LECTURE 3

ON DIGITIZING A FUNCTION/SIGNAL

(or how to represent a signal—a continuous function over an interval
using bits).

3.1 Given a function \( f(t) \) over an
interval say \([0, 1]\), we assume that
\( f(t) \) takes values between \( F_\downarrow \) and \( F_\uparrow \);
Hence we write:

\[
    f(t) : [0, 1) \rightarrow [F_\downarrow, F_\uparrow) \subset \mathbb{R}
\]

The function \( f(t) \) is a "signal", i.e.,
it describes the variations in time,
or w.r.t. any other continuous, ordered
variable that increases from 0 to 1,
(denoted by \( t \)), of some quantity of
interest.
The signal can be

- MUSIC or SPEECH in which case \( p(t) \) denotes temporal variations in air pressure
- ALTITUDE of an airplane during flight
- DISTANCE to THE OBSTACLES around a robot which senses the environment with a rotating scanner (as a function of angle - \( 2\pi t \))
- The TEMPERATURE versus height on a mountain etc. etc. etc...

Clearly the object \( p(t) \) is continuous and cannot be fully/precisely "described" with any finite set of elements, yet this is precisely what we want to do: We want to "represent", "describe", "approximate" the function \( p(t) \) with a finite number of bits \( B \).
In order to represent \( G(t) \) with bits we shall have to overcome the fact that both the domain over which it take values, i.e \([0,1]\) and the range of (real) values that \( G(t) \) may assume, i.e \([0,\infty)\), are intervals on the real line \( \mathbb{R} \), hence have uncountably large sets of values. Therefore both of them will have to be represented with finite sets.

To do so we can proceed as follows: map the interval \([0,1]\) to, say, \( N \) values by dividing it into equal intervals of length \( \frac{1}{N} \), as follows

\[
\Delta_i \equiv \left[\left(i-1\right)\frac{1}{N}, i\frac{1}{N}\right] \rightarrow \left(i-1\right)\frac{1}{N} + \frac{1}{2N} = t_i \triangleq \hat{t}_i
\]

for \( i \in \{1, 2, \ldots, N\} \)

\( t_i \) takes \( N \) possible values.
Furthermore let us similarly divide the interval \([\epsilon_l, \epsilon_{l+1}]\) into \(K\) equal intervals as follows:

\[ \Lambda_l = \left[ \epsilon_l + (l-1) \frac{\epsilon_{l+1} - \epsilon_l}{K}, \epsilon_l + l \frac{\epsilon_{l+1} - \epsilon_l}{K} \right] \rightarrow \epsilon_l + (l-1) \frac{(\epsilon_{l+1} - \epsilon_l)}{K} + \frac{l}{2} \frac{(\epsilon_{l+1} - \epsilon_l)}{K} \]

\[ \leq \epsilon_l \leq l \]

for \(l \in \{1, 2, \ldots, K\}\)

\(\epsilon_l\) takes \(K\) possible values

Now we could propose to digitize the function \(\phi(t) : [0, T] \rightarrow [\epsilon_l, \epsilon_{l+1}]\) as follows:

Given \(\phi(t)\) represent it by \(\epsilon_l(t_i)\) (see figure)

\[
\begin{cases} 
\text{if } t \in \Delta_i \text{ and } \\
\text{if } \phi(t_i) \in \Lambda_l.
\end{cases}
\]

Clearly we then map \(\phi(t)\) to a vector

\[
[\epsilon_{l_1}(t_i), \epsilon_{l_2}(t_i), \ldots, \epsilon_{l_N}(t_i)].
\]

of \(N\) values in the finite set of \(K\) possible values for \(l\).
What we have done, in fact, is to "SAMPLE" the function $g$ at the points $\{t_i\}$ and to "QUANTIZE" the values $g(t_i)$ with a $K$-level uniform quantizer, $Q_k()$.

This is a straightforward way to "represent" $g(t)$ with a vector of numbers that can assume only $K$ possible values. There are, clearly, only $K^N$ such vectors and we need $N \log_2 K$ bits to specify each one of them.

The questions that remain: How good is such a representation? Are there other choices? Can we design "optimal" digitizations with respect to some reasonable quality measures?
We shall try to answer these questions in the sequel.

1. Let us analyse the quality of the proposed digitization.

We have replaced \( \Phi(t) \) by \( \Phi^u(t) \) so that

\[
\Phi^u(t) = Q[u(\Phi(t))] \quad \text{for } t \in \Delta_i
\]

Let us therefore compute the mean squared error in this approximation.

