Lecture 1

On Averages and Expectations

(or: How to find a "representative" number for a group of objects - numbers, vectors, functions etc - and measure how good it is).

1.1 Given a finite list of numbers, we write them in an ordered sequence

\[ x_1, x_2, \ldots, x_N \]

so that \((x_{i+1} \geq x_i) \quad x_i \in \mathbb{R}\)

What is a number that best "represents" the numbers in our list? Clearly we have to understand what do we mean by representing the given numbers, and how do we measure the quality of the representation.
The dictionary defines "to represent" as follows: to stand for, to typify, to act as a substitute for, to depict/describe. Hence when we select a number say \( \Theta \) "to stand for" the numbers on the list, we mean that we shall replace all the \( x_i \)'s by \( \Theta \). Clearly this is not done without inducing some errors of representation. These errors are

\[
x_1-\Theta, \ x_2-\Theta, \ \ldots, \ x_i-\Theta, \ \ldots, \ x_N-\Theta
\]

and we would like to make all these errors \( \varepsilon_i \equiv x_i-\Theta \) as small as possible.

To define mathematically what we want, we have to define a function of the size of these errors that somehow measures the quality of \( \Theta \) by combining the values \( |\varepsilon_i| \), for all \( i = 1, 2, \ldots, N \). We shall call such a function a "Quality Measure" function.
These multivariate "quality measure" functions will be denoted by $\Psi(\cdot, \cdot, \cdot)$, as follows:

$\Psi(1 \xi_1, 1 \xi_2, \ldots, 1 \xi_N) = \text{measures the quality/size of the set of positive numbers } 1 \xi_1, 1 \xi_2, \ldots, 1 \xi_N$.

We require to have for $\Psi(\cdot, \cdot, \cdot)$ the following properties:

$\Psi(0, 0, 0, \ldots, 0) = 0$ (the best possible quality = the lowest value)

$\frac{\partial \Psi}{\partial \xi_i}(\xi_1, \xi_2, \ldots, \xi_N) > 0$ for all $[\xi_1, \xi_2, \ldots, \xi_N] \in \mathbb{R}^N$

IE, for all positive points in the multivariate range of $\Psi$ at all points in $\mathbb{R}^N$

$\Psi(\xi_1, \xi_2, \ldots, \xi_N) > 0$

and if all errors are "equally important", the function $\Psi(\xi_1, \xi_2, \ldots, \xi_N)$ should also be permutation invariant, i.e. its value should not change if we permute the values assigned to $\xi_1, \xi_2, \ldots$ and $\xi_N$. 
A function \( \Psi(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) \) which obviously satisfies the required properties is:

\[
\Psi(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) = \sum_{i=1}^{N} |\varepsilon_i|^2 = \sum_{i=1}^{N} (\alpha_i - \theta)^2
\]

This function, for a given set of numbers \( x_1, x_2, \ldots, x_N \), becomes a univariate function of the variable \( \theta \), and we shall call

\[
\Psi_{\text{MSE}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \theta)^2
\]

the "mean squared error" loss function or quality measure. Since clearly \( \Psi \)-here also measures the squared-length of the vector \( (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N) \), we can also call \( \Psi \) proposed the \( L_2 \)-norm loss function.

This loss function is particularly nice since we have the possibility to find the unique minimizing \( \theta \) for the given
set of values \( x_1, x_2, \ldots, x_N \), as follows:

\[
\frac{d}{d \theta} \Psi_{\text{MSE}}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} 2(x_i - \theta) = 0
\]

and this yields

\[
\theta_{\text{min}} = \arg \min_{\theta} \Psi_{\text{MSE}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

Hence the optimal representative value for \( x_1, x_2, \ldots, x_N \), when the quality is measured by the \( l_2 \)-norm of the error vector, or by the mean squared errors induced is the average value of the given numbers, or the DATA. What is the mean squared error induced by the optimal choice?

