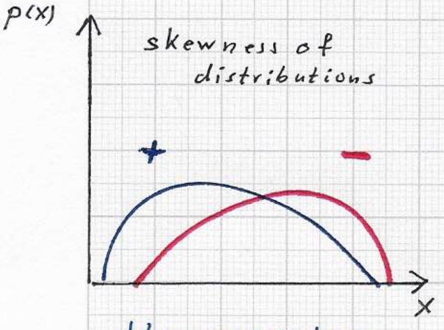
■ GENERATING OF SYNTHETIC DATA – Cont'd.

• **Review – Moments & other characteristics:** A variety of statistical properties of feature point/vector distributions can be used to establish meaningful definitions for the difference of two distributions or error of an approximation of a distribution. Several properties are based on MOMENTS; important ones are the following (defined for discrete, finite data sets $\{x_i\}_{i=1}^N$):



blue p(x): + skewness
red p(x): – skewness

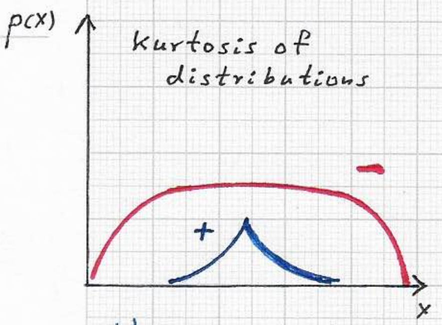
• mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

• variance: $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

• standard deviation: $\sigma = \sqrt{\sigma^2}$

• skewness: $\frac{1}{N} \sum_{i=1}^{N} \left((x_i - \bar{x})/\sigma \right)^3$

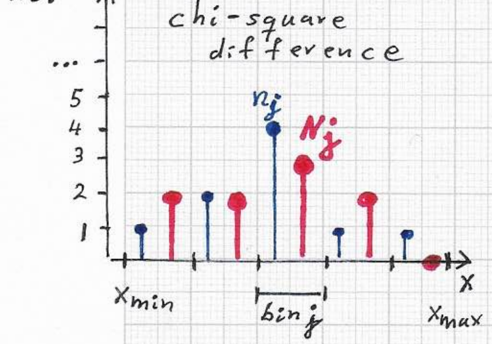
• kurtosis: $-3 + \frac{1}{N} \sum_{i=1}^{N} \left((x_i - \bar{x})/\sigma \right)^4$



blue p(x): + kurtosis
red p(x): – kurtosis

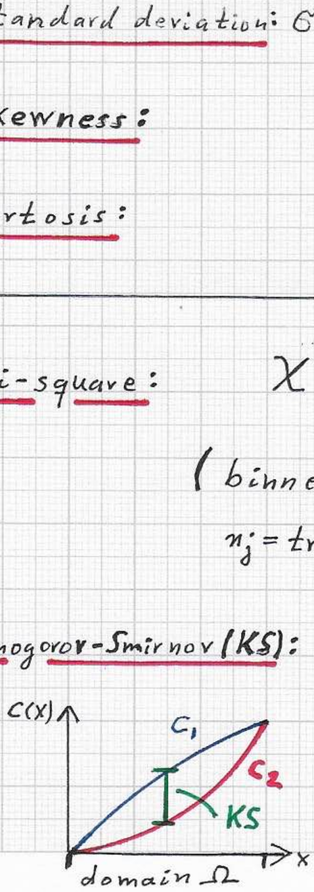
• chi-square: $\chi^2 = \sum_{j=1}^{B} (N_j - n_j)^2 / n_j \quad , n_j \neq 0$

(binned data, bins 1 … B; n_j = true, expected no., N_j = observed no.)



• true, expected
• measured, observed

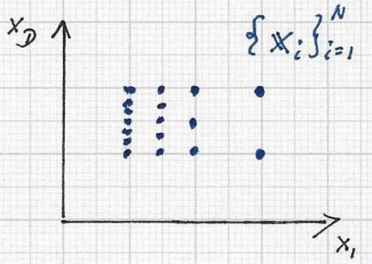
• Kolmogorov–Smirnov (KS): $KS = \max_{\Omega} | c_1(x) - c_2(x) |$

