# Estimating Probability Densities from Numeric Samples

Hoa Nguyen

School of Computational Science
& Department of Mathematics
Florida State University
nguyen@csit.fsu.edu

*Collaborators*: Max Gunzburger, Yuki Saka, John Burkardt

# The problem

- A fundamental problem of statistics is to estimate an unknown probability density function (PDF) that has generated a given set of sample points.
- The sample pointset can be equally spaced percentiles of the PDF, or just a set of random deviates generated by the PDF.
- We propose a simple method to construct an approximate model of the unknown PDF, based ONLY on the given set of sample points.

# Outline

- Review Basic Statistics in 1D.
- Propose the method.
- Present numerical results.
- Extensions.

# Cumulative Distribution Function (CDF)

Given a random variable $X : \Omega \longrightarrow R$ defined on the probability space $\Omega$, the cumulative distribution function (CDF) is a function $F$ giving the probability that the random variable $X$ is less than or equal to some value $a$, for every $a \in R$.

$$\begin{aligned} F(a) &= P(X \leq a) \\ &= \int_{-\infty}^{a} f(z)dz \end{aligned}$$

equivalent to

$$f(z) = \frac{d}{dz}F(z)$$

where $f$ is the Probability Density Function (PDF).

# Probability Density Function (PDF)

The probability density function (PDF) of a random variable $X$ is a function $f$ *that can be integrated*, and satisfies 2 conditions:

1) $f(z) \geq 0, \forall z$

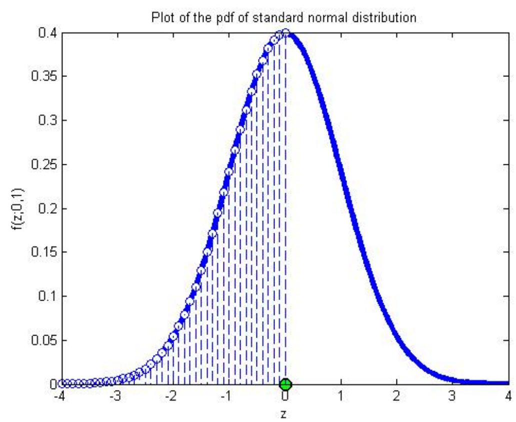
2) $\int_{-\infty}^{\infty} f(z) dz = 1$

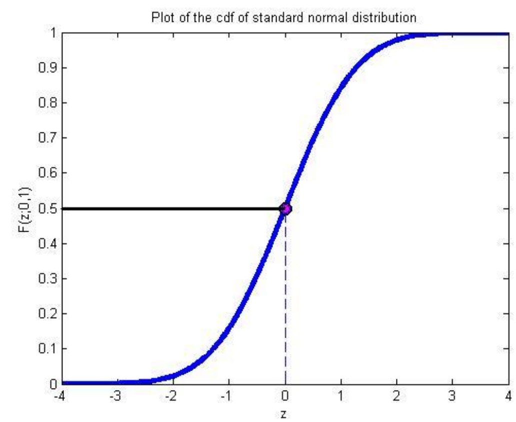
# PDF and CDF of standard normal distribution

The PDF of the standard normal (Gaussian) distribution of mean $\mu = 0$ and variance $\sigma = 1$ is $f(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$.

The CDF of standard normal/ Gaussian distribution is

$$F(z) = \int_{-\infty}^{z} f(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-t^2}{2}} dt = \frac{1}{2}[1 + erf(\frac{z}{\sqrt{2}})]$$



# Inverse Cumulative Distribution Function (ICDF)

- CDF $F(z) = P(X \leq z)$ is a function from R to $(0, 1)$, continuous from the right, and monotone increasing.

- $F$ has a well-defined inverse $F^{-1}$, called Inverse Cumulative Distribution Function (ICDF).

