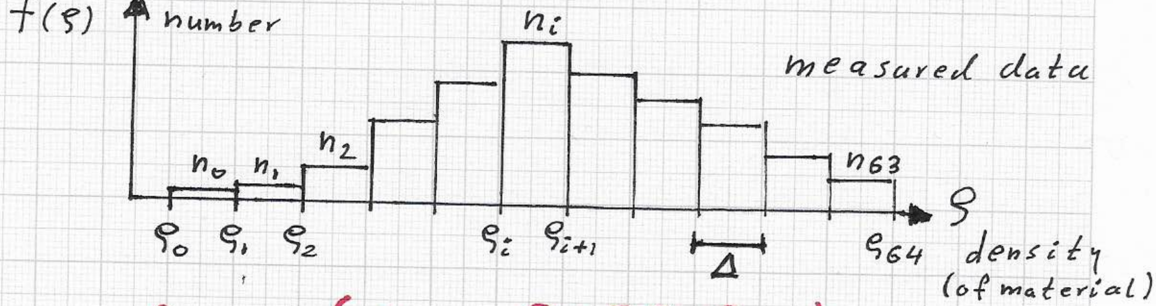
BEST APPROXIMATION OF DISTRIBUTIONS

- **IDEA:** Use a LOW-DEGREE polynomial spline with FEW POLYNOMIAL SEGMENTS to approximate optimally measured histogram data (=piecewise constant functions). The degrees of freedom of the spline, defined as a best approximation, should "suffice" to capture the relevant and needed characteristics of the histogram data (e.g., mean, variance, skewness, kurtosis, other "moments"). These characteristics of the spline density function will define the features used for data classification purposes.

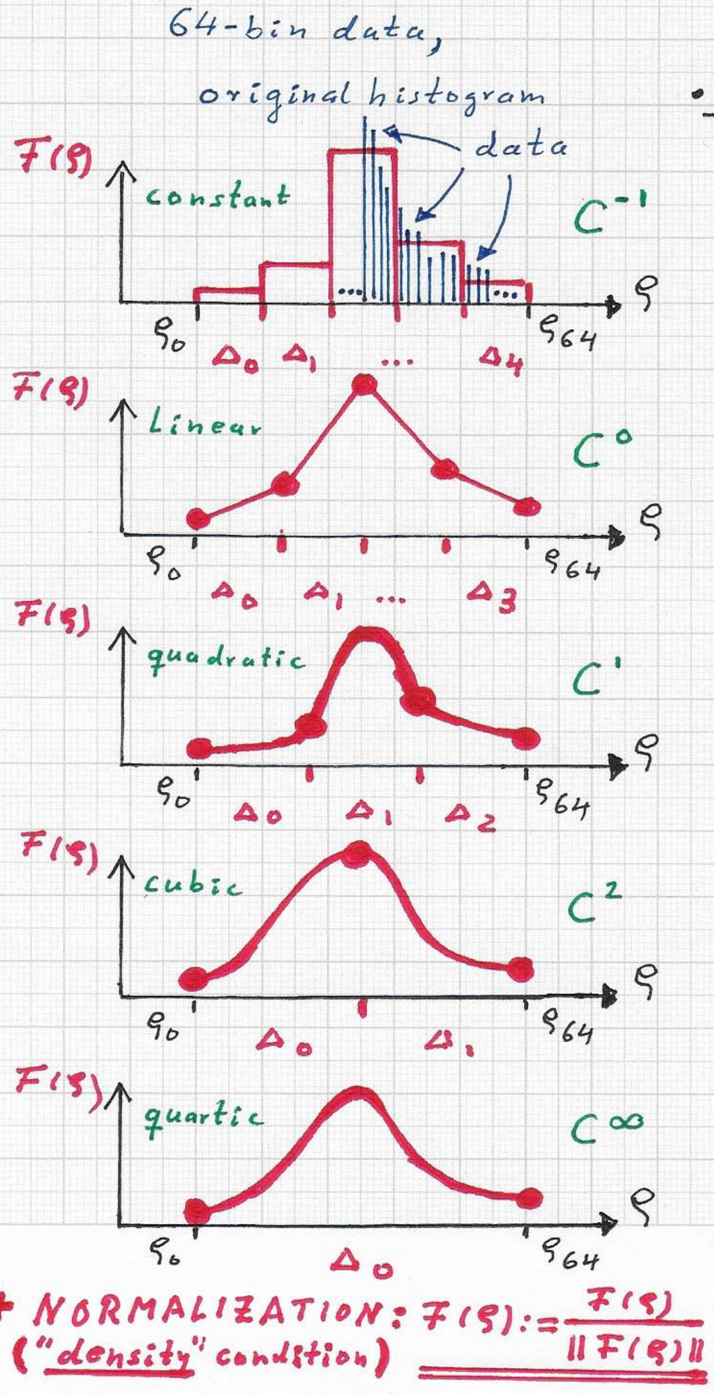
- **EXAMPLE:** We consider an example to explain the general principles and methods. First, we assume that histogram data with $K = 64$ bins, of equal width, is given.