A very simple calculation gives:

\[
\Psi_{\text{MSE}}(\theta_{\text{min}}) = \frac{1}{N} \sum_{i=1}^{N} (x_i^2 - 2x_i \theta_{\text{min}} + \theta_{\text{min}}^2) = \\
= \frac{1}{N} \sum_{i=1}^{N} x_i^2 - 2 \theta_{\text{min}} \frac{1}{N} \sum_{i=1}^{N} x_i + \theta_{\text{min}}^2\theta_{\text{min}}
\]
\[
\frac{1}{N} \sum_{i=1}^{N} x_i^2 - \Theta_{\text{min}}^2 = \frac{1}{N} \sum_{i=1}^{N} x_i^2 - \left( \frac{1}{N} \sum_{i=1}^{N} x_i \right)^2 \geq 0
\]

In some sense, we see that this choice of quality/or loss measure was ideal. We obtained from it that the average of a set of numbers is their best representative, an intuitively pleasing result. What if we chose a different \( \Psi \)?

1.3 Another natural choice would be

\[
\Psi(1|\varepsilon_1|, |\varepsilon_2|, .. |\varepsilon_N|) = \sum_{i=1}^{N} |\varepsilon_i| = \sum_{i=1}^{N} |x_i - \Theta|
\]

This function too, for a given set of \( x_i \)'s, becomes a univariate function of the variable \( \Theta \), and we shall call...
\[ \text{MAE} (\theta) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \theta| \]

the "mean absolute deviation", or the MAD-loss function. This quality measure is clearly defined in terms of the so-called \(l_1\)-norm of the vector \((\xi_1, \xi_2, \ldots, \xi_N)\), the sum of absolute values of its components.

This loss function is not so straightforward to optimize since the function \(|x|\) is not so nice to differentiate, especially at \(x=0\).

We can write, formally, that (at \(\theta \neq x_i\))

\[ \frac{d}{d\theta} \text{MAE} (\theta) = -\frac{1}{N} \sum_{i=1}^{N} \text{sign} (x_i - \theta) \]

and seek the values of \(\theta\) making this sum to zero. In fact we shall have a result of zero when \(\theta\) will be such that the signs of the expressions \(|x_i - \theta|\) in the sum will cancel out, i.e. we'll have the same number of \(x_i\)'s to the right
of θ as the number of x_i's to the left. This is achieved when θ is a median of the set x_1, x_2, ..., x_N (i.e., we order the set as we have done

x_1 \leq x_2 \leq \cdots \leq x_{[N/2]-1} \leq x_{[N/2]} \leq \cdots \leq x_N

and select θ to be a value in the middle of the list.

- if \( N = 2k \) then \( \theta \in [x_k, x_{k+1}] \)
- if \( N = 2k+1 \) then \( \theta = x_{k+1} \)

The above calculation can be made rigorous as follows:

Note that we can write:

\[
\text{MAD}(\theta) = \frac{1}{N} \sum_{i=1}^{N} |x_i - \theta| = \frac{1}{N} \sum_{i=1}^{N} (x_i - \theta) \text{sign}(x_i - \theta)
\]
If $\theta \in [x_k, x_{k+1}]$ then:

$$V_{\text{MAD}}(\theta) = \frac{1}{N} \left( \sum_{i=1}^{k}(\theta - x_i) + \sum_{i=k+1}^{N}(x_i - \theta) \right) =$$

$$= \frac{1}{N} \left( k\theta - (N-k)\theta + \sum_{i=k+1}^{N} x_i - \sum_{i=1}^{k} x_i \right) =$$

$$= \frac{2k-N}{N} \theta + \frac{1}{N} \left( S(N) - 2S(k) \right) =$$

$$= \left( \frac{2k}{N} - 1 \right) \theta + \frac{1}{N} \left( S(N) - 2S(k) \right)$$

where we define:

$$S(k) \triangleq \sum_{i=1}^{k} x_i \quad \text{and hence} \quad S(N) \triangleq \sum_{i=1}^{N} x_i$$

Now we see that the function $V_{\text{MAD}}(\theta)$ is piecewise linear and decreases when

$$\frac{2k}{N} < 1 \quad \text{or} \quad k < \frac{N}{2}$$

is constant if

$$\frac{2k}{N} = 1 \quad \text{or} \quad k = \frac{N}{2}$$

and increases if

$$\frac{2k}{N} > 1 \quad \text{or} \quad k > \frac{N}{2}.$$
Note that we obtained the result that the minimum of $\Psi_{\text{MAD}}(\theta)$ is attained at the value $x_{k+1}$ if $N=2k+1$ and we have $k=[N/2]$ and $k+1=[N/2]+1$, or in the entire interval $[x_k, x_{k+1})$ if $k = \frac{N}{2}$, i.e. $N=2k$.