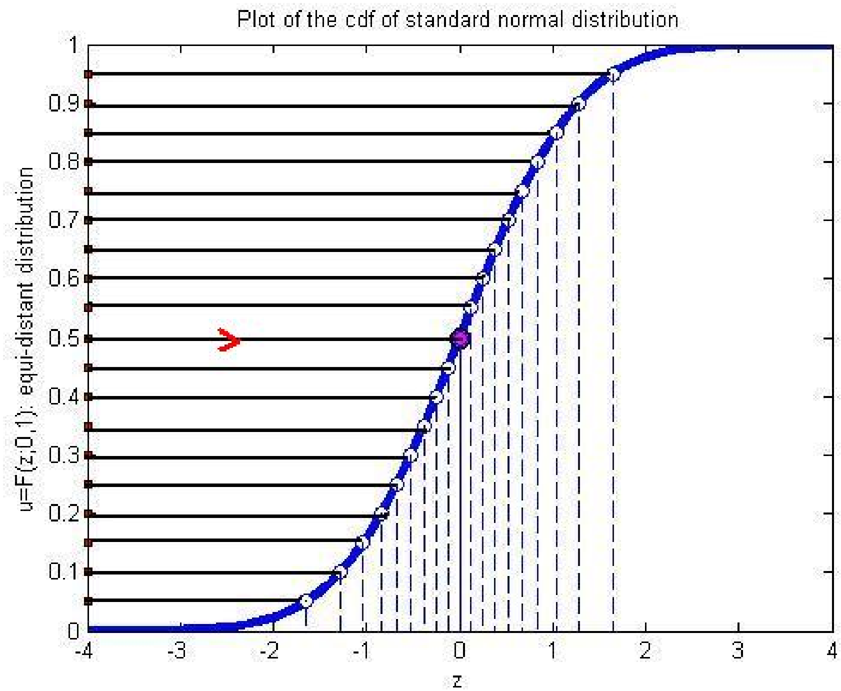
- ICDF $F^{-1}(u) = inf\{z|F(z) = u, 0 < u < 1\}$

- If $u$ is a value in $(0, 1)$, then $z = F^{-1}(u)$ follows the CDF $F$, i.e., $P(X \leq z) = F(z) = u$.

# Use ICDF to generate a pointset of the associated PDF

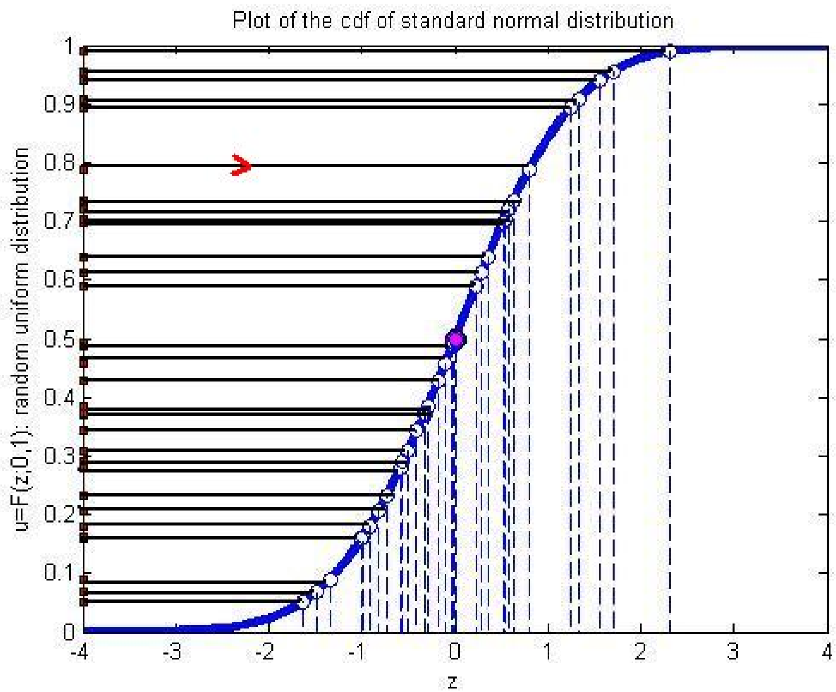
Generate a pointset of equally spaced percentiles of the PDF of standard normal distribution.

"Equally-spaced" Type



Generate a set of random deviates from the PDF of standard normal distribution.

## "Random" Type



# The inverse problem in statistics

- PDF $\longrightarrow$ A pointset
- A sample pointset $\longrightarrow$ PDF ?

# A simple method

Given

- A sample pointset $P = \{z_i\}_{i=1}^N$ in $(a, b) \subset R$ (assume P is in the increasing order)

