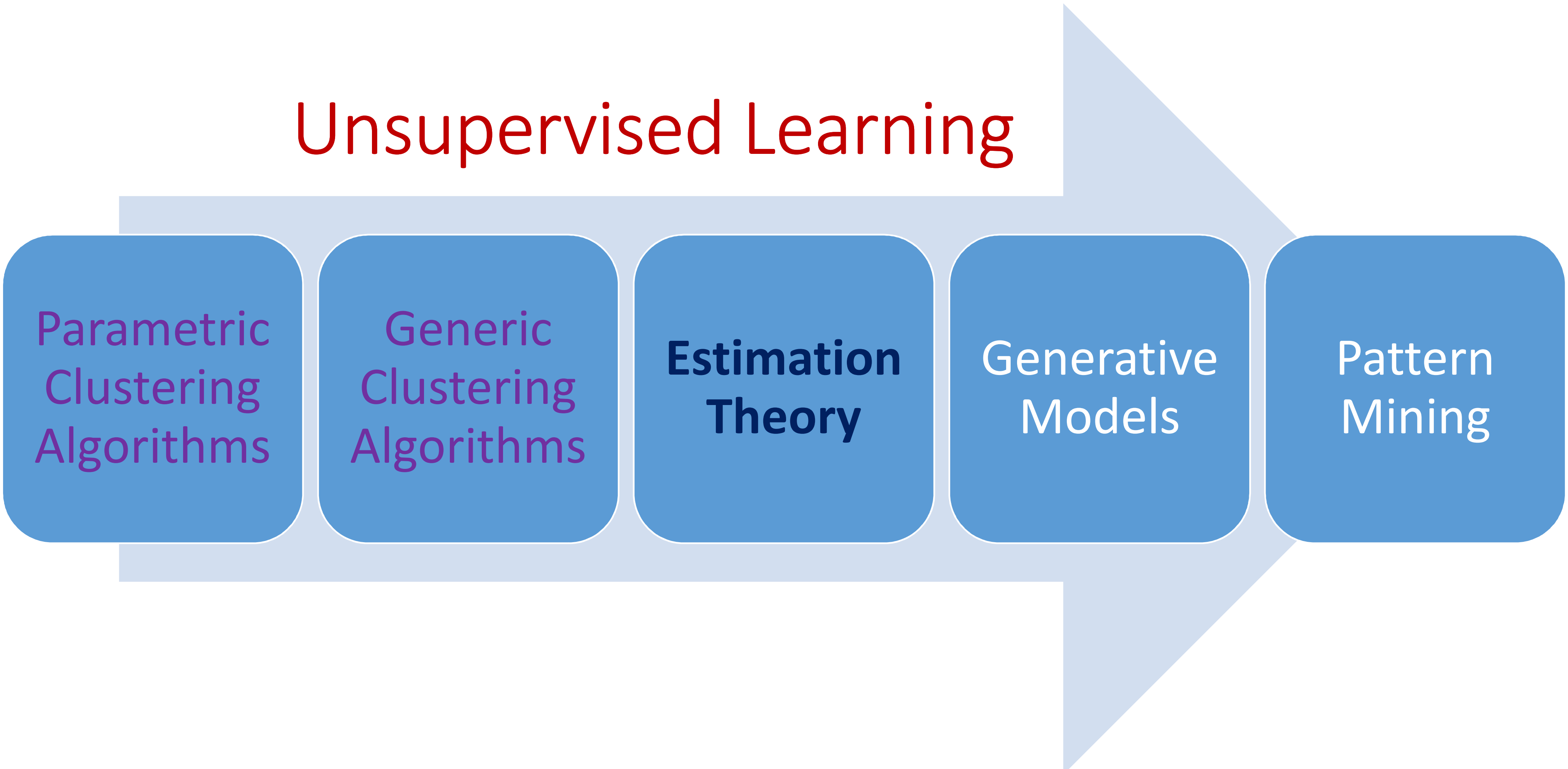
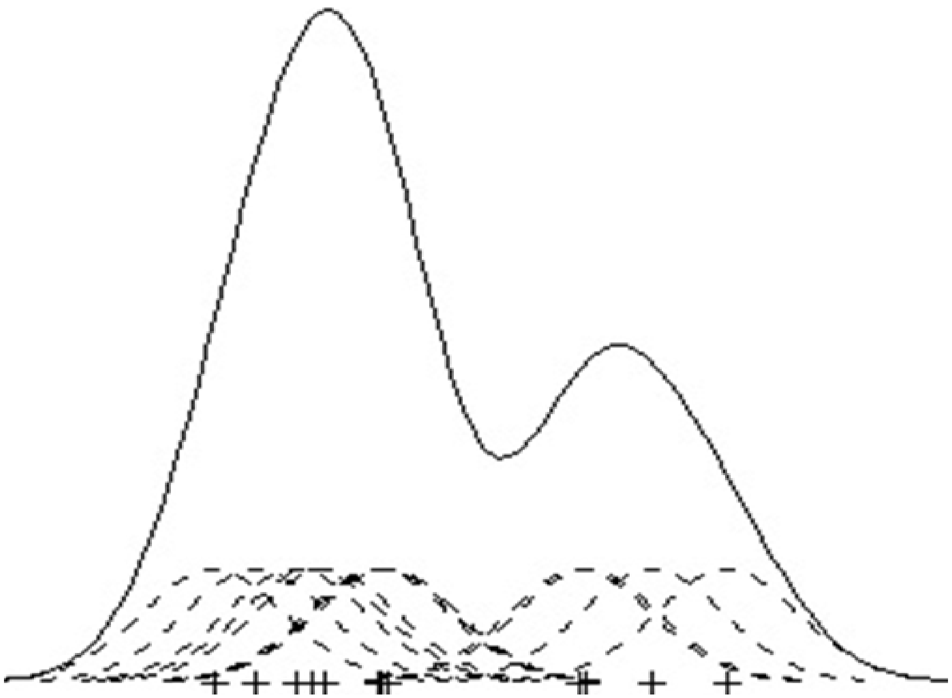
Nonparametric Estimation & Mean-Shift Clustering

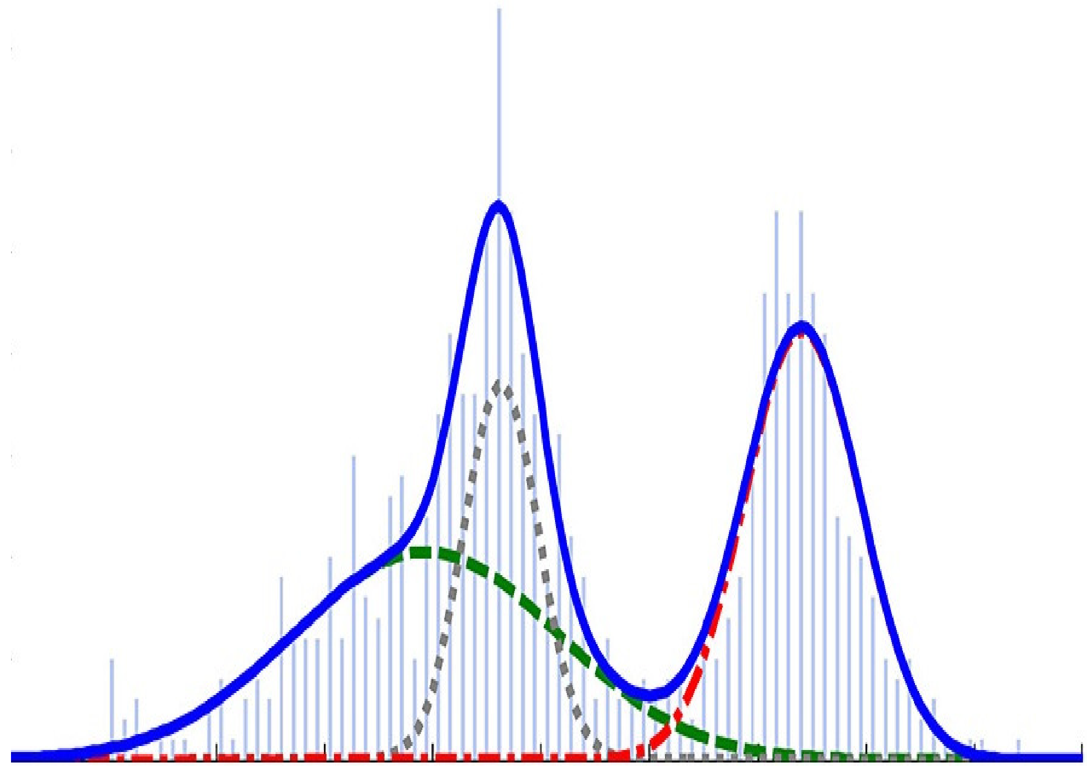
Prithwijit Guha
Dept. of EEE, IIT Guwahati


Unsupervised Learning
Parametric Clustering Algorithms
Generic Clustering Algorithms
Estimation Theory
Generative Models
Pattern Mining