Hence we define the median of the set $x_1, x_2, \ldots, x_N$ as the value of $x_{\text{med}}$ interpreted as:

$$x_{\text{med}} = \begin{cases} \frac{1}{2}(x_{[N/2]} + x_{[N/2]+1}) & \text{if } N \text{ is an integer} \\ x_{[N/2]+1} & \text{if } N \text{ is not an integer} \end{cases}$$

What is the Mean Absolute Deviation for $\theta = x_{\text{med}}$?

We have

$$\Psi_{\text{MAD}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} |x_i - x_{\text{med}}|$$
We have

\[ \psi_{\text{MAD}}(\theta = x_{\text{median}}) = \]

\[ = \frac{2k-N}{N} x_{\text{median}} + \frac{1}{N} S'(N) - \frac{2}{N} S'(k) \]

hence if \( N=2k \) we get

\[ \psi_{\text{MAD}}(\theta = x_{\text{median}}) = \frac{1}{N} S'(N) - \frac{2}{N} S'(\lfloor \frac{N}{2} \rfloor) \]

if \( N=2k+1 \) we get

\[ \psi_{\text{MAD}}(\theta = x_{k+1}) = \frac{1}{N} S'(N) - \frac{2}{N} S'(\lfloor \frac{N}{2} \rfloor) - \frac{1}{N} x_{k+1} \]

\[ = \frac{1}{N} S'(N) - \frac{2}{N} \left( S'(\lfloor \frac{N}{2} \rfloor) + \frac{1}{2} x_{k+1} \right) \]

Note here that

\[ \frac{1}{N} S'(N) \] is the average value of \( x_1, x_2, \ldots, x_N \)

and that

\[ \frac{1}{\left( \frac{N}{2} \right)} S'(\lfloor \frac{N}{2} \rfloor) \] is the average value of \( x_1, x_2, \ldots, x_{\lfloor \frac{N}{2} \rfloor} \)

if \( N=2k \)

and

\[ \frac{1}{2k+1} \left( S'(k) + \frac{1}{2} x_{k+1} \right) \] is the average of the numbers

with weights \( \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2} \right) \)
So for we analyzed the "best" way to represent the list of numbers $x_1, x_2, \ldots, x_n$ by a single representation value, and we recovered the mean and the median as being optimum in different ways. 

Note that both representation values are invariant to permutations of the numbers in our list (or DATA): clearly

$$(x_1, x_2, \ldots, x_n)$$ and $$(x_{\pi_1}, x_{\pi_2}, \ldots, x_{\pi_N})$$ will have the same mean and median for any permutation $\{\pi_i\}$ of the numbers $1, 2, \ldots, N$.

We can generalize the process described above with weighted lists of numbers. Suppose that we are given a list of pairs of numbers $(x_i, w_i)$ where
$x_i \in \mathbb{R}$, and $w_i > 0$ are weights that quantify the "importance" or the "precision" of the data point $x_i$.

Then we could ask for the number that minimizes a weighted loss function, like, for example:

$$\mathcal{V}_{\text{MSE}}^W(\theta) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i (x_i - \theta)^2$$

yielding via:

$$\frac{\partial}{\partial \theta} \mathcal{V}_{\text{MSE}}^W(\theta) = 0 \Rightarrow \sum_{i=1}^{N} 2 w_i (x_i - \theta) = 0$$

$$\theta_{\min}^W = \arg\min \mathcal{V}_{\text{MSE}}^W(\theta) = \frac{1}{\sum_{i=1}^{N} w_i} \left( \sum_{i=1}^{N} x_i w_i \right)$$

and this generalizes the mean to the weighted mean of the numbers $x_i$.
The mean squared error for $\Theta_{\min}^w$ is then

$$\Sigma_{MSE}^w (\Theta_{\min}^w) = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i (x_i - \Theta_{\min}^w)^2 =$$

$$= \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i x_i^2 - \frac{2}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i x_i \Theta_{\min}^w + \frac{1}{\sum_{i=1}^{N} w_i} (\Theta_{\min}^w)^2 \sum_{i=1}^{N} w_i =$$

$$= \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i x_i^2 - 2(\Theta_{\min}^w)^2 + (\Theta_{\min}^w)^2 =$$

$$= \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i x_i^2 - \left(\frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i x_i\right)^2 \geq 0$$

