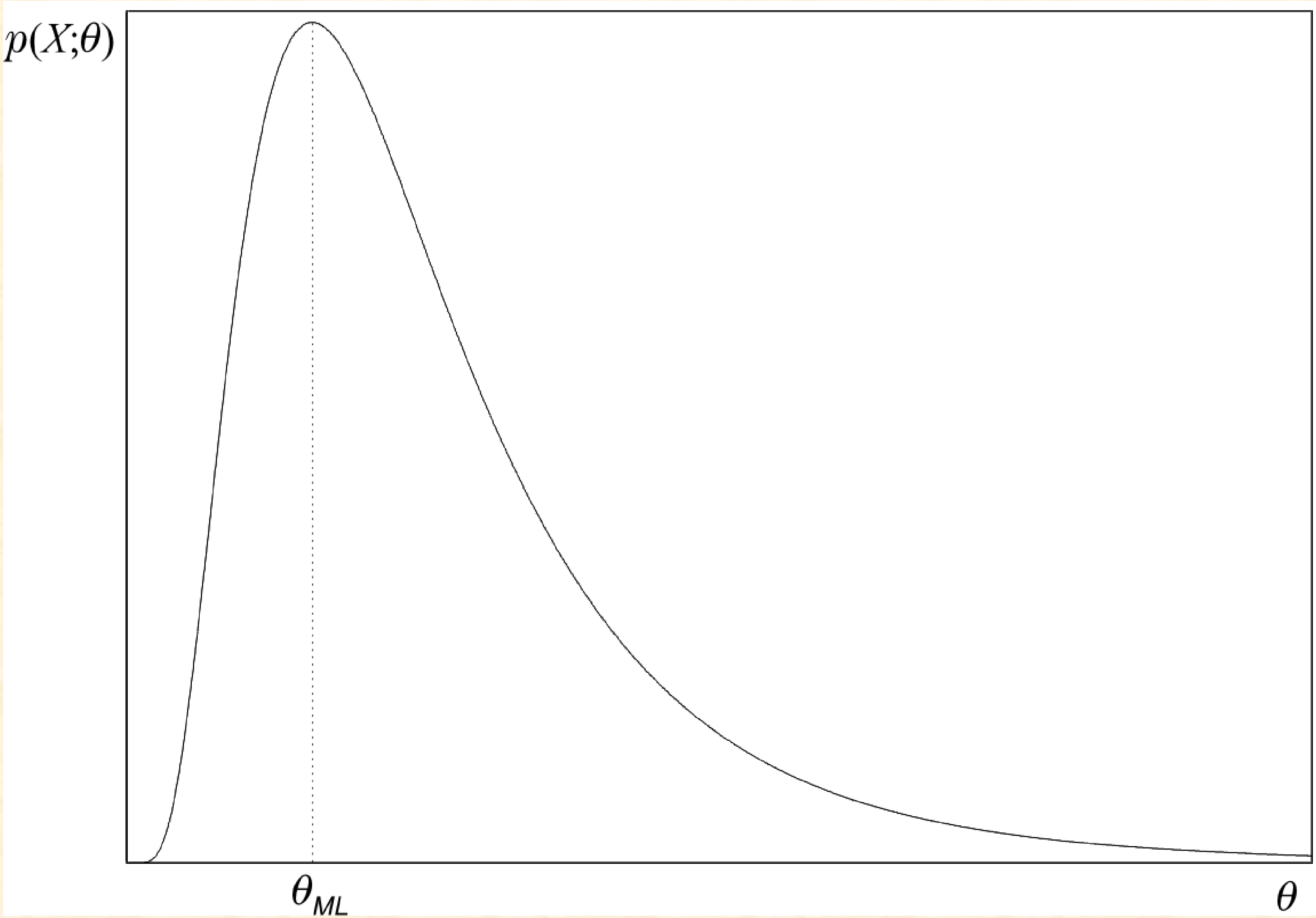
which is known as the Likelihood of $\underline{\theta}$ w.r. to X

The method:

- $\boxed{\hat{\underline{\theta}}_{\text{ML}} : \arg\max_{\underline{\theta}} \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})}$

- $L(\underline{\theta}) \equiv \ln p(X;\underline{\theta}) = \sum_{k=1}^{N} \ln p(\underline{x}_k;\underline{\theta})$