Estimation Theory



Non-Parametric Estimation



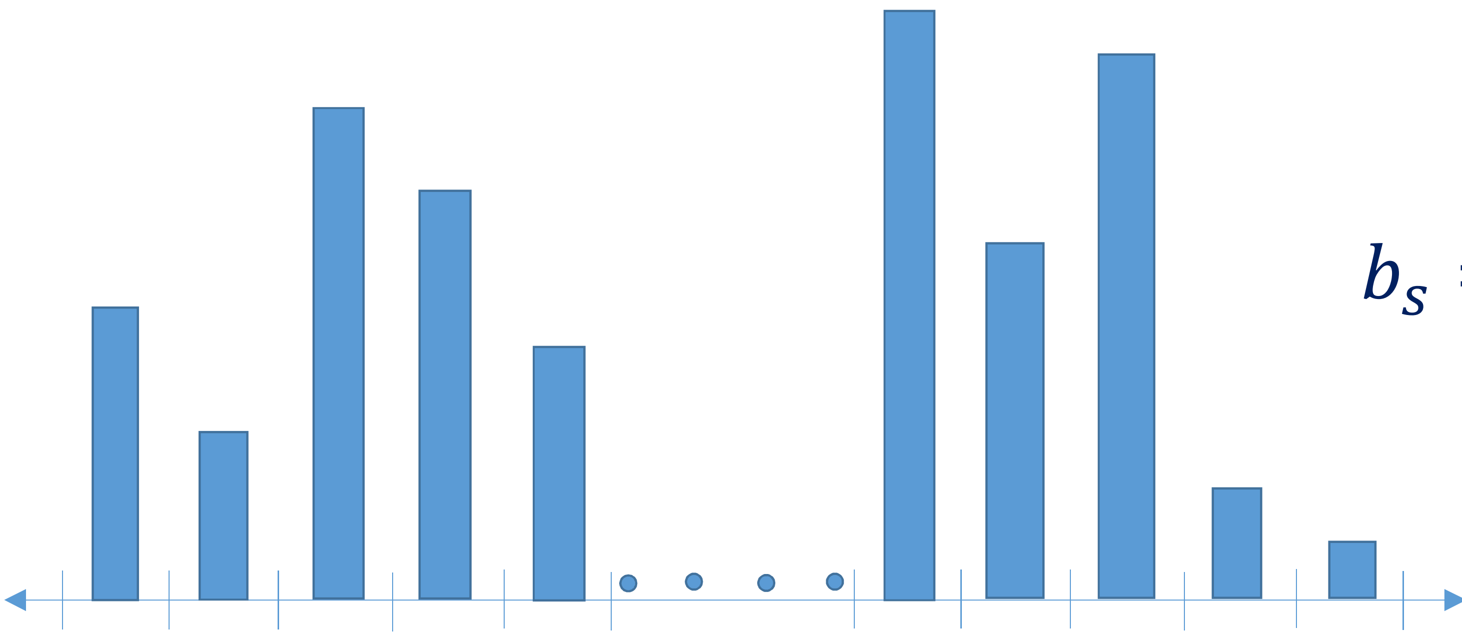
Parametric Estimation

Frequency Distribution

$$\boldsymbol{S} = \{x_1, \ldots x_i, \ldots x_n\}$$

$$min(\boldsymbol{S}) = x_{min}$$

$$max(\boldsymbol{S}) = x_{max}$$



$$b_s = \frac{x_{max} - x_{min}}{m}$$

Frequency Distribution: Pseudo-code

- INPUT: $\boldsymbol{S} = \{x_1, \ldots x_i, \ldots x_n\}$
- Intervals: $x_{max} = max(\boldsymbol{S}); x_{min} = min(\boldsymbol{S})$
- Bin: $\boldsymbol{b_s} = \frac{x_{max} - x_{min}}{m}$
- INITIALIZE: $\boldsymbol{H}[j] = 0; j = 1, \ldots m$
- FOR $i = 1 \rightarrow n$
  1. Get Bin Index: $b_i = \left\lfloor \frac{x_i - x_{min}}{b_s} \right\rfloor + 1$
  2. UPDATE: $\boldsymbol{H}[b_i] = \boldsymbol{H}[b_i] + \frac{1}{n}$
- END FOR

Frequency Distribution: Mathematical Expression

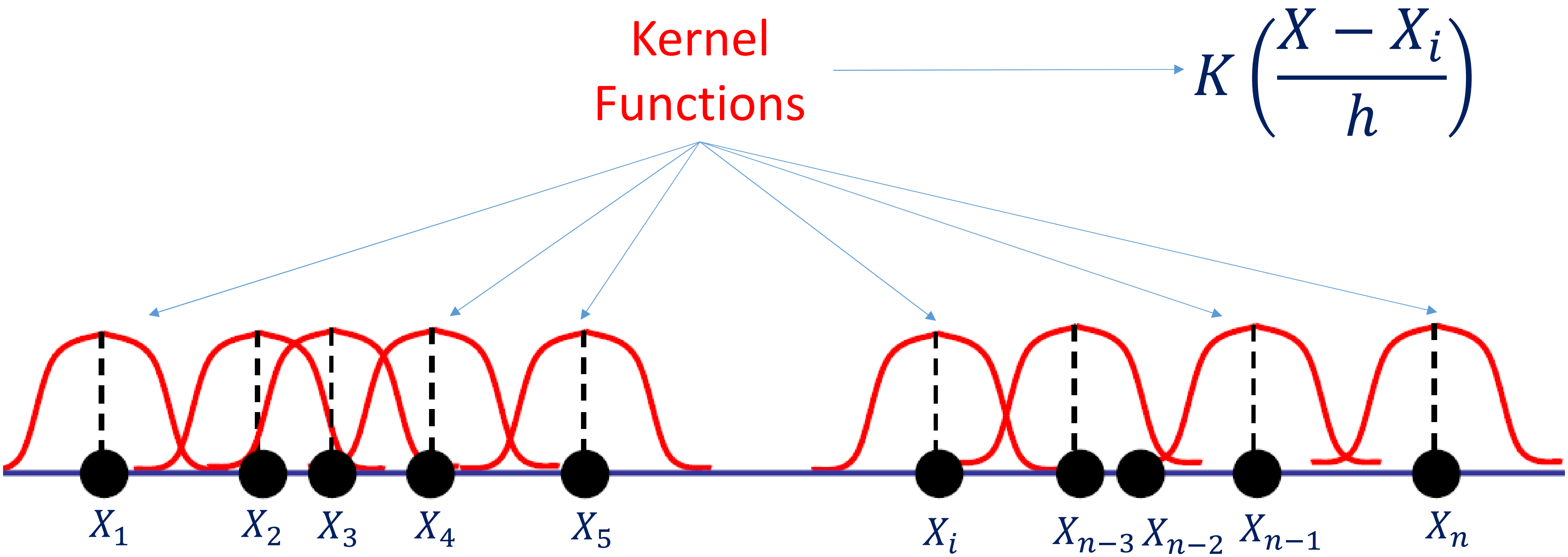
$$\boldsymbol{H}[j] = \frac{1}{n}\sum_{i=1}^{n} \delta\left[\left\lfloor \frac{x_i - x_{min}}{b_s} \right\rfloor + 1 - \mathrm{j}\right]$$

