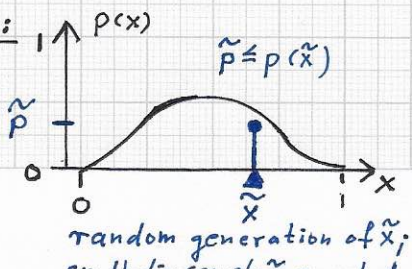
GENERATION OF SYNTHETIC DATA WITH RANDOM NUMBERS AND REJECTION SAMPLING

• **GOAL:** In many applications, the amount of data (e.g., images, 3D scans) is rather limited – Limited in the sense that a machine learning-based data classification method usually requires much more "sample data" than available, to perform highly reliable classification. Thus, there is a need for automatic techniques that can synthetically generate "artificial sample data" that can serve as viable data for augmenting an original, relatively small sample data set.

We are concerned with multi-material classification where a material class is represented by its "fingerprint" in high-dimensional feature space. Given a set of samples of the same material (class), the samples together define a distribution of feature values over the underlying D-dimensional feature space.

The purpose of an acceptance-rejection-based stochastic method is the generation of more, synthetic, sample data, by using random numbers to create feature value distributions that greatly resemble those of the given, actual data.

• **Ex:**



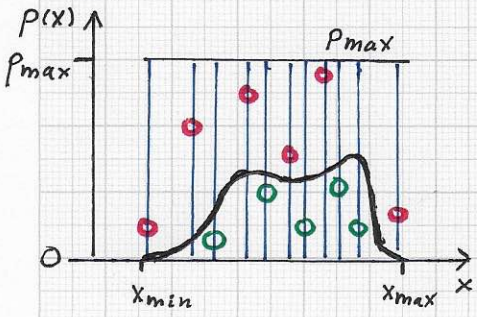
random generation of $\tilde{x}$; synthetic sample $\tilde{x}$ accepted.

We consider the 1D, univariate case. Knowing a distribution function $p(x)$, $x \in [0,1]$ and $p(x) \in [0,1]$, for example, randomly generate a value $\tilde{x} \in [0,1]$ and a value $\tilde{p} \in [0,1]$; if $\tilde{p} \le p(\tilde{x})$ then accept $\tilde{x}$.

Stratovan

■ GENERATION OF SYNTHETIC DATA – Cont'd.

• Stochastic method:



random generation of tuples $(\tilde{x}, \tilde{p})$; $\tilde{x}$-values with an associated tuple $(\tilde{x}, \tilde{p})$ are accepted as "samples."

→ We consider the 1D, univariate case.

→ The x-space defines the 1D feature space; feature values x lie in $[x_{min}, x_{max}]$.

→ A distribution function $p(x)$, $p(x) \in [0, p_{max}]$, is known. (For example, $p(x)$ could be a piecewise, low-degree polynomial function constructed from finite, discrete data.)

→ It is possible to randomly generate $\tilde{x}$-values between x_{min} and x_{max} and $\tilde{p}$-values between 0 and p_{max} with UNIFORM PROBABILITY.

→ The region $[x_{min}, x_{max}] \times [0, p_{max}]$ is subdivided into two sub-regions:

$$A = \{(\tilde{x}, \tilde{p}(\tilde{x})) \mid \tilde{x} \in [x_{min}, x_{max}] \wedge \tilde{p}(\tilde{x}) \in [0, p(\tilde{x})]\},$$

$$R = \{(\tilde{x}, \tilde{p}(\tilde{x})) \mid \tilde{x} \in [x_{min}, x_{max}] \wedge \tilde{p}(\tilde{x}) \in [p(\tilde{x}), p_{max}]\}.$$

(In other words, tuples $(\tilde{x}, \tilde{p})$ are ACCEPTED when they lie under the curve $(x, p(x))$ and are REJECTED otherwise.)