$$f(\rho) = \begin{cases} n_i \,, & \rho \in [\rho_i, \rho_{i+1}) \\ 0 \,, & \rho < \rho_0 \text{ or } \rho \geq \rho_{64} \end{cases}$$

■ BEST APPROXIMATION OF DISTRIBUTIONS

- **GOAL:** Compute a BEST spline approximation of a given 64-bin measured distribution – using a spline that is "as simple as possible" and that "preserves" the classification-relevant distribution characteristics!



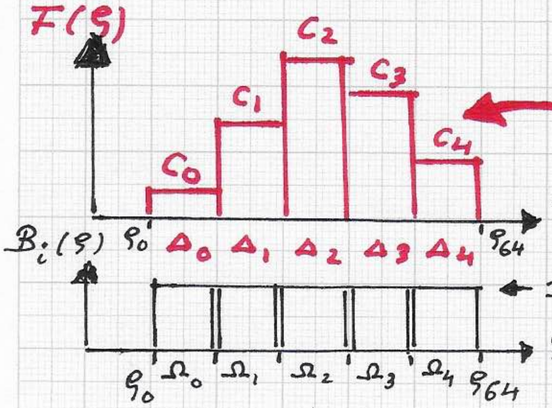
- Example: Assume that "5 degrees of freedom," "5 coefficients," suffice to capture all relevant and needed characteristics of a 64-bin histogram. Thus, one can consider these spline degrees: constant, linear, quadratic, cubic and quartic; one obtains splines with 5, 4, 3, 2 and 1 segments, respectively. (The lengths of the ϱ-intervals over which spline segments are defined can vary.) The resulting splines are C^{-1}-, C^0-, C^1-, C^2- and C^∞-continuous, respectively. Exactly 5 degrees of freedom exist for each spline, due to continuity constraints.

+ NORMALIZATION: $F(\varrho) := \dfrac{F(\varrho)}{\|F(\varrho)\|}$ ("density" condition)

BEST APPROXIMATION OF DISTRIBUTIONS

- **Computation:** The unknown spline coefficients are computed by solving the normal equations that result when minimizing the (squared) differences between the given histogram (= piecewise constant function) and the spline.

- **Examples:** Linear equation systems to be solved for the constant, linear and (single-segment) quartic splines (see previous page):



BOX basis functions

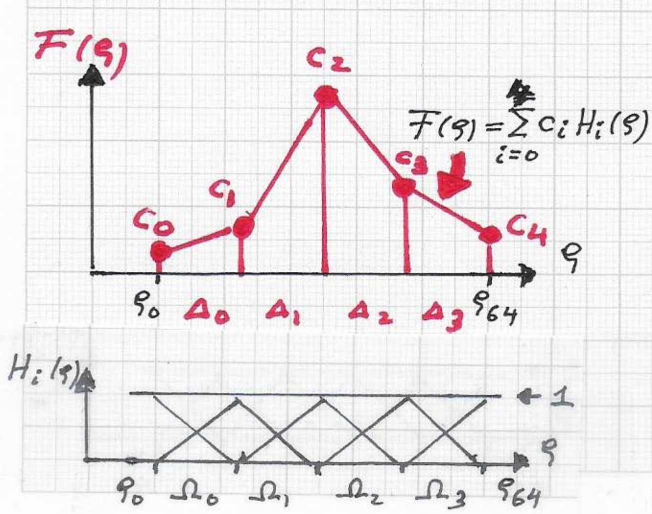
→ best approximation:

$$F(\rho) = \sum_{i=0}^{4} c_i B_i(\rho)$$

→ Linear equation systems:

$$\begin{bmatrix} \Delta_0 & & & & 0 \\ & \Delta_1 & & & \\ & & \Delta_2 & & \\ & & & \Delta_3 & \\ 0 & & & & \Delta_4 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} = \begin{bmatrix} \int_{\Omega_0} f(\rho)\,d\rho \\ \vdots \\ \int_{\Omega_4} f(\rho)\,d\rho \end{bmatrix}$$