$j = 1, \ldots m$

$$b_s = \frac{x_{max} - x_{min}}{m}$$

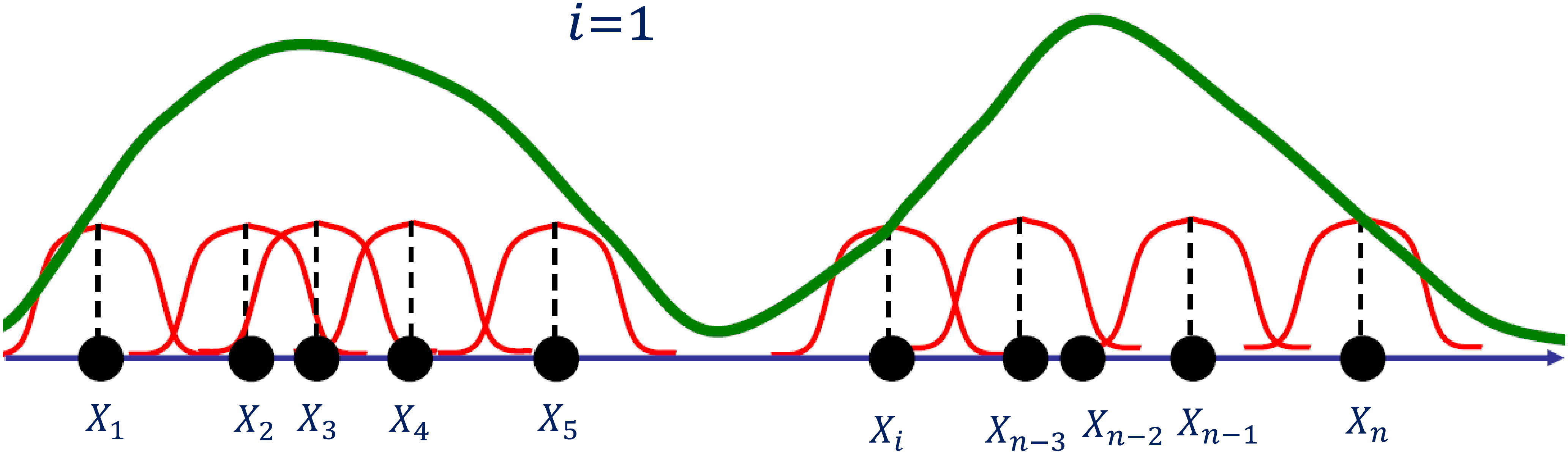
Distribution of Multivariate Data

$$\boldsymbol{S} = \{X_1, \dots X_i, \dots X_n\}$$

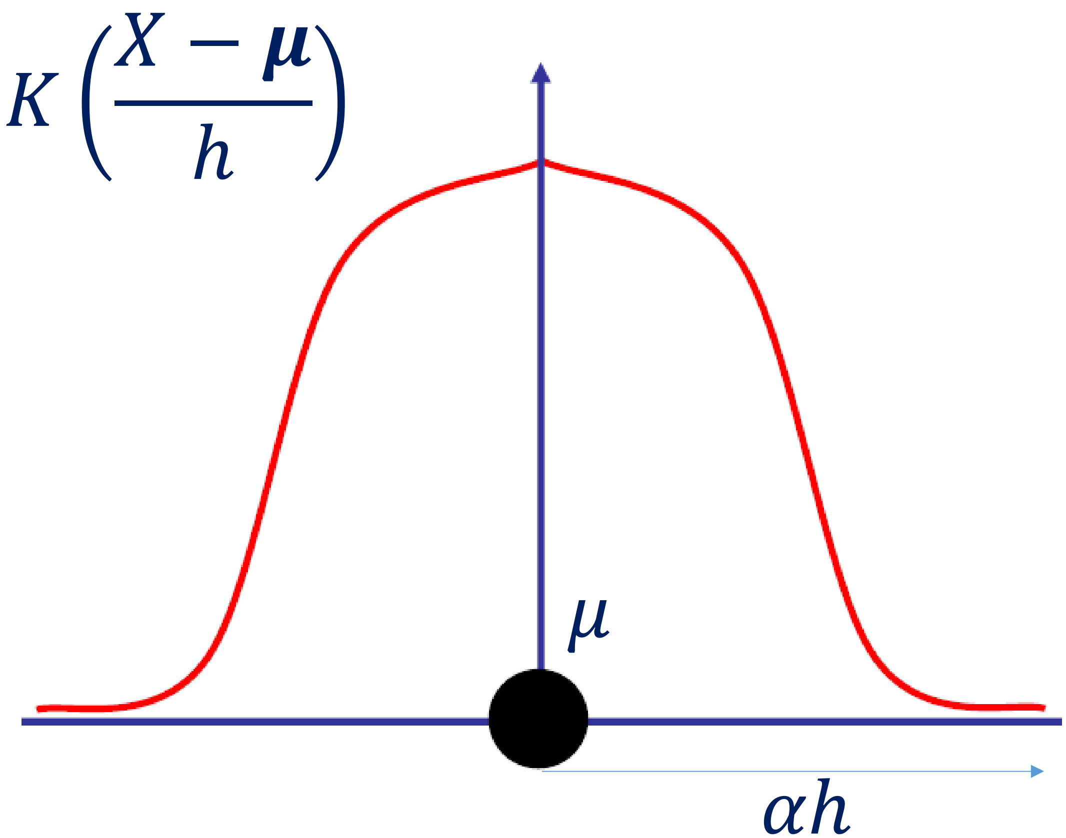


Distribution of Multivariate Data

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



Kernel Functions



Decays to ZERO As Moves Away From $\boldsymbol{\mu}$

Rate of Decay Controlled by Bandwidth h

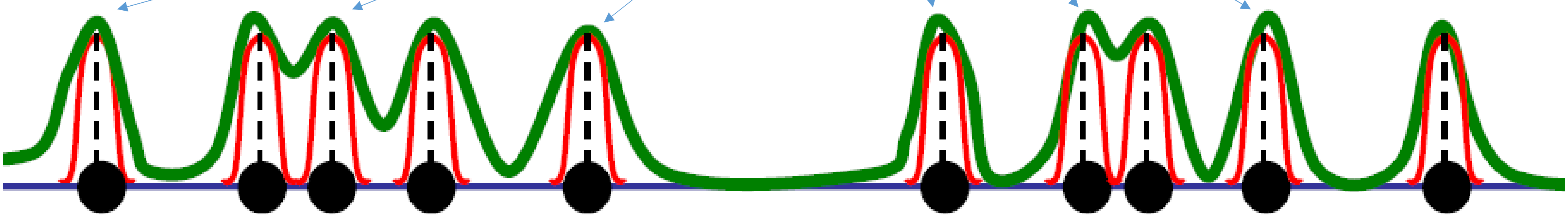
Maximum at $\boldsymbol{\mu}$

Symmetric around $\boldsymbol{\mu}$

$$\int_{\mathbb{R}^N} K\left(\frac{X-\mu}{h}\right) dX = 1$$

Effect of Low Bandwidth (h)

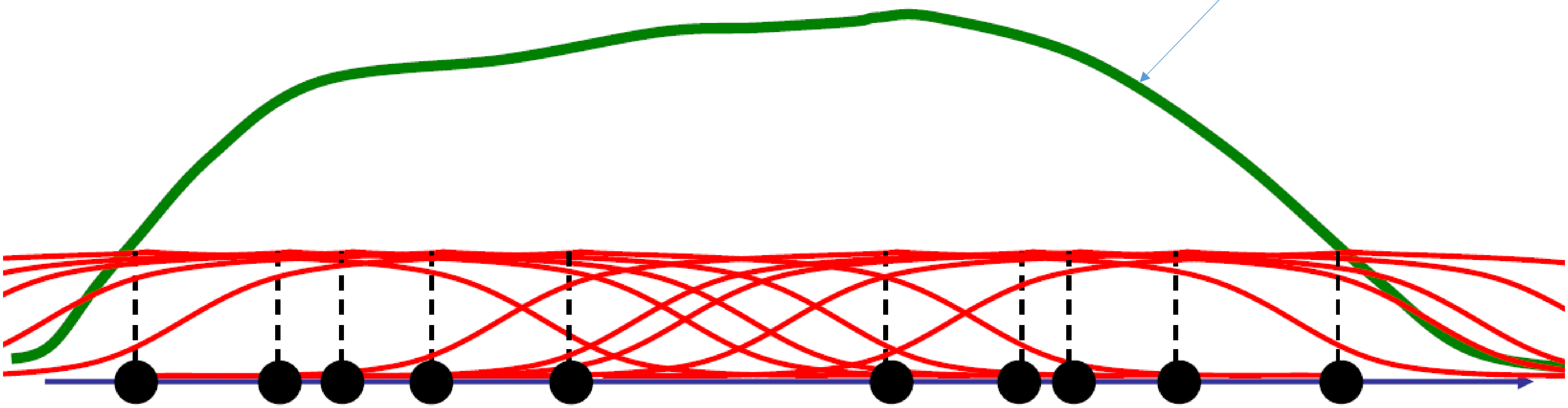
$$P(X) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{X - X_i}{h}\right)$$



Too Many Modes Loose the Interpolability

Effect of High Bandwidth (h)

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



High Bandwidth Smoothens Out the Information

Kernel Profile Function

Kernel Profile Function $\boldsymbol{k(x)}$

$$\boldsymbol{K}\left(\frac{X-\mu}{h}\right) = c\boldsymbol{k}\left(\left\|\frac{X-\mu}{h}\right\|^2\right)$$

Kernel Function

Normalization Constant

Negative Derivative of Kernel Profile

$$\boldsymbol{g}(x) = -\frac{\boldsymbol{d}}{\boldsymbol{dx}}\{\boldsymbol{k}(\boldsymbol{x})\}$$

$$\nabla_{\mathrm{X}}\boldsymbol{K}\left(\frac{X-\mu}{h}\right) = c\nabla_{X}\boldsymbol{k}\left(\left\|\frac{X-\mu}{h}\right\|^{2}\right)$$

Negative Derivative of Kernel Profile

$$\nabla_X \boldsymbol{K}\left(\frac{X-\mu}{h}\right) = c\nabla_X \boldsymbol{k}\left(\left\|\frac{X-\mu}{h}\right\|^2\right)$$

$$= c\boldsymbol{k}'\left(\left\|\frac{X-\mu}{h}\right\|^2\right) 2\left(\frac{X-\mu}{h^2}\right)$$

$$= \frac{2c}{h^2}(\mu - X)\left\{-\boldsymbol{k}'\left(\left\|\frac{X-\mu}{h}\right\|^2\right)\right\}$$

Negative Derivative of Kernel Profile

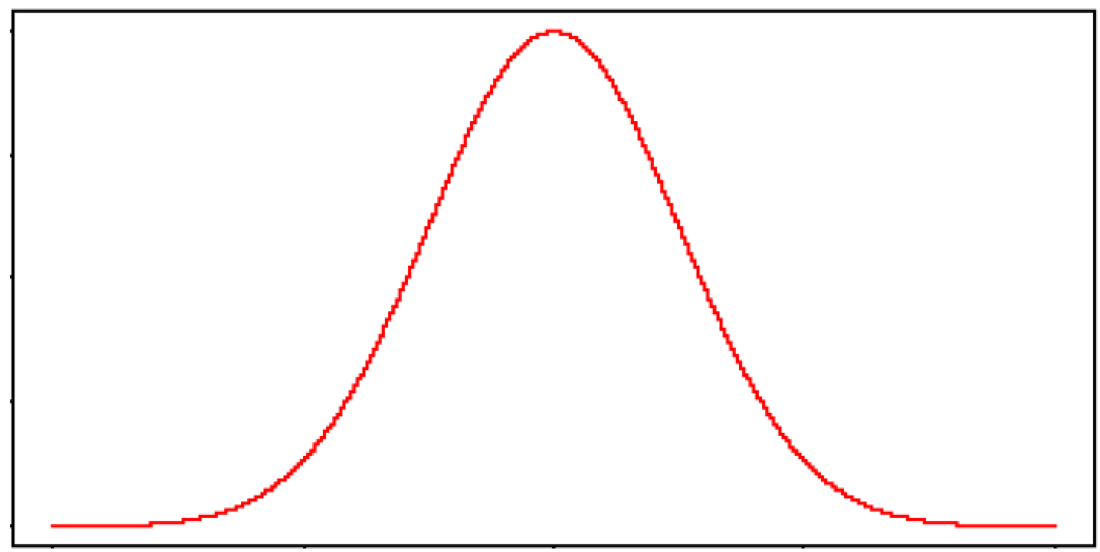
$$\nabla_{\mathrm{X}} \boldsymbol{K}\left(\frac{X-\mu}{h}\right) = \frac{2c}{h^2}(\mu - X)\left\{-\boldsymbol{k}'\left(\left\|\frac{X-\mu}{h}\right\|^2\right)\right\}$$

$$\nabla_{\mathrm{X}} \boldsymbol{K}\left(\frac{X-\mu}{h}\right) = \frac{2c}{h^2}(\mu - X)\, g\left(\left\|\frac{X-\mu}{h}\right\|^2\right)$$

Gaussian Kernel Function

$$\boldsymbol{K_N}\left(\frac{X-\mu}{h}\right) = c_N exp\left(-\frac{1}{2}\left\|\frac{X-\mu}{h}\right\|^2\right)$$

$$\boldsymbol{k_N}(x) = exp\left(-\frac{x}{2}\right)$$



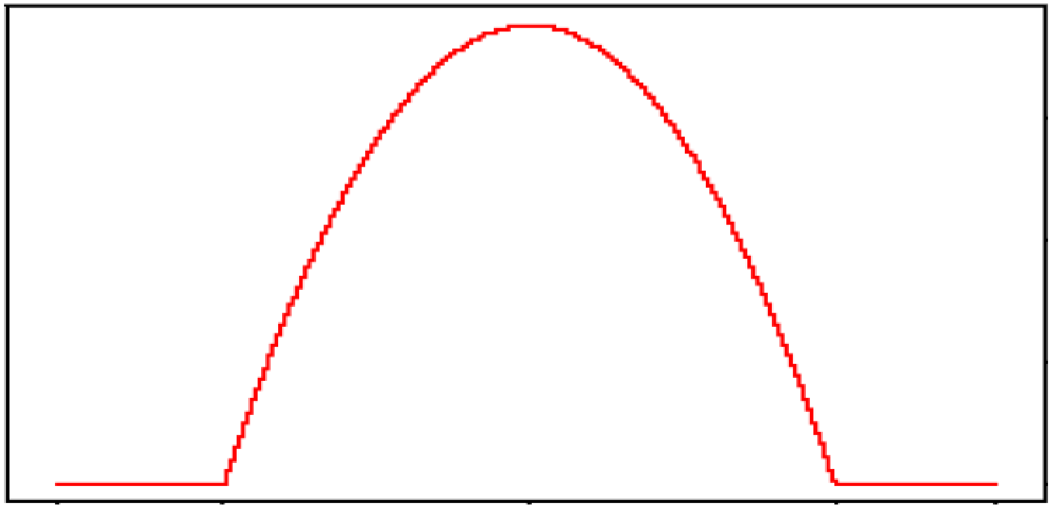
$$\boldsymbol{g_N}(x) = \frac{1}{2} exp\left(-\frac{x}{2}\right)$$