- $\hat{\underline{\theta}}_{ML} : \dfrac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} = \sum_{k=1}^{N} \dfrac{1}{p(\underline{x}_{k;}\underline{\theta})} \dfrac{\partial p(\underline{x}_{k;}\underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$


$p(X;\theta)$
θ_{ML}
θ

❖ Example:

$$p(\underline{x}):\ N(\underline{\theta},\Sigma):\ \underline{\theta}\ \text{unknown},\quad \underline{x}_1,\underline{x}_2,\dots,\underline{x}_N\quad p(\underline{x}_k)\equiv p(\underline{x}_k;\underline{\theta})$$

$$L(\underline{\theta})=\ln\prod_{k=1}^{N}p(\underline{x}_k;\underline{\theta})=C-\frac{1}{2}\sum_{k=1}^{N}(\underline{x}_k-\underline{\theta})^T\Sigma^{-1}(\underline{x}_k-\underline{\theta})$$

$$p(\underline{x}_k;\underline{\theta})=\frac{1}{(2\pi)^{\frac{l}{2}}\left|\Sigma\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\underline{x}_k-\underline{\theta})^T\Sigma^{-1}(\underline{x}_k-\underline{\theta})\right)$$

$$\frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})}\equiv\begin{bmatrix}\frac{\partial L}{\partial\theta_1}\\ \cdot\\ \cdot\\ \cdot\\ \frac{\partial L}{\partial\theta_l}\end{bmatrix}=\sum_{k=1}^{N}\Sigma^{-1}(\underline{x}_k-\underline{\theta})=\underline{0}\Rightarrow\underline{\theta}_{ML}=\frac{1}{N}\sum_{k=1}^{N}\underline{x}_k$$

Remember: if $A=A^T\Rightarrow\dfrac{\partial(\underline{\alpha}^T A\underline{\alpha})}{\partial\underline{\alpha}}=2A\underline{\alpha}$

❖ Maximum Aposteriori Probability Estimation

- In ML method, $\underline{\theta}$ was considered as a parameter
- Here we shall look at $\underline{\theta}$ as a random vector described by a pdf $p(\underline{\theta})$, assumed to be known
- Given

$$X = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$$

Compute the maximum of

$$p(\underline{\theta} | X)$$

- From Bayes theorem

$$p(\underline{\theta}) p(X | \underline{\theta}) = p(X) p(\underline{\theta} | X) \text{ or}$$

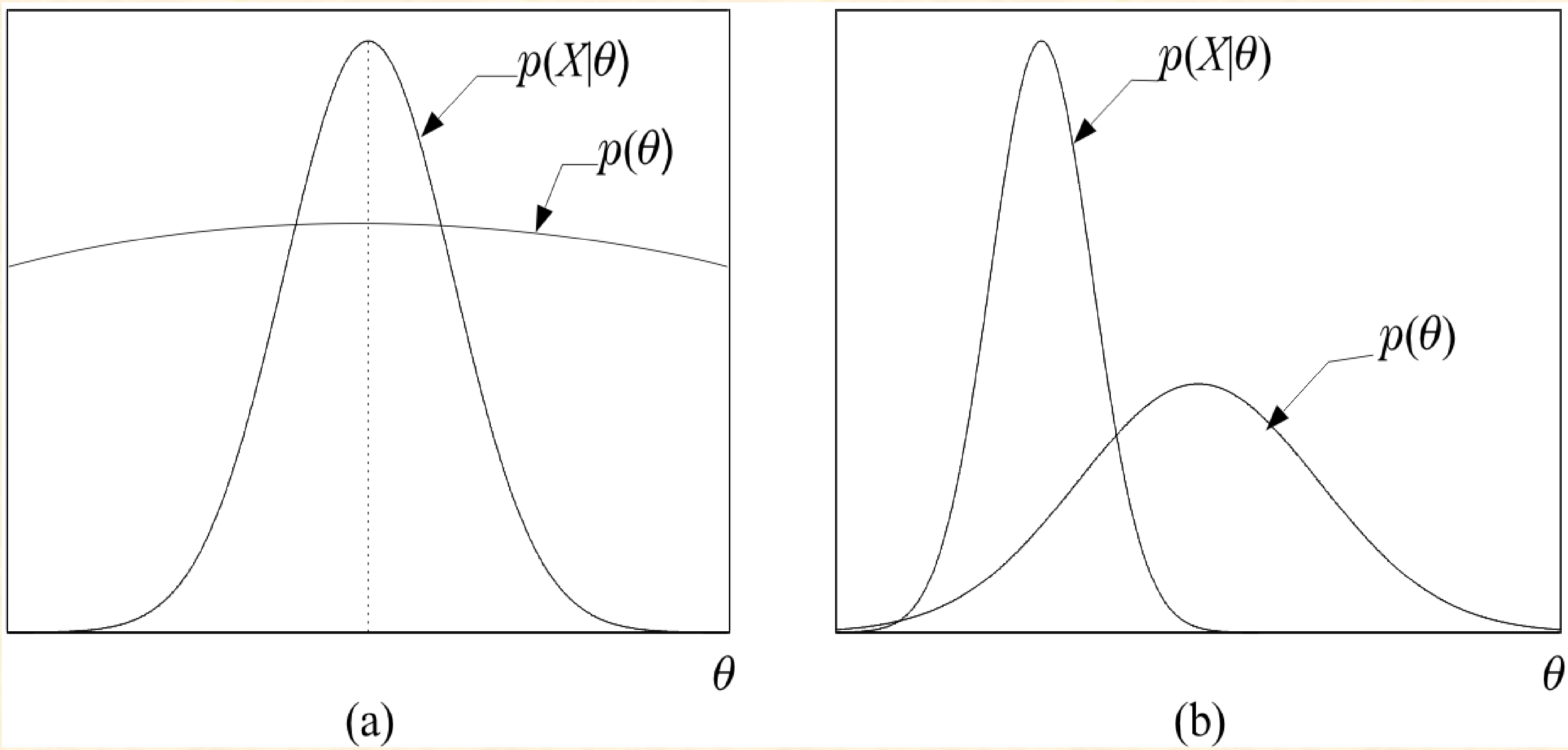
$$p(\underline{\theta} | X) = \frac{p(\underline{\theta}) p(X | \underline{\theta})}{p(X)}$$