We have:

\[
\text{NSE} = \sqrt{\int_0^1 (\Phi(t) - \Phi^u(t))^2 \, dt} = \\
= \sum_{i=1}^{N} (\Phi(t) - Q[u(\Phi(t))]^2)
\]
\[
\sqrt{\text{HSE}} = \int_0^t \rho^2(t) \, dt - 2 \sum_{i=1}^N Q(\rho(t)) \int_{\Delta_i} \rho(t) \, dt + \sum_{i=1}^N [Q(\rho(t))]^2 / \Delta_i
\]

This is the expression for the error. We cannot go much farther, but we feel that if \( N \) is a big number and if \( K \) is also big, the error will be small, if the function \( \rho(t) \) is "nice", i.e. smooth enough. Indeed if \( N \) is large we can write for each \( \Delta_i \) that

\[
\rho(t) \approx \rho(t_i) + \rho'(t_i)(t-t_i) + O(\epsilon^2)
\]

Then

\[
\int_{\Delta_i} \rho(t) \, dt \approx \rho(t_i) \cdot \Delta_i + \rho'(t_i) \int_{\Delta_i} (t-t_i) \, dt
\]

and

\[
Q(\rho(t_i)) \int_{\Delta_i} \rho(t) \, dt = Q(\rho(t_i)) \rho(t_i) \cdot \Delta_i
\]
Then we get that:

\[ \gamma(p \to \hat{\phi}) = \int_0^1 p^2 \, dt - 2 \sum_{i=1}^N Q_i (p(t_i) \rho(t_i) \mid \Delta_i) + \sum_{i=1}^N Q_i (p(t_i))^2 \mid \Delta_i | = \]

(by adding and subtracting \( \sum_{i=1}^N p(t_i) \mid \Delta_i | )

\[ = \int_0^1 p^2 \, dt - \sum_{i=1}^N p(t_i) \mid \Delta_i | + \sum_{i=1}^N \left[ p(t_i) - Q_i(p(t_i)) \right]^2 \mid \Delta_i | = \]

\[ = \sum_{i=1}^N \int (p(t_i) + \Delta(t_i))^2 \, dt - \sum_{i=1}^N p(t_i)^2 \mid \Delta_i | + \sum_{i=1}^N \left[ p(t_i) - Q_i(p(t_i)) \right]^2 \mid \Delta_i | = \]

\[ = \sum_{i=1}^N p(t_i)^2 \mid \Delta_i | + 2 \sum_{i=1}^N p(t_i) \Delta(t_i) \int_{\Delta_i} - \sum_{i=1}^N \Delta(t_i)^2 \mid \Delta_i | + \sum_{i=1}^N \left[ p(t_i) - Q_i(p(t_i)) \right]^2 \mid \Delta_i | = \]
\[
= \sum_{i=1}^{N} (\varphi(t_i))^2 \left( t - t_i \right)^3 \left| \frac{t_i + \frac{1}{2} \frac{1}{N}}{t_i - \frac{1}{2} \frac{1}{N}} \right| + \\
\sum_{i=1}^{N} \left[ \varphi(t_i) - Q_n(\varphi(t_i)) \right]^2 \Delta_i
\]

Let us compute the two terms in the expression above. We have first:

\[
\sum_{i=1}^{N} (\varphi(t_i))^2 \frac{1}{3} \left[ \left( \frac{1}{2} \frac{1}{N} \right)^3 - \left( -\frac{1}{2} \frac{1}{N} \right)^3 \right] = \\
\sum_{i=1}^{N} (\varphi(t_i))^2 \frac{1}{3} \cdot \frac{1}{4} \left( \frac{1}{N} \right)^3 = \frac{1}{12} \sum (\varphi(t_i)) \frac{1}{N^3} = \\
= \frac{1}{12} \cdot \frac{1}{N^2} \sum_{i=1}^{N} (\varphi(t_i))^2 \Delta_i = \\
= \frac{1}{12} \cdot \frac{1}{N^2} \int_{0}^{1} (\varphi(t))^2 dt
\]
For the second term we have:

\[
\sum_{i=1}^{\infty} [p(c_i) - Q_n(c_i)]^2 |\Delta c_i| = \\
= \frac{1}{N} \sum_{i=1}^{N} [p(c_i) - Q_n(c_i)]^2 \in \mathbb{E}_w
\]

Now \([p(c_i) - Q_n(c_i)]\) is the error in replacing a value from the interval \([c_L, c_H]\) with a representation value given by the K-level uniform quantizer of the range \([c_L, c_H]\).