Epanechnikov Kernel Function

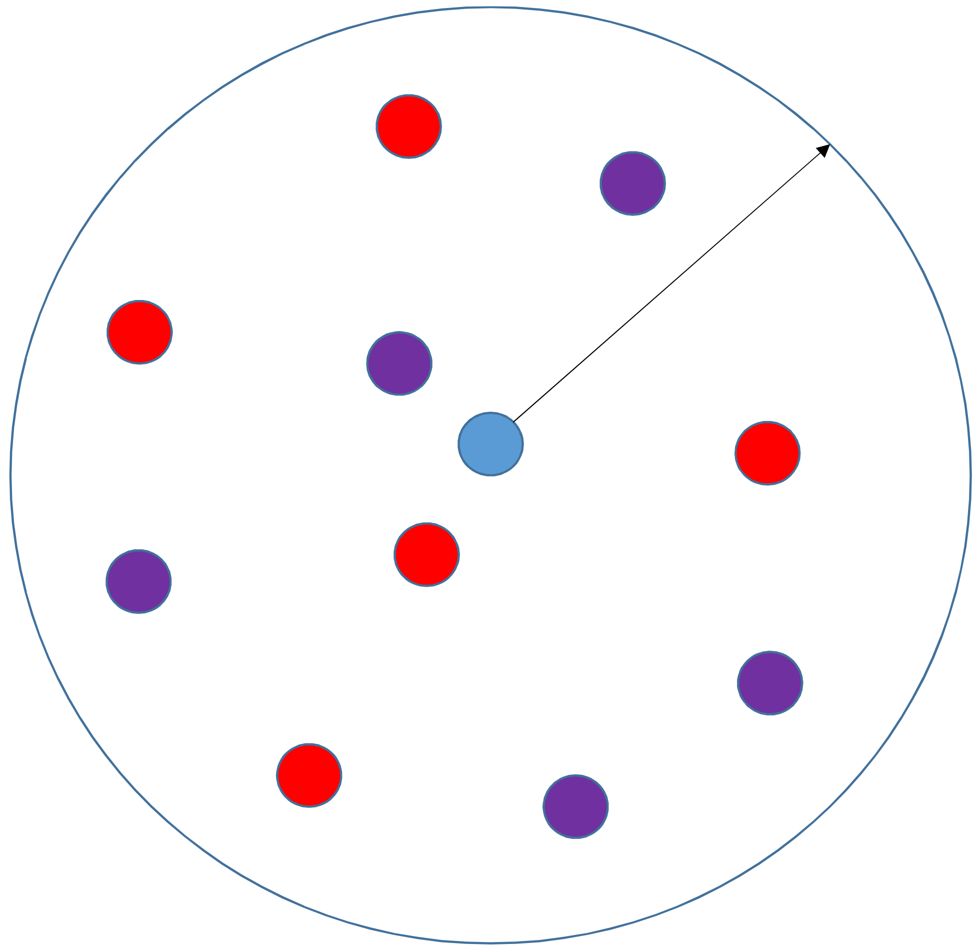
$$\boldsymbol{K_E}\left(\frac{X-\mu}{h}\right) = \begin{cases} c_E\left(1 - \left\|\frac{X-\mu}{h}\right\|^2\right), & \|X-\mu\| \le h \\ 0, & \|X-\mu\| > h \end{cases}$$

$$\boldsymbol{k_E}(x) = \begin{cases} 1-x, & x \le 1 \\ 0, & x > 1 \end{cases}$$

$$\boldsymbol{g_E}(x) = \begin{cases} 1, & x \le 1 \\ 0, & x > 1 \end{cases}$$



Classification & Regression in the Neighborhood



Application: Regression

$$\boldsymbol{S} = \{(X_i, Y_i); i = 1, \ldots n\}$$

$$\boldsymbol{X} \in \mathbb{R}^N$$

$$\boldsymbol{Y} \in \mathbb{R}^M$$

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

Application: Classification

$$\boldsymbol{S} = \{(X_i, Y_i); i = 1, \dots n\}$$

$$\boldsymbol{X} \in \mathbb{R}^N$$

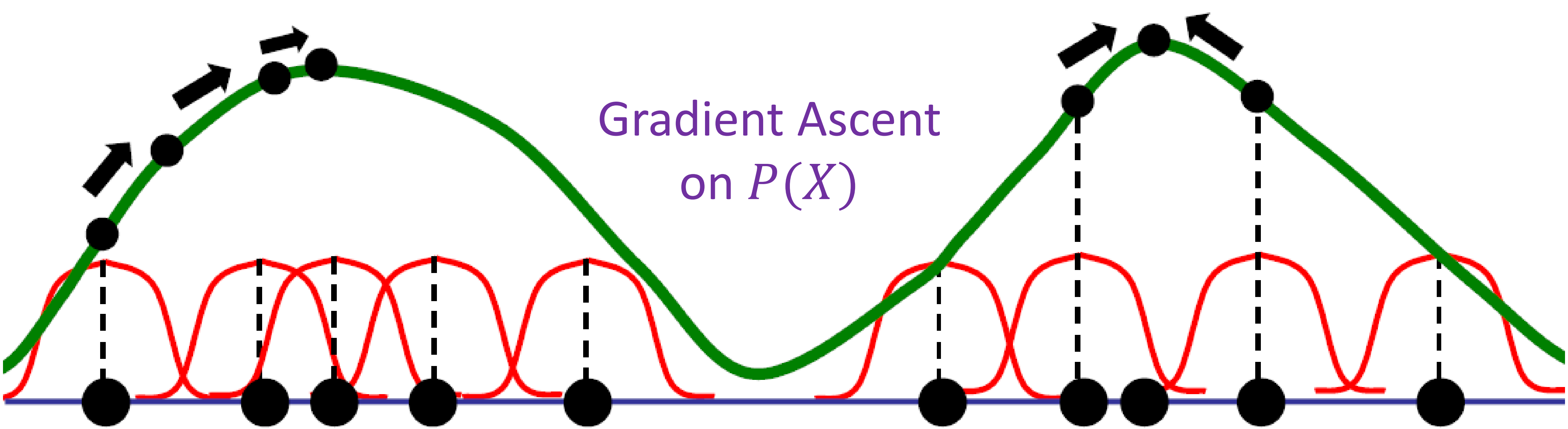
$$\boldsymbol{Y} \in \{0,1\}^M$$

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

$$Y[j] = \frac{V[j]}{\sum_{r=1}^{m} V[r]}$$

Seeking the Modes of $P(X)$

$$X^{(t+1)} = X^{(t)} + \eta_t \nabla_X P\left(X^{(t)}\right)$$



Evaluating Gradient of $P(X)$

$$P(X) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{K}\left(\frac{X - X_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} c\boldsymbol{k}\left(\left\|\frac{X - X_i}{h}\right\|^2\right)$$

$$\nabla_X P(X) = \frac{1}{n}\sum_{i=1}^{n} c\nabla_X \boldsymbol{k}\left(\left\|\frac{X - X_i}{h}\right\|^2\right)$$

$$\nabla_X P(X) = \frac{1}{n}\sum_{i=1}^{n} c\boldsymbol{k}'\left(\left\|\frac{X - X_i}{h}\right\|^2\right) 2\left(\frac{X - X_i}{h^2}\right)$$

Evaluating Gradient of $P(X)$

$$\nabla_X P(X) = \frac{1}{n}\sum_{i=1}^{n} c\boldsymbol{k}'\left(\left\|\frac{X - X_i}{h}\right\|^2\right) 2\left(\frac{X - X_i}{h^2}\right)$$

$$\nabla_X P(X) = \frac{2c}{nh^2}\sum_{i=1}^{n} (X_i - X)\left\{-\boldsymbol{k}'\left(\left\|\frac{X - X_i}{h}\right\|^2\right)\right\}$$

$$\boxed{\nabla_X P(X) = \frac{2c}{nh^2}\sum_{i=1}^{n} (X_i - X)\boldsymbol{g}\left(\left\|\frac{X - X_i}{h}\right\|^2\right)}$$

Seeking Modes of $P(X)$: Gradient Ascent

$$X^{(t+1)} = X^{(t)} + \eta_t \nabla_X P\left(X^{(t)}\right)$$

$$X^{(t+1)} - X^{(t)} = \eta_t \left\{ \frac{2c}{nh^2} \sum_{i=1}^{n} \left(X_i - X^{(t)}\right) \boldsymbol{g}\left(\left\| \frac{X^{(t)} - X_i}{h} \right\|^2 \right) \right\}$$

$$= \frac{2c\eta_t}{nh^2} \left\{ \sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\| \frac{X^{(t)} - X_i}{h} \right\|^2 \right) - \sum_{i=1}^{n} X^{(t)} \boldsymbol{g}\left(\left\| \frac{X^{(t)} - X_i}{h} \right\|^2 \right) \right\}$$

Seeking Modes of $P(X)$: Mean Shift Vector

$$X^{(t+1)} - X^{(t)} = \frac{2c\eta_t}{nh^2}\left\{\sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right) - \sum_{i=1}^{n} X^{(t)} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)\right\}$$

$$X^{(t+1)} - X^{(t)} = \underbrace{\frac{2c\eta_t}{nh^2}\left\{\sum\nolimits_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)\right\}}_{\text{scalar}} \underbrace{\left\{\frac{\sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)} - X^{(t)}\right\}}_{\text{Mean Shift Vector}}$$

Seeking Modes of $P(X)$: Mean Shift Iteration

$$X^{(t+1)} - X^{(t)} = \frac{2c\eta_t}{nh^2}\left\{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)\right\}\left\{\frac{\sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)} - X^{(t)}\right\}$$

$$\eta_t = \frac{nh^2}{2c}\left\{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)\right\}^{-1} \Rightarrow \frac{2c\eta_t}{nh^2}\left\{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)\right\} = 1$$

$$X^{(t+1)} - X^{(t)} = \frac{\sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)} - X^{(t)}$$