We have obtained that the weighted average of the squared data minus the squared weighted average of the data is the value of the loss function at the optimal choice for $\Theta$. 
We see, here again that

while \( \left( \sum_{i=1}^{K} W_i - \sum_{i=k+1}^{N} W_i \right) \) is negative

the function \( \psi_{\text{MAD}}(\theta) \) decreases and attains

a minimal value when the total

weights of points to the right of \( \theta \)

balance the total weights of the points
to the left, when weights are arranged
to correspond to the nondecreasing set

of numbers \( x_1 \leq x_2 \leq x_3 \ldots \leq x_N \).

\[ W_1 + W_2 + W_3 + \cdots + W_N \]

\[ \theta \]

\[ \Theta \]

**Homework #1:**

Work out this example in detail and
find the expression for the \( \psi_{\text{MAD}}(\Theta) \)
for \( \Theta_{\text{min}} \) with respect to the MAD loss function.
15 Suppose now that we put the numbers $x_1, x_2, \ldots, x_N$ in a box and we select one of them "at random". Formally, we define a variable $w$ which takes values in the set $\{1, 2, 3, \ldots, N\}$ so that

$$\Pr\{i\}\equiv \text{Probability}\{w=i\}=\frac{1}{N} \quad \forall i \in \{1, 2, \ldots, N\}$$

We call $X_w$ a random variable and have clearly

$$\Pr\{X_w=x_j\}=\frac{1}{N}$$

In this setting we may ask: what is a value $\mu$ which "best" represents the values the random variable $X_w$ takes, accounting for the probabilities of taking the various possible values?
In the probabilistic setting, we define the "expected" error in "replacing" each \( x_i \) by \( \mu \) as follows:

\[
\text{Expected } \{ \varepsilon(w) = x_w - \mu \} \triangleq \mathbb{E}[x_w - \mu^2] = \\
\frac{1}{N} \sum_{i=1}^{N} \sum_{x_w = x_i^3} (x_i - \mu) \Pr\{X_w = x_i^3\} = \\
\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu) = \frac{1}{N} \sum_{i=1}^{N} x_i - \mu
\]

We see that by selecting

\[
\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i \Pr\{X_w = x_i^3\}
\]

we get zero expected error.

The number

\[
\mu_x = \sum_{i=1}^{N} x_i \Pr\{X_w = x_i^3\}
\]

is called the expected value, or expectation of the random variable \( X_w \).
What about the process of minimizing a functional of the expected errors, different from the sum of them weighted by their probabilities, which allows for cancellations of positive and negative errors?

Let us consider the expected mean squared error, i.e.

\[
E\{ \mathcal{E}_w^2(\mu) \} = \text{Expected} \{ l(x_w - \mu)^2 y \} = \\
\sum_{i=1}^{n} (x_i - \mu)^2 \Pr\{ x_w = x_i \} = \\
\sum_{i=1}^{n} (x_i^2 - 2\mu x_i + \mu^2) \Pr\{ x_w = x_i \} = \\
\sum_{i=1}^{n} x_i^2 \Pr\{ x_w = x_i \} - 2\mu \sum_{i=1}^{n} x_i \Pr\{ x_w = x_i \} + \mu^2 \sum_{i=1}^{n} \Pr\{ x_w = x_i \}
\]
Now we can use the fact that
\[ \sum_{i=1}^{N} \Pr \{ X_w = x_i \} = \sum_{i=1}^{N} \Pr \{ w = i \} = 1 \]
and this is true not only for the "uniform distribution" we assumed before (where \( \Pr \{ w = i \} = 1/N \) for all \( i \)), but for any set of nonnegative probabilities
\[ \Pr \{ w = i \} = p_i \geq 0, \sum_{i=1}^{N} p_i = 1. \]
Therefore we have:
\[ E \{ \varepsilon^2_w(\mu) \} = \sum_{i=1}^{N} x_i^2 p_i - 2 \mu \cdot \mu_x + \mu^2 \cdot 1 \]
This enables us to find \( \mu \) which minimizes the expected mean squared error, and we get
\[ \frac{d}{d\mu} E \{ \varepsilon^2_w(\mu) \} = -2 \mu_x + 2 \mu = 0 \]
hence
\[ \mu_{\text{optimal}} = \mu_x = \sum_{i=1}^{N} x_i p_i \leq E \{ X_w^2 \} \]
The above discussion introduced (informally) the concepts of "random variable" as a box from which a genie selects "objects" (possible values) according to a variable $\omega \in \Omega$ (a set of possible indices) on which a probability distribution, i.e., a "measure" of weights defined as the probabilities for $\omega$ to take various values in $\Omega$, has been postulated/defined.