Suppose we have a random variable \(w\) distributed over \([c_L, c_H]\) according to a distribution \(p_c(v)\). The expected squared error in this case is

\[
E((w - Q_n(w))^2) = \sum_{c_L}^{c_H} (v - Q_n(v))^2 p_c(v) dv = \\
= \sum_{l=1}^{K} \int (v - \phi_l)^2 p_c(v) dv
\]
Now assuming that $K$ is large we can approximate $p_{q}(v)$ over $\Lambda_{l}$ by its midpoint value and write

\[ E(\mathcal{Z}_{w}^{\ell})^{2} \equiv \sum_{\ell=1}^{K} p_{q}(\epsilon_{\ell}) \int_{\Lambda_{l}} (v - \epsilon_{\ell})^{2} dv = \]

\[ = \sum_{\ell=1}^{K} p_{q}(\epsilon_{\ell}) \frac{1}{3} (v - \epsilon_{\ell}) \left|^{\epsilon_{L} + (l-1)\frac{\epsilon_{H} - \epsilon_{L}}{K}}_{\epsilon_{L} + l\frac{\epsilon_{H} - \epsilon_{L}}{K}} \right. \]

\[ = \sum_{\ell=1}^{K} p_{q}(\epsilon_{\ell}) \cdot \frac{1}{3} \left[ \frac{1}{2^3} \left( \frac{\epsilon_{H} - \epsilon_{L}}{K} \right)^{3} + \frac{1}{2^3} \left( \frac{\epsilon_{H} - \epsilon_{L}}{K} \right)^{3} \right] \]

(Since $\epsilon_{\ell} = \epsilon_{L} + (l-1)\frac{\epsilon_{H} - \epsilon_{L}}{K} + \frac{1}{2} \left( \frac{\epsilon_{H} - \epsilon_{L}}{K} \right)$

and the endpoints of $\Lambda_{l}$ are equidistant from $\epsilon_{\ell}$ at a distance of $\frac{1}{2} \left( \frac{\epsilon_{H} - \epsilon_{L}}{K} \right)$)

Hence we get

\[ E(\mathcal{Z}_{w}^{\ell})^{2} = \frac{1}{12} \frac{(\epsilon_{H} - \epsilon_{L})^{2}}{K^{2}} \sum_{\ell=1}^{K} p_{q}(\epsilon_{\ell}) \cdot |\Lambda_{l}| \]
Therefore we obtained here that
\[
E(\varepsilon_w^2) \approx \frac{1}{12} \left( \frac{E^+ - E^-}{K^2} \right)^2 \int p_e(v) dv
\]

\[
\approx \frac{1}{12} \left( \frac{E^+ - E^-}{K^2} \right)^2 \cdot 1
\]

Now looking at the sum
\[
\frac{1}{N} \sum_{i=1}^{N} [E(t_i) - Q_u(E(t_i))]^2
\]
we realize that this sum approximates well, for large \( N \) the expected mean squared error we computed above, hence we can write
\[
\frac{1}{N} \sum_{i=1}^{N} [E(t_i) - Q_u(E(t_i))]^2 \approx \frac{1}{12} \left( \frac{E^+ - E^-}{K^2} \right)^2
\]

Indeed we have that from the \( N \) samples there will be \( \frac{p(E)}{N} \Delta T \) that will have values in the interval mapped to the value of \( Q_u(E(t)) \).
3.2 We see therefore that:

\[
\sqrt{\text{MSE}_{\text{\{\hat{c}\}\rightarrow c}}} = \frac{1}{12} \left\{ \frac{1}{N^2} \int_0^1 (\dot{c}(t))^2 \, dt + \frac{1}{K^2} (\hat{c}_0 - c_0)^2 \right\}
\]

\[
= \frac{1}{12} \left\{ \frac{C_1}{N^2} + \frac{C_2}{K^2} \right\}
\]

with

- \( C_1 = \int_0^1 (\dot{c}(t))^2 \, dt \) measuring the energy of the derivative of \( c(t) \).

and

- \( C_2 = (\hat{c}_0 - c_0)^2 \) measuring the "energy" of the span of values of \( c(t) \) over \([0, 1]\).

Now, using this result, valid for large \( N \) and \( K \)'s, for the mean squared error in digitizing the signal \( c(t) \) by the simple "DOUBLE QUANTIZATION" process, we can address a problem of BIT ALLOCATION.
Assume that $B_{\text{tot}}$ is the number of bits that we can use to represent/approximate the signal $S(t)$. Then we can write that

$$B_{\text{tot}} = N \cdot \log_2 K$$

Suppose we write $K = 2^b$ then $B = N \cdot b$.

The problem is: how to allocate the bits between $N$ - the temporal/spatial resolution parameter and $b/K$ the level quantization or quantization depth parameter, in order to minimize the mean squared error of the digitized representation.

