Truncation and Rounding

I.Nelson
AP, ECE
SSN College of Engineering

Errors resulting from Rounding and Truncation

- Rounding and truncation introduces an error depending on the number of bits in the original number relative to the number of bits after quantization.

- The characteristics of the errors introduced through either truncation or rounding depend on the particular form of number representation.

(i) Fixed – point representation:

- Let us consider a fixed – point representation in which a number x is quantized from 'b_u' bits to 'b' bits. Thus the number x = 0.11101……..01, consisting of 'b_u' bits prior to quantization is represented as, x=0.11101…..1, containing 'b' bits, where $b < b_u$.

- The Quantizer truncates the value of x and the truncation error is defined as,

$$E_t = Q_t(x) - x$$

	Sign magnitude	2's Complement
Positive fixed point numbers	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$
Negative fixed point numbers	$0 \leq E_t \leq (2^{-b} - 2^{-bu})$	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$
Finally	$-(2^{-b} - 2^{-bu}) \leq E_t \leq (2^{-b} - 2^{-bu})$	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$

Example:

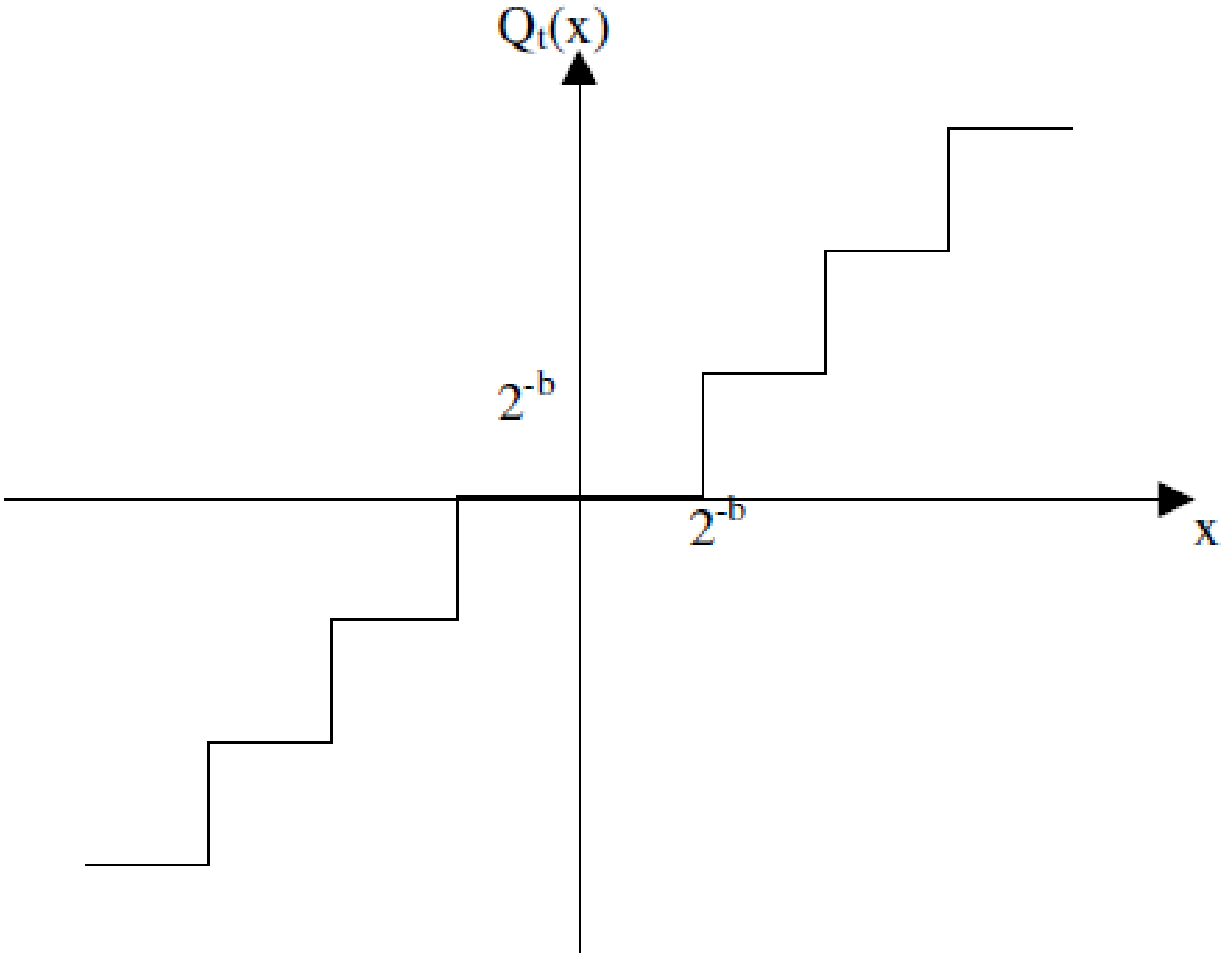
	Sign magnitude	2's Complement
Positive fixed point numbers	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$	$-(2^{-b} - 2^{-bu}) \leq E_t \leq 0$
0.8125	0.1101 After truncation, 0.110 = 0.75 Error = 0.75-0.8125 = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$	0.1101 After truncation, 0.110 = 0.75 Error = 0.75-0.8125 = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$
0.9375	0.1111 After truncation, 0.111 = 0.875 Error = 0.875-0.9375 = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$	0.1111 After truncation, 0.111 = 0.875 Error = 0.875-0.9375 = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$