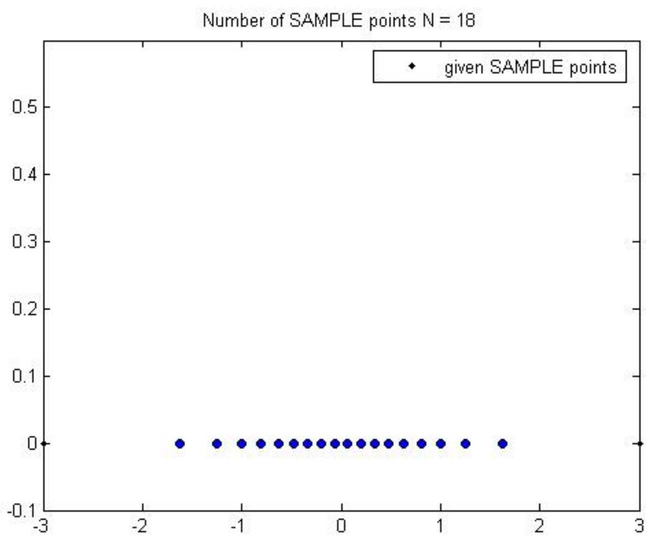
- The partition $V$ of $(a, b)$ into $N$ Voronoi regions $\{V_i = [t_{i-1}, t_i)\}_{i=1}^N$ such that $t_0 = a$ , $t_i = \frac{z_i + z_{i+1}}{2}$ $(i = 1, ..., N-1)$ , $t_N = b$.

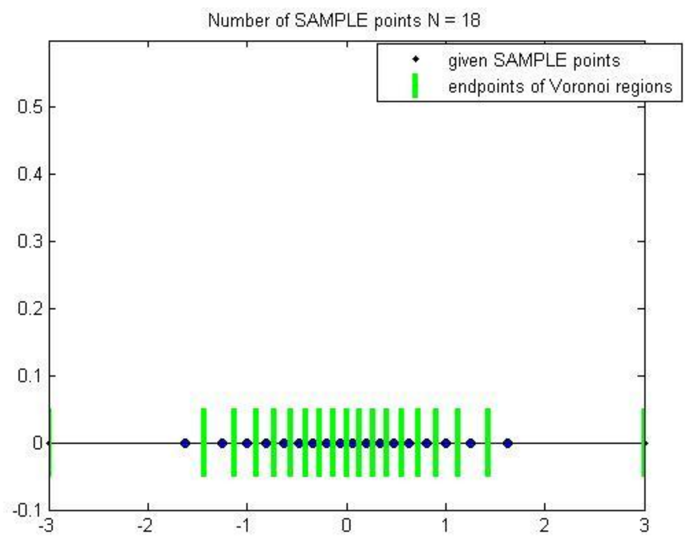
Define the point density at any point $x \in V_i = [t_{i-1}, t_i), i = 1, ..., N$ as

$$\lambda(x) = \frac{1}{N h_i}$$

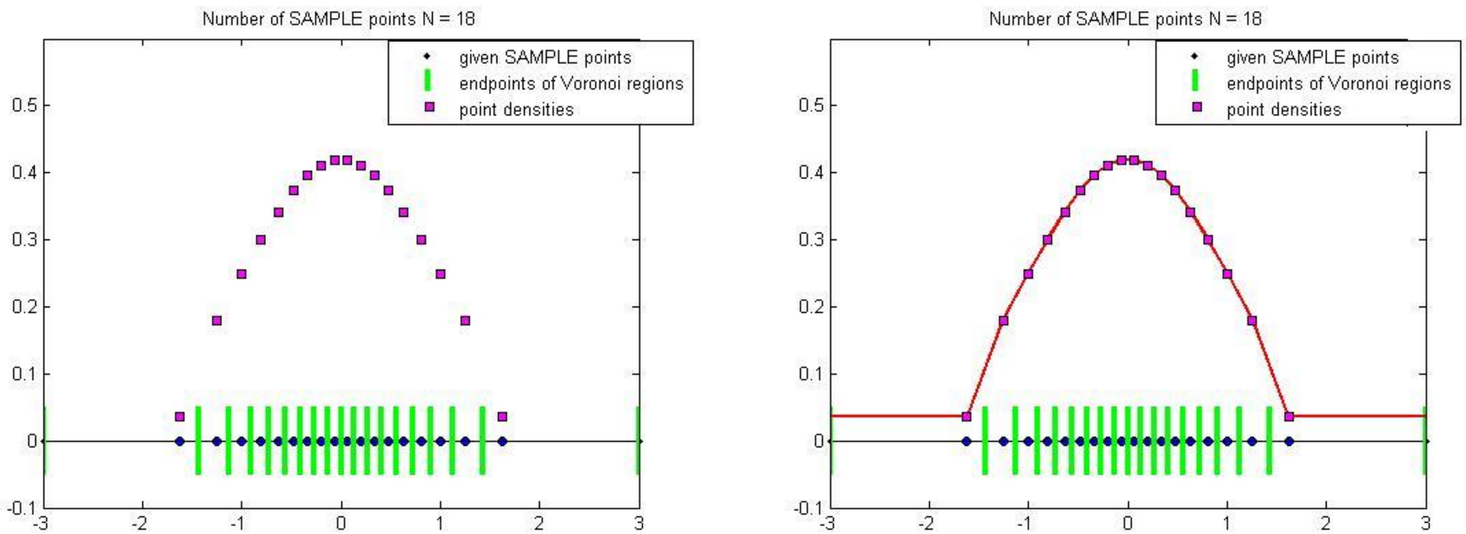
where $h_i = t_i - t_{i-1}$ is the bin width of $V_i, i = 1, .., N$.