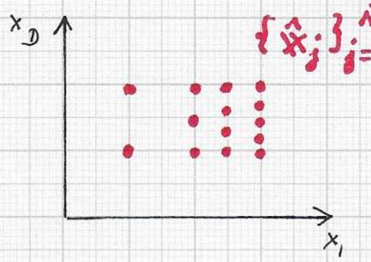


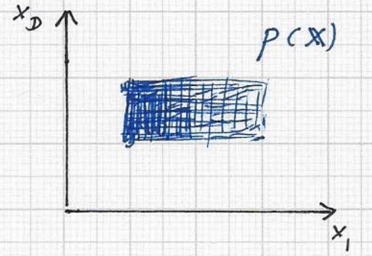
(analytically defined data; consider absolute difference of cumulative distributions $c_1(x)$ and $c_2(x)$)

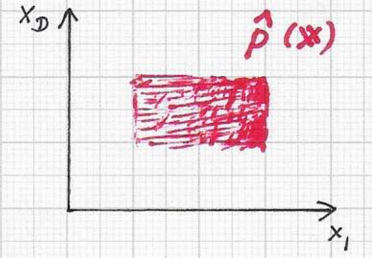
■ GENERATION OF SYNTHETIC DATA – Cont'd.

- **Review – Difference measures for distributions:**









given feature point samples $\{\mathbf{x}_i\}$ and $\{\hat{\mathbf{x}}_j\}$; constructed distribution functions $p(\mathbf{x})$ and $\hat{p}(\mathbf{x})$, used to determine similarity of $\{\mathbf{x}_i\}$ and $\{\mathbf{x}_j\}$.

Functions $p(\mathbf{x})$ and $\hat{p}(\mathbf{x})$ should be normalized over their common domain Ω with volume $|\Omega|$.

One must often define and compute meaningful distances/differences between two distributions of feature point/vector data over some (finite) domain Ω in D-dimensional feature space. For example, one might have to compare two discrete, finite feature value distribution sets $\{\mathbf{x}_i\}_{i=1}^{N}$ and $\{\hat{\mathbf{x}}_j\}_{j=1}^{\hat{N}}$. The two sets each "imply" an underlying analytical function p, $\hat{p}$ that represents the respective sample $\{\mathbf{x}_i\}$, $\{\hat{\mathbf{x}}_j\}$. (Piecewise polynomial best approximations could be thought of.) Regardless of the specific method chosen to construct a function $p(\mathbf{x})$ from $\{\mathbf{x}_i\}$, and $\hat{p}(\mathbf{x})$ from $\{\hat{\mathbf{x}}_j\}$, one can consider a variety of measures for the comparison of $\{\mathbf{x}_i\}$ and $\{\mathbf{x}_j\}$. When evaluating $p(x)$ in a uniformly integrated sense, point sets of "class $\{\mathbf{x}_i\}$" result; similarly $\hat{p}(x)$ produces point sets of "class $\{\hat{\mathbf{x}}_j\}$" when evaluated uniformly in integrated fashion. Thus, one can define and compute difference measures for p and $\hat{p}$ to define difference of $\{\mathbf{x}_i\}$, $\{\hat{\mathbf{x}}_j\}$:

- root-mean-square (RMS): $\left(\frac{1}{|\Omega|}\int_{\Omega}\left(p(\mathbf{x})-\hat{p}(\mathbf{x})\right)^2 d\mathbf{x}\right)^{1/2}$

- maximal difference: $\max_{\Omega} |p(\mathbf{x})-\hat{p}(\mathbf{x})|$

- discrete RMS: $\left(\frac{1}{N}\sum_{i=0}^{N}\left(p(\mathbf{x}_i)-\hat{p}(\mathbf{x}_i)\right)^2\right)^{1/2}$,
$\left(\frac{1}{\hat{N}}\sum_{j=1}^{\hat{N}}\left(p(\hat{\mathbf{x}}_j)-\hat{p}(\hat{\mathbf{x}}_j)\right)^2\right)^{1/2}$.

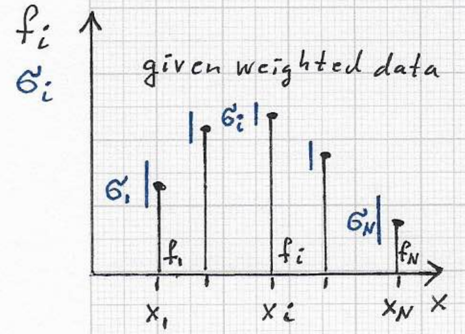
■ GENERATION OF SYNTHETIC DATA – Cont'd.