- The method:

$$\underline{\hat{\theta}}_{MAP} = \arg\max_{\underline{\theta}} p(\underline{\theta}|X) \text{ or}$$

$$\boxed{\underline{\hat{\theta}}_{MAP} : \frac{\partial}{\partial \underline{\theta}} (P(\underline{\theta}) p(X|\underline{\theta}))}$$

If $p(\underline{\theta})$ is uniform or broad enough $\underline{\hat{\theta}}_{MAP} \cong \underline{\theta}_{ML}$



$p(X|\theta)$
$p(\theta)$
θ
(a)
$p(X|\theta)$
$p(\theta)$
θ
(b)

❖ Example:

$$\theta \ \text{ unknown, } \ let \ p(x|\theta) \rightarrow N(\theta, \sigma_\eta^2), X = \{x_1, ..., x_N\}$$

$$p(\theta) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma_0} \exp\left(-\frac{\|\theta - \theta_0\|^2}{2\sigma_0^2}\right)$$

$$\hat{\theta}_{MAP}: \quad \frac{\partial}{\partial\theta} \ln\left(\prod_{k=1}^{N} p(x_k|\theta) p(\theta)\right) = 0 \quad \text{or} \quad \sum_{k=1}^{N} \frac{1}{\sigma_\eta^2}(x_k - \theta) - \frac{1}{\sigma_0^2}(\theta - \theta_0) = 0 \Rightarrow$$

$$\hat{\theta}_{MAP} = \frac{N\overline{x} + \theta_0 \frac{\sigma_\eta^2}{\sigma_0^2}}{N + \frac{\sigma_\eta^2}{\sigma_0^2}} \quad For \quad \frac{\sigma_\eta^2}{\sigma_0{}^2} << 1, \text{ or for N} \rightarrow \infty, \quad \hat{\theta}_{MAP} \cong \hat{\theta}_{ML} = \overline{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$For \quad \frac{\sigma_\eta^2}{\sigma_0{}^2} >> 1, \quad \hat{\theta}_{MAP} \cong \theta_0$$

❖ Bayesian Inference

➢ $\text{ML}, \text{MAP} \Rightarrow$ a single estimate for $\underline{\theta}$.

Here a different root is followed.