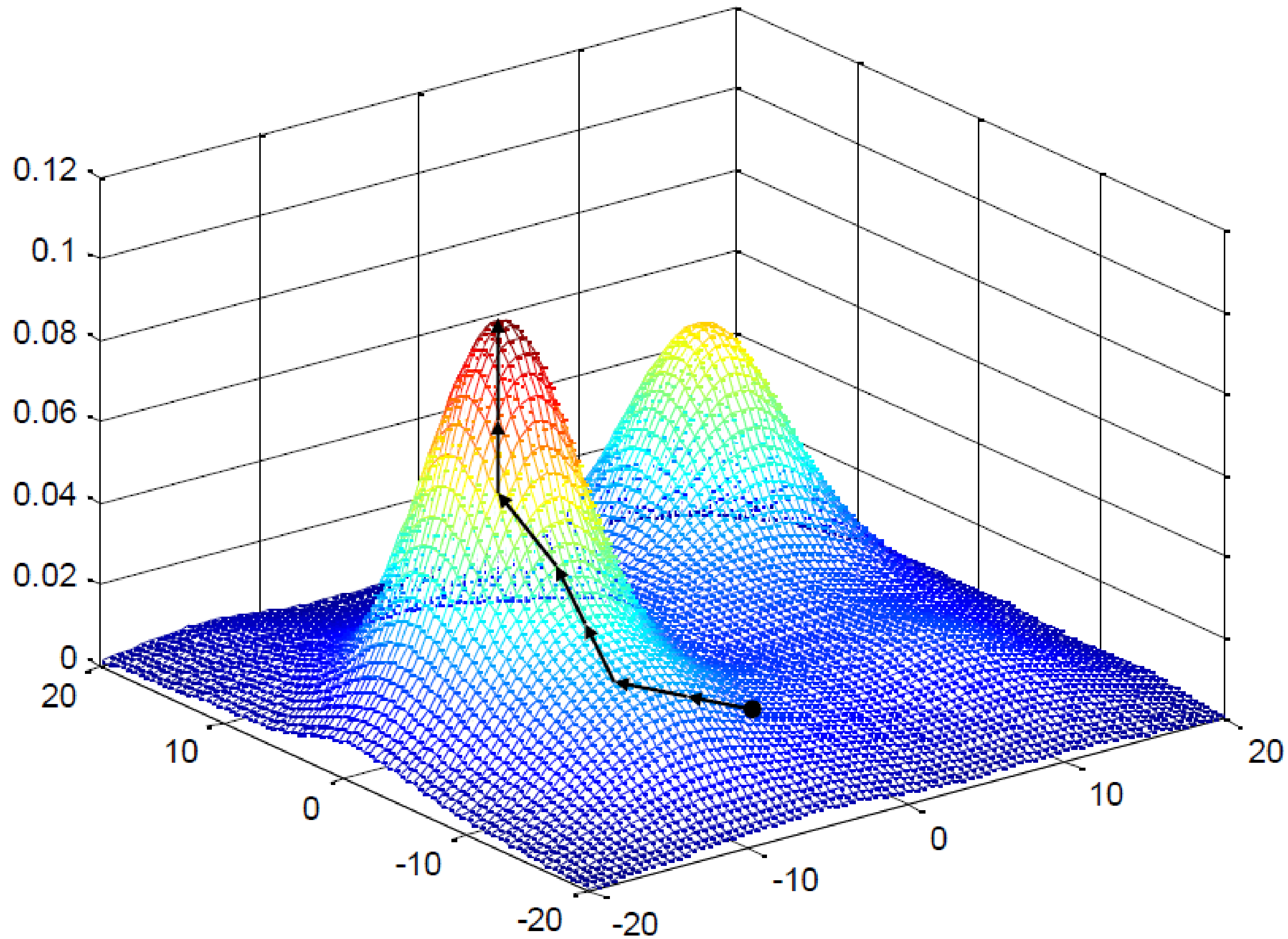
Seeking Modes of $P(X)$: Mean Shift Iteration

$$X^{(t+1)} = \frac{\sum_{i=1}^{n} X_i \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \boldsymbol{g}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}$$

Input: $\boldsymbol{S} = \{X_i; i = 1, \ldots n\}$

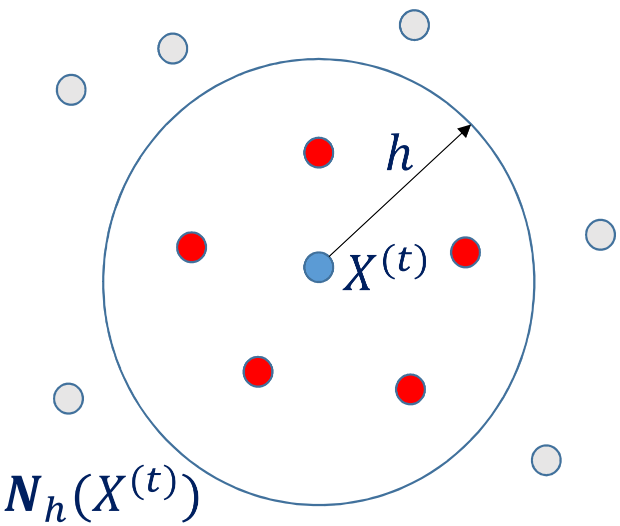
$X^{(0)} \in \boldsymbol{S}$

Seeking Modes of $P(X)$: Visualization

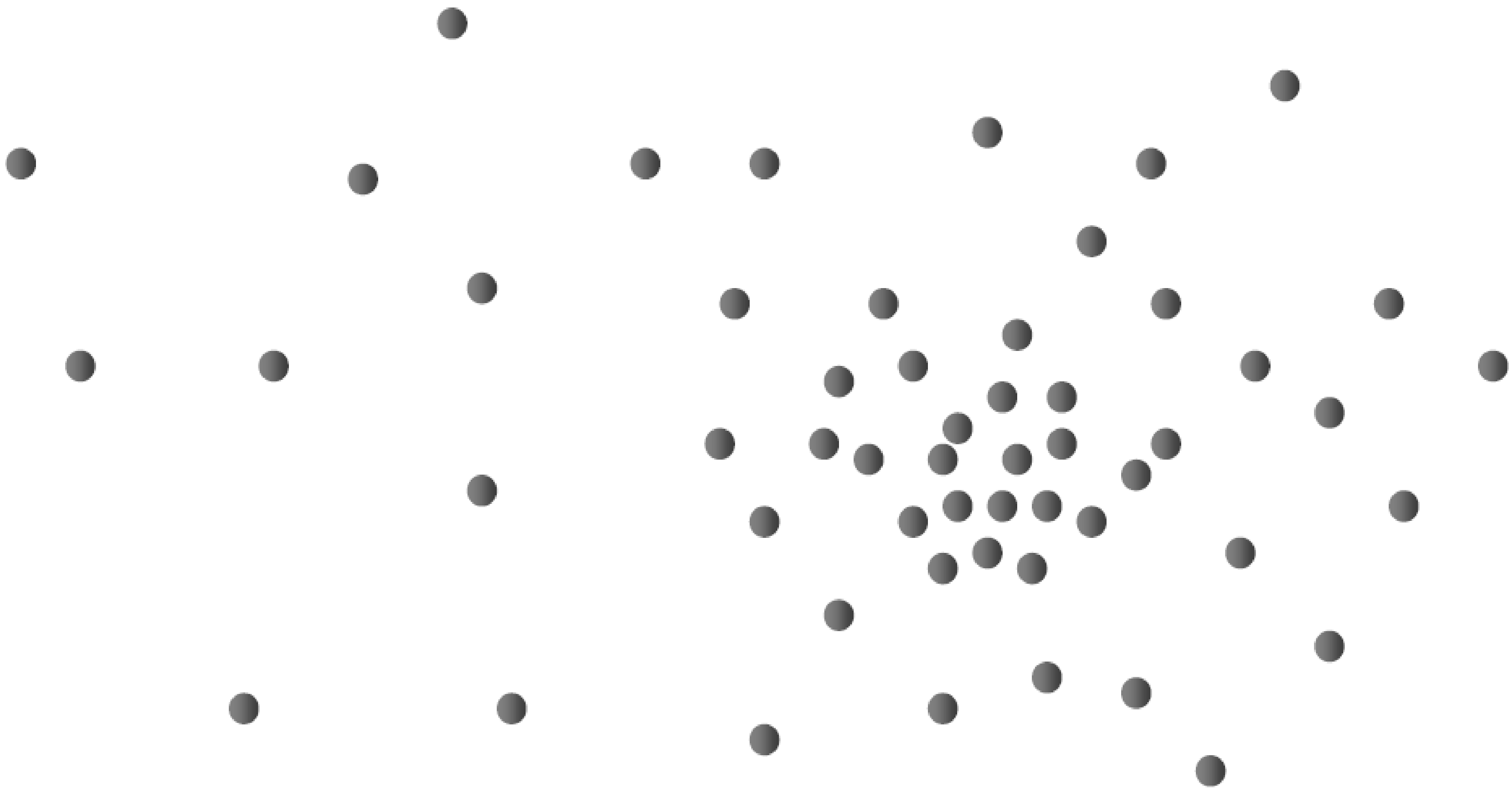


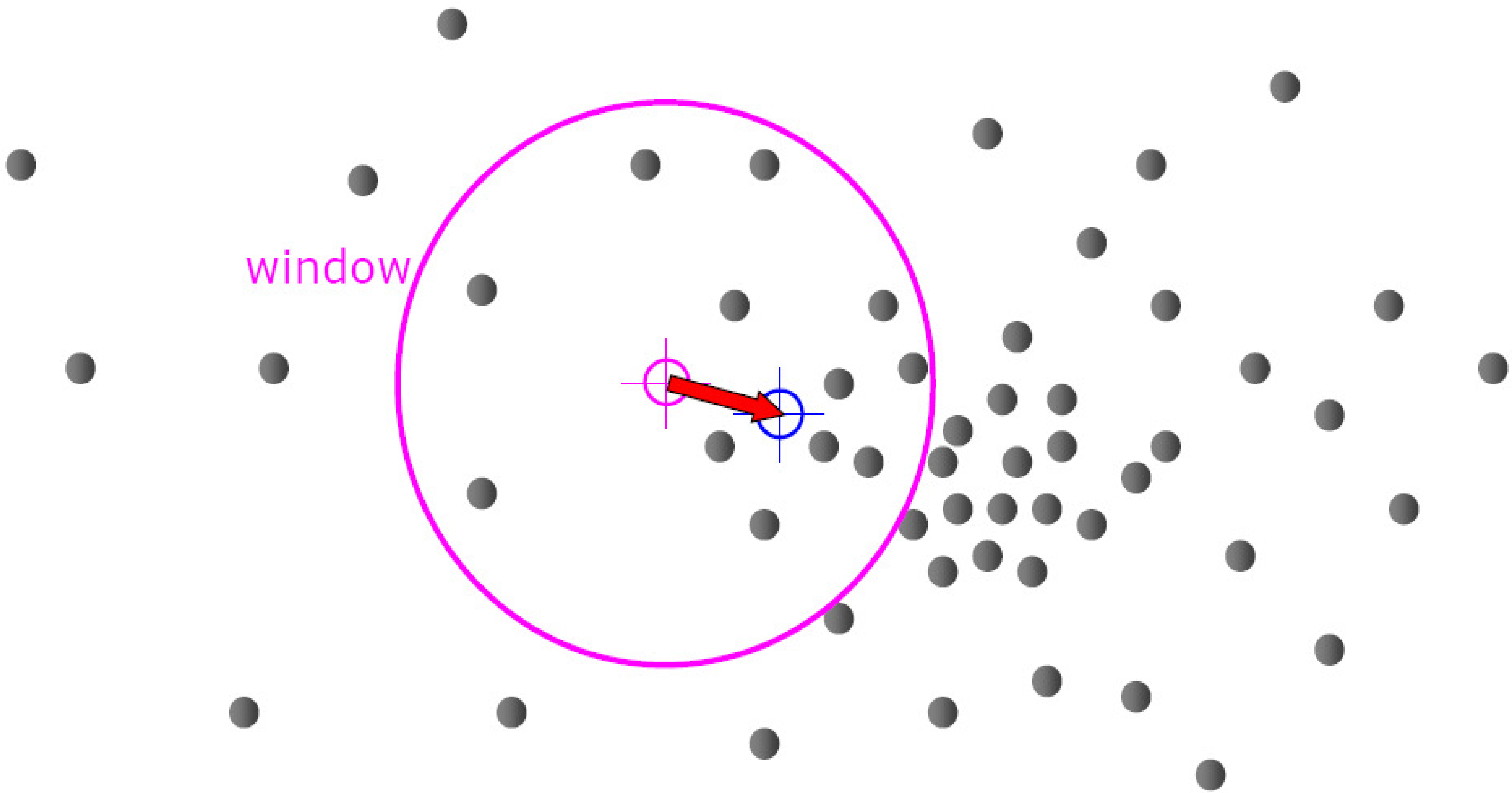
Seeking Modes of $P(X)$: Choice of Kernel