The "genie" draws $\omega$ at random from the distribution $\Pr\{\omega=i\}$. The variable $x_i$ is a "realization" of $X_\omega$. 

$$x_i \triangleq X_\omega \{\omega=i\}$$
As we have seen the expected value or the mean of a random variable $X_w$ defined as:

$$
\mu_x = E\{X_w\} = \sum_{i} x_i \Pr(X_w=x_i)
$$

does the following:

1) it zeroes (by definition) the expected error

$$
E\{\varepsilon_w(\mu)=(X_w-\mu)\} = 0 \text{ for } \mu=\mu_x
$$

2) it minimizes the mean (expected value) of the squared error.

$$
E\{\varepsilon_w^2(\mu)\} \rightarrow \text{min} \text{ for } \mu=\mu_x.
$$

What is the minimum mean squared error? Well, we have:

$$
E\{\varepsilon_w^2(\mu_x)\} = \sum_{i} x_i^2 p_i - \mu_x^2 = \\
= E\{X_w^2\} - (E\{X_w\})^2
$$
The quantity
\[ E \{ X_w \} \]
is called the expected value, or the mean, or the first moment of the distribution of \( X_w \).

\[ E \{ X_w^2 \} \]
is called the second moment of the distribution of \( X_w \).

and
\[ E \{ (X_w - \mu_x)^2 \} \]
is called the variance of the random variable \( X_w \).

Sometimes we denote the variance (which is the minimum mean squared error in representing \( X_w \) by its mean, by definition) by \( \sigma_x^2 \) (and it's square root \( \sigma_x \) is called the standard deviation.)
A note about the expectation operator $E$.

We have seen that $E$ averages, with weights given by the probabilities, over the values of random variable realizations determined by $w$. The realization values of random variables can be combined and operated upon to define new random variables that become functions of two or more random selections $w, \tilde{w}$... by their respective 'genies'.

When we consider $X_w$ and $Y_{\tilde{w}}$, two random variables that are independent, i.e. their 'genies' randomly select $w$ and $\tilde{w}$ independently, i.e $Pr\{w, \tilde{w}\} = Pr\{w\} \cdot Pr\{\tilde{w}\}$ (meaning that the probability of the genies selecting the pair $(w, \tilde{w})$ is just the product of the probabilities that the first selects $w$ and the second $\tilde{w}$).
We shall have:

\[ E\{X_\omega Y_\omega\} = E\{X_\omega\}E\{Y_\omega\} \]

(since 

\[ E\{X_\omega Y_\omega\} = \sum_{(i,j)} x_i y_j \Pr\{\omega = i, \tilde{\omega} = j\} = \sum_{(i,j)} x_i y_j \Pr\{\omega = i\} \Pr\{\tilde{\omega} = j\} = \sum_i x_i \Pr\{\omega = i\} \sum_j y_j \Pr\{\tilde{\omega} = j\} = E\{X_\omega\} E\{Y_\omega\} \]

---

Note that we always have

\[ E\{aX_\omega + bY_\omega + c\} = \]

\[ = a E\{X_\omega\} + b E\{Y_\omega\} + c \quad \text{for } a, b, c \text{- scalar constants.} \]

even if the random variables are dependent, i.e. the "genies" select \( \omega \) and \( \tilde{\omega} \) in such a way that they influence each other, i.e.

\[ \Pr\{\omega, \tilde{\omega}\} \neq \Pr\{\omega\} \cdot \Pr\{\tilde{\omega}\} \]

We say that the AVERAGING/EXPECTATION operator is **LINEAR**!
6. Suppose next that we consider a univariate function over an interval $\Delta \subset \mathbb{R}$.

$$
\phi(t) : \Delta \to \mathbb{R}
$$

What would be the constant function that best approximates / represents the values of $\phi(t)$ over $\Delta$?