Given: $X = \{\underline{x}_1, \ldots, \underline{x}_N\}$, $p(\underline{x}|\underline{\theta})$ and $p(\underline{\theta})$

The goal: estimate $\mathrm{p}(\underline{x}|X)$

How??

$$p(\underline{x}|X) = \int p(\underline{x}|\underline{\theta})\, p(\underline{\theta}|X)\, d\underline{\theta}$$

$$p(\underline{\theta}|X) = \frac{p(X|\underline{\theta})\, p(\underline{\theta})}{p(X)} = \frac{p(X|\underline{\theta})\, p(\underline{\theta})}{\int p(X|\underline{\theta})\, p(\underline{\theta})\, d\underline{\theta}}$$

$$p(X|\underline{\theta}) = \prod_{k=1}^{N} p(\underline{x}_k|\underline{\theta})$$

A bit more insight via an example

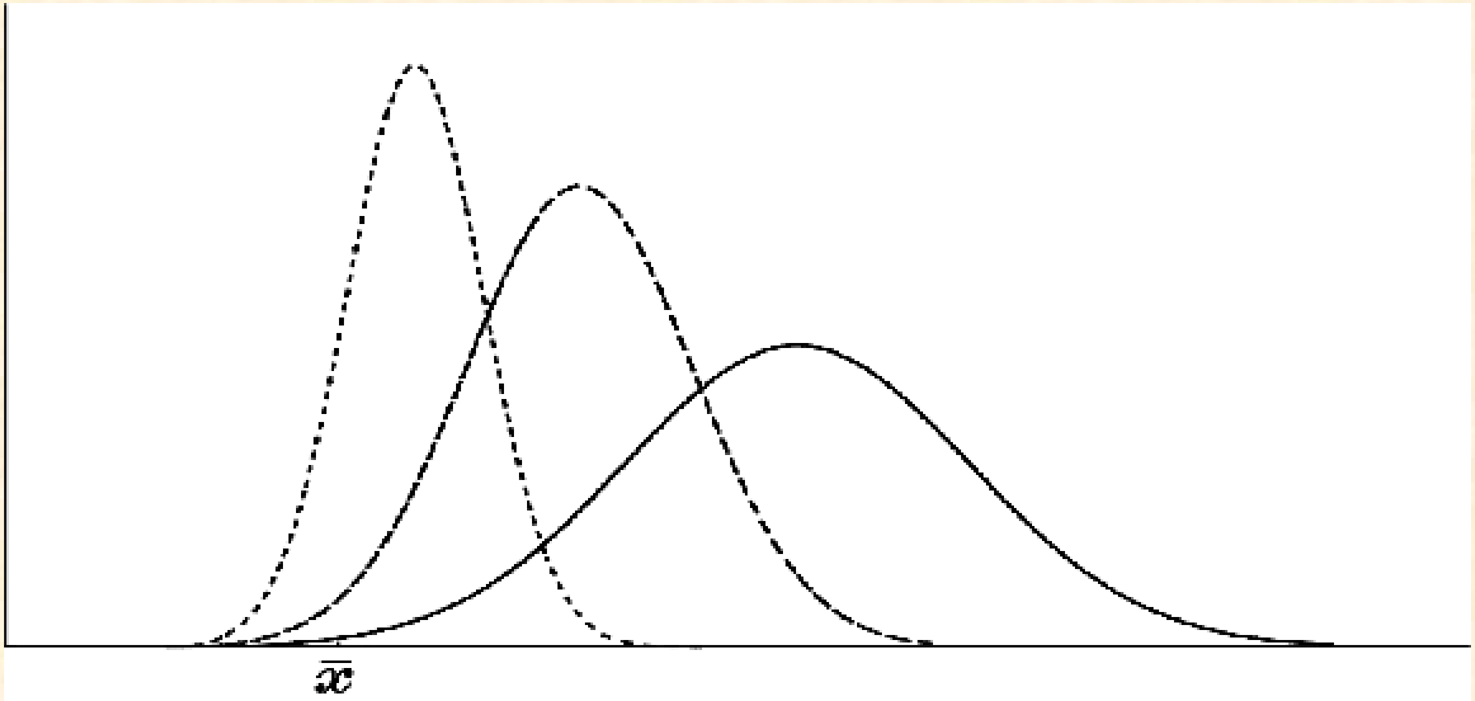
- $Let\ p(x|\theta) \to N(\theta, \sigma_\eta^2)$
- $p(\theta) \to N(\theta_0, \sigma_0^2)$
- It turns out that: $p(\theta|X) \to N(\theta_N, \sigma_N^2)$

$$\theta_N = \frac{N\sigma_0^2 \bar{x} + \sigma_\eta^2 \theta_0}{N\sigma_0^2 + \sigma_\eta^2}, \qquad \sigma_N^2 = \frac{\sigma_\eta^2 \sigma_0^2}{N\sigma_0^2 + \sigma_\eta^2}, \qquad \bar{x} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