We have to solve:

$$\text{Minimize} \ \frac{1}{12} \left\{ \frac{C_1}{N^2} + \frac{C_2}{K^2} \right\}$$

subject to $N \log_2 K = B$

or

$$\text{Minimize} \ \frac{1}{12} \left\{ \frac{C_1}{N^2} + \frac{C_2}{2^{2b}} \right\}$$

subject to $N \cdot b = B$
To solve the Bit Allocation problem we may proceed as follows:

Set \( N = \frac{B}{b} \) and write

\[
\frac{1}{12} \left\{ C_1 \frac{b^2}{B^2} + C_2 \frac{1}{2} b \right\} =
\]

\[
= \frac{1}{12} \left\{ C_1 \frac{b^2}{B^2} + C_2 e^{-2b/\nu_2} \right\} = \Psi_{\text{MSE}}(b)
\]

\[
\frac{d\Psi_{\text{MSE}}}{db} = \frac{1}{12} C_1 2b - 2C_2 e^{-2b/\nu_2} = 0
\]

Yielding

\[
\frac{C_1}{B^2} b = C_2 \nu_2 e^{-2b/\nu_2}
\]

and this equation yields the optimal \( b^* \).

We can see that the bit allocation problem has a unique optimal point from the behaviour of the two terms comprising \( \Psi_{\text{MSE}}(b) \).
3.3. At this point we might wonder whether the problem of signal/waveform digitization is completely solved. However, several questions arise naturally.

- Is the digitization process proposed the best we can do?
- Could we approximate $Q(t)$ better via a different type of digitization?

As a first improvement we could suggest the following: after we divide the time interval (the range) into $N$ segments of equal length, yielding the intervals

$$
\Delta_i = \left[ \frac{(i-1)}{N}, \frac{i}{N} \right] \quad i = 1, 2, \ldots, N
$$

we are facing a problem of best describing the function $Q(t)$ over each of these intervals.
We know that the best description of $\phi(t)$ over an interval $\Delta i$ is not the value of $\phi(t)$ at the midpoint of the interval, but rather the average value of $\phi(t)$ over that interval. This, from the fact that the mean squared error is minimized by averages, as we recall, we have that:

$$\begin{align*}
\text{MSE}_{\text{Ai}}(\theta) &= \frac{1}{|\Delta i|} \int_{\Delta i} (\phi(t) - \theta)^2 dt \\
\text{is minimized by} \\
\hat{\phi}_i &= \frac{1}{|\Delta i|} \int_{\Delta i} \phi(t) dt \\
\text{with mean squared error given by} \\
\text{MSE}_{\text{Ai}}(\theta = \hat{\phi}_i) &= \frac{1}{|\Delta i|} \int_{\Delta i} \hat{\phi}_i^2 dt - \left(\frac{1}{|\Delta i|} \int_{\Delta i} \phi(t) dt \right)^2
\end{align*}$$
Therefore we can modify the above-discussed digitization process as follows:

1) Partition the range, \( t \in [0,1] \), into \( N \) intervals defined as before:
\[
\Delta_i = \left[ (i-1) \frac{1}{N}, i \frac{1}{N} \right) \text{ for } i = 1, 2, \ldots, N.
\]

2) For each interval \( \Delta_i \) determine the optimal approximation of \( \phi(t) \) i.e.
\[
\phi_i^* = \frac{1}{|\Delta_i|} \int_{\Delta_i} \phi(t) \, dt.
\] (Optimal in the MSE sense.)

3) Quantize the values \( \phi_i^* \), \( i = 1, 2, \ldots, N \) that are all in the interval \([\phi_L, \phi_H]\) i.e. \( \phi_i^* \in [\phi_L, \phi_H] \) and distributed according to some \( \phi^*(\cdot) \), with a \( K \)-level quantizer, optimized for the distribution \( \phi^*(\cdot) \), as opposed to the use of the uniform quantizer before!
Let us consider the quality of the above-described
digitization process. We have the following

\[ \Theta(t) : [0,1] \rightarrow [\Theta_L, \Theta_H] \text{ is mapped to} \]

a vector of values

\[ [\Theta_1^*, \Theta_2^*, \ldots, \Theta_N^*] \]

where

\[ \Theta_i^* = \frac{1}{\Delta_i} \int_{\Delta_i} \Theta(t) dt \]

The values \( \Theta_i^* \) are quantized with an
"optimal" quantizer adapted to the distribution
of the values \( \Theta^* \), i.e., \( \Theta_i^* (\tau) \).