# Approximate Density based on a sample pointset of "Equally-spaced" Type



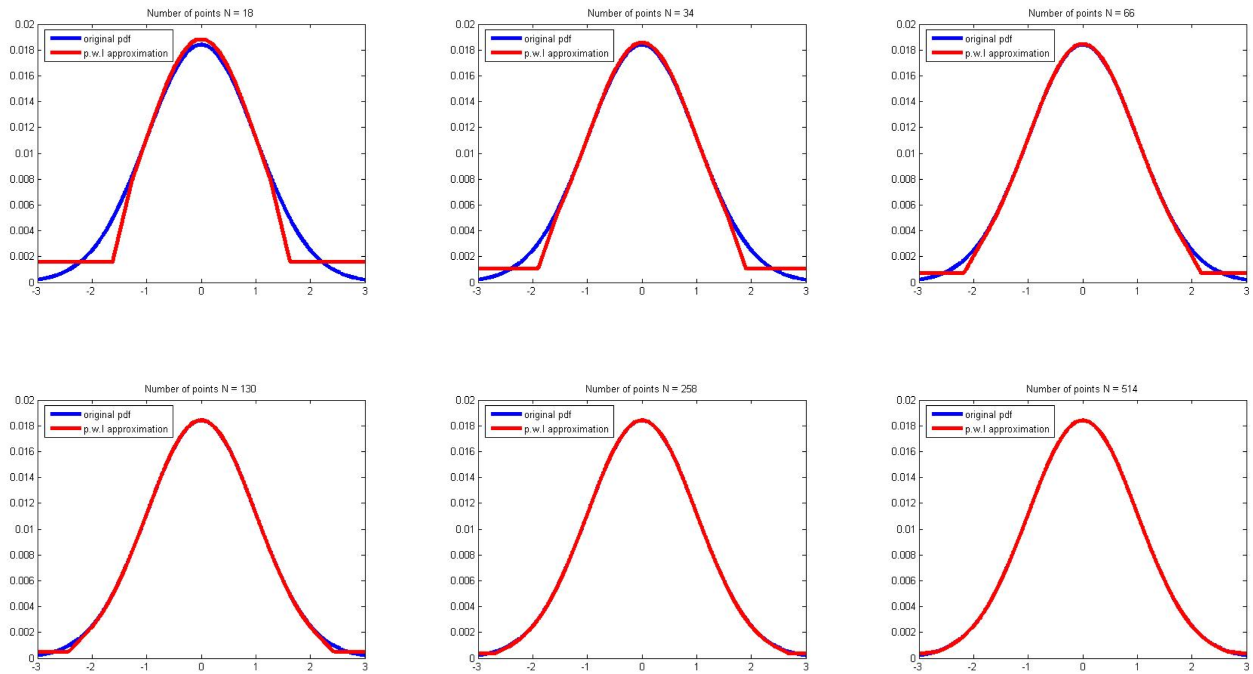




*18 given sample points of "Equally-spaced" Type in blue circles, the associated Voronoi endpoints in green bars*




*point densities* $\{\lambda(z_i)\}_{i=1}^N$ *in magenta diamonds,*
*the piecewise linear approximation in red line segments*