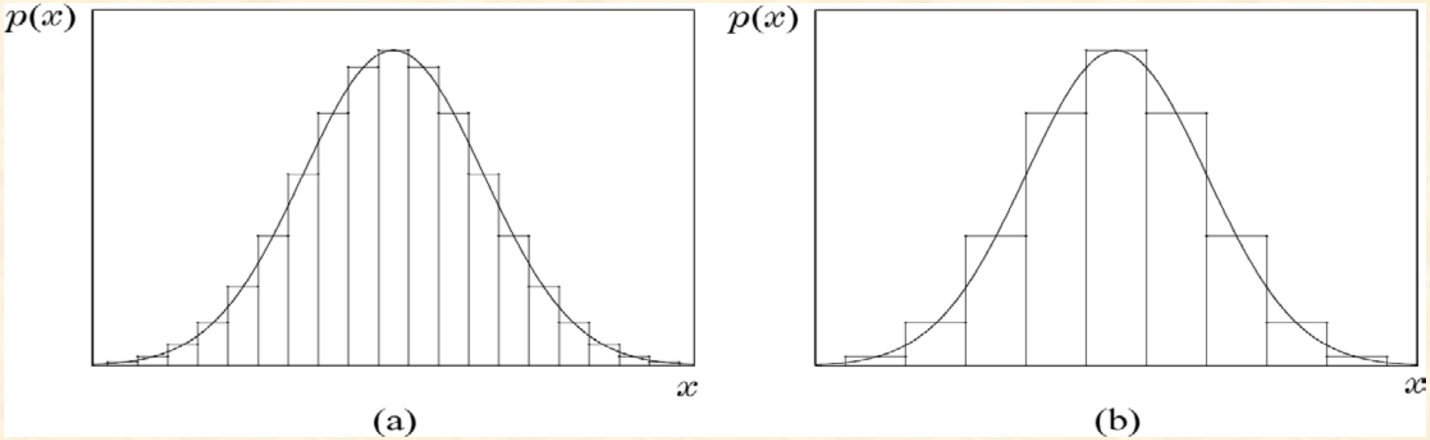
- Also: $p(x|X) \to N(\theta_N, \sigma_x^2), \quad \sigma_x^2 = \sigma_\eta^2 + \frac{\sigma_\eta^2 \sigma_N^2}{\sigma_N^2 + \sigma_\eta^2}$

- Bayesian inference has given us a result similar to MAP
- Same as MAP concerning the estimation of the parameter
- But: Additional information as regards the uncertainty of the estimate

- The above is a sequence of Gaussians as $N \to \infty$



❖ Parzen Windows

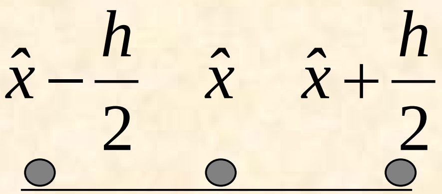


$$P \approx \frac{k_N}{N}$$

k_N in h

N total

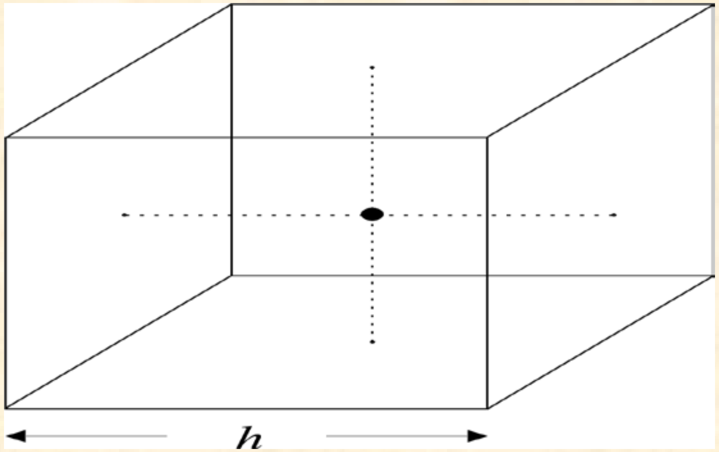
$$\hat{p}(x) \equiv \hat{p}(\hat{x}) = \frac{1}{h}\frac{k_N}{N}, \left|x - \hat{x}\right| \le \frac{h}{2}$$