Hence writing

\[ \text{MSE} \]

\[ \sum_{i=1}^{N} (\Theta(t) - Q(\Theta_i^*))^2 dt = \]

\[ = \int_0^1 (\Theta(t) - Q(\Theta_i^*))^2 dt = \]

\[ = \int_0^1 \Theta^2(t) dt - 2 \sum_{i=1}^{N} Q(\Theta_i^*) \cdot \int_{\Delta_i} \Theta(t) dt + \sum_{i=1}^{N} Q(\Theta_i^*) \cdot \Delta_i \]
Since we know that, in this case
\[ \mathcal{E}_i \cdot \Delta i = \int_{\Delta i} \mathcal{E}(t) \, dt \]
we can write that,

\[
\mathcal{V}^{\text{MSE}}(\hat{\mathcal{E}}_{\text{obs}}) = \int_0^1 (\mathcal{E}(t) - \sum_{i=1}^N (\mathcal{E}_i^*)^2) \, dt + \sum_{i=1}^N (\mathcal{Q}(\mathcal{E}_i^*))^2 \, \Delta i + \sum_{i=1}^N (\mathcal{E}_i^* - \mathcal{Q}(\mathcal{E}_i^*))^2 \, \Delta i. 
\]

Now the first term here can be written as

\[
\int_0^1 (\mathcal{E}(t) - \sum_{i=1}^N (\mathcal{E}_i^*)^2) \, dt =
\sum_{i=1}^N \left[ \int_{\Delta i} \mathcal{E}(t) \, dt - (\mathcal{E}_i^*)^2 \, \Delta i \right] 
\]
Considering that
\[ \psi_{\text{MSE}} (\phi \rightarrow \phi_i \text{ over } \Delta_i) = \]
\[ = \frac{1}{|\Delta_i|} \int_{\Delta_i} \phi_i'^2 (t) \, dt - (\phi_i^*)^2 \]

We see that
\[ \int_0^1 \phi_i'^2 (t) \, dt - \sum_{i=1}^N (\phi_i^*)^2 \Delta_i = \]
\[ = \sum_{i=1}^N \left| \Delta_i \right| \psi_{\text{MSE}} (\phi \rightarrow \phi_i^*) = \frac{1}{N} \sum_{i=1}^N \psi_{\text{MSE}} (\phi - \phi_i^*) \]
\[ = \sum_{i=1}^N \int_{\Delta_i} (\phi(t) - \phi_i^*)^2 \, dt = \]
\[ = \text{MEAN SQUARED ERROR IN} \]
\[ \text{REPLACING over each interval } \Delta_i \]
\[ \phi(t) \text{ by } \phi_i^* \]

Hence the first term is the error in 
the general/optional sampling process.
We can consider that the digitization process first replaces \( p(t) \) by a piecewise constant function

\[
\mathcal{P}_{\text{SAMPLED}}(t) = \frac{1}{\Delta_i} \int_{\Delta_i} p(t') dt' = \mathcal{P}^*_i
\]

for \( t \in \left[ i\Delta N, i\Delta N + \frac{1}{N} \right) \)

and the mean squared error in this process (using the optimal choices for \( \mathcal{P}^*_i \)) is

\[
\mathcal{P}_{\text{MSE}} = \int_0^1 (p(t) - \mathcal{P}_{\text{SAMPLED}}(t))^2 dt = \int_0^1 \mathcal{P}^2(t) - \frac{1}{N} \sum_{i=1}^N (\mathcal{P}^*_i)^2
\]

Now using the fact that \( N \) is large we can again evaluate the sampling MSE by considering that

\[
p(t) = p(\text{midpoint of } \Delta_i) + p(\text{midpoint of } \Delta_i) \cdot (t - \text{midpoint of } \Delta_i)
\]

\[+ O(\varepsilon^2) \quad \text{over } \Delta_i\]

and in this case we have that:

\[
\mathcal{P}^*_i = \frac{1}{\Delta_i} \int_{\Delta_i} (p(mp\Delta_i) + p(mp\Delta_i)(t - mp\Delta_i)) dt =
\]

\[= p(mp\Delta_i) \quad \text{(the midpoint of the interval } \Delta_i \text{ like before!)}
\]
Furthermore (like before) we have that:

\[
\int_{\Delta_i} \varphi^2(t) dt = \int_{\Delta_i} (\varphi(mp\Delta_i) + \varphi(mp\Delta_i)(t-mp\Delta_i))^2 dt = \\
= \varphi^2(mp\Delta_i) |\Delta_i| + O \left( \varphi(mp\Delta_i)^2 \frac{1}{3} (t-mp\Delta_i)^3 \right)_{t=mp\Delta_i} \\
= \varphi^2(mp\Delta_i) |\Delta_i| + \varphi(mp\Delta_i)^2 \frac{1}{12} |\Delta_i|^3
\]

Hence

\[
\chi_{MSE} = \sum_{i=1}^{N} \left( \varphi(mp\Delta_i) |\Delta_i| + \varphi(mp\Delta_i)^2 \frac{1}{12} |\Delta_i|^3 \right) \\
= \frac{1}{12} \frac{1}{N^2} \sum_{i=1}^{N} \varphi(mp\Delta_i)^2 |\Delta_i| = \\
= \frac{1}{12} \frac{1}{N^2} \int \varphi(t)^2 dt
\]