Example:

	Sign magnitude	2's Complement
Negative fixed point numbers	$0 \le E_t \le (2^{-b} - 2^{-bu})$	$-(2^{-b} - 2^{-bu}) \le E_t \le 0$
-0.8125	1.1101 After truncation, 1.110 = -0.75 Error = (-0.75)-(-0.8125) = 0.0625 $(2^{-b} - 2^{-bu}) = (2^{-3} - 2^{-4}) = 0.0625$	1.0011 After truncation, 1.001 = (after 2's complement) = -0.875 Error = (-0.875)-(-0.8125) = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$
-0.6875	1.1011 After truncation, 1.101 = -0.625 Error = (-0.625)-(-0.6875) = 0.0625 $(2^{-b} - 2^{-bu}) = (2^{-3} - 2^{-4}) = 0.0625$	1.0101 After truncation, 1.010 = (after 2's complement) = -0.75 Error = (-0.75)-(-0.6875) = -0.0625 $-(2^{-b} - 2^{-bu}) = -(2^{-3} - 2^{-4}) = -0.0625$

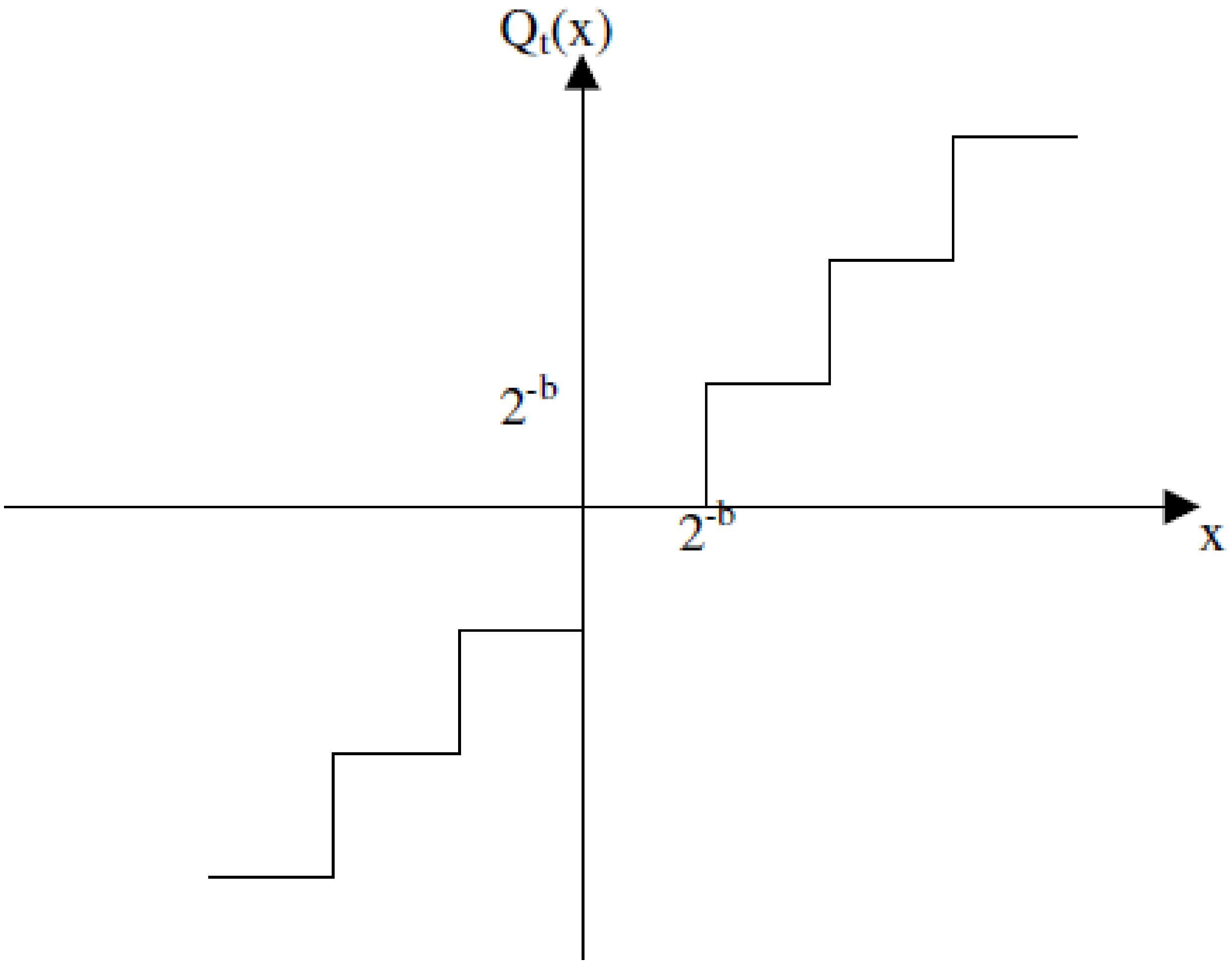
1. The truncation error for the sign – magnitude representation is symmetric about zero and falls in the range