$$\Rightarrow c_i = \frac{1}{\Delta_i} \int_{\Omega_i} f(\rho)\,d\rho$$



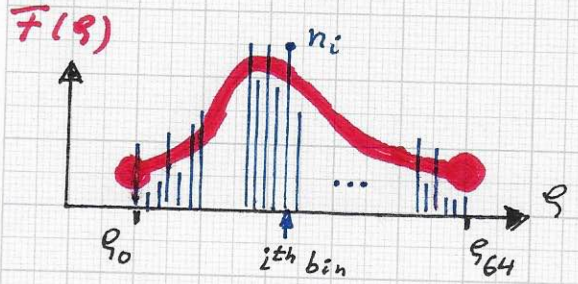
HAT basis functions

$$\begin{bmatrix} 2\Delta_0 & \Delta_0 & & & 0 \\ \Delta_0 & 2(\Delta_0+\Delta_1) & \Delta_1 & & \\ & \Delta_1 & 2(\Delta_1+\Delta_2) & \Delta_2 & \\ 0 & & \Delta_2 & 2(\Delta_2+\Delta_3) & \Delta_3 \\ & & & \Delta_3 & 2\Delta_3 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix} =$$

$$= 6 \left[\int_{\Omega_0} H_0(\rho) f(\rho)\,d\rho,\ \int_{\Omega_0 \cup \Omega_1} H_1(\rho) f(\rho)\,d\rho, \ldots \right]^{Transp}$$

BEST APPROXIMATION OF DISTRIBUTIONS

- ... The SINGLE-SEGMENT quartic spline can be computed "directly" via the normal equations used to determine the optimal least squares approx.:



→ 64 bin values: $n_0, \dots, n_{63}$

→ 64 intervals: $[\varrho_0, \varrho_1), \dots, [\varrho_{63}, \varrho_{64})$

→ wanted BEST quartic polynomial:

$$F(\varrho) = \sum_{i=0}^{4} c_i \varrho^i$$

⇒ Equations: $F(\bar{\varrho}_j) = \sum_{i=0}^{4} c_i \bar{\varrho}_j^{\,i}$, $j = 0 \dots 63$

where $\bar{\varrho}_j = (\varrho_j + \varrho_{j+1})/2$, $j = 0 \dots 63$

⇒ Matrix form of linear equation system:

$$\underbrace{\begin{bmatrix} 1 & \bar{\varrho}_0 & \bar{\varrho}_0^{\,2} & \bar{\varrho}_0^{\,3} & \bar{\varrho}_0^{\,4} \\ \vdots & & & & \vdots \\ 1 & \bar{\varrho}_{63} & \bar{\varrho}_{63}^{\,2} & \bar{\varrho}_{63}^{\,3} & \bar{\varrho}_{63}^{\,4} \end{bmatrix}}_{R} \cdot \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}}_{\mathbb{C}} = \underbrace{\begin{bmatrix} n_0 \\ \vdots \\ n_{63} \end{bmatrix}}_{\mathbb{n}}$$

solve: $\mathbb{C} = (R^T R)^{-1} R^T \mathbb{n}$ (+ normalize $F(\varrho)$)

→ These splines "encode and compress" the distribution characteristics via the spline coefficients $\{c_i\}$. Question: How can one optimally use c_i-values for classification?

▪ BEST APPROXIMATION OF DISTRIBUTIONS

- Considering the example (64-bin histogram density data approximated optimally with constant, linear, quadratic, cubic, quartic splines, these splines provide "differential characteristics" of the given data:

DIFFERENTIAL PROPERTIES

- constant spline ⇒ 5 function values
- Linear " ⇒ 4 1st deriv. "
- quadratic " ⇒ 3 2nd " "
- cubic " ⇒ 2 3rd " "
- quartic " ⇒ 1 4th " "

- One can also compute statistically relevant data for these splines, e.g., means and variances (and skewness, kurtosis) — or other "moment-based" characteristics... Example (with $F(\varrho)$ representing normalized spline):

STATISTICAL PROPERTIES

- mean $\mu = \int_{\varrho_0}^{\varrho_{64}} \varrho \cdot F(\varrho)\, d\varrho$
- variance $\sigma^2 = \int_{\varrho_0}^{\varrho_{64}} (\varrho - \mu)^2 \cdot F(\varrho)\, d\varrho$

- One can compute these "normalized spline density functions" for ALL SEGMENTS representing samples of the same material class. Thus, for each material class, one obtains DISTRIBUTIONS OF DIFFERENTIAL AND STATISTICAL CHARACTERISTICS. These "DISTRIBUTIONS OF DISTRIBUTIONS" can be used for classification. ~ BH