Well, we are by now well equipped to answer this question.

Suppose we set

$$
\hat{\phi}(t) = \bar{\phi} = \text{constant} : \Delta \to \mathbb{R}
$$

and consider, over $t \in \Delta$:

$$
\varepsilon(t) = \phi(t) - \hat{\phi}(t) = \phi(t) - \bar{\phi}
$$

Clearly we may request to minimize the average squared error over $\Delta$, i.e.

$$
\Psi_\Delta(\bar{\phi}) = \frac{1}{\text{Vol}(\Delta)} \int_\Delta (\phi(t) - \bar{\phi})^2 dt
$$
We can determine the optimal $\bar{e}$ as follows:

$$
\Psi_\Delta(\bar{e}) = \frac{1}{|\Delta|} \int_\Delta (\bar{e}(t)^2 - 2\bar{e}(t)\bar{e}(t) + \bar{e}(t)^2) dt = \\
= \frac{1}{|\Delta|} \left( \int_\Delta \bar{e}(t)^2 dt - 2\bar{e} \int_\Delta \bar{e}(t) dt + \bar{e}(t)^2 \right) dt = \\
= \frac{1}{|\Delta|} \int_\Delta \bar{e}(t)^2 dt - 2\bar{e} \frac{1}{|\Delta|} \int_\Delta \bar{e}(t) dt + \bar{e}(t)^2 \\

\frac{d\Psi_\Delta(\bar{e})}{d\bar{e}} = -2 \frac{1}{|\Delta|} \int_\Delta \bar{e}(t) dt + 2\bar{e} = 0
$$

Yields $\bar{e}_{\text{optimal}} = \frac{1}{|\Delta|} \int_\Delta \bar{e}(t) dt$

and the value of the loss function for $\bar{e}_{\text{optimal}}$ is

$$
\Psi_\Delta(\bar{e}_{\text{optimal}}) = \frac{1}{|\Delta|} \int_\Delta \bar{e}(t)^2 dt - \left( \frac{1}{|\Delta|} \int_\Delta \bar{e}(t) dt \right)^2 > 0
$$

(Here $|\Delta|$ is the size of the interval, i.e. $\int_\Delta dt = |\Delta|$)
1.7 Example (HW)

Consider \( f(t) = mt + n \).

In this case

\[
\overline{\text{Optimal}} = \frac{1}{|\Delta|} \int_{\Delta} (mt + n)\,dt =
\]

\[
= \frac{1}{|\Delta|} \int_{\Delta} mt\,dt + \frac{1}{|\Delta|} \int_{\Delta} n\,dt =
\]

\[
= m \frac{1}{|\Delta|} \int_{\Delta} t\,dt + n
\]

If \( \Delta = [c_A - \frac{|\Delta|}{2}, c_A + \frac{|\Delta|}{2}] \) we obtain

\[
\overline{\text{Optimal}} = m \frac{1}{|\Delta|} \left( \frac{1}{2} t^2 \right)_{c_A - \frac{|\Delta|}{2}}^{c_A + \frac{|\Delta|}{2}} + n =
\]

\[
= m \frac{1}{|\Delta|} \frac{1}{2} \left( (c_A + \frac{|\Delta|}{2})^2 - (c_A - \frac{|\Delta|}{2})^2 \right) + n =
\]

\[
= m \frac{1}{|\Delta|} \frac{1}{2} \left( c_A^2 + |\Delta| + \frac{|\Delta|^2}{4} - c_A^2 - |\Delta| - \frac{|\Delta|^2}{4} \right) + n =
\]

\[
= m \frac{1}{|\Delta|} \frac{1}{2} 2c_A |\Delta| + n = m c_a + n
\]