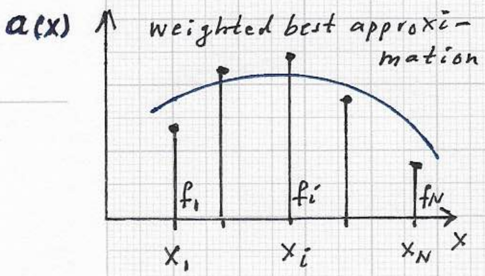
• Chi-square/weighted best approximation:

The statistical properties variance (σ^2) and standard deviation ($\sqrt{\sigma^2}$) are often considered as "WEIGHTS" in the context of constructing a best approximation for a given distribution of data (given either as discrete/binned data or as a distribution function). If it is possible to determine meaningful variance/standard deviation values to discrete/binned data, or analytically defined variance/standard deviation functions to distribution functions, then one can perform WEIGHTED BEST APPROXIMATION (or CHI-SQUARE APPROXIMATION). Often the value of χ^2 is minimized when performing best approximation of discrete, binned distribution data. Specifically, and considering a more general discrete data approximation setting, one is given values x_i in the domain, dependent values f_i at the locations x_i, and standard deviations σ_i associated only with the f_i values. Chi-square approximation minimizes the quantity

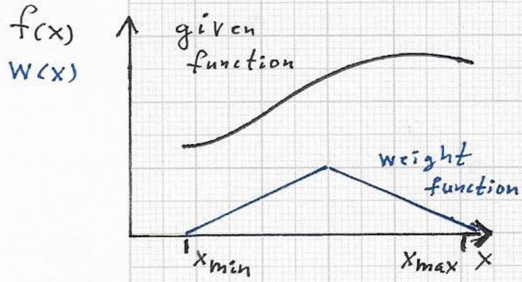
$$\chi^2 = \sum_{i=1}^{N} \left((a(x_i) - f_i)/\sigma_i \right)^2 .$$

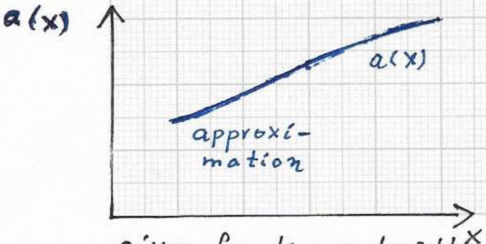
The function $a(x)$ is the best approximation to be computed; $1/\sigma_i^2$ serves as a "weight" for f_i.





given discrete data (x_i, f_i), with standard deviations σ_i (="weights") for each f_i value; weighted best approximation emphasizing low-σ data





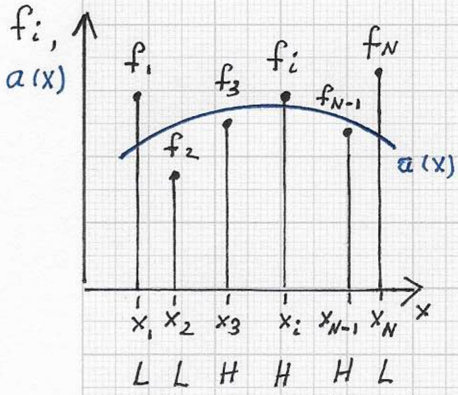
given function and weight function; weighted best approximation emphasizing "middle part" of f(x)

GENERATION OF SYNTHETIC DATA – Cont'd.