$-(2^{-b} - 2^{-bu}) \leq E_t \leq (2^{-b} - 2^{-bu})$



2. The truncation error for two's complement representation is always negative and falls in the range

$-(2^{-b} - 2^{-b_u}) \le E_t \le 0$



- Next, let us consider the quantization errors due to rounding of a number.

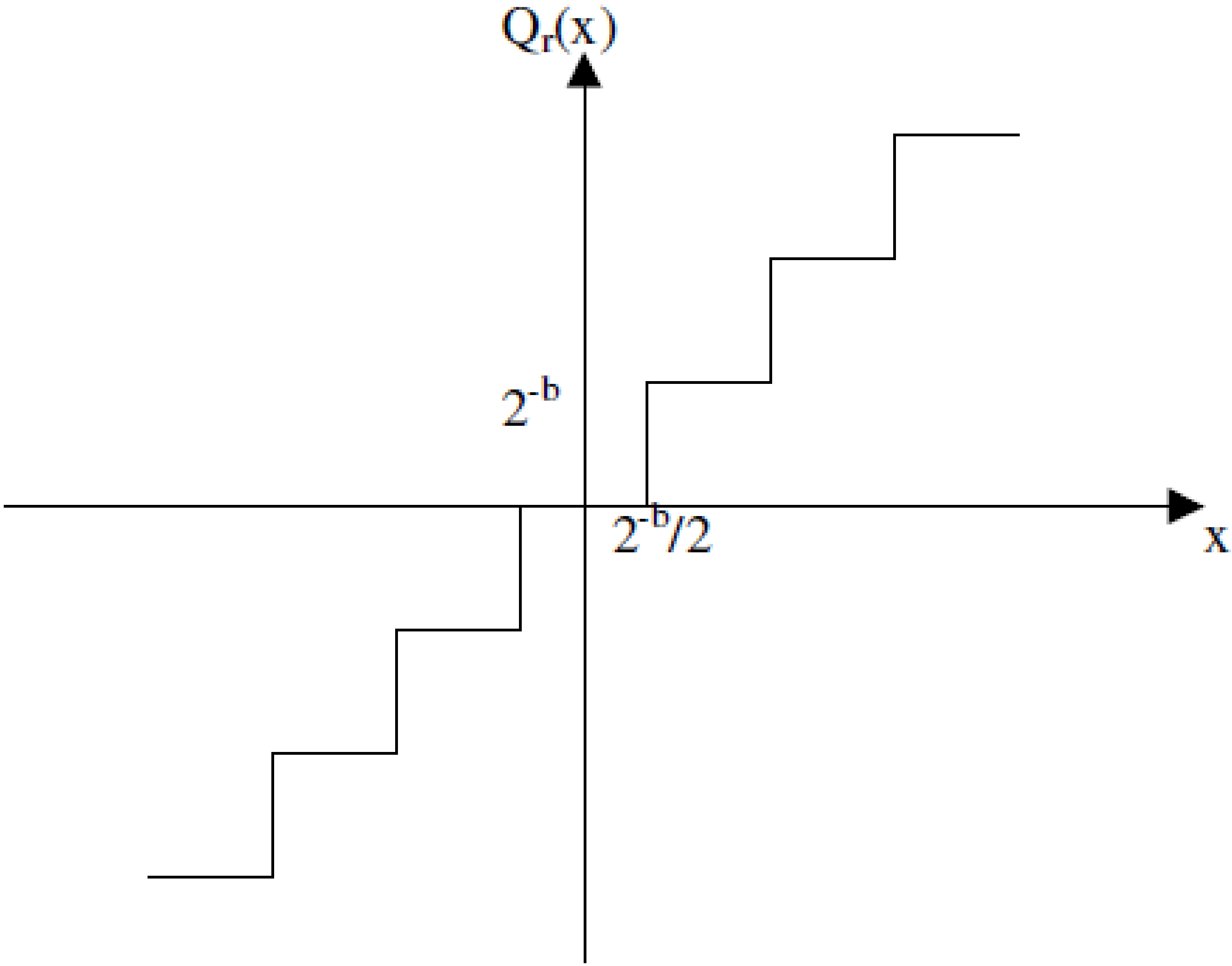
- A number x, represented by b_u bits before quantization and b bits after quantization, includes a quantization error
$E_r = Q_r(x) - x$