(The value of \( f(t) \) at the "center" of \( \Delta \).)
and the mean squared error is:

$$
\Psi_\Delta (\theta_{optimal} = m C_\Delta + \eta) = \\
= \frac{1}{|\Delta|} \int_\Delta (m t + n)^2 dt - (m C_\Delta + \eta)^2 = \\
= \frac{1}{|\Delta|} \int_{C_\Delta - \frac{|\Delta|}{2}}^{C_\Delta + \frac{|\Delta|}{2}} (m^2 t^2 + 2 m n t + n^2) dt - (m C_\Delta + \eta)^2 = \\
= \frac{1}{|\Delta|} \left( m^2 \frac{1}{3} t^3 \bigg|_C_{C_\Delta - \frac{|\Delta|}{2}}^{C_\Delta + \frac{|\Delta|}{2}} + 2 m n \frac{1}{2} t^2 \bigg|_C_{C_\Delta - \frac{|\Delta|}{2}}^{C_\Delta + \frac{|\Delta|}{2}} + n^2 \cdot |\Delta| \right) - \\
- m^2 C_\Delta^2 - 2 m n C_\Delta - n^2 = \\
= \frac{1}{|\Delta|} m^2 \frac{1}{3} \left[ (C_\Delta + \frac{|\Delta|}{2})^3 - (C_\Delta - \frac{|\Delta|}{2})^3 \right] + \frac{2 m n}{|\Delta|} \left[ (C_\Delta + \frac{|\Delta|}{2})^2 - (C_\Delta - \frac{|\Delta|}{2})^2 \right] = \\
- m^2 C_\Delta^2 - 2 m n C_\Delta - n^2 = \\
= \frac{m^2}{|\Delta|} \frac{1}{3} \left[ C_\Delta^3 + 3 C_\Delta^2 \frac{|\Delta|}{2} + 3 C_\Delta \frac{|\Delta|^2}{4} + \frac{|\Delta|^3}{8} - \\
- C_\Delta^3 + 3 C_\Delta^2 \frac{|\Delta|}{2} - 3 C_\Delta \frac{|\Delta|^2}{4} + \frac{|\Delta|^3}{8} \right] + \frac{2 m n}{|\Delta|} C_\Delta = \\
- m^2 C_\Delta^2 - 2 m n C_\Delta =
\[ \begin{align*}
= \frac{m^2}{|\Delta|} \cdot \frac{1}{3} \left[ 3C\Delta^2|\Delta| + |\Delta|^3 \right] - m^2C\Delta^2 = \\
= \frac{1}{12} |\Delta|^2 \cdot m^2 + m^2C\Delta^2 - m^2C\Delta^2 = \\
= \frac{1}{12} |\Delta|^2 \cdot m^2
\end{align*} \]

Therefore if \( \Phi(t) = mt + m \) over \( \Delta = [C\Delta - \frac{|\Delta|}{2}, C\Delta + \frac{|\Delta|}{2}] \) we have that

\[ \overline{C}_{\text{optimal}} = mC\Delta + m \]

and

\[ \Psi_{\Delta} (\overline{C}_{\text{optimal}}) = \frac{1}{|\Delta|} \int (\Phi(t) - \overline{C}_{\text{optimal}})^2 dt = \]

\[ = m^2 \cdot \frac{1}{12} |\Delta|^2 = \frac{(\Phi_{\text{max}} - \Phi_{\text{min}})^2}{12} \]

(*) Note here that \( m|\Delta| \) is in fact equal to:

\[ \Phi(t = mC\Delta + \frac{|\Delta|}{2}) - \Phi(t = mC\Delta - \frac{|\Delta|}{2}) = \]

\[ = m \cdot \Phi(mC\Delta) + m \frac{|\Delta|}{2} + \alpha - m \cdot \Phi(mC\Delta) + m \frac{|\Delta|}{2} - \alpha = m |\Delta| \]
(HW) Work out these problems, i.e. best representations are the corresponding mean errors for loss functions that are based on absolute deviations for both random variables and function over an interval.

---

Review also:

- representing values in an interval
  \[ x \in [x_L, x_H] \]