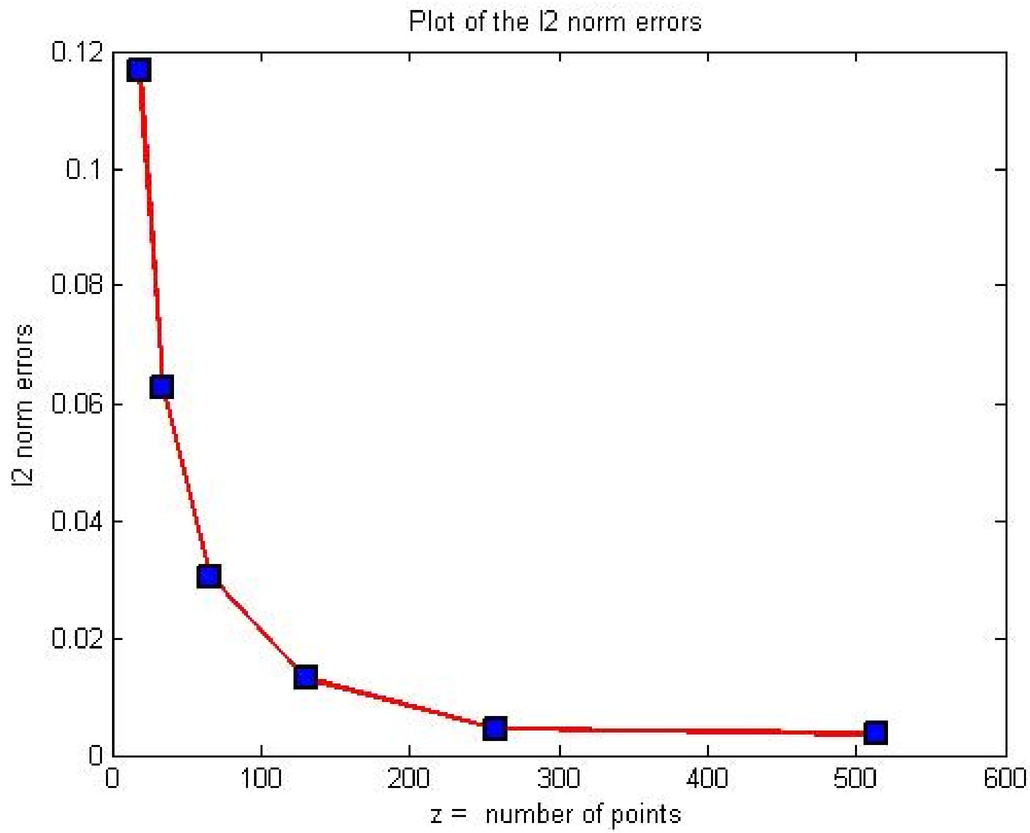
We need to *nomalize* the approximate density and the PDF of the standard normal distribution to compare them.




*PDF of standard normal distribution (blue) and the approximate density (red)*

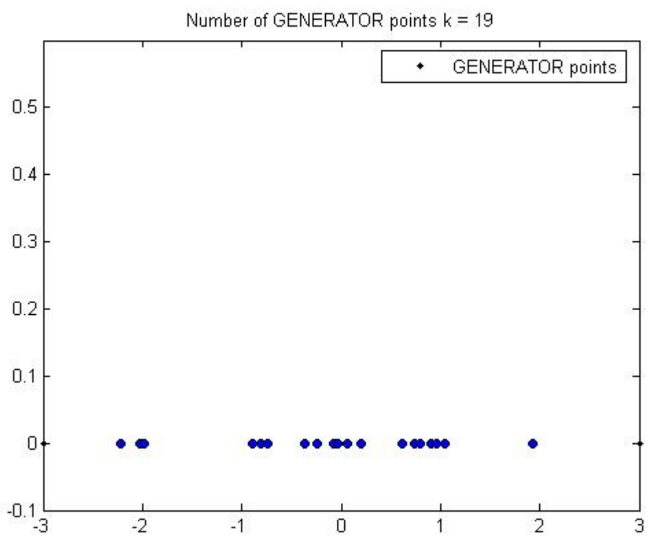
| N | 18 | 34 | 66 | 130 | 258 | 514 |
|---|---|---|---|---|---|---|
| $E_{l2norm}$ | 0.1167 | 0.0625 | 0.0304 | 0.0131 | 0.0045 | 0.0036 |

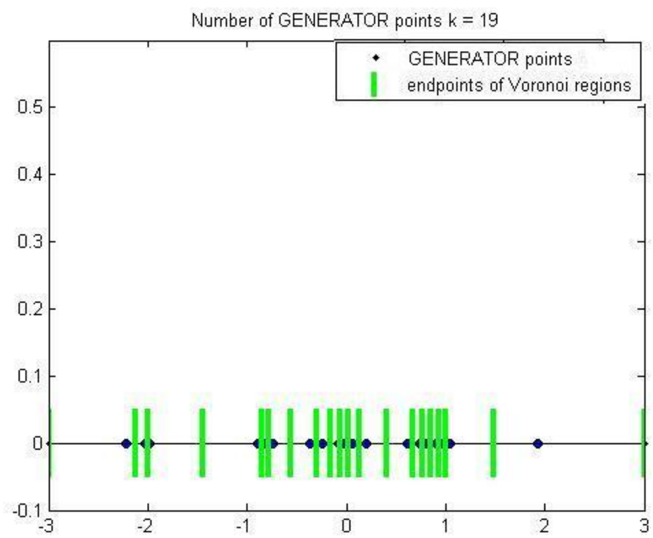
where $E_{l2norm} = \sqrt{\Sigma_{i=1}^{10000}}(f(x_i) - f_{app}(x_i))^2$, $\{x_i\}_{i=1}^{10000}$: equi-distant pointset in $[-3, 3]$.