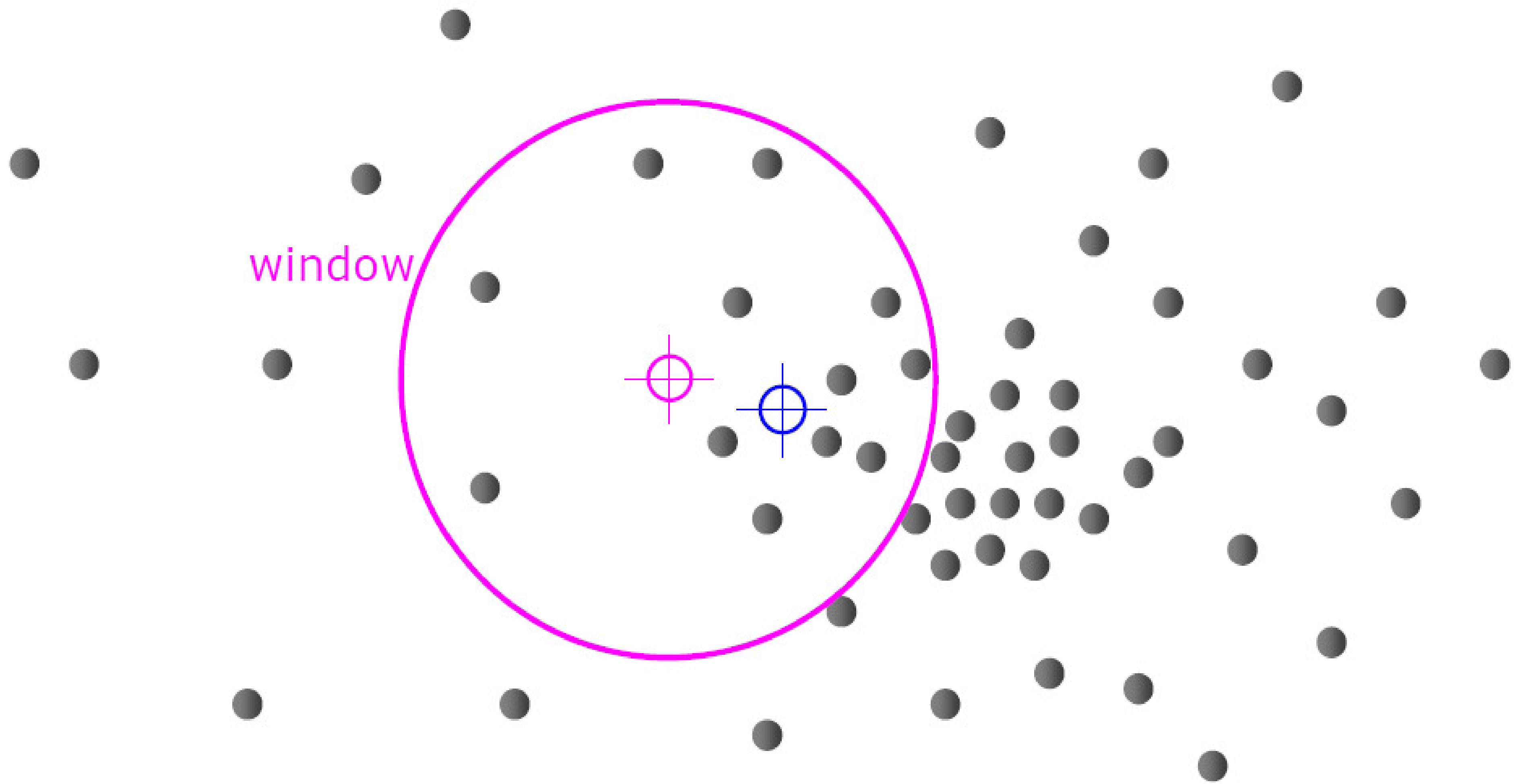
$$X^{(t+1)} = \frac{\sum_{i=1}^{n} X_i \boldsymbol{g_E}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \boldsymbol{g_E}\left(\left\|\frac{X^{(t)} - X_i}{h}\right\|^2\right)}$$