(as we have seen before - not surprising given that with large \(N\) the linearity makes midpoints optimal sampling points!)
The second term in $\sqrt{\text{NSEE} (\rho - \rho^*)}$ is

$$
\sum_{i=1}^{N} (\rho_i^* - Q(\rho_i^*))^2 |\Delta i| =
$$

$$
= \frac{1}{N} \sum_{i=1}^{N} (\rho_i^* - Q(\rho_i^*))^2
$$

This expression is seen to be the average of the quantization errors of the values $\rho_i^*$ (which are approximately the values of $\rho(t)$ at the midpoints of the intervals $\Delta i$). If we had a distribution density $p_E(r)$ describing the values that the function $\rho(t)$ has in the range $[\rho_L, \rho_H]$ we
could design the optimal quantizer for $\Phi \in \Phi_L \cup \Phi_{L+1}$ distributed according to $p_0(v)$. The expected mean square error of quantization would then be, for large $K$

$$E(\hat{\Phi} - \Phi^0)^2 = \frac{1}{12} \frac{1}{K^2} \left[ \int_{\Phi_L} \frac{1}{2^L} \int \left. \frac{1}{2^L} \int \right] dr \right]^3.$$

Since $N$ is large we expect the average of the quantization errors for $\Phi_i$'s to be well approximated by the expectation above, i.e.

$$\frac{1}{N} \sum_{i=1}^{N} (\Phi_i^* - \Phi_i^{optimal})^2 = \frac{1}{12} \frac{1}{K^2} \left[ \int_{\Phi_L} \frac{1}{2^L} \int \left. \frac{1}{2^L} \int \right] dr \right]^3.$$

This yields

$$\text{MSE} (\hat{\Phi} - \Phi^0) = \int_0^1 (\hat{\Phi}(t) - \sum_{i=1}^{N} \Phi_i^*)^2 + \frac{1}{N} \sum_{i=1}^{N} (\Phi_i^* - \Phi_i^{optimal})^2 =$$

$$= \frac{1}{12} \frac{1}{N^2} \int_0^1 (\Phi^0(t))^2 dt + \frac{1}{12} \frac{1}{K^2} \left[ \int_{\Phi_L} \frac{1}{2^L} \int \left. \frac{1}{2^L} \int \right] dr \right]^3.$$
Note here the striking (but in hindsight quite expected) similarity of the end result to the one we obtained with the straightforward point-sampling and uniform quantization approach. Indeed in the large $N$ and $K$ case the optimal sampling is very well approximated by mid-point sampling and if we use a uniform quantizer rather than an optimized one (for $P_{0}(\nu)$ which is estimated from $P(\tau)$) we lose the difference in the coefficient of $1/K^2$ (i.e. the error will be bigger by

$$\left[\epsilon_0^H - \epsilon_0^L\right]^2 - \left[\int_{\epsilon_0^L}^{\epsilon_0^H} P_{\nu}(\nu) d\nu\right]^2 angle 0$$

and if $P_{\nu}(\nu)$ is uniform over $[\epsilon_0^L, \epsilon_0^H]$ the difference vanishes.

For bit allocation, here too we have to find the solution of:

$$\min \frac{1}{12} \left\{ \frac{1}{N^2} \int (P_{0}(\tau))^2 d\tau + \frac{1}{K^2} \left[ \int_{\epsilon_0^L}^{\epsilon_0^H} P(\nu) d\nu \right]^2 \right\}$$

s.t. $N \log K = B_{\text{TOTAL}}$
So far we have seen two processes of digitization that map a function \( \mathcal{C}(t) \) in \( [0, 1] \) into a finite set of "representations", which enable encoding the function with a given number of bits \( B = N \log_2 K \), where the parameters \( N \) and \( K \) describe the sampling density or resolution, and the number of levels approximating the samples in a quantization process. We discussed ways to create the representations which were also functions over \([0, 1]\) of the form:

\[
\hat{C}(t) = Q(\mathcal{C}_i^*) \quad \text{for } t \in \Delta_i
\]

i.e. piecewise constant functions. The quality of representing \( \mathcal{C}(t) \) by a suitable \( \hat{C}(t) \) which depends on \( N \) parameters taking
K possible values (hence allowing for a total of $K^N$ different $\mathcal{E}(t)$'s) was measured via the mean squared error, i.e. by
\[ \Psi(\mathcal{E} \circ \mathcal{E}) = \frac{1}{1} \int_0^1 (\mathcal{E}(t) - \mathcal{E}(t))^2 dt. \]
Let us write the function $\mathcal{E}(t)$ as follows:
\[ \mathcal{E}(t) = \sum_{i=1}^N \mathcal{G}(\mathcal{E}_i) \cdot 1_{\Delta_i}(t). \]
Where we define $1_{\Delta}(t)$ as a function which is equal to 1 over $\Delta$ and zero elsewhere
\[ 1_{\Delta}(t) = \begin{cases} 1 & \text{if } t \in \Delta \\ 0 & \text{if } t \notin \Delta \end{cases} \]
We clearly have
\[ \int_{-\infty}^{+\infty} 1_{\Delta}(t) dt = |\Delta| = \text{size(measure) of the interval } \Delta. \]
Consider the "vector space" of all "nice" functions over \([0, 1]\) \(\{\varphi(t)\}\) and \(\varphi: [0, 1] \rightarrow \mathbb{R}\). Consider the functions of the form

\[
\varphi^R(t) = \sum_{i=1}^{n} \delta_i \Delta_i(t) \quad \delta_i \in \mathbb{R}
\]