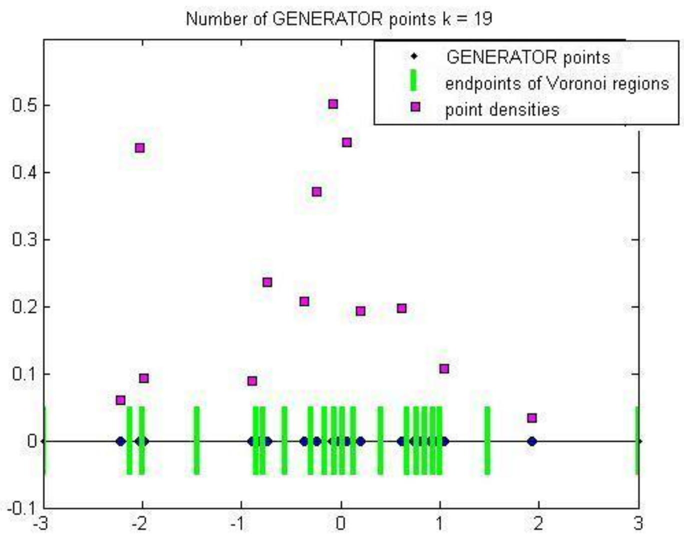

$E_{l2norm}$ *errors when number of sample points = 18, 34, 66, 130, 258 and 514.*

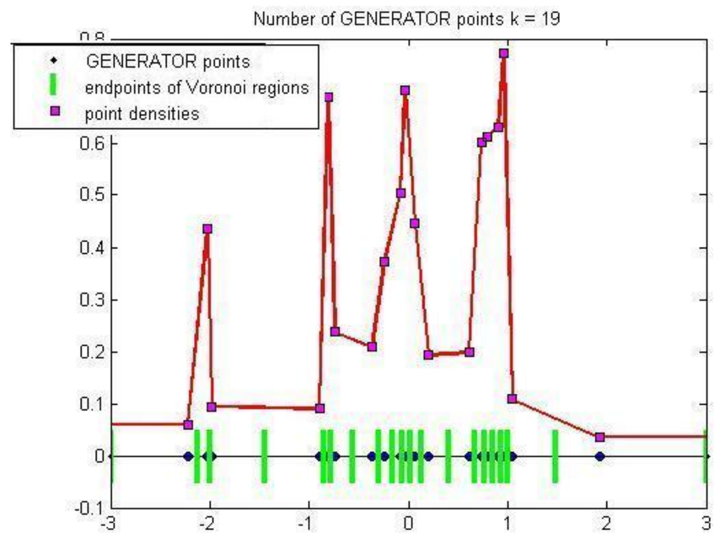
# Approximate Density based on a sample pointset of "Random" Type




*19 given sample points of "Random" Type in blue circles, the associated Voronoi endpoints in green bars*




*point densities* $\{\lambda(z_i)\}_{i=1}^{N}$ *in magenta diamonds,*
*the piecewise linear approximation in red line segments*

$\Longrightarrow$ Need to generate the NEW pointset which is equally spaced percentiles of the same PDF as the one of the given sample pointset (regularization method).

The NEW pointset is called generator pointset.

# Regularization Method = Optimization Problem

Given

- A sample pointset $P = \{z_i\}_{i=1}^{N}$ in $(a, b) \subset R$
- The probability density $\Phi(x)$ defined on $(a, b)$

Want to find the generator pointset $G$ of $k$ points $\{c_j\}_{j=1}^{k}$ , $(k << N)$, on $(a, b)$ to minimize the 2nd-power distortion:

$$D = \frac{1}{k}\Sigma_{j=1}^{k}\Sigma_{z_i \in V_j \cap P}\Phi(z_i)|z_i - c_j|^2 dx \tag{1}$$