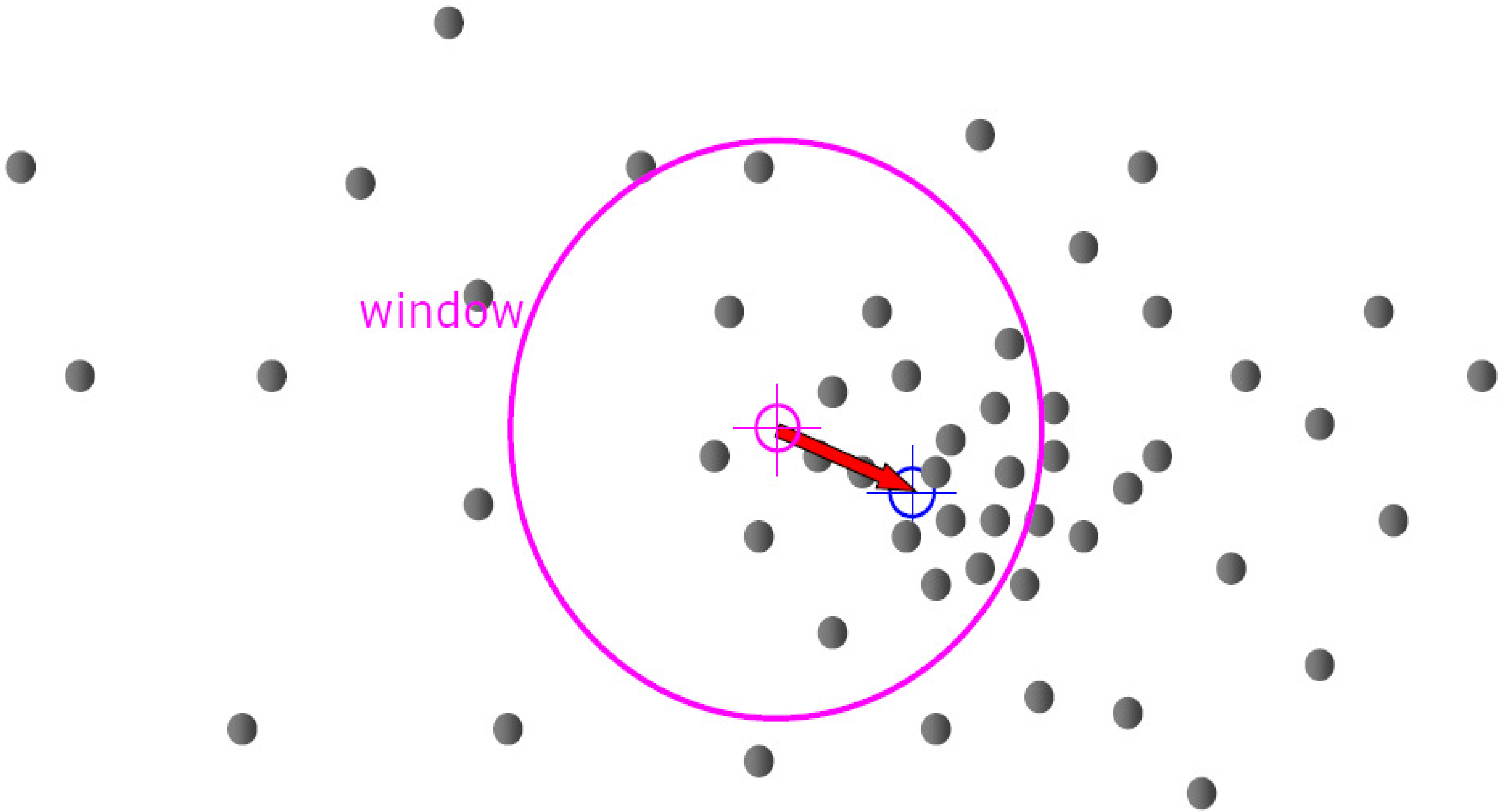


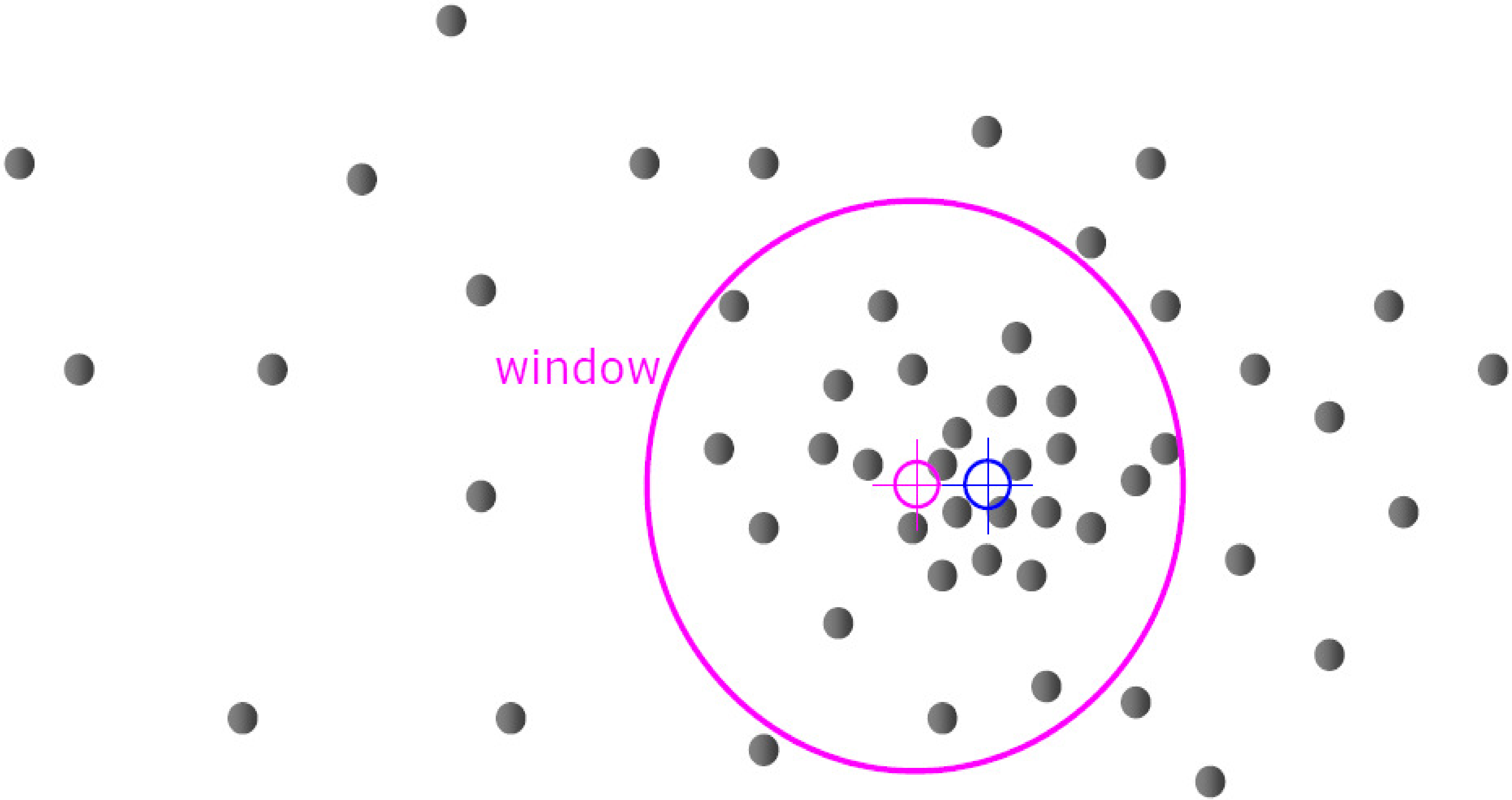
$$X^{(t+1)} = \frac{\sum_{X \in \boldsymbol{N}_h(X^{(t)})} X \cdot 1}{\sum_{X \in \boldsymbol{N}_h(X^{(t)})} 1}$$

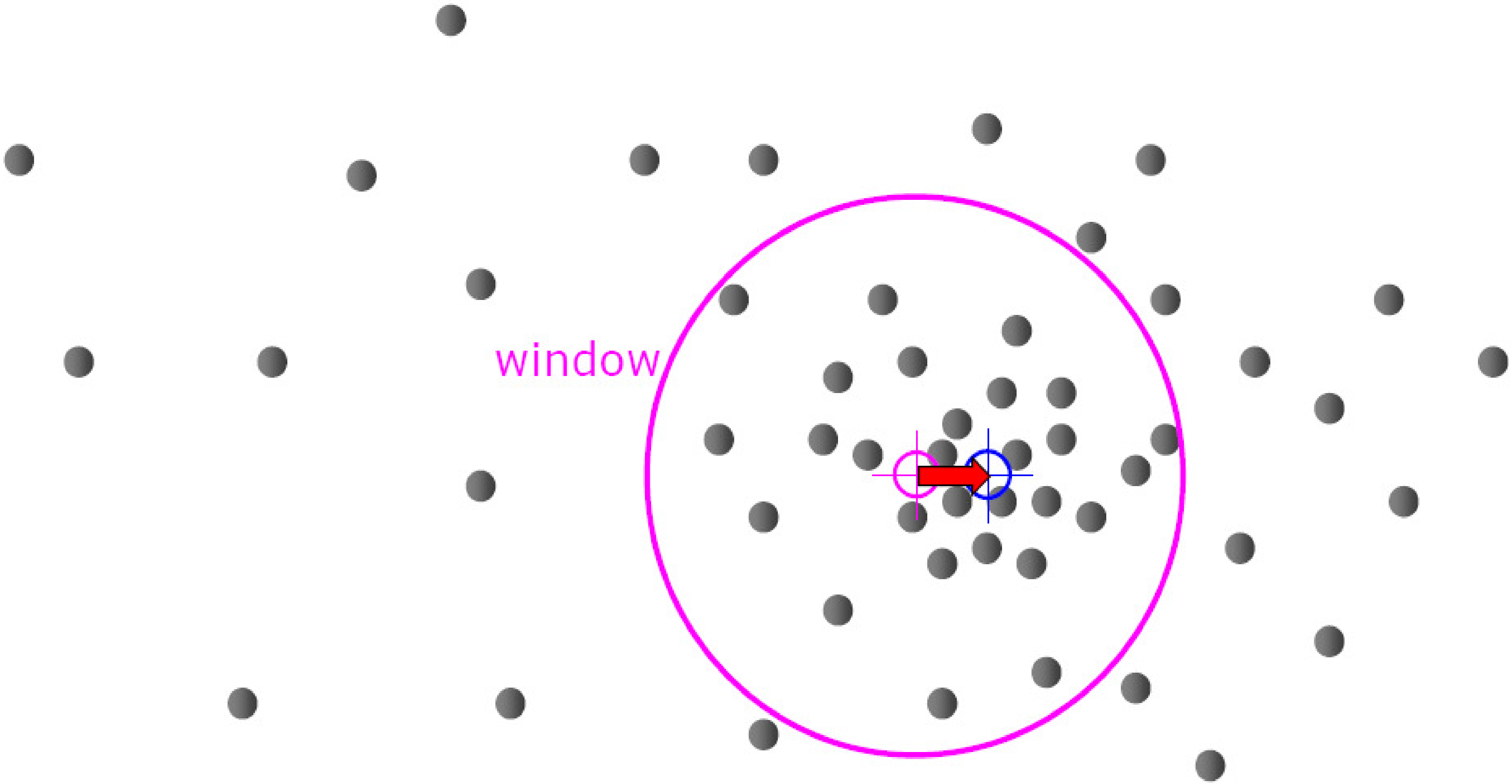


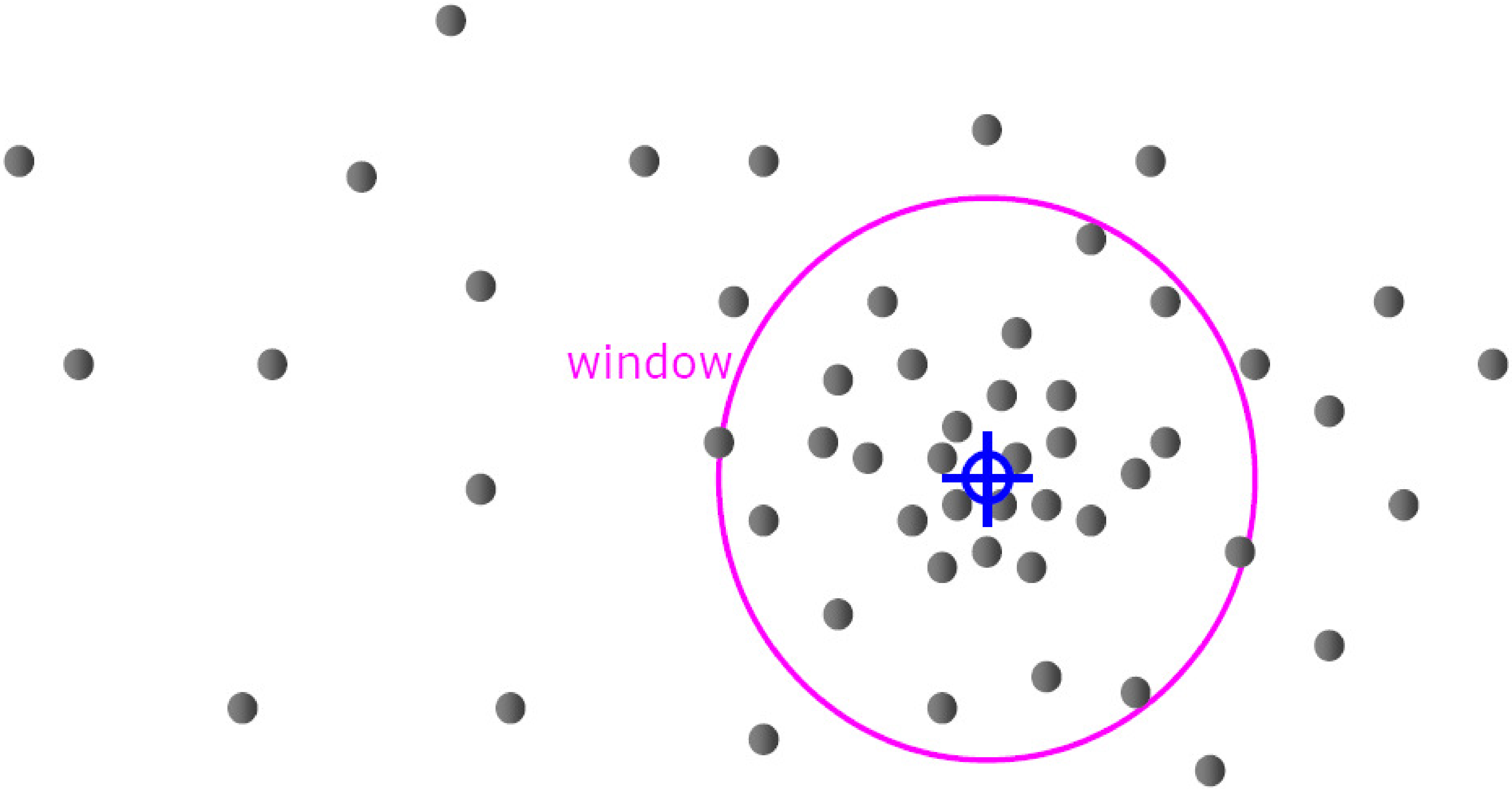

window


window


window


window


window


window

Mean-Shift Clustering: Problem

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$	$\mathbf{X}_7$	$\mathbf{X}_8$	$\mathbf{X}_9$	$\mathbf{X}_{10}$
4.2	2.5	4.7	3.4	2.9	4.3	6.0	2.6	3.1	1.9
1.7	0.4	2.0	4.6	1.5	0.9	2.8	2.2	4.8	3.9