We have over these spaces a natural inner product

\[
\langle \varphi_1(t), \varphi_2(t) \rangle = \int_{0}^{1} \varphi_1(t) \varphi_2(t) \, dt
\]

inducing a norm

\[
\| \varphi(t) \| = \langle \varphi(t), \varphi(t) \rangle = \left( \int_{0}^{1} \varphi(t)^2 \, dt \right)^{1/2}
\]

and a "distance between" \(\varphi_1\) and \(\varphi_2\)

\[
\| \varphi_1 - \varphi_2 \| = \left( \int_{0}^{1} (\varphi_1(t) - \varphi_2(t))^2 \, dt \right)^{1/2}
\]

Clearly with these concepts we can do a bit of geometry on the space of "nice" functions over \([0, 1]\).
If the space of nice functions is considered to include the functions (with jumps) $P_{\mathcal{C}}(t)$, we can consider the problem: given a function $P(t)$, what is the function $P_{\mathcal{C}}(t)$ which is closest to $P(t)$? In other words, we want to "project" $P(t)$ into the subspace of piecewise constant functions $\{P_{\mathcal{C}}(t)\}$.

Let us proceed to determine $P_{\mathcal{C}}(t)$ that is "nearest" to $P(t)$.

The distance between $P(t)$ and $P_{\mathcal{C}}(t)$ is

\[ \|P(t) - P_{\mathcal{C}}(t)\|_2^2 = \int_0^1 (P(t) - P_{\mathcal{C}}(t))^2 \, dt = \int_0^1 \left( \sum_{i=1}^N \delta_{i \Delta_i}(t) \right)^2 \, dt = \int_0^1 \left( P(t) - \sum_{i=1}^N \delta_{i \Delta_i}(t) \right)^2 \, dt = \int_0^1 (P(t)^2 - 2P(t)\sum_{i=1}^N \delta_{i \Delta_i}(t) + \sum_{i=1}^N \delta_{i \Delta_i}(t)^2) \, dt \]
\[ \| \phi(t) - \phi^p(t) \|^2 = \int_0^1 \phi^2(t) dt - 2 \sum_{i=1}^N \delta_i \int_0^1 \phi(t) \phi_{\Delta_i}(t) dt + \sum_{i=1}^N \delta_i^2 \int_{\Delta_i}^1 \phi(t) dt + \sum_{i=1}^N \sum_{j=1}^N \delta_i \delta_j \int_{\Delta_i}^1 \int_{\Delta_j}^1 \phi(t) \phi(t') dt' dt = 0 \text{ since } \Delta_i \cap \Delta_j = \emptyset \]

\[ = \int_0^1 \phi^2(t) dt - 2 \sum_{i=1}^N \delta_i \int_{\Delta_i}^1 \phi(t) dt + \sum_{i=1}^N \delta_i^2 |\Delta_i| \]

We found a formula for the distance as a function of the \( N \) parameters \( \delta_1, \delta_2, \ldots, \delta_N \).

We want to determine them to minimize the (squared) distance. Hence we do

\[ \frac{\partial \| \phi(t) - \phi^p(t) \|^2}{\partial \delta_i} = -2 \int_{\Delta_i} \phi(t) dt + 2 \delta_i |\Delta_i| = 0 \]

yielding

\[ \delta_{i, \text{optimal}} = \frac{1}{|\Delta_i|} \int_{\Delta_i} \phi(t) dt \]