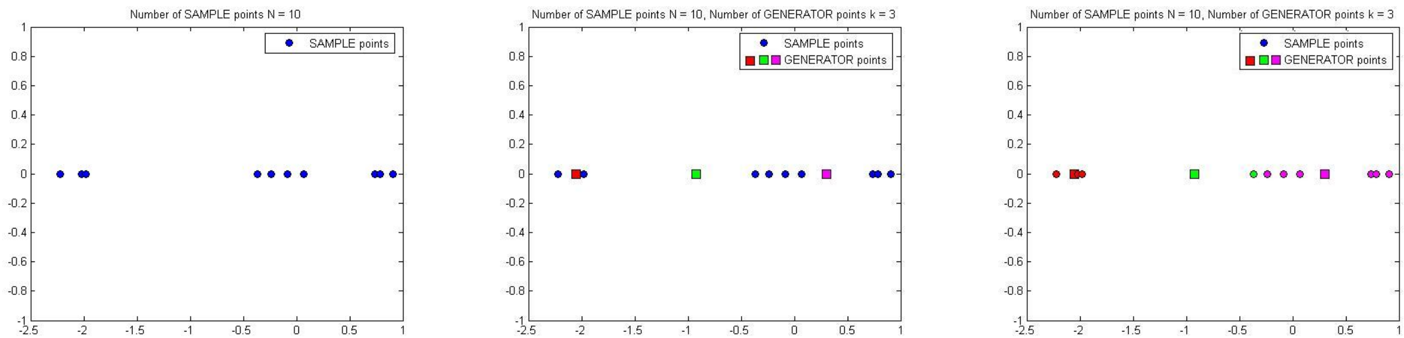
where $V = \{V_j = [t_{j-1}, t_j)\}_{j=1}^{k}$ are the Voronoi regions of $G = \{c_j\}_{j=1}^{k}$.

The quantizer (G,V) that minimizes the 2nd-power distortion (1) has

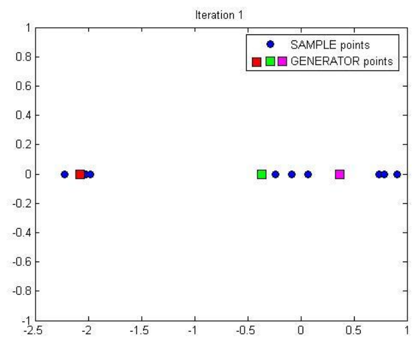
$$\lambda(x) \cong \frac{\Phi(x)^{1/3}}{\int \Phi(x)^{1/3} dx}$$

This property is called Optimum Quantizer Point Densities (OQPD).

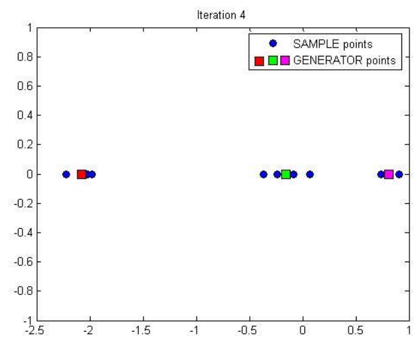
# Regularization by K-means Algorithm




$\Rightarrow$ one step




$\Rightarrow$ many steps




*Use K-means algorithm to regularize 10 sample points by 3 generator points*

## Approximate Density based on the generator pointset

- The generator pointset $G$ from the K-means algorithm minimizes the 2nd-power distortion (1).

- According to OQPD, the generator pointset $G$ has the point density $\lambda(x)$ proportional to the original PDF $\Phi(x)$ of the sample pointset $P$ raised to the power $\frac{1}{3}$.

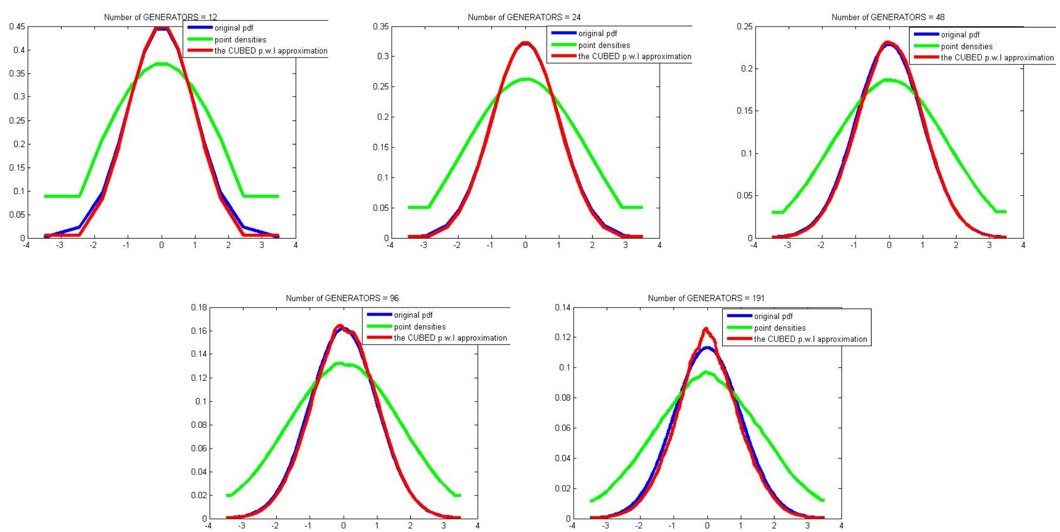
  $\Longrightarrow$ The original PDF of the sample pointset P is:

$$\Phi(x) \cong c\lambda(x)^3$$

  where $c$ is some constant.