- Since the rounding involves only on the magnitude of the number and hence it is independent of the type of fixed – point representation.

- The maximum error that can be introduced through rounding is $2^{-b}/2$ and this can be either positive or negative, depending on the value of x. The round-off error is symmetric about zero and falls in the range $-(2^{-b})/2 \le E_r \le (2^{-b})/2$

Example:

	Sign magnitude	2's Complement
	$-(2^{-b})/2 \le E_r \le (2^{-b})/2$	$-(2^{-b})/2 \le E_r \le (2^{-b})/2$
0.8125	0.1101 After rounding, 0.111 = 0.875 Error = (0.875)-(0.8125) = 0.0625 $(2^{-b})/2 = 0.0625$	0.1101 After rounding, 0.111 = 0.875 Error = (0.875)-(0.8125) = 0.0625 $(2^{-b})/2 = 0.0625$
-0.6875	1.1011 After rounding, 1.110 = -0.75 Error = (-0.75)-(-0.6875) = -0.0625 $-(2^{-b})/2 = -0.0625$	1.0101 After rounding, 1.011 = (after 2's complement) = -0.625 Error = (-0.625)-(-0.6875) = 0.0625 $(2^{-b})/2 = 0.0625$



$Q_r(x)$
2^{-b}
$2^{-b}/2$
x

(ii) Floating Point Representation:

- Here, the mantissa is either rounded or truncated.
- Due to non-uniform resolution, the corresponding error in a floating – point representation is proportional to the number being quantized.
- The quantized value is represented as,

  $Q(x) = x + e.x$

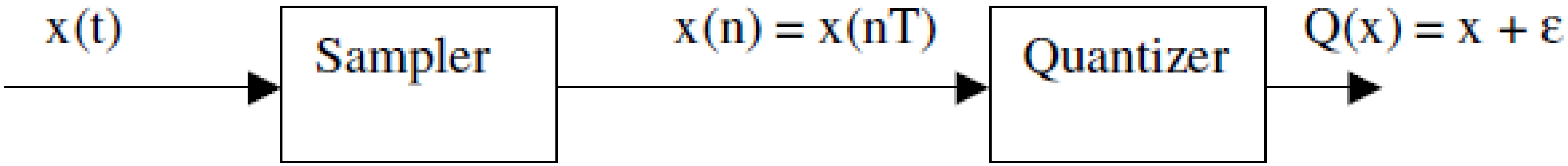
  where 'e' is called the relative error.
- Now, $e.x = Q(x) - x$