If $p(x)$ continuous , $\hat{p}(x) \to p(x)$ as $N \to \infty$, if

$$h_N \to 0, \quad k_N \to \infty, \quad \frac{k_N}{N} \to 0$$

- Divide the multidimensional space in hypercubes



- Define

$$\phi(\underline{x}) = \begin{cases} 1, & |x_j| \le \frac{1}{2} \ \forall j \\ 0, & \text{otherwise} \end{cases}$$

- That is, it is 1 inside a unit side hypercube centered at 0
- $\boxed{\hat{p}(\underline{x}) = \frac{1}{h^l}\left(\frac{1}{N}\sum_{i=1}^{N}\varphi\left(\frac{\underline{x}_i - \underline{x}}{h}\right)\right)}$
- $\frac{1}{\text{volume}} * \frac{1}{N} *$ number of points inside an h-side hypercube centered at $\underline{x}$
- The problem: $p(\underline{x})$ continuous

  $\varphi(.)$ discontinuous
- Parzen windows-kernels-potential functions

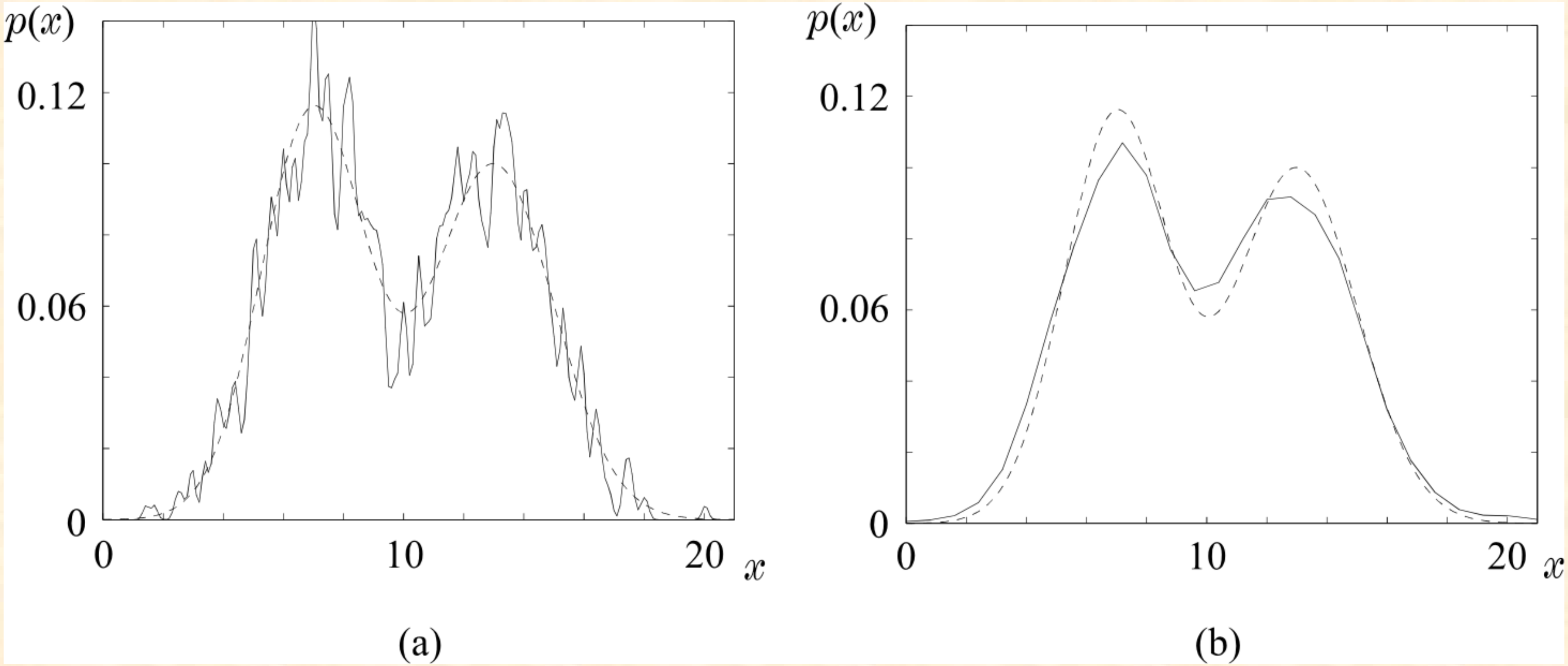
  $\varphi(\underline{x})$ is smooth

$$\varphi(\underline{x}) \ge 0, \ \int_{\underline{x}} \varphi(\underline{x})\, d\underline{x} = 1$$