ACCEPTANCE – REJECTION – method/sampling

• Algorithm: Input: $x_{min}, x_{max}, p_{max}, p(x), K$

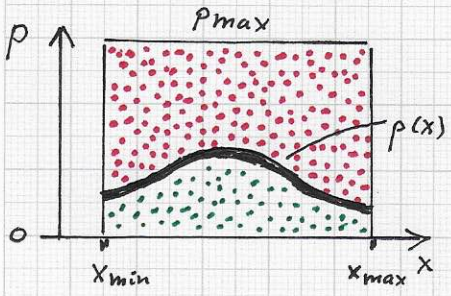
Output: $\{x_i\}_{i=1}^{K}$

Method:
- randomly generate $\tilde{x} \in [x_{min}, x_{max}]$;
- " " $\tilde{p} \in [0, p_{max}]$;
- if $\tilde{p} \leq p(\tilde{x})$ then add $\tilde{x}$ to set $\{x_i\}$;

/* stop when K $\tilde{x}$-values have been accepted. */

■ GENERATION OF SYNTHETIC DATA - Cont'd.

• Note: → The discrete synthetic sample data set $\{x_i\}_{i=1}^{K}$ has an associated "discrete distribution function" that resembles the analytical given distribution function $p(x)$ – since $p(x)$ is used in the acceptance-rejection method. Via the use of proper kernels / kernel functions, one can construct also an analytical distribution function based on the set $\{x_i\}_{i=1}^{K}$. For $K \to \infty$ this function converges to $p(x)$.



accepted values $\tilde{x}$ defining a new distribution that converges to $p(x)$.

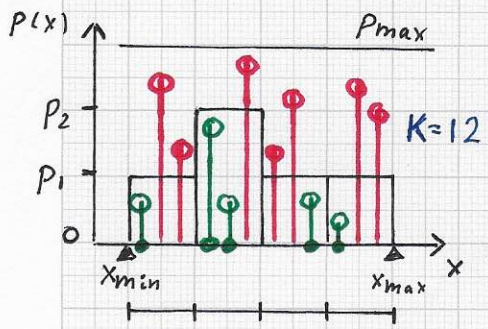
STOCHASTIC AUGMENTATION

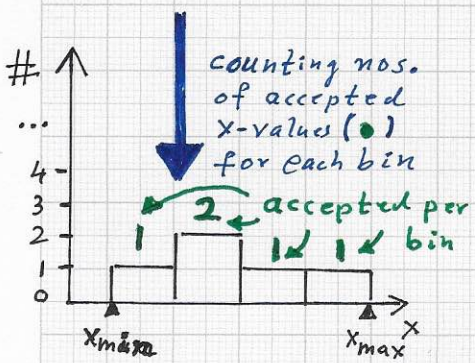
→ This stochastic, random number-based approach does not require one to define or use a cumulative distribution. It also avoids the need for a complex optimization method (with complicated feature space boundary conditions).

→ Depending on the needs of a specific machine learning-based classification method, one must generate specific numbers of "synthetic augmentation data sets" $\{x_i\}_{i=1}^{K}$; further, the value of K can be used as a parameter to generate synthetic feature value distributions that resemble those of the given, real data "more or less accurately." In addition, it is possible to use $\{x_i\}_{i=1}^{K}$ "as is," i.e., as a discrete distribution representation, or one can compute an analytical approximation of it.

■ GENERATION OF SYNTHETIC DATA - Cont'd.

• Ex.: We consider a 1D, univariate example where the analytical definition of the distribution function p(x) is piecewise constant - represented by 4 bins in x-space and bin-specific p-values (p_1 and p_2 in the figure). The p(x) function can now be used to randomly generate synthetic tuples with x- and p-values — where x-values are based on random number generation between x_{min} and x_{max} (uniform probability); and p-values are between 0 and p_{max} (also generated with uniform probability). An x-value is accepted when its associated random p-value is smaller than or equal to the given p-function's value at x. The example shows the random generation of synthetic histogram data where 5 x-values pass the "acceptance test" and are placed into their corresponding bins. (In the shown special case, the randomly generated histogram data represents the same 4-bin analytical distribution that is originally given. Generally, the synthetic histogram data do not replicate exactly the given distribution p(x).)