$X^{(0)} = (3.7, 2.5)$ $\quad\quad$ $X^{(1)} = (\quad , \quad)$

$h = 1.73$

Mean-Shift Clustering: Solution

$$\begin{bmatrix} 4.2 & 1.7 \\ 2.5 & 0.4 \\ 4.7 & 2.0 \\ 3.4 & 2.6 \\ 2.9 & 1.5 \\ 4.3 & 0.9 \\ 6.0 & 2.8 \\ 2.6 & 2.2 \\ 3.1 & 4.8 \\ 1.9 & 3.9 \end{bmatrix} \quad [3.5 \quad 2.7] \rightarrow \|\cdot\|_2 \rightarrow \begin{bmatrix} \boxed{0.94} \\ 2.42 \\ \boxed{1.12} \\ 2.12 \\ \boxed{1.28} \\ \boxed{1.71} \\ 2.32 \\ \boxed{1.14} \\ 2.38 \\ 2.28 \end{bmatrix} \xrightarrow{\text{Neighbors of } (3.5,2.7)} \begin{bmatrix} \boxed{4.2} & \boxed{1.7} \\ 2.5 & 0.4 \\ \boxed{4.7} & \boxed{2.0} \\ 3.4 & 2.6 \\ \boxed{2.9} & \boxed{1.5} \\ \boxed{4.3} & \boxed{0.9} \\ 6.0 & 2.8 \\ \boxed{2.6} & \boxed{2.2} \\ 3.1 & 4.8 \\ 1.9 & 3.9 \end{bmatrix}$$

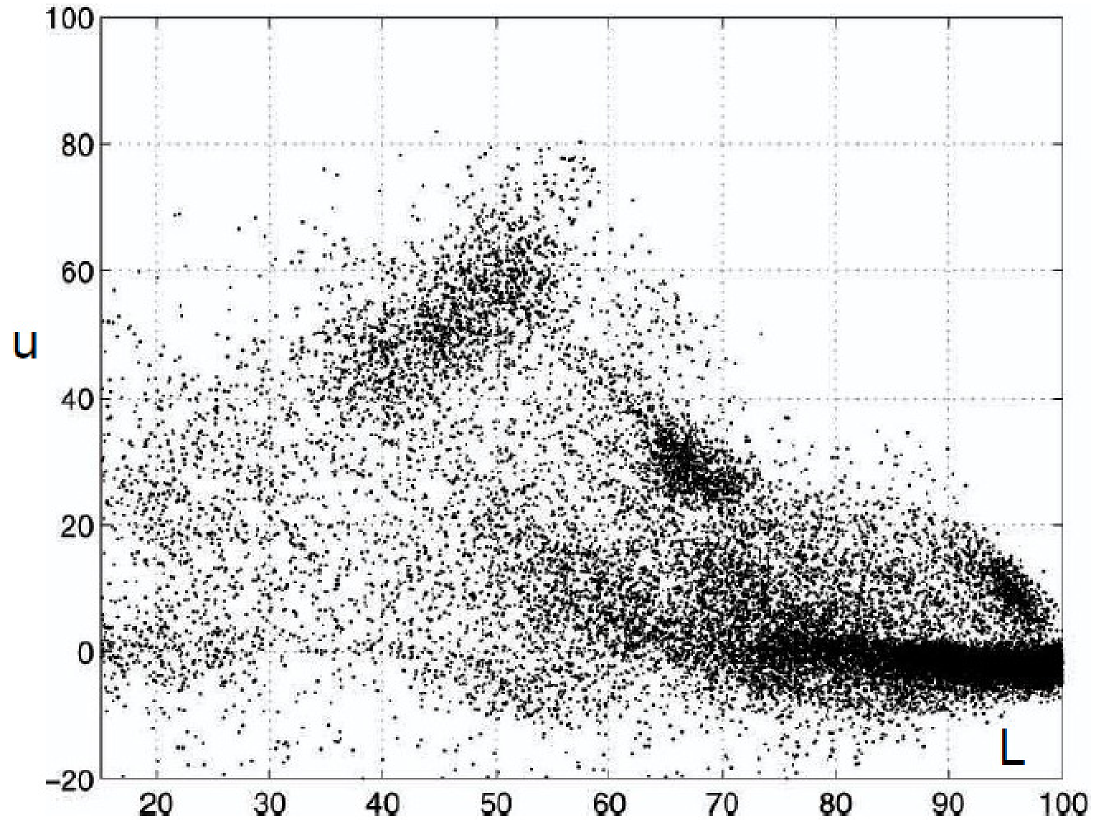
$$h = 1.73$$

Mean-Shift Clustering: Solution

$$X^{(1)} = \frac{\sum_{X \in N_{1.73}(3.5,2.7)} X \cdot 1}{\sum_{X \in N_{1.73}(3.5,2.7)} 1}$$

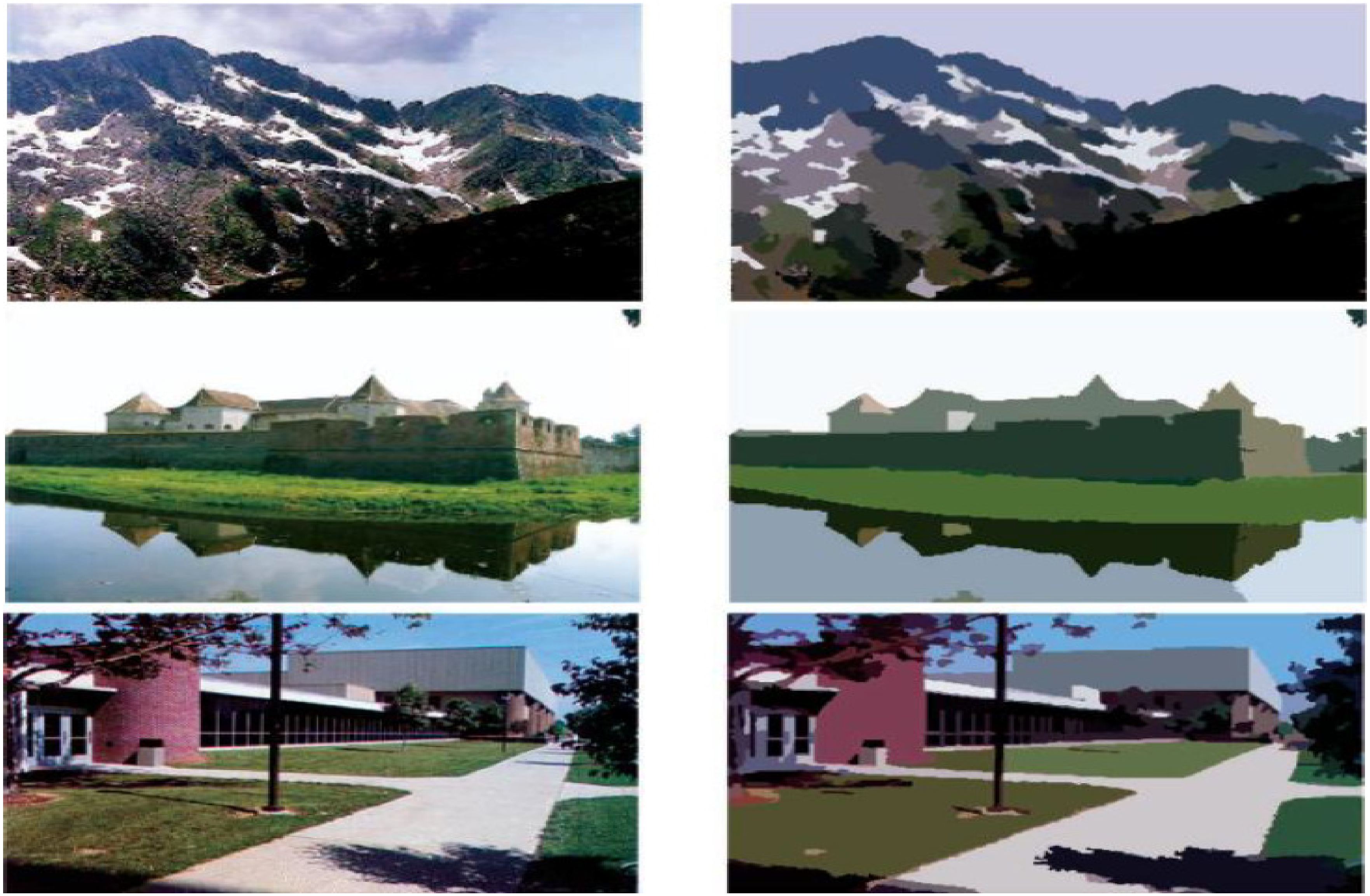
$$X^{(1)} = \frac{\begin{bmatrix} 4.2 \\ 1.7 \end{bmatrix} + \begin{bmatrix} 4.7 \\ 2.0 \end{bmatrix} + \begin{bmatrix} 2.9 \\ 1.5 \end{bmatrix} + \begin{bmatrix} 4.3 \\ 0.9 \end{bmatrix} + \begin{bmatrix} 2.6 \\ 2.2 \end{bmatrix}}{5} = \begin{bmatrix} 3.74 \\ 1.66 \end{bmatrix}$$

Application: Image Segmentation



Data Vector Formed by Using (L, u, v, x, y)

Separate Bandwidth Along Each Dimension



Advantages

- Seeks the Mode of Distributions
- Ideal for Most Practical Cases
- Provides Model Free Nonlinear Regression
- Minimal Parameter Tuning

Disadvantages

- Needs All Data Points
- Often Computation Intensive

Summary

- Nonparametric Representation of Distributions
- Model Free Regression
- Nonparametric Classification
- Mean-Shift Clustering

Thank You