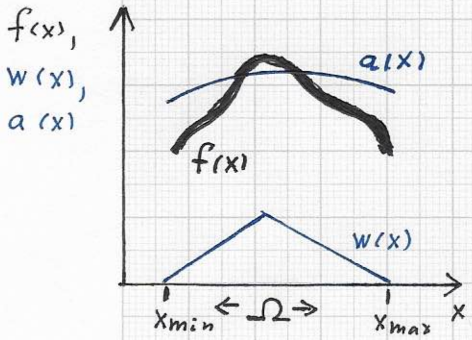
- **Weighted best approximation:**

We briefly review the method of weighted best approximation. It is important when constructing an analytical representation of a given distribution of feature point/vector data via an approximation over a region in D-dimensional feature space. In some cases it is possible to assign variance/standard deviation values to an individual feature point $\mathbf{x}_i = (x_1^i, \ldots, x_D^i)^T$ and/or to the associated probability values p_j that characterize the distribution implied by $\{\mathbf{x}_i\}_{i=1}^N$ via a set of p_j-values at certain locations $\mathbf{x}_j$. The p_j-values must be approximated with weights w_j. The weights w_j are inversely proportional to variance/standard deviation: A datum with low variance/standard deviation has high weight and vice versa. Thus, we can use weighted best approximation to compute the needed distribution function $p(\mathbf{x})$ from p_j- and w_j-values given at points $\mathbf{x}_j$. For example, one could use a polynomial (or piecewise polynomial) function as approximation.

Weighted best approximation can also be used when the given data are a function $f(\mathbf{x})$ (to be approximated) with an associated weight function $w(\mathbf{x})$. The approximation $a(\mathbf{x})$ will "emphasize" high-weight regions.



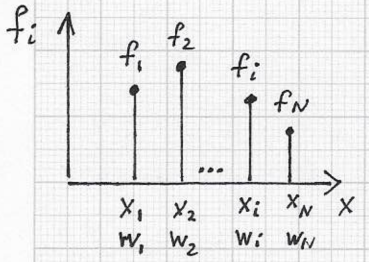
data (x_i, f_i) with low (L) or high (H) weights to be approximated with $a(x)$ – "DISCRETE CASE"



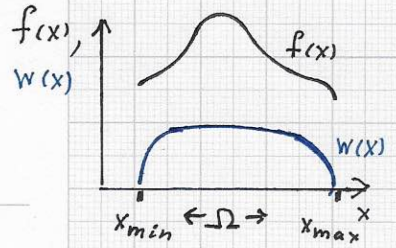
function $f(x)$ and weight function $a(x)$ given, weighted best approximation $a(x)$ computed – "CONTINUOUS CASE"

■ GENERATION OF SYNTHETIC DATA – Cont'd.

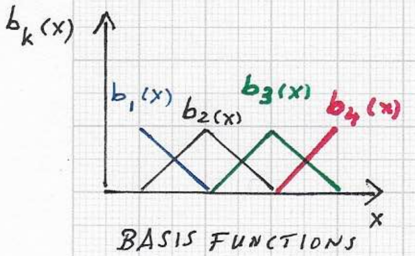
• **Weighted best approximation:**



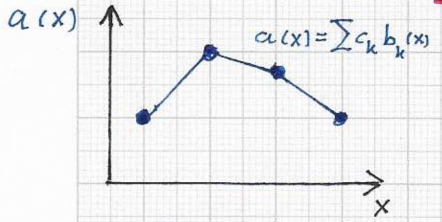
DISCRETE CASE



CONTINUOUS CASE



BASIS FUNCTIONS $b_k(x)$, $k=1...4 (=K)$



WEIGHTED BEST APPROXIMATION = LINEAR COMBINATION OF $b_i(x)$, $i=1...K$

One usually employs least-squares approximation to compute the unique weighted best approximation in the discrete and continuous cases. Thus, one minimizes the following quantities:

$$\left(\sum_{i=1}^{N} w_i \left(a(\mathbf{x}_i) - f_i \right)^2 \right), \quad w_i > 0,$$

$$\left(\int_{\Omega} w(\mathbf{x}) \left(a(\mathbf{x}) - f(\mathbf{x}) \right)^2 \right), \quad w(\mathbf{x}) > 0 .$$

The weights $w_i > 0$ (or the weight function $w(\mathbf{x})$) should be "normalized" in such a way that a w-value represents the probability of a function to have the value f_i ($f(\mathbf{x})$) at location $\mathbf{x}_i$ ($\mathbf{x}$).

The weighted best approximation $a(\mathbf{x})$ itself is a linear combination of appropriate basis functions $b_k(\mathbf{x})$ (e.g., polynomials, spline basis functions or radial basis functions):

$$a(\mathbf{x}) = \sum_{k=1}^{K} c_k \, b_k(\mathbf{x}) \quad .$$

The unknown coefficients c_k result from solving the normal equations defined by inserting $\sum_{k=1}^{K} c_k b_k(x)$ into the expressions for the disivete or continuous case to be minimized.

~ BH