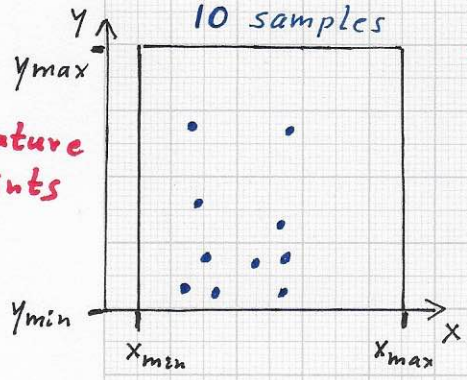


random generation of 12 x- and associated p-values; accepting 5 x-values based on given piecewise constant distribution function p(x).

Data augmentation: Generate several synthetic histogram data, each histogram obtained via a random process generating K tuples of x- and p-values

■ GENERATION OF SYNTHETIC DATA - Cont'd.

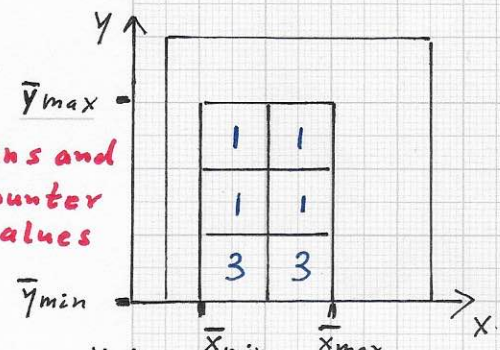
• Stochastic method for multi-dim. case:



feature points

In the general high-dimensional case, one is given a sample distribution of feature points / vectors over a finite D-dimensional domain. The figures illustrate the case of a 2D feature space.

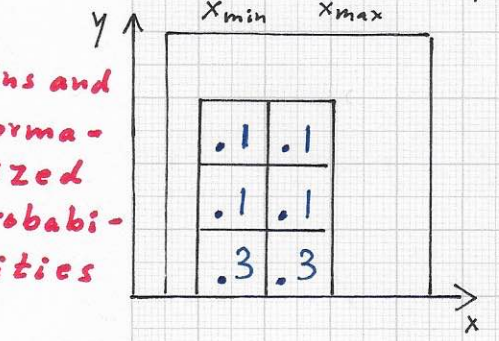
First, assume that the generally allowable ranges for x- and y-values are the intervals $[x_{min}, x_{max}]$ and $[y_{min}, y_{max}]$, respectively.



bins and counter values

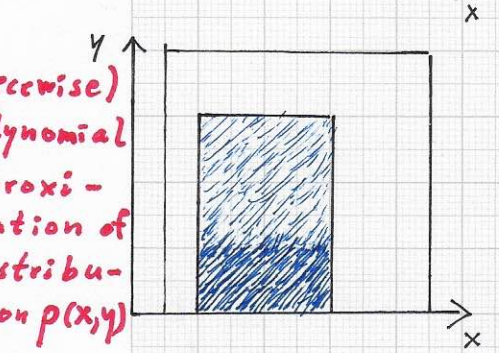
A specific sample data set generally has smaller "effective" ranges, i.e., $[\bar{x}_{min}, \bar{x}_{max}]$ and $[\bar{y}_{min}, \bar{y}_{max}]$. A sample data set of feature points $(x,y)^T$ define an original scattered data set - without connectivity data or bins.

Second, one can construct a set of bins covering the domain in feature space with samples; each bin stores the number of samples inside the bin.



bins and normalized probabilities

Third, it is advantageous to normalize the distribution values for all bins - by normalizing with respect to the total number of samples. (The random number-based synthetic sample generation procedure benefits from normalization.)



(piecewise) polynomial approximation of distribution p(x,y)

Fourth, one can also compute, via best approximation, a (piecewise) low-degree polynomial representation $p(x,y)$ to generate synthetic samples.

BH