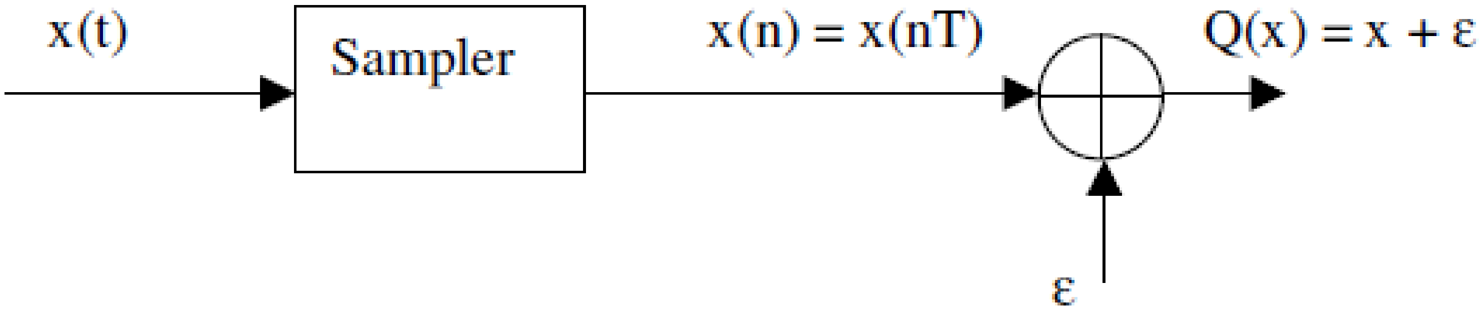
- In the case of truncation for positive numbers, based on 2's complement representation of the mantissa, we have
$-2^{E}\,2^{-b} < e_t\,x < 0$
- Since $2^{E-1} \le x < 2^{E}$, it follows that $-2^{-b+1} < e_t \le 0$
- For negative numbers, $0 \le e_t\,x < 2^{E}\,2^{-b}$
- Since $2^{E} \le x < 2^{E-1}$, it follows that $0 \le e_t < 2^{-b+1}$
- In the case of rounding the mantissa, the resulting error is symmetric relative to zero and has a maximum value of $\pm 2^{-b}/2$.
- Therefore, the round off error becomes $-2^{-b}/2 \le e_r \le 2^{-b}/2$

Note:

- It is convenient to adopt a statistical approach to the characterization of quantization errors.
- The quantizer can be modeled as introducing an additive noise to the unquantized value x.
- Thus we can write, $Q(x) = x + \varepsilon$.

where $\varepsilon = E_r$ for rounding and $\varepsilon = E_t$ for truncation.