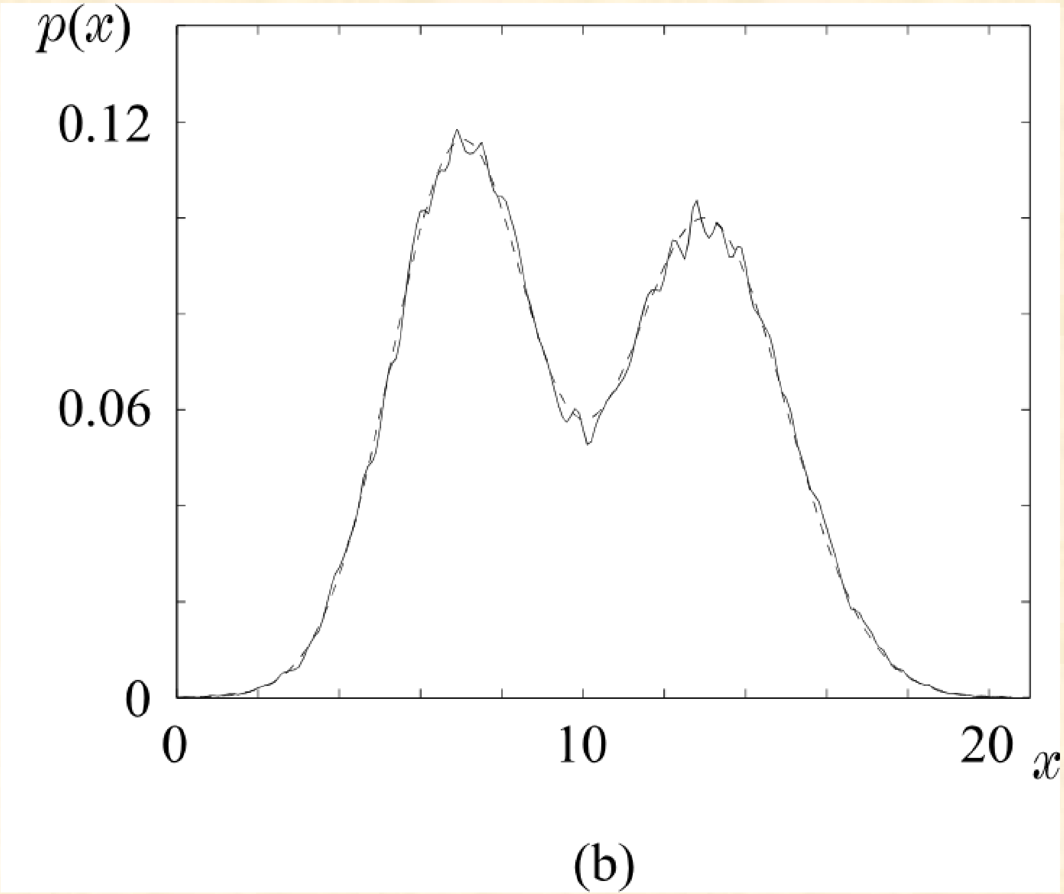
➢ Variance

- The smaller the h the higher the variance

h=0.1, N=1000 *h=0.8, N=1000*



h=0.1, N=10000



- The higher the N the better the accuracy

❖ Maximum Entropy

- Entropy

$$H = -\int p(\underline{x}) \ln p(\underline{x}) d\underline{x}$$

$$\hat{p}(x):\ \text{maximum}\ H\ \text{subject to the available constraints}$$

- Example: x is nonzero in the interval $x_1 \le x \le x_2$ and zero otherwise. Compute the ME pdf

  - The constraint:

$$\int_{x_1}^{x_2} p(x)dx = 1$$

  - Lagrange Multipliers

$$H_L = H + \lambda(\int_{x_1}^{x_2} p(x)dx - 1)$$

  - $\hat{p}(x) = \exp(\lambda - 1)$

$$\hat{p}(x) = \begin{cases} \dfrac{1}{x_2 - x_1} & x_1 \le x \le x_2 \\ 0 & \text{otherwise} \end{cases}$$