Hardly a surprising result, since we knew that we expect \( \delta_{i, \text{optimal}} \) to be \( \phi^*_i \)!
We can say that the projection of \( p(t) \) on the subspace of piecewise constant functions is given by

\[
\hat{p}(t) = \sum_{i=1}^{N} \hat{p}_i 1_{\Delta_i}(t)
\]

and the error

\[
(p(t) - \hat{p}(t)) = \sum_{i=1}^{N} (p(t) - \hat{p}_i) 1_{\Delta_i}(t)
\]

should be orthogonal to any piecewise constant function. Let us see:

\[
\langle (p(t) - \hat{p}(t)), \hat{p}(t) \rangle =
\]

\[
= \langle p(t), \hat{p}(t) \rangle - \langle \hat{p}(t), \hat{p}(t) \rangle =
\]

\[
= \int_{0}^{1} p(t) \sum_{i=1}^{N} \delta_i 1_{\Delta_i}(t) - \int_{0}^{1} \sum_{i=1}^{N} \left( \frac{1}{\Delta_i} \int_{\Delta_i} p(x) dx \right) \delta_i 1_{\Delta_i}(t) =
\]

\[
= \sum_{i=1}^{N} \delta_i \int_{\Delta_i} p(t) dt - \sum_{i=1}^{N} \delta_i \int_{\Delta_i} p(x) dx \frac{1}{\Delta_i} \int_{\Delta_i} (t) =
\]

\[
= 0.
\]
However, the projection onto the piecewise constant (over the $\Delta_i$-intervals) functions was just one step of the digitization process (which may be called the discretization process mapping $\mathcal{P}(t)$ to $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n \in \mathbb{F}$).

One more step is necessary: that of "projecting" $\mathcal{P}_n(t)$ further into a set of functions where the possible range of $\delta_i$'s (or $\mathcal{P}_n$'s) is mapped to a finite set of $K$ representation levels—the step we called quantization. Let us see it here:

We have the subset of $\mathcal{P}_n(t)$ of functions of the form $\sum_{i=1}^{n} \delta_i \chi_{\Delta_i}(t)$ where $\delta_i \in \{\theta_0, \theta_2, \ldots, \theta_k\}$, a finite set of numbers indexed by $1, 2, \ldots, K$.

Now the $(\text{distance})^2$ between an arbitrary $\mathcal{P}_n(t)$ and a function of the desired form is
\[ \| \phi^P(t) - \sum_{i=1}^{N} \delta_i^q l_{\Delta_i}(t) \|^2 = \]
\[ = \int_0^1 \left( \sum_{i=1}^{N} \delta_i^q l_{\Delta_i}(t) - \sum_{i=1}^{N} \delta_i^q l_{\Delta_i}(t) \right)^2 dt = \]
\[ = \int_0^1 \left( \sum_{i=1}^{N} (\delta_i^q - \delta_i^q) l_{\Delta_i}(t) \right)^2 dt = \]
\[ = \sum_{i=1}^{N} (\delta_i^q - \delta_i^q)^2 \int_0^1 l_{\Delta_i}(t) dt + \sum_{i \neq j} \int_0^1 l_{\Delta_i} l_{\Delta_j} dt \]
\[ = \sum_{i=1}^{N} (\delta_i^q - \delta_i^q)^2 |\Delta_i| \]

Hence the optimal assignment of \( \delta_i^q \)'s

\( \delta_i^q \) is the value from \{\theta_1, \theta_2, \ldots, \theta_k\} that is nearest to \( \delta_i \):

\[ \begin{array}{cccc}
\theta_1 & \theta_2 & \theta_3 & \theta_4 & \ldots & \theta_{k-1} & \theta_k \\
\end{array} \]

The decision regions for mapping \( \delta_i \) to \( \theta_i \)'s

This we know from quantization or in fact we determined the best \( \theta_i \)'s for minimizing
the expected (mean) squared error.

Then we had that in mapping
\[ \hat{\Phi}(t) \to \Phi_{PC}(t) \to \hat{\Phi}(t) \]
we have an orthogonal projection of \( \Phi(t) \)
into the subset (subspace) of PC functions.
Then we "projected" \( \Phi_{PC}(t) \) into a subset
of \( K^N \) possible functions by moving inside
the PC set to the closest of the \( K^N \)
"quantized" PC functions.

The total squared distance is, we saw
\[
\| \Phi(t) - \hat{\Phi}(t) \|^2 = \| \Phi(t) - \Phi_{PC}(t) \|^2 + \|
\Phi_{PC}(t) - \hat{\Phi}(t) \|^2
\]

SAMPLING MSE

The picture

explaining this is the following
\[
\| \Phi(t) - \Phi_{PC}(t) \|^2 = \int_0^t \hat{\Phi}(t') dt - \sum_{i=1}^N \frac{1}{A_i} \left( \sum_{n=1}^N (\hat{q}(t'_n) - q(t'_n))^2 \right)
\]

\[
\| \Phi_{PC}(t) - \hat{\Phi}(t) \|^2 = \sum_{i=1}^N (\hat{q}(t_i) - q(t_i))^2 A_i
\]
The $K^n_{PC}$ functions corresponding to the different choices of coefficients selected from the set of quantization levels.

Pythagoras says:

$$\mathcal{E}^2_{TOTAL} = \mathcal{E}^2_{SAMPLING} + \mathcal{E}^2_{QUANTIZATION}$$

If fact true for any selection of quantized PC functions.