# Approximate Density based on the generator pointset, given 9,995,297 sample points of "Random" Type

From left to right, top to bottom: the number of generator points increases as 12, 24, 48, 96 and 191.



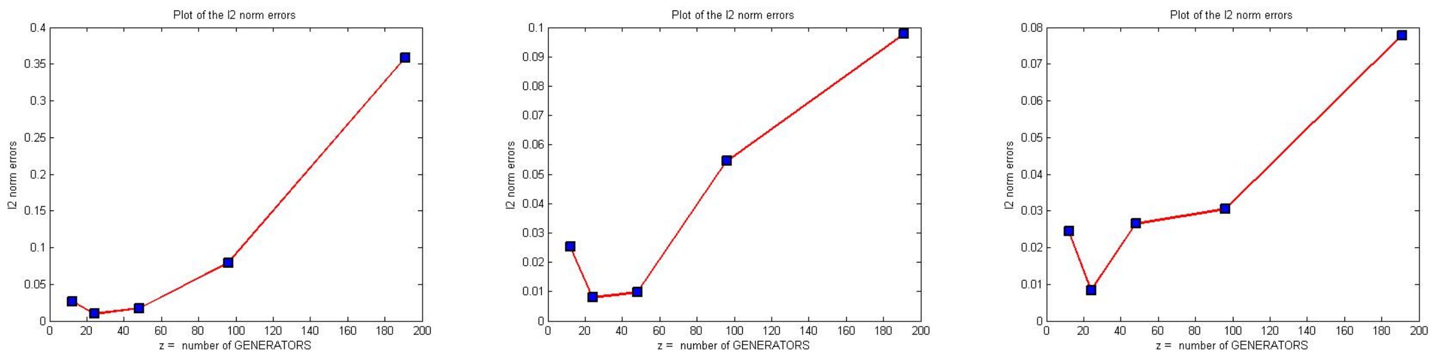

*PDF of standard normal distribution (blue),*
*the point densities of the generator pointset (green),*
*the CUBED piecewise linear approximate density (red)*

*Table of $E_{l2norm}$ errors*

| k | N=1 million | N=5 millions | N=10 millions |
|---|---|---|---|
| 12 | 0.0264 | 0.0253 | 0.0244 |
| 24 | 0.0095 | 0.0080 | 0.0083 |
| 48 | 0.0169 | 0.0097 | 0.0265 |
| 96 | 0.0792 | 0.0544 | 0.0304 |
| 191 | 0.3584 | 0.0977 | 0.0778 |

$k$: number of generator points, $N$: number of sample points

where $E_{l2norm} = \sqrt{\Sigma_{i=1}^{10000}}(f(x_i) - f_{app}(x_i))^2$, $\{x_i\}_{i=1}^{10000}$: equi-distant pointset in $[-3.5, 3.5]$.




*From left to right:* $E_{l2norm}$ *errors (k* $=$ *12, 24, 48, 96, 191) when given 1 million sample points, 5 million sample points, 10 million sample points.*

$\Longrightarrow$ The number of generators that minimizes the $E_{l2norm}$ errors is $k = 24$, which approximately follows Sturges' rule: $k = 1 + log_2 N$.

# Extensions

- Experiment the method on different types of PDF and on the PDF's of higher dimesnsions.

- Utilize various tests in statistics (such as Kolmogorov - Smirnov, Anderson - Darling, etc.) to compare the approximate density function with the original PDF's.

- Compare the results of our method with the ones of the most widely used density estimators such as Histograms, the Kernel estimator, etc.

- Use the approximate density function to generate the pointsets of any number of points; then compare them with the pointsets generated from the original PDF.

# References

[1] Sangsin Na and David L. Neuhoff; Bennett's Integral for Vector Quantizers, *IEEE Trans. Inform. Theory* **41(4)**, July 1995.

[2] Q. Du, M. Gunzburger, and L.-L. Ju; Meshfree, probabilistic determination of point sets and support regions for meshless computing, *Comput. Meths. Appl. Mech. Engrg* **191** 2002, pp. 1349-1366.