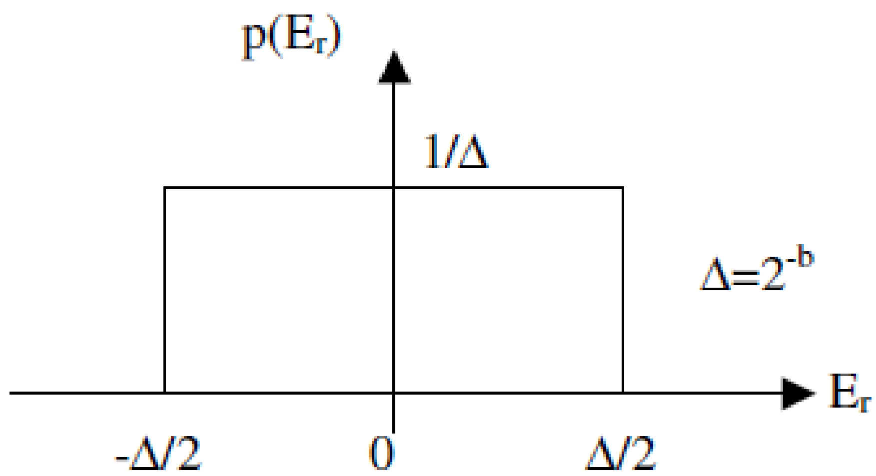




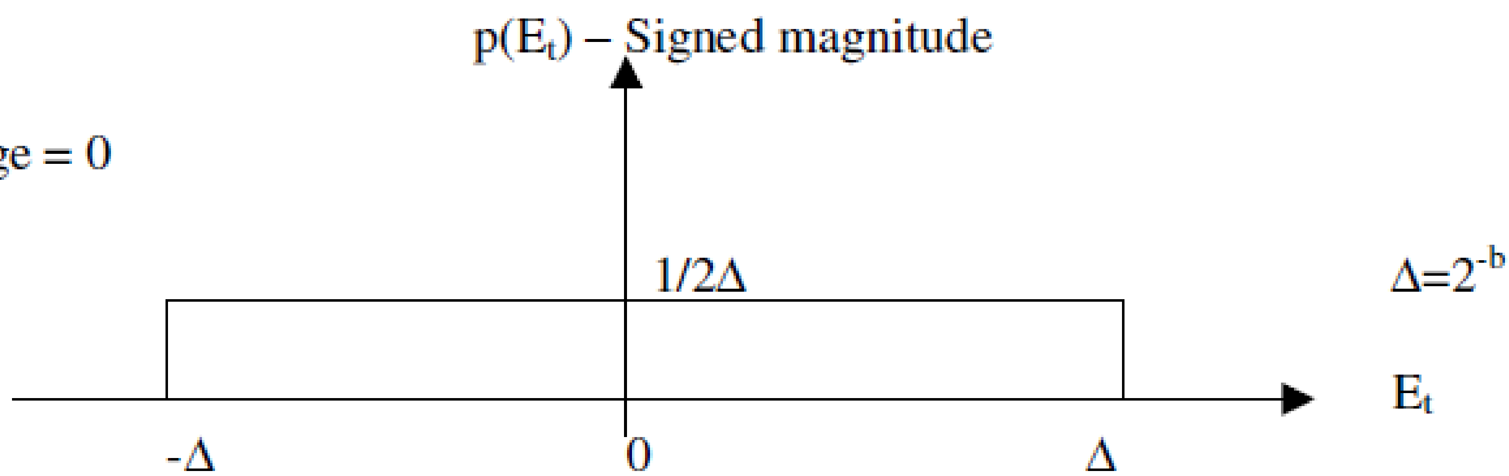
- The Quantization error can be modeled as Random variable, since falls within any of the levels of the quantizer.

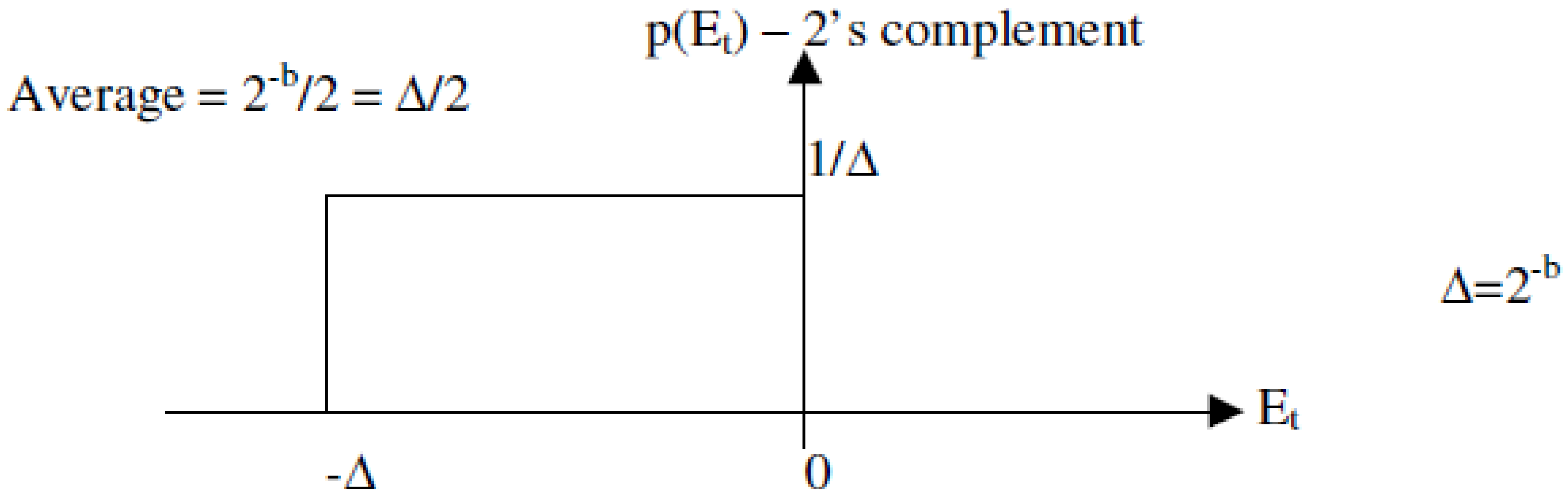
- This random variable is assumed to be uniformly distributed over the ranges specified for the fixed-point representation.

Average = 0



Average = 0





p(E_t) – 2's complement
Average = 2^-b/2 = Δ/2
1/Δ
Δ=2^-b
E_t
-Δ
0