- random variables with values in an interval \( \sim \) distributed uniformly and also \( \sim \) according to some probability density function \( p(x) \).
a. Suppose we have values in an interval, i.e. \( x \in [x_L, x_U] \) and each value \( x \) carries a weight \( w(x) > 0 \). We have for a representation level \( \theta \) the Mean Squared Error

\[
\psi_{MSE}(\theta) = \frac{1}{\int_{x_L}^{x_U} w(x) dx} \int_{x_L}^{x_U} (x - \theta)^2 w(x) dx
\]

\[
\Rightarrow \psi_{MSE}(\theta) = \frac{1}{\int_{x_L}^{x_U} w(x) dx} \left[ \int_{x_L}^{x_U} (x^2 - 2x\theta + \theta^2) w(x) dx \right] = \\
= \frac{1}{\int_{x_L}^{x_U} w(x) dx} \left[ \int_{x_L}^{x_U} x^2 w(x) dx - 2\theta \int_{x_L}^{x_U} x w(x) dx + \theta^2 \int_{x_L}^{x_U} w(x) dx \right]
\]

yielding

\[
\theta_{opt} = \frac{1}{\int_{x_L}^{x_U} w(x) dx} \int_{x_L}^{x_U} x w(x) dx
\]

and

\[
\psi_{MSE}(\theta_{opt}) = \frac{1}{\int_{x_L}^{x_U} w(x) dx} \int_{x_L}^{x_U} x^2 w(x) dx - \left( \frac{1}{\int_{x_L}^{x_U} w(x) dx} \int_{x_L}^{x_U} x w(x) dx \right)^2 \geq 0
\]
Suppose we have a random variable that takes values in an interval, i.e. $X_w \in [x_L, x_H]$ and $X_w = x$ with probability distribution $p(x)$, i.e. $p(x) dx = \Pr\{X_w \in (x,x+dx)\}$.

We have here the expected mean squared error

$$\mathcal{E}(\theta) = \mathbb{E}(x-\theta)^2 = \mathbb{E}_{\text{MSE}}$$

$$= \int_{x_L}^{x_H} (x-\theta)^2 p(x) dx = \int_{x_L}^{x_H} x^2 p(x) dx - 2\theta \int_{x_L}^{x_H} x p(x) dx + \theta^2 \int_{x_L}^{x_H} p(x) dx$$

Since clearly here $\int_{x_L}^{x_H} p(x) dx = 1$ we obtain

$$\theta_{opt} = \int_{x_L}^{x_H} x p(x) dx = \mathbb{E}(X_w) = \mu_x$$

and

$$\mathcal{E}_{opt}(\mu_x) = \int_{x_L}^{x_H} x^2 p(x) dx - \mu_x^2 = \int_{x_L}^{x_H} x^2 p(x) dx - \left(\int_{x_L}^{x_H} x p(x) dx\right)^2 \Rightarrow \sigma_x^2$$

If $p(x) = \frac{1}{x_H-x_L} = \text{constant}$ we have

$$\mu_x = \frac{x_H + x_L}{2} \quad \text{and} \quad \sigma_x^2 = \left[\int_{x_L}^{x_H} x^2 dx\right] \frac{1}{x_H-x_L} - \left(\frac{x_H + x_L}{2}\right)^2 = \frac{1}{12} (x_H - x_L)^2$$
The calculation of $\sigma_x^2$ for the uniform distribution over $[x_L, x_H]$ is

\[
\sigma_x^2 = \frac{1}{3} \left( \frac{1}{x_H - x_L} \right) \left( \frac{x_H^3 - x_L^3}{4} \right) - \left( \frac{x_H + x_L}{2} \right)^2 =
\]

\[
= \frac{1}{3} \left( \frac{1}{x_H - x_L} \right) (x_H - x_L) \left( \frac{x_H^2 + x_H x_L + x_L^2}{4} \right) - \frac{x_H^2 + 2x_H x_L + x_L^2}{4} =
\]

\[
= \frac{1}{3} \left( \frac{x_H^2 + x_H x_L + x_L^2}{x_H - x_L} \right) - \frac{1}{4} \left( \frac{x_H^2 + 2x_H x_L + x_L^2}{4} \right) =
\]

\[
= \frac{1}{12} \left( 4x_H^2 + 4x_H x_L + 4x_L^2 - 3x_H^2 - 6x_H x_L - 3x_L^2 \right) =
\]

\[
= \frac{1}{12} \left( x_H^2 - 2x_H x_L + x_L^2 \right) = \left( \frac{x_L - x_H}{2} \right)^2
\]

So far we know how to replace/represent optimally (mostly in the mean squared sense)

- a finite set of values / possibly weighted
- a range of values / possibly weighted
- a random variable with given distribution
  - discrete values
  - continuous values
- a function over an interval with a single representative value